Université de Poitiers

# Thèse de Doctorat

*Présentée par :*

Adrien Raison

## Analyse et interprétation intelligentes des sémantiques de descripteurs en imagerie médicale

Préparée à XLIM, UMR CNRS n°7252

Frédéric Alexandre, Directeur de Recherche, INRIA Bordeaux ..... Rapporteur
Marc Plantevit, Professeur des Universités, Université Lyon 1, LIRIS ..... Rapporteur
Jean-Yves Ramel, Professeur des Universités, Université de Tours, LIFAT ..... Examinateur
Mathieu Naudin, Ingénieur de Recherche, CHU de Poitiers, I3M ..... Examinateur
Pascal Bourdon, Maître de Conférences, Université de Poitiers, XLIM ..... Examinateur
David Helbert, Professeur des Universités, Université de Poitiers, XLIM .... Directeur de thèse

*Remerciements*

Tout d'abord, je tiens à remercier toutes les personnes qui ont porté le projet de recherche dans lequel ma thèse s'inscrit. Je pense aussi à mes encadrants, David et Pascal, qui m'ont fait pleinement confiance pour mener à bien ces travaux. Je les remercie aussi pour l'apport de leurs expériences et expertises. Je remercie toutes les personnes qui ont contribué de près ou de loin à la réalisation de ces trois années, à ceux qui m'ont donné le goût des sciences. À l'ensemble de ma famille, de mes amis, de mes camarades de recherche, mais aussi aux personnels du laboratoire XLIM qui m'ont aidé dans certaines de mes démarches. Une pensée particulière, et en désordre, pour Alexandre, Sylvain, Clément, Thibault, Damien, Soukayna, Alexis, Maugan, Guillaume, David, Gean, André, Nicolas, et tous les autres qui ont croisé mon chemin. Ce fut vraiment agréable. Sincèrement, merci.

# Table des matières

# Chapitre 1

# Introduction

Ces travaux se sont déroulés entre le 1 octobre 2020 et le 30 septembre 2023. Ils ont été financé par le projet de recherche **MEDIA-IRM-3D** (***ME****dical* ***D****ata* ***I****ntelligent* ***A****nalysis* – ***I****nterpretation,* ***R****econstruction and* ***M****anipulation in 3D*). Ce projet **MEDIA-IRM-3D** s'appuie sur l'avènement de nouvelles approches de modélisation à l'intersection entre la physique de l'IRM, la simulation numérique à large échelle à partir de modèles mathématiques, l'apprentissage machine et le calcul à haute performance qui permettront à un IRM de nouvelle génération (3 Tesla et 7 Tesla) de devenir un véritable outil de diagnostic, de prédiction, de simulation et d'aide aux décisions médicales par la réalisation d'examens totalement non invasifs. Ce projet s'inscrit dans le cadre de la visualisation de données médicales en vue de leur analyse physiologique et métabolique. Nous souhaitons utiliser les données issues des IRM de différentes champs magnétiques (1.5 Tesla à 7.0 Tesla) pour favoriser une reconstruction 3D plus fine des organes. Il implique 3 thèses, dont celle-ci, et 1 post-doc. Il a été porté par Sébastien Horna (I3M, Laboratoire XLIM – UMR CNRS 7252 – Université de Poitiers).

Les partenaires académiques sont :

— Laboratoire de Mathématiques et Applications – UMR CNRS 7031 – Université de Poitiers

— Université de Balamand (Liban)

— Université Heriot-Watt (Edimbourg, Royaume-Uni)

Les partenaires socio-économiques sont :

— SIEMENS HEALTHINEERS

— Centre Hospitalier et Universitaire de Poitiers
— Centre Hospitalier et Universitaire de Limoges

Le laboratoire I3M (**I**magerie **M**étabolique **M**ulti-noyaux **M**ulti-organes) est un laboratoire commun CNRS entreprise, associant SIEMENS HEALTHINNERS, le CHU et l'Université de Poitiers. Le laboratoire I3M (Imagerie Métabolique Multi-noyaux Multi-organes), centre transdisciplinaire unique en France, et même en Europe, basé sur l'imagerie métabolique des organes, dédié à la recherche fondamentale et clinique. Ce premier laboratoire commun avec une entreprise pour l'institut INSMI a le privilège d'exploiter les images de l'IRM 7 Tesla et de travailler avec le laboratoire XLIM.

Le laboratoire XLIM UMR CNRS 7252, est un Institut de Recherche pluridisciplinaire, localisé sur plusieurs sites géographiques, à Limoges sur les sites de la Faculté des Sciences et Techniques, de l'ENSIL, d' Ester-Technopole, sur le Campus Universitaire de Brive et à Poitiers sur le site de la Technopole du Futuroscope. Il est centré sur l'électronique et les hyperfréquences, l'optique et la photonique, les mathématiques, l'informatique et l'image, la Conception Assistée par Ordinateur (CAO), dans les domaines spatial, des réseaux télécom, des environnements sécurisés, de la bio-ingénierie, des nouveaux matériaux, de l'énergie et de l'imagerie. Parmis les axes du laboratoire XLIM, il y a l'axe ASALI. L'axe ASALI est une structure de recherche dédiée à la synthèse et à l'analyse d'images, sur deux sites géographiques (Limoges et Poitiers). Les défis scientifiques sont la conception d'objets complexes structurés en dimension arbitraire, la modélisation et le traitement des informations couleurs et spectrales des images et des vidéos, la synthèse d'image réaliste à base de modèles procéduraux qu'ils soient statistiques ou physiques. La dynamique scientifique de l'axe repose sur plusieurs éléments structurants :

— le développement de modèles statistiques pour la définition et la mesure d'attributs visuels, exploités à la fois pour l'analyse d'images, la synthèse de textures, ou encore la simulation d'éclairage :
  — la conception de modèles liés à la perception humaine pour la prise en compte de l'observateur dans les outils d'analyse, de traitement, de recherche d'images ou dans l'estimation de la qualité perçue ;
  — le développement d'outils théoriques pour l'analyse et la prise de décision

(apprentissage) dans le contexte des images multivaluées ;
  - la définition de modèles et logiciels robustes dédiés à la génération et à la manipulation de la géométrie, discrète ou continue, en dimension quelconque ;
  - la conception et la mise en œuvre de modèles de haut niveau dédiés à la simulation de phénomènes naturels et à l'ajout de détails de mondes virtuels réalistes.
- les activités de l'axe ASALI forment une chaîne complète, depuis des recherches fondamentales jusqu'au développement d'applications industrielles, au travers de 3 équipes dans les 2 thématiques suivantes :
  - l'informatique graphique et la synthèse d'image : Cette thématique correspond à la communauté scientifique Computer Graphics, avec une partie importante des aspects centrés autour de la modélisation géométrique, du rendu réaliste et de la simulation d'éclairage. Deux équipes de recherche composent ce thème, IG et SIR, chacune a ses spécificités scientifiques.
  - le traitement et l'analyse (de séquences) d'images : Cette thématique a une identité scientifique centrée sur la chaine de vie des images multivaluées. Initialement centrée sur l'image couleur, la thématique s'est étendue à l'ensemble des images multivaluées et plus récemment sur l'adaptation des approches d'apprentissage profond et de l'Intelligence Artificielle à ce contexte multivalué. L'ensemble correspond à la communauté Image Processing and Computer Vision. Cette thématique correspond aux travaux de recherche de l'équipe ICONES, équipe à laquelle nous appartenons.

Nous présentons maintenant le contexte scientifique de nos travaux.

Depuis les trois derniers siècles, l'automatisation a permis, dans le domaine industriel, de déléguer des tâches de traitement de la matière à des machines, tâches qui étaient précédemment effectuées par l'Homme. Cette automatisation a permis de drastiquement augmenter les volumes de productions, de pérenniser et de fiabiliser les processus de productions. Cependant, il y a encore quelques décennies, le traitement automatisé et efficace de l'information n'était lui pas transférable aux machines et était une capacité foncièrement humaine. Aujourd'hui, dus à différents facteurs, nous pouvons traiter l'information efficacement avec des machines. Comme

pour la matière, cette automatisation via des machines a permis d'augmenter significativement le volume de traitement de l'information par rapport à celui réalisable par des humains. Cependant, l'étude de la matière a commencé bien plus tôt que celle de l'information. Bien que le traitement automatique de l'information reçoive aujourd'hui une grande attention de la part de la communauté scientifique, la fiabilisation des processus de traitement de l'information par des machines reste un vaste sujet d'étude, sûrement plus large que leurs homologues traitant la matière. Ces entités calculatoires, ou encore, ces modèles, permettant de traiter automatiquement et efficacement de l'information, sont alors qualifiés d'intelligents.

Tout comme l'impact qu'a eu le traitement automatisé de la matière sur le développement des sociétés et de leurs économies, celui du traitement automatisé de l'information joue aujourd'hui un rôle prépondérant sur la définition des sociétés d'aujourd'hui et de demain. Il existe aujourd'hui une multitude de domaines où le traitement automatisé de l'information est déterminant pour la réussite de certaines tâches notamment, celle relative à la société, par exemple dans le domaine médical, politique, ou encore dans celui du transport autonome.

Cependant, pour des raisons éthiques ou encore juridiques, ces traitements doivent être fiables et ne doivent pas avoir d'impact négatif sur nos sociétés. Il est alors nécessaire d'avoir une instance de contrôle pour ces modèles. Cela permet alors de rendre transparents leurs fonctionnements et ainsi de décider si leurs utilisations sont bénéfiques ou non pour nos sociétés. En parallèle du développement des modèles intelligents toujours plus performants et applicables à de plus en plus de problèmes, les scientifiques se sont alors nécessairement intéressés aux questions sur l'utilisabilité et la transparence de ces modèles dans et pour notre société. Cela est l'objet des travaux présentés dans ce document.

Les quatre premiers chapitres de ce document se réfèrent à l'existant concernant ces modèles dits intelligents et sur les méthodes permettant leurs contrôles. Les trois derniers chapitres porteront sur les travaux, effectués pendant ces trois années de recherche au sein de l'axe ASALI du laboratoire XLIM, s'intéressant au contrôle des modèles intelligents.

Dans le premier chapitre, nous introduisons ces modèles, leur genèse, leurs limites et ce qui leur est nécessaire pour être qualifiés d'intelligents.

Dans le second chapitre, nous nous intéressons à une classe de modèle ayant des propriétés particulièrement intéressantes pour le traitement de l'information contemporaine, il s'agit des réseaux de neurones.

Dans le troisième chapitre, nous nous intéresserons à une classe de problèmes spécifiques , dit à a priori géométriques, pour lesquels les réseaux de neurones peuvent être adaptés, donnant naissance à une sous partie de l'ensemble des réseaux de neurones : les réseaux de neurones a priori géométriques.

Dans le quatrième chapitre, nous évoquons les travaux existants concernant les méthodes de contrôle pour les modèles intelligents incluant un a priori géométrique dans leurs définitions ou non.

Dans le cinquième chapitre, nous décrivons nos premiers travaux dans le domaine du contrôle des modèles intelligents. En effet, ils portent sur une analyse des méthodes de contrôle existantes dans le domaine de la neurologie. Ce travail a permis de retrouver des résultats présents dans la littérature biologique via des arguments numériques produits par des modèles intelligents. Ce travail illustre, de manière favorable, la possibilité de l'utilisation de modèles intelligents dans le domaine de la neurologie.

Dans le sixième chapitre, nous détaillons la généralisation d'une méthode de contrôle efficace à une classe de modèles, à a priori géométrique, plus large que celle pour laquelle elle avait été créée initialement. Dans ce travail, nous illustrons en quoi cette méthode reste pertinente par rapport à l'état de l'art avec des mesures quantitatives et qualitatives usuelles.

Dans le septième chapitre, nous appliquons la méthode de contrôle précédemment évoquée à la caractérisation numérique d'émotions faciales. Nous montrons là aussi la pertinence de la méthode par rapport à l'état de l'art, renforçant ainsi la pertinence globale de la méthode par rapport aux méthodes existantes, bien que celle-ci ait quelques limitations.

Enfin, dans le huitième chapitre, nous levons une des principales limitations de la précédente méthode et intégrons un point manquant, à notre connaissance, dans toutes les méthodes de contrôle existantes. Dans ce chapitre, nous présentons alors une nouvelle méthode de contrôle ayant une grande compatibilité avec les modèles intelligents à a priori géométriques, et incluant la dépendance sociale auxquelles les

méthodes de contrôle sont par essence dépendantes comme nous le verrons.

# Chapitre 2

# Introduction à l'intelligence artificielle

## Sommaire

Aujourd'hui ce que l'on appelle communément les intelligences artificielles sont des modèles calculatoires reposant sur la théorie de l'apprentissage profond. Dans plusieurs domaines, ces modèles sont plus performants pour résoudre certains problèmes que les humains. De part leurs performances sur une multitude d'applications, ils sont aujourd'hui largement utilisés par le plus grand nombre. Ces personnes, surtout celles n'ayant pas de sensibilité scientifique particulière, ont souvent une vision floue de ce que sont réellement ces modèles d'intelligence artificielle. Dans ce chapitre nous allons présenter les modèles d'intelligence artificielle sous un angle constructiviste, intuitif, et historique, cadrant ainsi ce que sont ces modèles et ce qu'ils ne sont pas. On peut définir la notion d'intelligence artificielle, en accord avec la définition initiale que nous allons voir et celle que nous lui prêtons aujourd'hui comme étant des méthodes capables de faire un traitement automatique de l'information. Dans cette partie, nous introduirons du point de vue historique, informel, et sans rigueur mathématique et physique, cette notion de traitement automatique de l'information. Dans un premier temps, nous introduirons la notion d'information puis nous évoquerons par la suite comment nous pouvons la traiter de manière automatique [1].

## 2.1 Théorie de l'information

### 2.1.1 Formulation physique

Une définition possible pour la science [2] est la suivante : "la science est la somme de connaissances qu'un individu possède ou peut acquérir par l'étude, la réflexion ou l'expérience". La science regroupe plusieurs disciplines. Parmi elles, la science physique. La science physique est la science qui essaie de comprendre, de modéliser et d'expliquer les phénomènes naturels de l'Univers de la manière la plus formelle possible. Cette science est d'intérêt, car, en effet, l'Homme a toujours cherché à expliquer le monde dans lequel il évolue à savoir notre Univers. En sciences phy-

1. Les concepts présentés dans ce chapitre ont des définitions formelles et ont chacun un cadre théorique très rigoureux, dont la portée réelle est bien au-delà de ce document. La rédaction de ce chapitre a été fortement inspirée de [McCorduck, 2004, Jaynes et al., 2003, Tibshirani and Friedman, ,Bishop, 2006,Murphy, 2012,Carton, ,Mitchell, 1997,Mohri et al., 2018,Shalev-Shwartz and Ben-David, 2014].

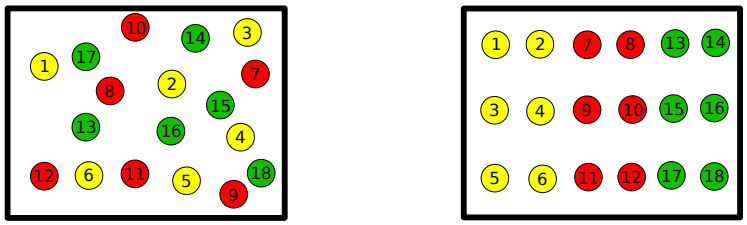
2. D'après le dictionnaire Larousse

siques, la compréhension des différents phénomènes se fait souvent par induction via différentes observations d'un système physique. Les grandeurs physiques caractérisent numériquement, par une mesure ou un calcul, relativement à une référence, ces systèmes. On dit que le système peut alors prendre plusieurs états, au sens où les grandeurs physiques peuvent varier selon les différentes observations du système. Ces grandeurs physiques s'expriment selon sept unités référencées dans le Système International (SI). Les grandeurs que les physiciens considèrent sont toutes dérivées des sept grandeurs canoniques ci dessous :

— la longueur qui s'exprime dans le SI en mètre ($M$)
— le temps qui s'exprime dans le SI en seconde ($S$)
— le masse qui s'exprime dans le SI en kilogramme ($KG$)
— la température qui s'exprime dans le SI en kelvin ($K$)
— l'intensité d'un courant électrique qui s'exprime dans le SI en ampère ($I$)
— l'intensité lumineuse qui s'exprime dans le SI en candela ($J$)
— la quantité de matière qui s'exprime dans le SI en mole ($N$)

Parmi les grandeurs dérivées existantes, une va retenir particulièrement notre attention : il s'agit de l'entropie. Initialement, la notion d'entropie a été introduite dans des études portant sur les machines à vapeur, au XVIIIe siècle lors de la révolution industrielle. Ces études ont donné naissance à tout un pan des sciences physiques : la thermodynamique. Dans sa formulation physique, l'entropie est une grandeur physique caractérisant le degré de désorganisation d'un système physique. Cette grandeur s'exprime, selon le SI, en $KG \cdot M^2 \cdot S^{-2} \cdot K^{-1} \cdot I^0 \cdot J^0 \cdot N^0$. Il s'avère que l'entropie d'un système physique, ou son degré de désorganisation, quantifie le manque d'information contenue dans le système. Illustrons ce propos avec un exemple. Considérons le système physique suivant : une piscine remplie d'une multitude de boules de couleurs rouges, vertes et jaunes. Considérons deux observations de ce système et supposons qu'un physicien souhaite caractériser ce système selon ces deux observations :

— Considérons maintenant une première observation (observation A) de ce système dans un état désordonné : les boules sont éparpillées dans la piscine et les couleurs des boules sont mélangées. Disons que l'observation A caractérise le système physique {piscine + boules colorées} dans une version très désor-




(a) Observation A - haute entropie

(b) Observation B - basse entropie

FIGURE 2.1 – Les différents états du système physique {piscine + boules colorées}

donnée de celui-ci, c.-à-d., dans une version avec une haute entropie (voir Figure 2.1a).

— Considérons maintenant une deuxième observation (observation B) de ce système dans un état ordonné : les boules d'une même couleur sont regroupées ensemble dans un groupe de couleur et les différents groupes de couleur ne se mélangent pas entre eux. Disons que l'observation B caractérise le système physique {piscine + boules colorées} dans une version très ordonnée de celui-ci, c.-à-d. dans une version avec une basse entropie (voir Figure 2.1b).

Maintenant, voici les descriptions formelles que le physicien pourrait faire en fonction de ces deux niveaux d'entropie du même système :

— dans le cas de l'observation A , il pourrait décrire le système par :
*"la boule jaune n°1 se trouve à gauche de la boule verte n°17 qui elle-même se trouve en haut à gauche de la boule rouge n°8 située au-dessus de la boule verte n°13 qui elle-même est située en haut à droite de la boule rouge n°11 qui ..."* ; il donne alors une description microscopique du système.

— dans le cas de l'observation B , il pourrait décrire le système par :
*"le groupe des boules jaunes se trouve à gauche du groupe des boules rouges qui se trouve lui-même à gauche du groupe des boules vertes."* ; il donne alors une description macroscopique du système.

Imaginons maintenant que le physicien a accès à une autre observation A' du même niveau de désordre que celle de l'observation A, et pour simplifier, disons que l'observation A' est construite de telle manière que les boules sont juste permutées entre elles par rapport à celles de l'observation A et que l'espace de la piscine est donc occupé de la même manière entre l'observation A et A'. De même, supposons

que le physicien ait accès à une observation B' analogue du point de vue de l'entropie à l'observation B. Alors dans les deux cas, il décrirait le système selon l'observation B' de la même manière que selon l'observation B. Et même chose pour l'observation A' et A. Cependant, ces descriptions sont beaucoup plus représentatives du système dans le cas des observations de type B que dans celles de type A. En effet, des observations de type B, il en existe $18! \approx 6.4 \times 10^{15}$, alors que pour l'observation de type A , il existe $3! = 6$. Autrement dit, les observations de type $B$ ont un pouvoir informatif pour le physicien de l'ordre de $10^{-15}$ alors que les observations de type A ont un pouvoir informatif de $\frac{1}{6}$, qui est bien plus grand que $10^{-15}$, pour caractériser le système.

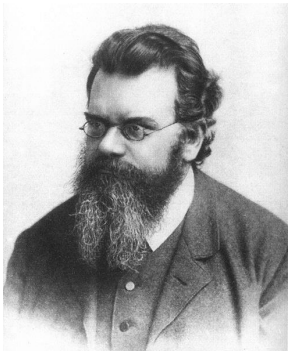


FIGURE 2.2 – Ludwig Boltzmann (1844-1906), Wikipedia

Et la quantité d'informations, au sens commun, qu'a extrait le physicien de ces observations, est corrélée à la quantité d'informations, au sens physique, contenue intrinsèquement dans le système. Et c'est cette notion d'ordre (c.-à-d. représentation ordonnée contre représentation chaotique) du système qui formalise cette notion d'information physique du système. Autrement dit, la notion de désordre d'un système physique, numériquement représenté par l'entropie, quantifie le manque d'informations entre dans les observations microscopiques du système sachant ces observations macroscopiques. C'est le physicien Ludwig Boltzmann (Figure 2.2) qui formalisa l'entropie sous cette forme probabiliste en 1877. Cependant la notion d'entropie ne s'est pas arrêtée au domaine de la physique.

### 2.1.2 Formulation mathématique

En 1948, le mathématicien américain Claude Shannon (Figure 2.3) redéfinit la notion d'entropie dans le domaine des télécommunications [Shannon, 1948]. Il définit l'entropie d'un message ou d'un état physique comme étant la quantité d'informations manquantes pour connaître entièrement le message ou l'état physique et constitue les prémices de la théorie de l'information. Initialement, cette théorie a été construite pour étudier comment envoyer un message informatif de structure optimale au travers d'un signal bruité, par exemple dans le cas d'échange

radiophonique. Quelle est l'intuition globale derrière cette structure optimale du message ? Prenons le cas d'une communication entre deux personnes pouvant s'envoyer des informations par rapport à des événements dont leurs occurrences sont plus ou moins probables. Si le message envoyé contient une information relative à un événement qui peut très probablement survenir, alors le message doit avoir une structure très simple par rapport à l'ensemble des structures admissibles, de manière à ce qu'il soit facilement transmissible au travers d'un réseau de communication (au sens de l'énergie dépensée pour le transmettre). En effet, il est probable (par définition) que nous envoyions ce message un bon nombre de fois tout au long de la communication, d'où la nécessité d'une structure contenue de celui-ci parmi celles admissibles. Un message concernant l'occurrence d'un évènement ayant une faible probabilité de réalisation apporte plus d'informations pour le destinataire. On peut donc lui donner une structure plus complexe parmi celles admissibles. En effet, du fait de la forte probabilité d'occurrence de l'événement, le message associé n'est alors que très peu informatif pour celui qui va le recevoir. Le destinataire savait déjà que l'occurrence de l'événement contenu dans le message était très probable. La réception de ce message n'apporte donc que peu d'informations, ou du moins peu d'informations utiles au destinataire. Il n'apporte pas de connaissance supplémentaire significative au destinataire. À l'inverse, un message qui concerne un évènement qui est très peu probable d'occurrer est très informatif pour le destinataire. On peut considérer que cela altère fortement l'habitude de recevoir des messages concernant des événements très probables de survenir. Ce genre de message est donc plus interpellant que ceux reçus habituellement. D'un point de vue des télécommunications, afin d'optimiser l'utilisation des bandes passantes au sein des réseaux de communication, il faut donc accorder une structure légère aux messages fréquemment transmis et laisser les structures plus complexes aux messages moins fréquemment transmis, de manière à utiliser, en moyenne, peu de bandes passantes. Shannon définit la quantité d'information utile d'un message relatif à un événement comme étant égale à moins la log-probabilité

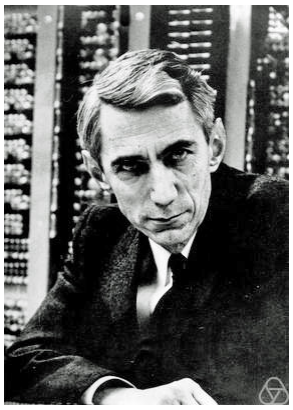


FIGURE 2.3 – Claude Shannon (1916-2001), Wikipedia

d'occurrence de cet événement. Cette quantité quantifie le volume d'information utile d'un message lorsque celui-ci est représenté de manière optimale. L'unité d'informativité est le bit de Shannon. Dans le cas où les messages sont représentés de manière binaire, le log associé est alors en base 2. Il définit aussi l'entropie de Shannon : elle quantifie la quantité moyenne d'information utile apportée par un ensemble de messages associés distribués selon une loi de probabilité $P$. Ainsi, si nous assimilons les messages à une variable aléatoire $X \sim P$, alors l'entropie de Shannon est définie par $\underset{X\sim P}{-\mathbb{E}}[\log(P(X))]$. Prenons un exemple concret : supposons que nous échangeons avec une station météo (via son écran digital) pour savoir s'il fera beau ou s'il pleut demain, sachant que ces événements sont équiprobables. Si la station météo nous indique qu'il fait beau demain, alors la station a réduit notre incertitude vis-à-vis de la météo de demain d'un facteur 2, car les événements sont équiprobables. Avant la réception du message, il y avait deux options possibles, maintenant il n'y en a plus qu'une : il fait beau demain. C'est un message apportant une grande quantité d'informations. En moyenne pour ce contexte, en calculant l'entropie de Shannon, les messages produits auront une quantité d'information utile égale à $-\frac{2}{2}\log(\frac{1}{2}) = 1$ bit de Shannon. Dans le cas présent, l'occurrence des évènements suit une loi de Bernoulli $\mathcal{B}(p)$ avec $p = \frac{1}{2}$. Pour $p \in ]0, 1[$, la quantité $\underset{X\sim\mathcal{B}(p)}{-\mathbb{E}}[\log(P(X))]$ évolue de la manière suivante :

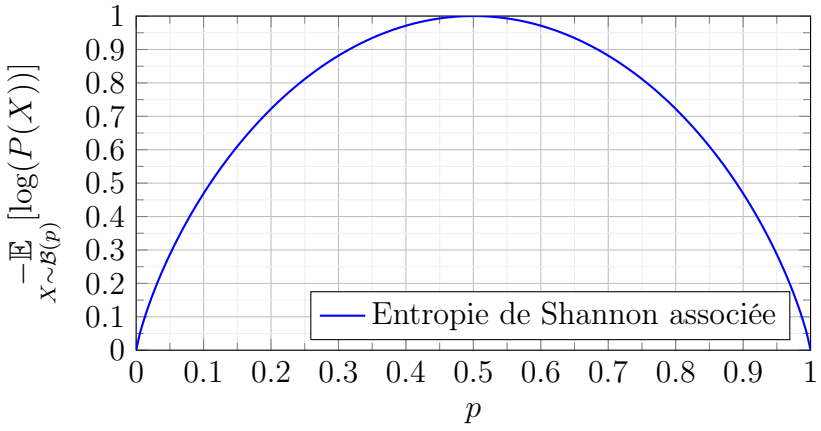



FIGURE 2.4

Plus précisément, la Figure 2.4 quantifie le comportement de l'entropie de Shannon lorsque $\mathcal{B}(p)$ s'approche d'une distribution déterministe ($p = 0$ ou $p = 1$) ou lorsqu'elle s'approche d'une distribution uniforme, c'est à dire $p = \frac{1}{2}$. Lorsque $p =$

0 ou $p = 1$, la quantité d'information utile est en moyenne égale 0, car il n'a plus d'aspect aléatoire dans la réalisation des événements. On sait déjà tout ce qui se passe à l'avance, tout est complètement déterminé à l'avance. Dans notre contexte météorologique, il fait alors soit toujours beau et jamais de pluie ou toujours de la pluie et jamais beau. Lire le message délivré par la station météo ne sert donc à rien dans les deux cas. Lorsqu'il y a de l'incertitude qui apparaît ($p \neq 0$ et $p \neq 1$), la quantité d'information utile augmente, en effet l'occurrence d'un événement plus ou moins aléatoire apporte une information plus ou moins utile, mais utile tout de même. L'entropie de Shannon augmente alors. De manière plus large, dans le cas avec plus de deux évènements où l'occurrence suit une loi uniforme, c'est à dire lorsqu'il est aussi probable d'avoir l'occurrence d'un évènement ou d'un autre, aucune occurrence n'est susceptible d'apporter plus d'information utile qu'une autre, car elles ont toutes la même probabilité de se produire : elles apportent donc toutes la même quantité d'information utile. L'entropie de Shannon est alors maximale et dans notre cas, vaut $-\log_2(\frac{1}{2}) = 1$ comme l'indique la Figure 2.4. Revenons maintenant sur le choix des structures du message. La valeur de l'entropie de Shannon caractérise pour, tout système de représentation (ou d'encodage) utilisable dans un problème donné, un seuil optimal moyen d'informativité nécessaire pour décrire un ensemble de réalisation d'événements. Si nous choisissons un système d'encodage type ASCII, le message "beau" est encodé sur 4 caractères, chacun encodé sur un octet. Le message envoyé sera donc de taille 32 bits. Le message "pluie" sera quant à lui encodé sur 40 bits, soit une taille de message moyenne tout au long de la communication de 36 bits. Ce système n'est alors pas vraiment économe puisque l'entropie de Shannon indique, pour ce problème, que seulement 1 bit est nécessaire pour rendre compte de toute l'information utile, décrite au sein des messages, fournie par la distribution des événements. Il y a donc en moyenne 35 bits qui circulent dans les réseaux de communication inutilement. Si nous choisissons un système d'encodage binaire du type : "0" pour le message concernant l'événement "il fait beau demain" et "1" pour le message concernant "il pleut demain", nous envoyons toujours un message encodé sur 1 bit (en moyenne aussi, trivialement). Ce système de représentation est alors optimal, car égal à l'entropie de Shannon. Si maintenant nous changeons les probabilités d'occurrence des deux événements : disons $\frac{1}{4}$ pour le beau temps et

$\frac{3}{4}$ pour le mauvais temps. La quantité d'information utile du message concernant l'événement "il fait beau demain" est 2 bits de Shannon. La quantité d'information utile du message concernant l'événement "il pleut demain" est 0,41 bit de Shannon. L'entropie de Shannon vaut alors 0,81 dans cette configuration. Ces valeurs ne sont pas aberrantes. Par exemple, dans le cas du beau temps, comme cet événement est très probable, l'information transportée n'apporte pas grand-chose, il n'y a pas beaucoup d'information utile compte tenu de la distribution de ces événements. On a le raisonnement opposé pour l'événement "il pleut demain".

Au travers de ces deux exemples, on observe intuitivement que la formulation physique et la formulation mathématique de l'entropie sont intimement liées. D'ailleurs, Shannon a aussi montré l'équivalence avec la formulation boltzmanienne de l'entropie dans ses travaux. Pendant des siècles, les objets d'études des physiciens ont été la quantification et la mise en relation d'une multitude de grandeurs comme la vitesse d'un corps céleste par exemple. Aujourd'hui, c'est la caractérisation de la grandeur information qui est l'objet d'étude d'une grande partie des scientifiques, et même plus précisément et pour certains scientifiques, c'est la caractérisation automatique de l'information qui est l'objet d'étude, autrement dit, c'est le traitement automatique de l'information qui est au centre d'une multitude de problématiques scientifiques.

## 2.2 Machine et intelligence

La portée des travaux de Shannon ne s'est pas seulement arrêtée à la formalisation de la théorie de l'information. Ce dernier a aussi travaillé avec George Boole, antérieurement aux travaux sur l'entropie, sur les circuits logiques et a montré dans [Shannon, 1938], que tout phénomène physique peut être encodé sous forme de bit de Shannon. Autrement dit, Shannon a formalisé d'un point de vue matériel tous les appareils permettant l'envoi, le stockage, le traitement et l'envoi d'information que nous connaissons de nos jours. Ce point est fondamental. En effet, il indique que tout ce qui constitue l'univers est descriptible par une suite de bit traitable par des circuits logiques, c'est à dire par des machines. Cependant, pouvons-nous traiter automatiquement de l'information par des machines ? Dans le paragraphe précédent,

nous nous sommes intéressés à la notion d'information. Mais dans le groupe de mots "traitement automatique de l'information", qu'entendons-nous par traitement ? Et qu'entendons-nous par automatique ?

### 2.2.1 La notion de traitement : théorie de la calculabilité

Dans cette partie, nous allons apporter des éléments de réponse à la question suivante : pouvons-nous "traiter" physiquement de l'information ? Étymologiquement, "traiter" signifie "agir sur une substance pour la modifier". Par extension, traitement signifie manière d'agir sur une substance pour la modifier. Ainsi, appliqué à la notion d'information, le traitement de l'information signifie prendre une information et selon un certain procédé, la transformer en une autre information différente. Mathématiquement, il s'agit alors d'une notion analogue à celle d'application, où une application $f$ construit une entité $f(x)$ à partir d'une entité initiale $x$ et où le processus d'action est défini par $f$. On dit aussi que l'on a calculé $f(x)$ selon le procédé $f$ à partir de l'entité initiale $x$.

FIGURE 2.5 – Alonzo Church (1903-1995), Wikipedia

Une question légitime émerge alors : qu'est ce qui est calculable dans notre Univers ? C'est-à-dire est que pour une entité $x$ donnée et un processus $f$, nous pouvons physiquement obtenir $f(x)$ ? Dans le cas du traitement de l'information, cela revient à se poser la question de la possibilité du traitement physiquement réalisable de certaines informations ? En effet, avant de vouloir traiter des informations, il est important de savoir si cela et possible et dans quelle mesure. C'est le mathématicien Alonzo Church (Figure 2.5), qui énonce la définition de la notion de calculabilité, dans ce que l'on appelle la thèse de Alonzo Church, ou thèse de Church [Church, 1936a], que voici (dans sa forme physique) :

*La notion physique de la calculabilité, définie comme étant tout traitement systématique réalisable par un processus physique ou mécanique, peut être exprimée par un ensemble de règles de calcul, défini de plusieurs façons dont on a pu démontrer mathématiquement qu'elles sont équivalentes.*

La notion de calculabilité, de ce qui est donc possible de calculer, est alors un

procédé $f$ réalisable par un processus physique ou mécanique. Si tel est le cas, on dit alors que $f$ est une fonction calculable. D'autres mathématiciens ont cherché eux aussi à formaliser la notion de calculabilité, comme Kurt Gödel et Jacques Herbrand avec la théorie des fonctions généralement récursives ; Alan Turing (Figure 2.6)avec les machines de Turing, ou bien encore Alonzo Church lui-même avec la théorie du $\lambda$-calcul. Dans sa publication [Turing, 1937a], Alan Turing montre que toutes ces formalisations de la calculabilité sont mathématiquement équivalentes. Un des avantages de la formulation de calculabilité, selon Turing, est qu'au travers des machines de Turing, Turing donne une approche constructive, ou mécanique, de la notion de calcul. Ainsi, et selon l'équivalence, il reformule la thèse de Church en la thèse de Church-Turing, que voici :

FIGURE 2.6 – Alan Mathison Turing (1912-1954), Wikipedia

*Toute fonction physiquement calculable est Turing calculable, c'est-à-dire calculable par une machine de Turing*

Cette seconde formulation est semblable à la première, où *"tout traitement systématique réalisable par un processus physique ou mécanique"* est synonyme de *"toute fonction physiquement calculable"*, sauf que celle-ci spécifie que *"l'ensemble des règles de calcul"* répond au formalisme des machines de Turing. En particulier, cela signifie que tout problème de calcul fondé sur une procédure mécanique peut être résolu par une machine de Turing.

Enfin, il existe un dernier postulat : il s'agit de la thèse de Church-Turing-Deutsch, que voici :

*Toute machine de Turing universelle peut simuler les processus physiques de notre Univers*

Ce postulat de David Deutsch fait en 1985 dans [Deutsch and Penrose, 1997] est très fort. En effet, il stipule que n'importe quel système physique de notre Univers est calculable par une machine de Turing (universelle). Cela signifie que pour tout phénomène physique, nous pouvons, grâce à une machine de Turing universelle, calculer chaque issue temporelle de ces phénomènes. Notons tout de même que ces thèses ne sont que des conjectures et n'ont pas été démontrées. Cependant, à ce jour,

il n'existe aucun contre-exemple pour les trois thèses. Maintenant, revenons plus en détail sur cette notion de machine de Turing universelle, et dans un premier temps, sur la notion de machine de Turing.

### Machine de Turing

Une machine de Turing est un objet théorique défini par Alan Turing dans [Turing, 1937b]. Formellement, une machine de Turing $M$ est un quintuplet $M = (E, A, e_0, \delta, F)$ [3] où :

- $E$ est un registre d'état fini qui constitue tous les états dans lesquels peut être la machine
- un ruban unidimensionnel infini à support dans $E$
- une tête de lecture/écriture sur le ruban se déplaçant soit à gauche, soit à droite
- $A$ est le registre de symboles de la machine, c'est le référencement de toutes les actions que peut faire la machine
- $e_0 \in E$ est l'état initial de la machine
- $\delta : (E \backslash F) \times A \to E \times A \times \{\to, \leftarrow\}$ est la fonction de transition de la machine, dans un état et une action effectuée. Celle-ci se retrouve dans un nouvel état, dans lequel elle va devoir effectuer une autre action, et ainsi de suite.
- $F \subset E$ est l'ensemble des états terminaux de la machine, c'est-à-dire que si la machine se retrouve dans un état de $F$, alors la machine s'arrête.

Par exemple, considérons la machine de Turing $M_0$ suivante, tel que :

- $E = \{0, 1, h\}$
- $A = \{s, n, i, e, c\}$
- $e_0 = s$
- une fonction de transition $\delta$ telle que :

| $\delta$ | $s$ | $c$ | $i$ | $e$ | $n$ |
|---|---|---|---|---|---|
| 0 | $(0, c, \to)$ | $(0, i, \to)$ | $(0, e, \to)$ | $(1, n, \to)$ | $(1, c, \to)$ |
| 1 | $(1, c, \to)$ | $(1, e, \to)$ | $(1, c, \to)$ | $h$ | $(1, c, \to)$ |

3. Ces objets, introduits très simplement ici, sont les objets de base de l'informatique théorique contemporaine.

— $F = \{h\}$

Par exemple, la machine de Turing décrite ci-dessus, écrit le mot `science` sur son ruban, puis s'arrête. Maintenant considérons une machine de Turing $M_1$, qui prend en entrée $(M_0, e_0)$ et dont le registre de symbole est dérivé intelligemment de celui de $M_0$, tel que les transitions de $M_1$ suivent celle de $M_0$, alors moyennant quelques arrangements, nous pouvons obtenir aussi `science` sur le ruban de $M_1$. Autrement dit, nous pouvons entièrement émuler une machine de Turing avec une autre machine de Turing. Il existe alors une machine de Turing canonique, dite universelle. Dans le vocabulaire ensembliste, la machine de Turing universelle joue alors le rôle de représentant de classe des machines de Turing. Autrement dit, faire un raisonnement ou démontrer un résultat par exemple, avec une machine de Turing universelle revient à le montrer pour toute machine de Turing. Dans notre exemple ci-dessus, on sent que la machine de Turing universelle $M_1$ semble peut être plus complexe que $M_0$, mais alors peut-on trouver une machine de Turing universelle la plus simple possible ? L'informaticien Stephen Wolfram (Figure 2.7) en a proposé une avec seulement 2 états et 5 symboles. Ainsi, avec seulement 10 descripteurs, nous pouvons réaliser tous les calculs physiquement réalisables de l'Univers.

FIGURE 2.7 – Stephen Wolfram (1959 - ·), Wikipedia

Nous avons vu que, si la thèse de Church est vraie, tout problème de calcul fondé sur une procédure mécanique peut être résolu par une machine de Turing. Autrement dit, que tout problème de calcul fondé sur une procédure algorithmique peut être résolu par une machine de Turing. De plus, si la thèse de Church-Turing-Deutsch est vraie, alors tout phénomène physique est simulable par une machine de Turing, et notamment par celle de Wolfram. Pour résumer, bien que les machines de Turing soient des objets abstraits permettant la formalisation de ce qu'est le traitement physiquement possible de l'information, elles sont toujours au centre de la recherche en informatique théorique, et elles représentent, en pratique, le fonctionnement des appareils de calcul, ce que l'on connaît plus communément aujourd'hui sous le nom d'ordinateur. Elles formalisent aussi la notion de "procédure mécanique", c'est-à-dire d'algorithme. Nous

reviendrons sur la notion d'algorithme plus tard, mais tout d'abord voici trois problématiques :

— Nous avons évoqué dans la partie ci-dessus le traitement de l'information à travers les machines de Turing, mais quid du traitement automatique de l'information par une machine de Turing ? Le mot "automatique" est dérivé du mot "automate". Le terme "automate" signifie "qui se meut soi-même" [4]. Ainsi le traitement automatique de l'information par une machine de Turing signifie que ce traitement de l'information se fait initialement par la machine de Turing et que la fin du traitement se fait par, et uniquement par, cette même machine de Turing. La machine de Turing dessine alors une "trajectoire" du traitement de l'information, mais quelle est la nature de cette trajectoire ? Quel est son point de départ et celui d'arrivée ? Comment sont construites les étapes intermédiaires liant le point de départ et le point d'arrivée ? Comment ces points interagissent-ils entre eux ?
— Initialement, nous étions intéressés par le traitement automatique de l'information. Cependant les travaux de Turing ou encore de Church, formalisent le traitement physiquement possible de l'information, ce qui est plus fort que le traitement de l'information au sens large. Mais alors, les machines de Turing ont-elles des limites ? Autrement dit, y a-t-il des choses qu'elles ne peuvent pas calculer ?
— De plus, en tant qu'humain, notre cerveau est le lieu de toutes nos pensées, de nos raisonnements : c'est le lieu où notre "intelligence" se matérialise. Cependant, notre cerveau est bien un système physique. Or, si la thèse de Church-Turing-Deutsch est vraie, alors notre cerveau est simulable par une machine de Turing. Mais alors, comment un procédé algorithmique peut-il "simuler de l'intelligence" ? Comment, à partir d'informations, pouvons-nous faire émerger une intelligence via une machine de Turing ?

Nous présentons ici quelques éléments de réponse à ces trois problématiques.

4. D'après le dictionnaire Larousse

### 2.2.2 Algorithmique, incomplétude et intelligence

La première notion d'algorithmique est apparue via Al-Khwârizmî (825), dans [Com, a]. Dans cet ouvrage, il pose les fondements de ce que nous connaissons aujourd'hui sous le nom d'algèbre, et il s'intéresse à la résolution des équations du premier et deuxième degré. Cependant, Al-Khwârizmî adopte une approche alternative pour la présentation des résultats. En effet, au lieu de donner simplement les résultats via une démonstration comme fait conventionnellement, il fournit plutôt les résultats sous forme d'un ensemble de méthodes de résolutions simple. Autrement dit, il fournit un procédé sûr permettant de résoudre ces équations en s'affranchissant de toutes les étapes de raisonnement difficile, que l'on trouve souvent dans les démonstrations écrites, et donne seulement des protocoles simples à respecter pour résoudre ces équations. C'est la naissance de l'approche dite "alkhwârizmîque" ou "algorithmique". Par exemple, le problème de scindage d'un trinôme peut paraître complexe de prime abord et semble nécessiter d'un certain niveau en mathématique pour être résolu. Or, la méthode du discriminant réduit la complexité de résolution de ce problème au calcul d'un simple discriminant, calculable par une personne n'ayant pas nécessairement les qualités de réflexion mathématique évoquée précédemment.

Il en va de même pour résoudre une équation diophantienne avec l'algorithme d'Euclide étendu, etc. Plus largement, d'un point de vue pratique , l'objet de tout théorème mathématique est de faciliter ou d'alléger l'effort de réflexion de la personne qui réfléchit à la résolution d'un problème mathématique. En effet, au lieu de dériver logiquement toutes les assertions, celle-ci doit juste s'intéresser à la validation des hypothèses du théorème et si tel est le cas, d'appliquer directement le résultat, s'affranchissant ainsi de la considération de toutes les étapes nécessaires à la construction de celui-ci. La pensée algorithmique a montré que de nombreux problèmes qui nécessitaient, à priori, d'une grande quantité d'intelligence pour être résolus, pouvaient en fait l'être en suivant simplement une succession précise d'opérations élémentaires, élémentaires du fait qu'individuellement, elles nécessitent une bien plus faible quantité d'intelligence pour être effectuées, alors qu'une fois combinées, elles permettent d'obtenir le même résultat final. Autrement dit, les individus les moins initiés peuvent résoudre des problèmes qui, d'apparence, nécessitent une quantité d'intelligence importante. Par individu "simple", nous entendons en parti-

culier les machines. L'idée de l'algorithmique, c'est de combiner des opérations très simples du point de vue individuel, pour résoudre des problèmes difficiles. L'intelligence aura alors émergé d'une succession d'opérations élémentaires simpliste. Ce dernier point est très important. En effet, l'intelligence n'émane que de l'ensemble de tous les symboles descripteurs de l'algorithme, c.-à-d. les opérations élémentaires simplistes et les procédés d'intéractions entre ces symboles. Si nous prenons localement deux symboles et ce qui les relie, il y a, à priori, beaucoup de chance pour qu'aucune intelligence n'émane de cette structure locale. Par exemple, prenons l'exemple d'une colonie de fourmis : deux fourmis ensemble ne peuvent pas résoudre des problèmes complexes, mais une colonie de fourmis, elle, peut résoudre des problèmes complexes, ainsi c'est dans l'ensemble des symboles, de leurs interactions et de leur enchevêtrement que l'intelligence émerge, et seulement sous ces conditions. Notons que lorsque le nombre d'interactions à considérer dans le système est plus grand que ce que le cerveau humain peut gérer, alors, nous, humains, considérons que ce système est une boite noire. On fait notamment cela pour le cerveau humain lui-même et aujourd'hui, aussi pour les programmes informatiques intelligents. Ainsi, l'émergence de l'intelligence d'un système algorithmique est conditionnée par le choix adéquat des symboles du système et de leurs interactions. Comment ce choix doit-il se faire et dans quel but ?

Pour que l'on puisse avoir accès en pratique à un algorithme produisant de l'intelligence artificielle, il faut que les descripteurs utilisés dans la trame algorithmique soient sous forme symbolique. C'est l'objet de la partie suivante.

### La nécessité du raisonnement symbolique pour une émanation de l'intelligence physiquement réalisable

Nous présentons ici une définition du mot "intelligence" :

*"Ensemble des fonctions mentales ayant pour objet la connaissance conceptuelle et rationnelle."*[5]

Cette notion, bien que floue, se définit donc notamment par rapport à une connaissance rationnelle des choses qui nous entourent. Nous présentons ici une définition de la notion de rationalité :

5. D'après le dictionnaire Larousse

*"Qui parait logique, raisonnable, conforme au bon sens, qui raisonne avec justesse."*[6]

In fine, on associe donc l'intelligence à la logique. Et dans le cas de la résolution intelligente de problème scientifique, cela veut dire que, pour bien résoudre un problème, il suffit de suivre la logique, d'effectuer des raisonnements logico-déductifs construits sur des concepts scientifiques logiquement et parfaitement définis de telle sorte que tous les enchevêtrements d'assertions logiques entre ces concepts soient logiquement et clairement établis. Cependant, effectuer des raisonnements purement logiques, notamment dans le contexte des lois de la physique, est parfois impossible à réaliser en pratique, en tout cas, dans un temps fini. Il est alors nécessaire de s'affranchir des descriptions microscopiques (c.-à-d. logiques) des choses et de raisonner avec des concepts définis macroscopiquement (c.-à-d. de manière symbolique), et cela s'applique dans le contexte de la calculabilité. Avant d'aborder la nécessité d'avoir une description "symbolique" de manière à ce que les choses soient calculables, il est nécessaire d'aborder les idées du théorème d'incomplétude de Kurt Gödel (Figure 2.8). On peut définir les mathématiques comme étant un ensemble de raisonnements fondés sur une base axiomatique. Un axiome est un énoncé que l'on choisit vrai, que l'on ne démontre pas. On utilise ces axiomes et on les combine en respectant le principe logico-déductif pour former les théorèmes. Il existe plusieurs bases axiomatiques : par exemple, pour l'arithmétique des entiers naturels, il existe les axiomes de Peano, ou pour les mathématiques modernes, la théorie ZFC qui décrit tous les objets mathématiques sous forme d'ensemble. Il y a aussi les axiomes d'Euclide pour la géométrie dans l'espace ou le plan. Lorsque l'on démontre un théorème, on le démontre et on établit sa véracité par rapport à la base axiomatique que l'on a choisie. Ainsi pour résoudre un problème de mathématiques, il faut bien choisir sa base axiomatique en fonction du problème à résoudre. Ainsi la notion de démontrabilité, donc de ce qui peut être démontré, est relative à la base axiomatique. En effet, si nous enlevons ou rajoutons

FIGURE 2.8 – Kurt Gödel (1906-1978), Wikipedia

6. D'après le dictionnaire Larousse

des axiomes dans une base axiomatique, dans le premier cas, il y a peut-être des théorèmes que nous ne serions plus capables de démontrer (bien qu'au fond, ils sont bel et bien toujours vrais sauf qu'on ne peut plus le démontrer avec la base axiomatique choisie car elle a été dégradée) ; dans l'autre cas , nous pourrions être dans une situation où l'on peut démontrer un prédicat et son contraire en même temps. La première notion (c.-à-d. enlever des axiomes de la base axiomatique entraînant l'impossibilité de démontrer certains résultats) c'est la notion de complétude de la base axiomatique. C'est à dire que dans une base axiomatique complète, tout prédicat vrai est démontrable et tout que prédicat faux est réfutable. Ainsi cette base est capable de déterminer la véracité ou non de n'importe quels prédicats, d'où sa qualification de complète. La deuxième notion (c.-à-d. rajouter des axiomes dans la base entraînant la possibilité de pouvoir démontrer une chose et son contraire à la fois), c'est la notion de cohérence de la base axiomatique. C'est a dire qu'une base axiomatique cohérente ne permet pas de démontrer une chose et son contraire en même temps. Une question naturelle survient alors : existe-t-il une base axiomatique qui soit à la fois complète, permettant de démontrer tous les résultats de toutes les mathématiques et cohérente, permettant de ne pas démontrer une chose et son contraire en même temps, de telle sorte que toutes les mathématiques pourraient être écrites selon cette base axiomatique, donc une sorte de base axiomatique "canonique" et "universelle". David Hilbert, en 1930, énonce le programme de Hilbert. Il indique son souhait de décrire les mathématiques selon cette manière canonique, mais un an plus tard, Kurt Gödel démontre que cela est impossible dans [Gödel, ]. C'est le théorème d'incomplétude de Gödel. Dans notre démarche introductive, l'idée de la démonstration est de procéder par l'absurde :

Considérons une base axiomatique $S$ que l'on suppose complète et cohérente, on exprime un prédicat mathématique $P$ en fonction de $S$, que l'on note $P(S)$. L'astuce de Gödel réside dans le fait qu'on peut toujours formuler dans ce contexte , l'équivalence suivante pour tout $P$ :

$$P(S) \text{ est démontrable} \Leftrightarrow P(S) \text{ est fausse} \tag{2.1}$$

Alors :

$$P(S) \text{ est vrai} \Leftrightarrow P(S) \text{ est indémontrable} \tag{2.2}$$

d'après 2.1. Or $S$ est complet par hypothèse, donc :

$$P(S) \text{ est vrai} \Leftrightarrow P(S) \text{ est démontrable} \tag{2.3}$$

Par définition de la complétude.

On obtient alors une contradiction ! En effet, on vient de montrer que :

$$P(S) \text{ est démontrable} \Leftrightarrow P(S) \text{ est indémontrable} \tag{2.4}$$

C'est-à-dire que l'on vient de montrer une chose et son contraire. Cela contredit l'hypothèse sur la cohérence de $S$. Ainsi , nous devons rejeter l'hypothèse initiale et conclure que $S$ n'est pas une base axiomatique complète, qualifiée donc d'incomplète. Ainsi les bases axiomatiques cohérentes sont nécessairement incomplètes, et donc les prédicats exprimés selon ces bases sont toujours classés selon, non pas, 2 catégories (vrai ou faux dans la base axiomatique considérée), mais dans 3 catégories (vrai ou faux ou indémontrable dans la base axiomatique considérée).

Il existe un résultat analogue dans le domaine de l'informatique théorique concernant le fait que certaines choses sont incalculables par des machines de Turing. Le lien avec les machines de Turing est le suivant : si on suppose la "cohérence" du registre (par analogie de la cohérence de la base axiomatique permettant d'exprimer les mathématiques), qui définit l'ensemble des descripteurs traitable par une machine de Turing, alors le registre est nécessairement "incomplet". Autrement dit, en mathématiques nous avons la notion "d'indémontrabilité", en informatique nous avons celle de "l'incalculabilité".

C'est à dire, sachant une machine de Turing, il y a certains calculs exprimables au travers de d'un registre "cohérent" qui ne sont pas effectuables, au sens où ces calculs ne se terminent pas en un temps fini : ils n'aboutissent pas. On dit que ces calculs sont physiquement irréalisables. Cette notion "d'incalculabilité" s'illustre avec le problème de l'arrêt. C'est Turing et Church qui ont démontré ce résultat dans [Church, 1936b].

Le problème de l'arrêt est le suivant : sachant un algorithme $A$ et une entrée $X$, est-ce que l'appel de $A$ avec $X$, noté $A(X)$, ou son "calcul" va s'arrêter ? Autrement dit, pouvons-nous trouver un algorithme $P$ qui prend en entrée un algorithme $A$ et une entrée $X$ et dont la sortie de $P$ retourne vrai si $A(X)$ s'arrête ou faux si $A(X)$ ne s'arrête pas, pour tout couple $(A, X)$ ? Turing et Church ont montré qu'un tel

$P$ n'existe pas ou du moins n'est pas calculable par une machine de Turing en un temps fini. La démonstration se fait aussi par l'absurde :

Supposons qu'un tel algorithme $P$ existe et que le registre pour le décrire ait une propriété analogue à celle de la cohérence et à celle de la complétude. Un registre "complet" est alors un registre permettant d'exprimer tout les calculs de l'Univers. Et un registre "cohérent" est un registre ne permettant pas de réaliser physiquement des calcul parfois et d'autre fois, que ces mêmes calculs soient physiquement irréalisables.

Considérons maintenant l'algorithme $Q$ tel que pour tout algorithme $A$ :

**Algorithme 1 :** Algorithme $Q$

**Entrées :** $A$

```
si P(A(A(X))) alors
    boucle à l'infini
sinon
    retourne
```

Maintenant, appelons $Q(Q)$. Selon l'algorithme $Q$, la première étape est de déterminer $P(Q(Q(X)))$, alors :

— si $P(Q(Q(X)))$ est vrai : alors d'une part $Q(Q)$ boucle à l'infini, car $P(Q(Q(X)))$ est vrai par hypothèse et par construction de $Q$ et d'autre part si $P(Q(Q(X)))$ est vrai alors cela veut dire que $Q(Q(X))$ s'arrête, par construction de $P$. Autrement dit on vient de montrer que :

$$Q(Q(X)) \text{ s'arrête} \Leftrightarrow Q(Q(X)) \text{ boucle à l'infini} \tag{2.5}$$

Autrement dit, nous venons de montrer que l'algorithme $Q$ s'arrête et que, en même temps, il boucle à l'infini, soit qu'il ne s'arrête pas. Autrement dit, nous avons illustré un cas où un calcul est physiquement réalisable mais aussi physiquement irréalisable.

— De manière parfaitement analogue, on peut aussi faire le cas où $P(Q(Q(X)))$ est faux : si $P(Q(Q(X)))$ est faux, alors $Q(Q(X))$ s'arrête par construction de l'algorithme $Q$ et d'une autre part si $P(Q(Q(X)))$ est faux alors $Q(Q(X))$ ne s'arrête pas. Autrement dit on vient de montrer là aussi que :

$$Q(Q(X)) \text{ s'arrête} \Leftrightarrow Q(Q(X)) \text{ boucle à l'infini} \tag{2.6}$$

Ce qui est de la même manière absurde par hypothèse de cohérence.

Donc il faut rejeter l'hypothèse sur l'existence de l'algorithme $P$. Ainsi, nous avons exhibé un calcul qui n'est pas effectuable par une machine de Turing. Ainsi les registres des machines de Turing n'ont pas cette propriété de "complétude". Elles ont cependant la propriété de "cohérence", du moins d'un point de vue physique, en effet, un calcul est soit réalisable ou non, de manière exclusif.

Dans cet exemple, nous avons illustré le lien informel entre Godel et Turing. Par abus de langage, nous venons donc de montrer informellement que si les machines de Turing ont des registres "cohérents", ce qui est le cas physiquement, alors ce même registre est nécessairement "incomplet" au sens de Gödel. Pour faire le lien avec les propos qui suivent, notamment sur les liens avec l'intelligence artificielle, nous allons donner quelques intuitions sur cette notion de d'incomplétude des machines de Turing, au sens de Gödel.

On peut faire le constat suivant : les machines de Turing semblent avoir quelques limitations, en effet, dans certains cas, elles sont incapables de faire aboutir certains calculs. Mais peut-être que la théorie des machines de Turing n'est pas assez sophistiquée, et que l'on peut trouver d'autres manières de rendre des calculs physiquement irréalisables, réalisables ? Il semblerait que malgré tout, la réponse soit non, et donc, qu'il ne s'agirait pas d'une simple limite théorique. En effet, dans certains cas, atteindre un temps infini, c'est la solution la plus rapide pour faire aboutir un calcul. Stephen Wolfram a formalisé cette idée avec la notion d'irréductibilité algorithmique. C'est l'idée qu'il peut exister des méthodes plus adéquates que d'autres algorithmiquement pour résoudre un problème, mais que pour certains de ces problèmes, les solutions les plus rapides sont celles qui retournent le calcul en un temps infini. Autrement dit, Wolfram formule ici l'idée que certains algorithmes sont inévitablement très longs pour aboutir et qu'il n'existe pas de solution algorithmique plus rapide. Nous présentons ici un exemple pratique de l'irréductibilité algorithmique. Supposons que nous voulions effectuer des prédictions concernant le comportement d'un phénomène physique d'ici à un an et que nous décrivions ce système à la résolution temporelle la plus fine théoriquement (c.-à-d. une description la plus microscopique, théoriquement, du problème) possible. Supposons également que la simulation entière du phénomène ne nécessite qu'un seul et unique calcul, ce qui est en soi une supposition déjà très forte. La puissance de calcul de nos ordinateurs d'aujourd'hui

est de l'ordre de $\mathcal{O}(10^9)$ calcul par seconde. La résolution temporelle physique la plus fine est définie par le temps de Plank. Le temps de Plank est une unité de temps : c'est le temps qu'il faudrait à un photon dans le vide pour parcourir une distance égale à la longueur de Planck. Comme celle-ci est la plus petite longueur mesurable théoriquement, et que la vitesse de la lumière dans le vide est la plus grande vitesse possible théoriquement, le temps de Planck est la plus petite mesure temporelle théorique ayant une signification physique dans le cadre des théories actuellement admises. La valeur du temps de Plank est de l'ordre de $10^{-43}$ seconde. Ainsi, pour nos ordinateurs actuels et d'après nos hypothèses, simuler ce phénomène à l'échelle du temps de Plank, c'est-à-dire le décrire à chaque intervalle de temps de $10^{-43}$ seconde, nécessite environ $\frac{10^7}{10^{-43}} = 10^{50}$ étapes de calcul. C'est à dire, à la vitesse de calcul de nos ordinateurs : $10^{41}$ secondes, soit $10^{34}$ années, soit $10^{23}$ fois l'âge de notre Univers. Ainsi, le moyen le plus rapide pour connaître le comportement d'un phénomène physique d'ici à un an, est simplement d'attendre un an et d'observer ledit phénomène. Cet exemple est une illustration pratique du constat suivant : Si nous voulons prédire (c-à-d. calculer), de manière physiquement réalisable le comportement d'un phénomène physique à l'échelle microscopique alors le calcul sera nécessairement long, voir infiniment long. Autrement dit, pour pouvoir prédire algorithmiquement le comportement du phénomène en un temps raisonnable, il faut nécessairement adopter une description du phénomène plus macroscopique (par exemple une résolution temporelle bien moins fine) pour le prédire, ou atteindre sa réalisation effective. Mais dans ce cas, nous ne sommes plus dans de la prédiction.

Ce constat vaut aussi, et en particulier, pour nous humains. En effet, si la thèse de Church-Turing-Deutsch est vraie alors nos cerveaux sont simulables par une machine de Turing, autrement dit nos cerveaux peuvent être assimilés à des machines de Turing. Or le traitement de l'information par nos cerveaux est physiquement réalisable, ou "calculable", car tous nos raisonnements sont faits en un temps fini, et si nous supposons que notre base réflexive possède une propriété analogue à celle de la cohérence, ce qui à priori est le cas, alors notre base réflexive est nécessairement incomplète au sens de Gödel, c'est-à-dire qu'il y a des pensées que nous ne pouvons pas effectuer sous peine de devoir attendre un temps infini pour que le processus de réflexion s'achève. Cela implique donc que les pensées que nous

construisons sont nécessairement d'une base réflexive constituée de représentation symbolique, plus légère à traiter. C'est-à-dire que l'ensemble de nos pensées, de nos raisonnements, ne peuvent pas être générés par un ensemble de briques élémentaires de trop bas niveau (analogie avec la description strictement formelle, logique, des choses, comme l'exemple précédent décrit à l'échelle du temps de Plank). Nos pensées ne peuvent être générées qu'à partir d'un ensemble de symboles de plus haut niveau. Par exemple, est-il possible de définir formellement et logiquement ce qu'est un chien ? Non, nous ne définissons le chien qu'à partir de concepts symboliques dits de haut niveau, comme par exemple le nombre de pattes, les caractéristiques de sa tête, de son poil, et ainsi de suite [7]. Ces concepts, bien que plus spécifiques que celui du "chien" sont eux-mêmes des concepts de haut niveau que l'on peut décomposer en sous concepts. Nous pouvons aller par exemple jusqu'au niveau de l'électron de chaque atome qui compose chaque molécule du chien. Mais la notion d'électron est elle-même encore une notion relativement de haut niveau. En effet des physiciens s'intéressant à la physique des particules, passent encore des années à caractériser cette entité physique, de diverses façons. Ils ne sont pas encore tous d'accord sur la manière de caractériser formellement ce qu'est "l'électron". Ils sont à peu près tous d'accord sur la définition de celui-ci, mais celle-ci n'est pas exprimée au sens purement logique. Et probablement que selon les différentes communautés des physiciens, cette définition varie légèrement. Mais déjà, une définition du chien à la granularité de l'électron, bien que encore de "haut niveau" d'un point de vue strictement formel, n'est pas possible à se représenter dans notre cerveau d'humain. En effet, un mammifère comme le chien est constitué d'environ $10^{27}$ atomes. De plus nous savons que notre cerveau est composé d'environ $10^{10}$ neurones. Si nous supposons que un neurone est capable de stocker l'information relative à un atome, ce qui est déjà beaucoup, alors il faudrait environ $10^{17}$ personnes pour se représenter atomiquement et de manière complète, ce qu'est un unique chien. Or nous sommes environ, que $\mathcal{O}(10^9)$ humains. Ainsi, la représentation atomique du chien, bien qu'elle soit toujours de haut niveau, n'est pas représentable humainement, et à fortiori tout traitement de cette quantité d'informations n'est pas réalisable humai-

7. Bien que certains chercheurs ont peut être réussi à "formaliser" statistiquement ce qu'est un chien, voir [Le et al., 2011]

nement à cette granularité. Et nous ne sommes même pas là, à une granularité de type formelle.

Ainsi, pour rendre les choses physiquement traitables il est nécessaire de s'affranchir des descriptions logiques et formelles des choses et se tourner vers des descriptions plus haut niveau, dites symboliques. Cela implique nécessairement, que pour nos cerveaux, il existe un processus qui conçoit ces représentations symboliques. Mais quelles sont les entrées de ce processus , comment fonctionne t-il ? Est-il variable selon les individus ? A priori c'est le cas, puisque nos pensées et nos représentations , issues d'une base réflexive qui, pour un concept donné, sont différentes mais probablement similaires. Nous avons donc en nous, de manière inconsciente, un choix à faire concernant la détermination et la construction de ce processus. En effet, nous sommes à peu près tous d'accord pour dire qu'un chien est un chien. Cela vient du fait que, bien que nos bases réflexives soient différentes, elles ne sont pas trop "éloignées" entre elles. Elles ont alors probablement été construites selon le même procédé, ou du moins selon des procédés similaires entre eux. Nous allons y revenir plus largement dans le prochain paragraphe concernant le phénomène de généralisation. Nous avons construit nos représentations symboliques du chien en observant ce qui était désigné (ou étiquetté) par autrui comme étant un chien et nous avons construit, à force de rencontrer ces entités ou bien "d'apprendre" leurs représentations, par créer une loi, une définition, décrite de manière symbolique, de ce qu'est le "chien". Et ce raisonnement est valable pour la majorité des entités physiques qui composent notre univers. Ainsi, nous avons vu que l'approche algorithmique permettait de faire émaner de l'intelligence d'une machine à calculer, et que pour que cette émanation soit physiquement réalisable, il fallait que les composants des algorithmes soient décrits sous forme symbolique. Mais alors, comment pouvons nous créer ces symboles ? Y'en a t'il certains plus adéquat que d'autres ? Autrement dit, comment pouvons-nous créer un programme informatique créant de l'intelligence artificiellement ?

### 2.2.3 L'apprentissage des machines

Historiquement, l'intelligence artificielle était un domaine de recherche qui s'intéresse à la résolution de problèmes par des machines. On pensait que pour bien

résoudre un problème, il fallait écrire le bon code, c'est-à-dire la bonne suite d'instructions à suivre pour les machines. D'une certaine manière, on avait une approche assez déontologique pour les machines. On leur disait précisément ce qu'elles devaient faire et ce qu'elles devaient ne pas faire. Aujourd'hui on préfère l'apprentissage par l'exemple : on donne aux machines un volume conséquent de données et on fait confiance aux machines pour apprendre toutes seules ce qu'elles doivent faire et ce qu'elles ne doivent pas faire. On parle parfois d'approche "guider par la donnée" ou encore d'apprentissage machine ou plus communément de *machine learning*. Les succès contemporains de l'apprentissage machine sont impressionnants et spectaculaires, à tel point que les les programmes d'intelligence artificielle d'aujourd'hui sont tous des programmes issus de l'apprentissage machine. Nous sommes forcés de constater empiriquement que aujourd'hui, l'apprentissage machine fonctionne en pratique, mais est-ce aussi le cas en théorie ?

En 1950, Turing a émis des postulats et insinue que oui, l'apprentissage machine ça marche en théorie et même que c'est nécessaire pour créer un programme d'intelligence artificielle. Nous présentons ici les grandes idées de son raisonnement qu'il a détaillées dans son article [TURING, 1950]. Pour Turing, une machine intelligente, c'est une machine qui passe le test de Turing. Ce test consiste à confronter verbalement et aveuglément un humain et une machine de Turing et un autre humain. Si la personne qui engage les conversations avec la machine et si l'autre humain n'est pas capable de dire lequel de ses interlocuteurs est une machine, alors on peut dire que la machine a réussi le test de Turing. Or nous, humains, nous réussissons à passer le test de Turing. De plus et selon Turing, le "code source" de notre cerveau représente environ $10^9$ bits d'informations. Or, toujours selon Turing, le cerveau humain produit environ $10^3$ bits d'informations par jour. Donc il faut environ $10^6$ jours pour produire les $10^9$ bits de notre cerveau. Or, $10^6$ jours représentent environ 2800 ans. Donc un seul humain n'a pas assez de sa vie pour reproduire le "code source" de notre cerveau. Mais alors, si plusieurs ingénieurs en parallèle tentent de produire ce code, disons 50 ingénieurs, alors il faudra un peu plus de 50 ans à cette équipe pour produire le "code source" de notre cerveau. Mais la réalisation de cette tâche va durer 50 ans si et seulement s'il n'y a aucune erreur dans la transcription du code source, si l'équipe est rigoureusement organisée et avance en groupe, si elle

a bien cadré et défini le "code source" en question. Or, de par la complexité des relations humaines, maintenir ce cadre de travail, chaque jour des 50 années, sans relâche, ne semble pas réalisable en pratique. Ainsi, Turing indique qu'il faut une méthode plus rapide pour réaliser cette tâche de création du "code source". Turing part du principe que le cerveau humain se complexifie avec le temps. Turing indique que le cerveau humain enfant est beaucoup plus facile à programmer que le cerveau humain adulte, et il indique qu'il est peut être moins difficile (du moins, suffisamment court pour pouvoir être écrit par une poignée d'ingénieurs en quelques années seulement) d'écrire un "code source" avec les mêmes facultés d'apprentissage que le cerveau d'un enfant et qui a la capacité de s'autocomplexifier avec le temps avec l'apprentissage. Il dit donc que, ce qui permettra d'atteindre l'intelligence artificielle par un programme, c'est d'écrire un programme capable de simuler le cerveau d'un enfant étant capable d'apprendre. Il faudra ensuite lui fournir le bon enseignement pour que, comme le cerveau d'un enfant, ce programme s'autocomplexifie et qu'il atteigne un jour l'intelligence artificielle. C'est alors la naissance du machine learning, ou en français, l'apprentissage machines. Il indique donc que l'apprentissage machine est une condition au moins nécessaire pour obtenir une intelligence artificielle. Ainsi, dans cet article, Turing postule que l'apprentissage machine est le point pivot pour créer des programmes d'intelligence artificielle. Cependant, comment fournir à une machine le "bon enseignement" nécessaire pour atteindre cette intelligence artificielle ?

## 2.3 Protocole d'apprentissage des machines

### 2.3.1 L'inéluctabilité physique de l'échantillon

Comme Turing l'a indiqué, la création de programmes d'intelligence artificielle doit passer par l'apprentissage machine. Mais comment fait-on apprendre intelligemment à une machine ? Nous humains, nous nous confrontons à des exemples pour apprendre. Ainsi, "bien apprendre" ne signifie t'il pas tout simplement de mémoriser tout les exemples nécessaire et d'ensuite restituer pour réussir la tâche d'apprentissage ? Malheureusement, cela est, dans beaucoup de cas, physiquement impossible. En effet, supposons que nous voulions apprendre à un programme comment distin-

guer des images de chien et de chat, alors il suffit de réunir toutes les images de chien et de chat et de construire la fonction qui associe les images de chien au mot "chien" et de même pour les chats. Or, nous n'avons physiquement accès qu'à une infime partie des images de chien et de chat possibles et réalisables. Mais surtout, nous, humains du 21ème siècle, n'avons rencontré qu'une petite partie des chiens et chats qui existent sur terre, et nous n'avons pas encore rencontré ceux qui naîtront demain. Donc à fortiori, nous n'avons pas la possibilité physique d'avoir accès à la représentation numérique des chiens et chats qui nous sont contemporains. Nous n'avons nécessairement qu'une infime partie de l'ensemble des images de chien et chat à notre disposition. Ainsi, la notion symbolique du "chien" ou du "chat" n'est constructible qu'à partir d'un échantillon de taille très petite devant la taille de l'ensemble de toutes les images de chien et de chat au sens absolu. Ainsi, en ce sens, nos représentations de chien ou de chat sont très probablement peu précises, et ne permettent donc pas de construire un "bon" enseignement de ce qu'est le chien ou le chat. Cependant, tous les chiens se ressemblent, ils partagent tous certaines caractéristiques discriminantes.

### 2.3.2 La similitude des éléments de l'échantillon

Ainsi il n'est peut être pas nécessaire d'avoir l'ensemble, au sens absolu, des images de chien et chat pour réaliser la tâche d'apprentissage. En effet, si nous avons quatre images de chien semblable et que nous étiquetons qu'elles sont toutes des "chiens", que nous prenons une cinquième image de chien, semblable aux quatre autres images précédentes, alors nous pouvons légitimement dire ou prédire que c'est un "chien" aussi pour cette cinquième image. Le principe que nous venons d'introduire s'appelle le principe d'interpolation. Autrement dit, si nous notons deux images de chien par $x$ et $y$, et que nous notons par $f$ notre programme déterminant s'il s'agit de chien ou non, le principe d'interpolation est le suivant : si $x$ est "proche" de $y$, alors $f(x) = f(y)$.

C'est le principe de l'algorithme du plus proche voisin : on mémorise des exemples étiquetés et par l'algorithme, on prédit la valeur de l'élément le plus similaire à ceux que l'on a mémorisés.

Cependant, cette approche pose deux problèmes majeurs :

— Si on a $\mathcal{O}(N)$ exemples, alors l'algorithme du plus proche voisin nécessite $\mathcal{O}(N)$ traitements. Donc le volume de traitement est linéaire avec le nombre d'exemples à mémoriser, en particulier, le volume de traitement augmente si la taille de l'ensemble des exemples augmente, ce qui est problématique lorsque $N$ est grand.

— Si les données mémorisées sont exprimées dans un espace de petite dimension et que nous avons "assez" de données, alors l'algorithme est efficace. Mais si la dimension des données est trop grande devant le nombre de données alors, l'algorithme échouera. Illustrons ces propos :

### 2.3.3 Le fléau de la dimension

Illustrons pourquoi la dimension, au sens vectoriel, des données joue un rôle central dans le succès de l'algorithme du plus proche voisin : pour que l'algorithme soit efficace, il faut que la densité des données dans leur l'espace soit haute, pour avoir une bonne estimation de la "similarité". Or, la densité décroît exponentiellement avec la dimension de l'espace des données. Supposons que nous avons un espace des données discret de dimension 1 avec $p \in \mathbb{N}$ données parmi $N \in \mathbb{N}$ possibles. La densité de données moyenne dans l'espace est donc de $\frac{p}{N}$. Si maintenant, notre espace est de dimension 2, alors il y a $N^2$ points possibles pour nos données, et si nous n'avons toujours que $p$ données, alors nous obtenons une densité de $\frac{p}{N^2}$. Ainsi, pour conserver une densité de $\frac{p}{N}$, il faut donc rajouter $N$ fois $p$ données dans notre ensemble, car si nous avons $N \times p$ données initialement et que nous conservons la densité de nos données, nous avons $\frac{N \times p}{N^2} = \frac{p}{N}$. De manière générale, si notre ensemble de données est de dimension $d \in \mathbb{N}^\star$, il faut alors augmenter de $N^{d-1}$ fois la taille d'ensemble de données pour conserver notre densité initiale, car $\frac{N^{d-1} \times p}{N^d} = \frac{p}{N}$. De plus, pour que la densité soit haute, il faut que le nombre de données auquel nous savons l'accès en pratique soit proche du nombre de données théoriques nécessaires. C'est-à-dire dans le cas où notre espace est de dimension $d$, il faut un volume de données proche de $N^d$.

Illustrons ce phénomène dans notre exemple de classification d'images de chien et chat : supposons que nos images sont des images à trois canaux de couleurs rouge, vert et bleu, échantillonnées sur un octet chacun et que la dimension $d$ de

notre espace (celui des images) est de taille $1000 \times 1000 = 10^6$. Alors le nombre de données à récolter pour que l'algorithme du plus proche voisin soit efficace est : $(256^3)^{10^6} \approx 10^{7000000}$ au sens absolue [8], soit bien plus grand que le nombre d'atomes dans l'univers (environ $10^{80}$). Il est donc physiquement impossible de réunir toutes les images nécessaires pour que l'algorithme du plus proche voisin soit efficace. De toute façon, il est encore moins physiquement possible de traiter ce volume de données en un temps raisonnable. Du fait que nous n'ayons accès qu'à un nombre limité de données, ne serait-ce qu'un million, elles seront très raréfiées en terme de présence spatiale dans l'espace des données, empêchant le succès de l'algorithme du plus proche voisin. Ce phénomène s'appelle le fléau de la dimension.

Ainsi, l'algorithme du plus proche voisin semblait être intéressant pour résoudre le problème de l'apprentissage machine, mais nous avons vu qu'il subit le fléau de la dimension. Existe-t-il des algorithmes basés sur la même idée d'interpolation, mais qui n'est pas autant affectés lorsque la dimension du problème augmente ? Oui, il s'agit de méthode de régression linéaire.

### 2.3.4 Les modèles linéaires et leurs limites

Dans la précédente partie, nous avons introduit ce qu'est l'apprentissage dit supervisé [9]. L'ensemble de nos exemples ou de nos données est constitué d'observations (par exemple les images de chiens et chats) et d'étiquettes (par exemple "chien" et "chat"). Ainsi, un modèle apprenant de manière supervisée construit une fonction de prédiction $f$ qui à chaque observation $\mathbf{X}$ lui associe son étiquette $y$, comme exemple nous avons vu l'algorithme du plus proche voisin qui permet de construire une telle fonction. Dans la suite, nous noterons par $D$ l'ensemble de nos données à $d$ dimensions :

$$D = \{(\mathbf{X}_1, y_1), (\mathbf{X}_2, y_2), \cdots, (\mathbf{X}_N, y_N)\} \tag{2.7}$$

8. En toute rigueur, cette quantité représente le nombre d'images possible encodé sur 1 octet de taille $1000 \times 1000$, et non la quantité de chien et de chat qui ne représente qu'une partie de ce volume totale. Mais, quitte à ne prendre qu'une partie du volume total, ne serait-ce que 0,001%, le volume (environ $10^{6999995}$ données) est toujours bien plus grand que le volume physiquement traitable

9. Il existe aussi l'apprentissage non-supervisé (voir [Tibshirani and Friedman, ]) et par renforcement (voir [Sutton and Barto, 1998]), nous ne l'évoquerons pas dans ce document

Tel que le nombre d'éléments de $D$, $|D| = N \in \mathbb{N}$ et où, sans perdre de généralités, $\mathbf{X}_i \in \mathbb{R}^d$ et $y_i \in \mathbb{R}$ pour tout $i \in \{1, \cdots, N\}$. $D$ est un échantillon i.i.d issue d'une variable aléatoire suivant une loi de probabilité, nommée la "vraie loi" sous-jacente [10]. La régression linéaire est un des premiers algorithmes d'apprentissage machine, conçue par Roger Joseph Boskovic en 1755. Nous présentons ici une illustration de la méthode dans le cas où $d = 1$, où l'on s'intéresse à la montée du volume d'eau en période de crue d'une rivière selon le temps :

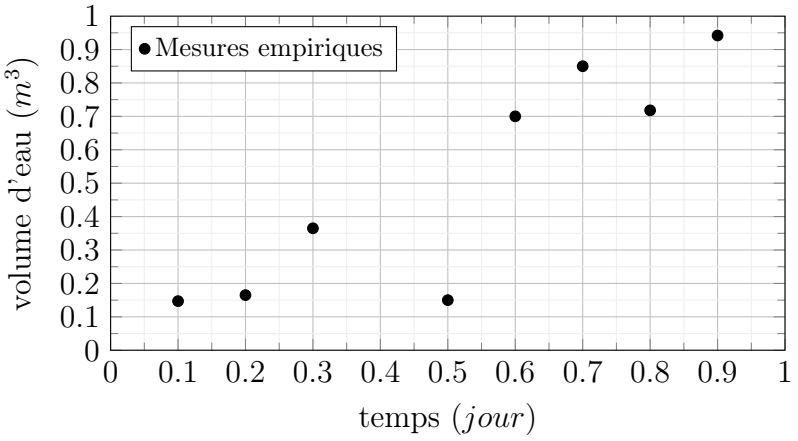



FIGURE 2.9 – Distribution des données

Mais un problème apparait alors : il est probable que les mesures effectuées tant par rapport à la mesure du niveau d'eau ou de celui de la mesure du temps soient légèrement erronées pour de multiples et variables raisons [11], par exemple, un capteur de mesure mal calibré. Boskovic avait fait le même constat dans ses expériences. Pour lui aussi, les mesures des données peuvent être plus ou moins erronées. Cependant, il remarqua que grossièrement, comme c'est le cas ici, ces données pouvaient être reliées par une droite et que l'écart vertical entre les mesures des données et cette droite représentait les erreurs de mesure, c'est-à-dire que pour tout $i$, l'erreur est $\epsilon_i = |aX_i + b - y_i|$ pour un certain $a, b \in \mathbb{R}$. L'idée est alors de trouver les $a$ et $b$ qui minimisent en moyenne ces erreurs. Et c'est Carl Friedrich Gauss (Figure 2.13) et Adrien-Marie Legendre (Figure 2.10), qui formalisèrent cette

10. Ces suppositions de structure i.i.d et l'existence même de la vraie loi de probabilité, sont faites ici mais peuvent être largement discutées.

11. On a tendance souvent à considérer les erreurs de mesure sur les données d'entraînement, mais les erreurs peuvent aussi provenir des données étiquettes pour la vérité terrain, voir [Becker and Hinton, 1992, Wang, 2001]

idée au travers de la méthode de moindre carré, notamment grâce au théorème central limite de Pierre Simon Laplace (Figure 2.12) et à la méthode du maximum de vraisemblance. Nous présentons ici les idées principales de leurs raisonnements.


FIGURE 2.10 – Adrien-Marie Legendre (1752 - 1833), Wikipedia

Laplace part du principe que les erreurs (localement) $\epsilon_i$ que nous observons en pratique sont dues à de petites erreurs indépendantes entre elles $\epsilon_{i,j}$, tel que $\epsilon_i = \sum_j \epsilon_{i,j}$ pour tout $i$. Par exemple, dans notre cas d'usage, un mauvais calibrage du capteur de mesure et une erreur d'arrondis de la personne qui a relevé la mesure, tel que une fois cumulées, produisent l'erreur $\epsilon_i$ que nous observons. Et Laplace, dans son théorème central limite, énonce que, la famille des $(\epsilon_i)_i$ suit une loi normale centrée réduite, c'est-à-dire que la probabilité d'avoir une erreur $\epsilon_i$ vaut :

$$\mathbb{P}(\mathcal{E}_i = \epsilon_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\epsilon_i^2} \tag{2.8}$$

Donc par hypothèse, d'indépendance et d'identique distribution des erreurs, on a :

$$\mathbb{P}(\mathcal{E}_1 = \epsilon_1, \cdots, \mathcal{E}_N = \epsilon_N) = \sqrt{2\pi}^{-N} e^{-\frac{1}{2}\sum_i^N \epsilon_i^2} \tag{2.9}$$

Ainsi, on souhaite maximiser la vraisemblance des paramètres, c'est-à-dire de trouver les paramètres $a$ et $b$ de la régression linéaire qui rendent les plus probables les erreurs observées sachant les paramètres courants. Autrement dit, on cherche à maximiser la vraisemblance des erreurs sachant les paramètres courants, c'est le principe du maximum de vraisemblance.

De manière plus formelle, on souhaite donc résoudre le problème suivant :

$$a^\star, b^\star = \arg\max_{a,b} \sqrt{2\pi}^{-N} e^{-\frac{1}{2}\sum_i^N \epsilon_i^2} \tag{2.10}$$

Nous décrivons dans l'Annexe C.2 les détails des calculs.

Nous présentons ici l'illustration de la méthode des moindres carrés dans notre cas d'usage :

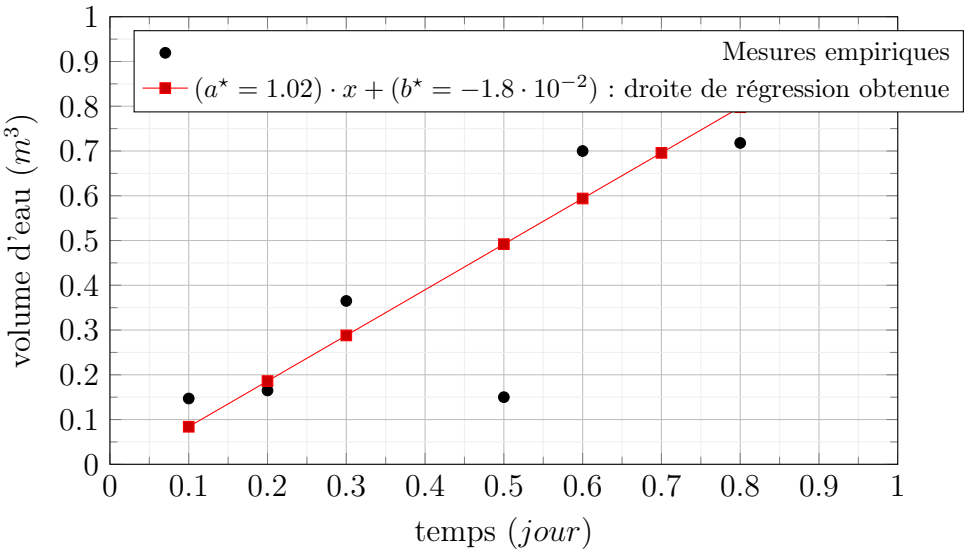



FIGURE 2.11 – Distribution des données

Donc, par la méthode des moindres carrés, on arrive à trouver par un calcul simple et physiquement réalisable, une régularité décrite par la droite de régression dans l'ensemble de données $D$. Cela permet donc par la suite de se servir du modèle construit pour prédire les étiquettes de données non encore observées. Mais cette confiance en la régularité construite doit être relative, nous y reviendrons plus largement. Ainsi la régression linéaire permet de résumer toute la complexité d'un ensemble de données , en une simple droite. Cette méthode a donc plusieurs propriétés intéressantes :

FIGURE 2.12 – Pierre-Simon Laplace (1749 - 1827), Wikipedia

— La régression linéaire est une matérialisation physiquement réalisable du processus de construction d'une représentation symbolique des données : en effet, par un algorithme calculable en temps raisonnable, la régression linéaire permet de résumer toute la complexité d'un jeu de données à une simple droite, soit deux réels ici.

— De manière générale, la régression linéaire est généralisable et conserve son optimalité en plus grandes dimensions : non pas comme l'algorithme du plus proche voisin, la méthode des moindres carrés se généralise

très bien lorsque la dimension de l'ensemble de données augmente. Ainsi, en dimension $d$, la régression consiste à trouver un hyperplan ajustant les données. Ce dernier est alors décrit par $d+1$ nombres. De plus, l'algorithme permettant de trouver les $d+1$ nombres optimaux sont calculable en un temps polynomial selon le $d$ et $N$, autrement dit , ces nombres sont calculables en un temps raisonnable.

— La régression linéaire peut aussi être adaptée à des problèmes de classification , on parle alors de classification linéaire : la classification linéaire repose sur le même principe que la régression linéaire, mais l'idée n'est plus de trouver l'hyperplan qui ajuste le mieux le jeu de données, mais plutôt celui qui sépare le mieux les données (selon les classes du problème) une fois transformé par une transformation non linéaire. On parle alors d'hyperplan séparateur. Les modèles de la classification linéaire partagent les mêmes bonnes et mauvaises propriétés que ceux de la régression linéaire.

Cependant, les modèles linéaires ont aussi plusieurs limites [12]. La plus évidente est que les modèles linéaires ne sont capables d'être optimaux que pour des phénomènes qui sont effectivement linéaires. La majeure partie des phénomènes physiques que nous rencontrons dans notre Univers sont majoritairement non linéaires. Comment pouvons-nous alors conserver les bonnes propriétés des modèles linéaires tout en pouvant modéliser des phénomènes non linéaires ? C'est l'astuce du noyau.

12. La méthode des moindres carrés a aussi des défauts en comparaison à la méthode des moindres déviations. En effet, la stricte convexité de la fonction carré entraîne que le poids de grande erreur, en valeur absolue, est plus important que le poids de petite erreur, ce qui peut mener le modèle donner plus d'importance a des mesures abbérantes par exemple, menant à un modèle erroné. Ce problème n'apparaît pas avec la méthode des moindres déviations car il est formulé par rapport à la norme $l_1$. Cependant, la méthode des moindre carrés a de meilleures propriétés d'un point de vue de l'optimisation, du fait de la plus grande régularité de la fonction carré par rapport à la fonction valeur absolue.

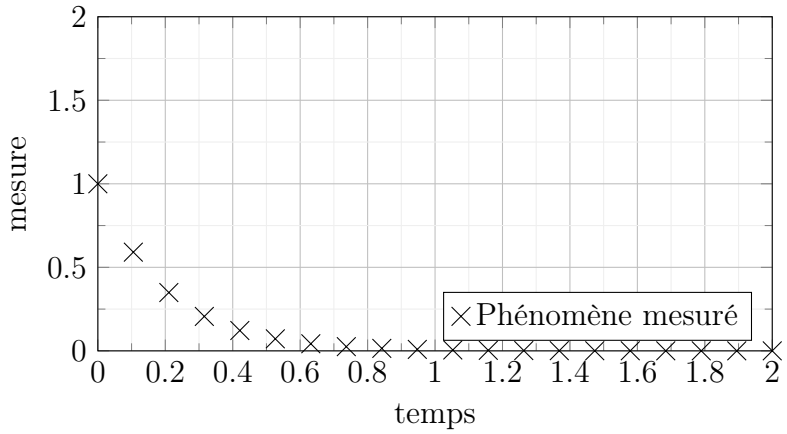



FIGURE 2.14

### 2.3.5 Extension du domaine applicatif des modèles linéaires : l'astuce du noyau

Comme évoqué, une des limites des modèles linéaires est de pouvoir modéliser, de manière optimale, que des phénomènes linéaires. Mais il existe tout de même un moyen d'utiliser les modèles linéaires et leurs bonnes propriétés dans le cas de la modélisation de phénomène non linéaire : c'est l'astuce du noyau. L'idée est d'appliquer une transformation non linéaire au jeu de données $D$, bien choisie, de manière à ce que dans ce nouvel espace, les données soient linéairement, ajustable ou séparable. Illustrons cela : considérons un problème de régression pour un phénomène physique non linéaire. Disons que les observations $y_i$ sont régies temporellement par le modèle suivant :

FIGURE 2.13 – Carl Friedrich Gauss (1777 - 1855), Wikipedia

$$y_i = e^{-5t_i} \tag{2.11}$$

Alors pour $N = 10$ relevé de mesures, nous avons, par exemple les observations suivantes, illustrées dans la Figure 2.14.

Dans la représentation actuelle du problème, la régression linéaire ne sera pas efficace, car le nuage de points ne décrit pas la relation linéaire entre les mesures et le temps. Cependant, si nous appliquons une transformation logarithmique aux mesures qui est bien une transformation non linéaire, alors nous obtenons un nouvel

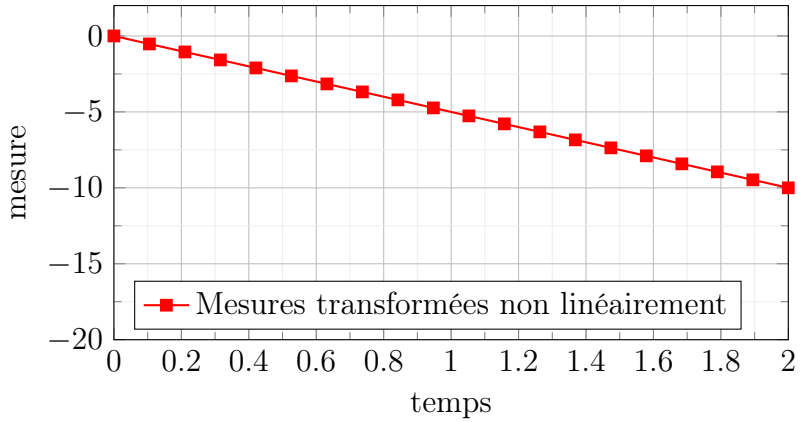



FIGURE 2.15 – Donnée projetée dans un espace linéairement ajustable

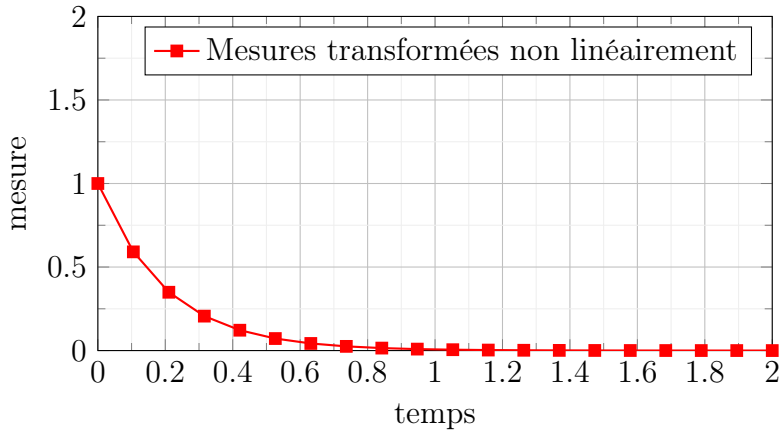



FIGURE 2.16 – Donnée projetée dans un espace linéairement ajustable

ensemble de données $D'$ tels que :

$$\ln(y_i) = -5t_i \tag{2.12}$$

pour tout $i \in \{1, \cdots, N\}$. Désormais, les données sont exprimées selon une relation linéaire avec leurs variables explicatives (c.-à-d. le temps), et nous pouvons, dans ce nouvel espace, effectuer une régression linéaire efficace. Et on observe dans la figure2.15 que ces données transformées sont linéairement ajustables par la droite d'équation $y = -5t$. Ainsi il suffit d'appliquer la transformation inverse pour obtenir la régression non linéaire initiale voulue, à l'aide de la régression linéaire, comme montré dans la figure2.16.

Ici les exemples sont triviaux, car nous connaissons à priori le modèle de construction des données, et donc nous pouvons plus facilement trouver la transformation non linéaire adéquate. Cette astuce du noyau fonctionne aussi pour les problèmes de

classification non linéaire, un exemple connu est celui du "ou exclusif". De plus, l'astuce du noyau permet de conserver les propriétés de représentativité symbolique de la régression linéaire, comme l'indique le théorème du représentant [Schölkopf et al., 2001]. Ce théorème indique que pour des données satisfaisant des hypothèses souvent rencontrées en pratique, il existe toujours un hyperplan optimal qui est décrit par un nombre de coefficients égal au nombre de données de l'échantillon considéré. Or comme nous avons un nombre fini d'éléments dans notre ensemble de données, alors notre hyperplan ajusteur ou séparateur est descriptible par un nombre fini de nombres réels, et donc l'ensemble de ces nombres sont calculables. Bien que le choix de la transformation non linéaire ne soit pas évident pour un problème quelconque, l'astuce du noyau permet cependant d'utiliser les bonnes propriétés de la régression linéaire pour un spectre de problèmes bien plus large. Cependant, il y a un dernier défaut majeur que nous n'avons pas évoqué concernant la régression linéaire. Ce problème est un problème statistique qui ne survient pas que pour les méthodes linéaires : il s'agit du compromis biais-variance.

### 2.3.6 La théorie ou la pratique : le compromis biais-variance

Comme nous l'avons vu, l'apprentissage a nécessairement, du point de vue physique, une structure statistique particulière. En effet, nous ne pouvons considérer qu'une infime partie des exemples nécessaires au bon apprentissage. Lorsque nous souhaitons modéliser un phénomène physique, nous n'avons alors nécessairement qu'une vue partielle de ce dernier, vue de laquelle nous devons "bien apprendre". Reprenons notre exemple avec la mesure du niveau d'eau d'une rivière en période de crue. Une mesure légitime pour savoir si nous avons bien modélisé notre phénomène est à priori la valeur de notre erreur moyenne entre les prédictions de notre modèle et la vérité terrain qui constitue notre échantillon. Supposons que cette erreur moyenne est proche de zéro : alors, de prime abord, nous pouvons dire que nous avons réussi à bien modéliser notre phénomène physique, et que les observations que nous avons observées sont les conséquences de ce phénomène physique modélisé ou de manière équivalente ce dernier est la cause des données observées. Bien que cela peut sembler légitime, c'est un raisonnement à ne pas faire. En effet, l'échantillon que nous avons observé n'est peut être pas représentatif de phénomène physique, et

cela pour différentes raisons que nous avons partiellement évoquées précédemment. En effet, il peut exister un autre échantillon menant à la modélisation du même modèle physique avec la même erreur moyenne. Nous avons construit ce type de cas dans la Figure 2.17. Des cas comme celui-ci, on peut en fabriquer une infinité. Autrement dit, il y a autant de raisons de conclure que le modèle physique modélisé dans chacun des cas est le modèle physique réel sous-jacent que l'on cherche à modéliser, qu'il y a d'échantillons qui permettent d'arriver à cette conclusion de validité. En d'autres termes, nous avons une illustration ici de la phrase "conséquence n'implique pas cause". Ainsi, donner trop de crédit à un certain échantillon par rapport à un autre qui pourrait être lui aussi valable pour modéliser le modèle physique, c'est ce que l'on appelle le phénomène de sur-interprétation. Dans ce cas, le crédit donné à notre échantillon se transfère alors au modèle construit, et c'est lors des phases d'inférence que l'on donne trop de crédit au modèle construit, à tort. Malheureusement dans ce contexte d'apprentissage supervisé, le seul objectif des machines est de réduire cette erreur moyenne. Par conséquent, elles peuvent facilement tomber dans cet écueil et nous fournir de mauvaise modélisation. Ce phénomène s'illustre bien en pratique car, lors de la phase de généralisation, la machine d'un point de vue prédictif peut dire tout et son contraire. On dit qu'il y a alors une grande variance prédictive. Une autre manière de faire une erreur prédictive est de ne pas modéliser le bon modèle et cela peut importe l'ensemble de données retenu pour construire le modèle. Ce phénomène est une illustration de la phrase, "cause n'implique pas conséquence". Ce phénomène s'appelle la sous-interprétation. Dans ce cas là, lors de la phase de généralisation, la machine d'un point de vue prédictif ne donnera jamais la bonne prédiction, peu importe l'exemple. On dit qu'il y alors un grand biais prédictif. D'ailleurs, on retrouve mathématiquement ces concepts dans l'expression de l'erreur moyenne. On peut montrer mathématiquement que l'erreur quadratique moyenne du modèle prédictif se scinde selon ces deux statistiques. Nous présentons ici les idées du calcul : l'idée est de quantifier pour un modèle prédictif $f$ donné, son erreur quadratique moyenne (asymptotique), sachant un ensemble de données $D$ obtenu par échantillonnage (noté que par construction, $f$ dépend de $D$), $D$ est alors considéré comme une variable aléatoire dont on va supposer qu'il existe sa loi $P$ qui a permis son échantillonnage, et que l'espérance existe. Ainsi, l'erreur quadratique

moyenne s'exprime pour $f$ par :

$$\mathop{\mathbb{E}}_{D\sim P}\Big[\mathop{\mathbb{E}}_{X\sim P_X(D)}\Big[\mathop{\mathbb{E}}_{Y\sim P_Y(D,X)}[[f(X)-Y]^2|X]]\Big] \tag{2.13}$$

Nous décrivons dans l'Annexe C.1 le détails des calculs.

Ainsi, l'erreur quadratique moyenne se décompose en la somme du biais au carré de $f$ et de sa variance. Lorsque la machine sur-interprète, alors elle produit des réponses bien différentes lorsque l'on change d'ensemble de données, en particulier pour les données encore jamais observées. A l'inverse si la machine produit un modèle qui est trop loin du modèle réel sous-jacent, c'est une manifestation du biais du modèle prédictif. Pour pouvoir éviter au maximum ces deux écueils , on utilise la méthode de la validation croisée. L'idée est de séparer l'échantillon $D$ en deux parties disjointes aléatoirement. Le premier ensemble est l'ensemble des données pour l'optimisation (on dit aussi pour l'entraînement, ou l'apprentissage), l'autre pour tester la pertinence du modèle construit lors de la phase d'optimisation, c'est l'ensemble de test. Le premier sous-échantillon sert à faire la construction du modèle et le deuxième sert à tester les capacités de généralisation du modèle construit, c'est à dire sa capacité à faire de bonne inférence sur des données non vues pendant son optimisation. C'est à dire en pratique, si il y a reçu un "bon enseignement". Ainsi une erreur faible sur l'ensemble d'entraînement et haute sur l'ensemble de test indique un phénomène de sur-interprétation. Le modèle s'est trop spécifié sur l'échantillon d'entraînement mais se trompe souvent sur les autres échantillons jamais rencontrés et notamment l'échantillon de test [13]. Lorsque l'erreur est haute sur l'échantillon d'entraînement et de test, alors c'est que le modèle courant n'est manifestement pas adapté pour modéliser le phénomène physique, il sous-ajuste. Cela signifie donc que les hypothèses structurelles du modèle ne sont tout simplement pas adaptées au phénomène physique à modéliser, typiquement un cas où nous utilisons un modèle linéaire pour modéliser un phénomène non-linéaire. Mais alors, dans le cas de sous-interprétation, comment faisons-nous pour estimer si le modèle est "loin" de celui-ci a modéliser ? C'est une question difficile. Le seul moyen pratique, sauf cas spécifique, est de bien comprendre le phénomène physique pour concevoir un modèle plus adapté.

13. Pourtant ces deux échantillons sont issus du même phénomène physique, par construction.

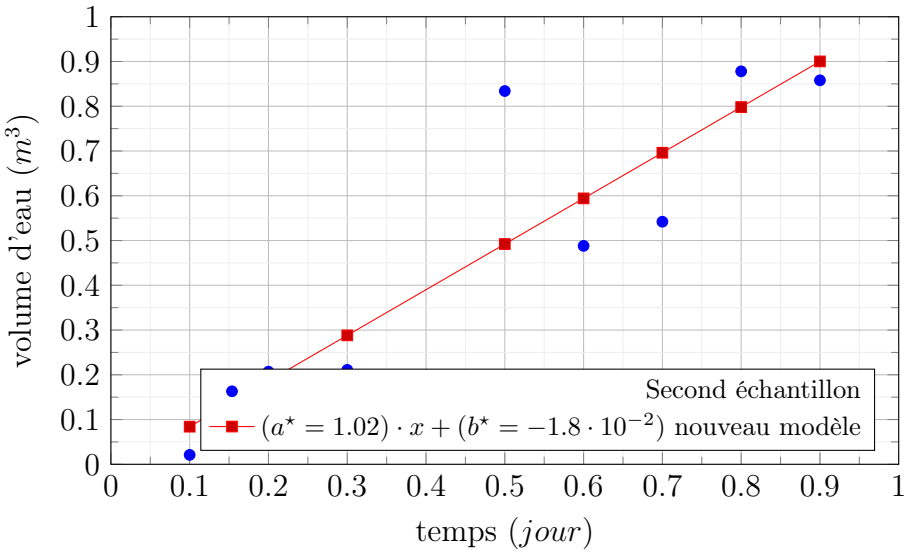



FIGURE 2.17 – Illustration d'un cas produisant la même modélisation mais avec un échantillon différent

Le but global est donc d'avoir une erreur basse sur l'ensemble d'entraînement et sur l'ensemble de test en même temps [14], évitant ainsi le phénomène de sur-interprétation et de sous-interprétation, indiquant alors que le modèle construit n'est probablement pas "loin" du phénomène physique à modéliser. Un stratégie que l'on rencontre en pratique, notamment dans le domaine de l'apprentissage profond car cela s'y prête bien, est d'autoriser l'algorithme à modéliser des phénomènes un peu plus complexe que nécessaire pour s'affranchir du problème de sous-interpréation et de contrôler le phénomène de sur-interprétation avec des méthodes que nous évoquerons. Ou alors, comme déjà évoqué, de mieux comprendre le phénomène physique.

Mais, il existe un résultat statistique nous donnant des garantir probabiliste pour éviter les phénomène de sur-interprétation. En effet, il existe un théorème de l'apprentissage machine qui indique que si la complexité du modèle est très petite devant le nombre de données utilisé pour le construire, alors on évitera probablement le phénomène de sur-interprétation. Ce théorème repose sur la notion de dimension Vapnik-Chervonenkis (dimension VC, ou *dim* VC) du modèle prédictif.

14. De plus, si notre méthode est dépendante par des hyperparamètres (pour faire varier le modèle courant dans une classe de modèles admissibles par exemple), alors on peut même utiliser un troisième sous-échantillon indépendant des deux autres, appelé ensemble de validation, pour mesurer l'erreur de généralisation par rapport à ces hyperparamètres.

### 2.3.7 La nécessité du grand volume de données pour la construction d'un programme d'intelligence artificielle : la dimension de Vapnik-Chervonenkis

Définit par Vladimir Vapnik (Figure 2.19) et Alexeï Chervonenkis (Figure 2.18) La dimension VC d'une fonction $f$ , sachant un échantillon $\{(x_1, y_1), \cdots, (x_N, y_N)\}$ , est le cardinal du plus grand ensemble de points de l'échantillon que $f$ peut interpoler/séparer parfaitement. En première approximation, la dimension VC peut être vue comme une mesure de la complexité , au sens commun, de $f$. Par exemple, pour un modèle $f$ étant un classifieur binaire linéaire agissant dans le plan, $dim$ VC$(f)$=3 car d'une part, une droite pourra toujours séparer trois points dont au moins un est de classe différente des deux premiers, d'où $dim$ VC$(f) \geq 3$, mais pour 4 points on peut trouver une configuration telle que $f$ ne sépare pas linéairement les 4 points , disons les points de l'ensemble $\{(0,0),(1,1)\}$ pour la classe 0 et les points de l'ensemble $\{(1,0),(0,1)\}$ pour la classe 1. L'union de ces deux ensembles n'est pas linéairement séparable. Donc $dim$ VC$(f) < 4$, d'où $dim$ VC$(f) = 3$. On peut même montrer de manière générale, la dimension VC d'un classifieur linéaire opérant dans un espace de dimension $d$ est égale a $d + 1$. Une des applications probabilistes de la dimension VC est de pouvoir quantifier pour un seuil donné $\eta \in ]0,1[$, la probabilité de l'erreur de généralisation, ou du "bon apprentissage" d'un modèle $f$ sachant la taille de l'ensemble de données $N$ et la dimension VC de $f$. C'est le théorème de Vapnik. [Van Der Vaart and Wellner, 1996, Vapnik, 2000], Il énonce dans sa version approximative que la variable aléatoire $Err_g$ représentant l'erreur de généralisation d'un modèle

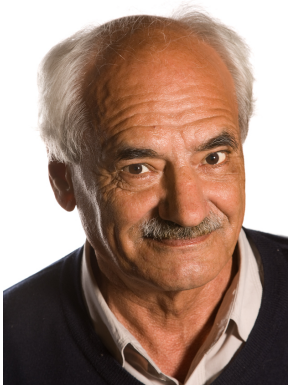


FIGURE 2.18 – Alexeï Chervonenkis (1938 - 2014), Wikipedia

$$\mathbb{P}\left[Err_g \leq \sqrt{\frac{dim\ \text{VC}(f)}{N} + g(\eta)}\right] = 1 - \eta \tag{2.14}$$

avec $g$ une fonction réelle dépendante de $\eta$, négligeable.

Dans le contexte de l'intelligence artificielle, rappelons que d'après Turing, la complexité d'un modèle nécessaire pour atteindre l'intelligence artificielle est mino-

rée donc la dimension VC de ces modèles intelligents aussi. Mais ces valeurs restent grandes, donc leurs valeurs de dimensions VC aussi. Ainsi, d'après l'équation 2.14 si on veut avoir une erreur de généralisation faible pour ces modèles intelligents, ou qu'ils aient fait un "bon apprentissage" avec forte probabilité, alors il est nécessaire d'avoir une quantité énorme de données, c'est à dire, au moins $N >> dim\ \mathrm{VC}(f)$.

FIGURE 2.19 – Vladimir Vapnik (1936 - ·), Wikipedia

Cependant, pour certaines tâches très complexes, le volume de données nécessaire peut ne pas être physiquement réalisable comme évoqué en début de ce document, notamment en grande dimension, un problème apparaît alors : nous sommes clairement limités par le nombre de données possiblement accessible et traitable. Mais nous devons malgré tout construire des modèles suffisamment complexes pour atteindre une modèle capable de produire de l'intelligence artificielle, nous sommes alors en grand risque de sur-interprétation constamment. De prime abord, il semblerait que oui. Cependant pour un problème donné, lorsque le modèle sur-interprète l'ensemble de données, il produit une "loi", celle permettant de lier entrée et sortie contenues dans l'ensemble de données. Ce même modèle aurait construit une autre "loi" s'il avait été confronté à un autre ensemble de données, puisque il est en régime de sur-interprétation. Et ce phénomène se serait reproduit s'il avait été confronté à un troisième ensemble de données, et ainsi de suite. Ainsi, pour un même problème, le modèle est tout de même capable de créer différentes lois, alors qu'à priori, il n'y a qu'une loi physique qui régit le problème et donc l'ensemble de données. Autrement dit, cela signifie que parmi les paramètres du modèle, certains sont utilisés pour approcher la vraie loi physique et les autres sont utilisés pour ajuster au mieux l'ensemble de données, mais que d'un point de vue numérique. Donc, seulement une partie des paramètres du modèle est réellement utile pour vraiment approcher la loi physique sous-jacente. La sélection parcimonieuse de ce sous-ensemble de paramètres du modèle est faisable grâce aux techniques de régularisation. C'est l'objet du prochain paragraphe.

### 2.3.8 Conséquences du manque de données pour les programmes d'intelligence artificielle : méthode de régularisation

Comme évoqué précédemment, pour éviter des situations de sur-interprétation pour les modèles nécessairement complexes comme ceux permettant de créer une intelligence artificielle, il faut les contraindre. Il ne faut autoriser le modèle qu'à utiliser une partie de l'ensemble de ses paramètres accessibles. C'est ce qu'on appelle une régularisation en norme $l_0$ [15]

L'idée de ces méthodes est que pour deux modèles ayant la même performance de classification, on préférera celui qui a le moins de paramètres non nuls. La régularisation peut être vue comme une implémentation du principe du rasoir d'Occam, ou principe de parcimonie. Les méthodes de régularisation permettent aussi d'une part d'éviter les phénomènes de sous-interprétation car elles permettent de "rapprocher" le modèle construit du modèle cible, limitant ainsi le biais de celui-ci ; et d'autre part, de limiter le phénomène de sur-interprétation car le modèle construit se rapprochant du modèle cible, la variance dans les échantillons n'est plus un problème. Par conséquent, on peut voir les méthodes de régularisation comme des méthodes automatisant la validation croisée et donc qui permettent de construire des modèles robustes.

Historiquement, pour obtenir de la robustesse, on perturbait les données d'entraînements par des petites transformations. On considérait que les données de l'échantillon pouvaient avoir des erreurs de mesure et donc en créant un voisinage pour chaque donnée, il était plus probable d'approcher la vraie donnée, celle dont la mesure n'est pas erronée. On considérait que le modèle était robuste si sa réponse ne changeait pas trop sur chacun de voisinages des données. C'est le principe d'augmentation de données. Ce principe permet aussi de créer des données valables et donc d'augmenter artificiellement le volume de données disponibles, ce qui peut être intéressant pour des problèmes contraints en volume de données. D'autres méthodes, comme celle de l'arrêt prématuré, existent aussi. Mais dans le cas de l'apprentissage profond, on peut aussi procéder autrement : au lieu de perturber les données, on per-

15. L'optimisation en norme $l_0$ mène à des problèmes d'optimisation non convexe, donc on préfère minimiser la norme $l_1$ ou $l_2$ des paramètres, qui permet d'optimiser des fonctions plus régulières, donc plus faciles à optimiser.

turbe les internes des modèles, et pour une certaine classe de modèle, sur lesquels on va largement revenir dans le prochain chapitre, on peut faire cela de manière automatique et rapide.

Aujourd'hui, le succès des modèles issu de l'apprentissage machine repose principalement sur une classe de modèles bien précis, ce sont les réseaux de neurones profonds. Cette classe de modèle a même ouvert tout un pan dans la littérature de l'apprentissage machine, ce sont les méthodes de l'apprentissage profond ou encore de "deep learning". Bien qu'il n'existe pas de méthode algorithmique suprême pour résoudre tous les problèmes, comme le démontre le théorème No Free Lunch [Wolpert and Macready, 1997], cette classe de méthodes comporte des propriétés qui en font des méthodes de choix pour résoudre des tâches d'apprentissage machine contemporains. C'est l'objet de notre prochain chapitre. Mais d'abord, récapitulons les arguments que nous avons évoqués pour construire un modèle capable d'intelligence artificielle et synthétisons-les.

## 2.4 Synthèse des arguments théoriques pour la réalisation d'un modèle d'intelligence artificielle

L'information est devenue la grandeur physique d'intérêt des dernières décennies. D'après Shannon, toute information peut être encodée dans des circuits logiques (par exemple, toute information est représentable dans nos ordinateurs d'aujourd'hui, qui sont matériellement, un circuit logique). Ainsi, nous avons un argument pratique pour le traitement de l'information. Pour un traitement automatique de l'information, physiquement réalisable, il faut avoir accès à des données, qui sont nécessairement incomplètes et il faut condenser l'information contenue dans ces données de manière à produire des symboles. Mathématiquement, on peut utiliser le principe d'interpolation. Cela revient donc à utiliser un modèle paramétrique dont on optimise le paramètre pour coller aux données utilisées. Cependant, des phénomènes de sur-interprétation ou de sous-interprétation peuvent subvenir du fait de l'incomplétude inhérente des données. Pour atteindre des modèles capables de faire des prédictions intelligentes, il faut qu'ils soient nécessairement complexes, c'est à dire composés de nombreux paramètres interagissant entre eux d'une certaine manière.

Mais en raison du manque crucial de données devant la complexité de certaines tâches, le risque d'avoir une grande erreur de généralisation est très grand. Il faut alors utiliser des méthodes de régularisation à priori ou à posteriori.

# Chapitre 3

# L'apprentissage machine moderne : l'apprentissage profond

## Sommaire

Depuis la dernière décennie, les modèles d'apprentissage machine reposent principalement sur le paradigme de l'apprentissage profond et sur les réseaux de neurones. En voici un point de vue [1]. Dans un premiers temps, nous les introduirons, avec leurs particularités. Nous illustrerons aussi les méthodes de l'apprentissage machine précédemment citées, et leurs adaptations pour les réseaux de neurones. Puis nous évoquerons comment nous pouvons spécifier ces modèles à des contextes statistiques particuliers.

## 3.1 Les réseaux de neurones : cas du perceptron multicouche

### 3.1.1 Contexte historique

La notion de *réseaux de neurones* est apparue au milieu du 20e siècle. Ils étaient définis via une vision cybernétique de l'apprentissage. Les réseaux de neurones [McCulloch and Pitts, 1943,Hebb, 1949] sont des modèles dont la conception était inspirée du processus d'apprentissage biologique fait par les cerveaux humains ou animals. Le premier modèle computationnel qui s'inscrivait dans ce paradigme était le perceptron simple [Rosenblatt, 1958]. Puis au milieu des années 1980, basé sur l'évolution des sciences cognitives, la vision des réseaux de neurones sous l'angle connexionniste est apparue. L'idée était de combiner plusieurs perceptrons ensemble pour résoudre des tâches d'apprentissage plus complexe. L'exploitation de ces modèles plus lourds a notamment été permise par le succès de l'algorithme de rétropropagation [Rumelhart et al., 1986]. Enfin, la où ces méthodes profondes ont connu un essor considérable et où elles ont été mises au centre d'un large panel de problématique scientifique, c'est à une compétition de vision par ordinateur appelée ImageNet Large Scale Visual Recognition Challenge (INLSVRC) 2012 où AlexNet [Krizhevsky et al., 2017], un réseau de neurones convolutionnel à réduit le taux d'erreur sur la tâche de classification de l'ensemble de données ImageNet (14 millions d'images annotées, 1000 classes) de presque un facteur 2 par rapport à l'état de l'art, ce qui était considé-

1. Les concepts présentés dans ce chapitre ont des définitions formelles et ont chacun un cadre théorique très rigoureux, dont la portée réelle est bien au-delà de ce document. La rédaction de ce chapitre a été fortement inspirée de [Courville et al., ,Roberts and Yaida, 2022]

rable à l'époque. Jusqu'au début des années 2000 [Hinton et al., 2006, Bengio et al., 2007, Ranzato et al., 2007], les réseaux de neurones n'étaient pas considérer comme viables pour résoudre des tâches d'apprentissage. Mais avec le résultat à INLSVRC 2012, les réseaux ont reçu un véritable engouement de la part de la communauté scientifique. Introduisons maintenant la brique élémentaire de ces modèles dits profonds.

### 3.1.2 Le perceptron simple

Le perceptron simple est un type de classifieur linéaire développée par [Rosenblatt, 1958] développée pour les tâches de classification binaire. Pour un tel problème de dimension $n \in \mathbb{N}$, le perceptron simple opère sur un vecteur d'entrée $\mathbf{x} \in \mathbb{R}^n$ issu de l'ensemble de données du problème, un vecteur de poids $\boldsymbol{\omega} \in \mathbb{R}^n$, un terme de biais $b \in \mathbb{R}$ et la fonction signe $s$ définit par :

$$\begin{array}{rccl} s: & \mathbb{R} & \to & \{-1,1\} \\ & x & \mapsto & \begin{cases} 1 & \text{si } x \geq 0 \\ -1 & \text{sinon} \end{cases} \end{array}$$

Originellement, le perceptron simple est alors la forme non linéaire définie par $g_{\boldsymbol{\omega},b}$ tel que :

$$\begin{array}{rccl} g_{\boldsymbol{\omega},b}: & \mathbb{R}^n & \to & \mathbb{R} \\ & \mathbf{x} & \mapsto & s(\boldsymbol{\omega}^T\mathbf{x} + b) \end{array}$$

Le perceptron simple est un cas particulier du perceptron quelconque. Aujourd'hui il est usuel de considérer les perceptrons quelconques, en remplaçant la fonction $s$ par une autre fonction $\sigma : \mathbb{R} \to \mathbb{R}$, non linéaire et différentiable [2], comme par exemple la fonction `relu`, `tanh` ou `sigmoïde`. Comme vu précédemment, l'élaboration de ces méthodes a été motivée initialement par des arguments d'imitations neurologiques, ainsi une telle fonction $\sigma$ est qualifiée "d'activation" car elle est l'analogie, dans le domaine neurologique, du principe d'activation synaptique. Ainsi on notera dans la suite un perceptron quelconque dépendant d'une telle fonction $\sigma$ par $g_{\boldsymbol{\omega}^\star,b^\star,\sigma}$. Comme évoqué, la capacité de séparation linéaire d'un perceptron quelconque est contrainte par la séparabilité linéaire du problème considéré.

2. presque partout

Nous illustrons dans la Figure 3.1 a) un problème de classification unidimensionnelle binaire, l'ensemble de données $D$ est de taille 40 et est l'union de deux tirages uniformes aléatoires de couple $(x, y) \in \mathbb{R} \times \{-1, 1\}$ (où $y$ est le label de $x$) sur le segment $[-2, 4] \subset \mathbb{R}$. On observe ici que dans le cas unidimensionnel non projeté (Figure 3.1 b)), un hyperplan séparateur (ici, le réel $\frac{-b^\star}{\omega^\star} \in \mathbb{R}$). La séparation n'est pas la plus optimale, au vu de $D$, cela était attendu. Nous ne l'avons pas illustré ici, mais en remplaçant le perceptron simple par un perceptron quelconque avec `relu` comme fonction d'activation, nous observons le même phénomène. Il est donc nécessaire de trouver une représentation de $D$ permettant sa séparabilité de manière linéaire. Ainsi, nous allons considérer deux autres perceptrons opérant indépendamment sur la sortie du perceptron initiale, nous obtenons alors une nouvelle représentation bidimensionnelle de $D$ comme étant l'image de $D$ par rapport à la composée de chaque perceptron additionnel avec la sortie du perceptron initiale (Figure 3.1 c). Ces deux perceptrons fournissent à leurs tours l'entrée d'un dernier perceptron. Nous avons illustré via la Figure 3.1 d) la sortie de ce dernier perceptron par rapport à ces entrées. Après avoir identifié des paramètres optimaux, nous pouvons séparer linéairement $D$ dans ce nouvel espace et ainsi trouver un hyperplan séparateur distinguant parfaitement les deux classes décrit par les éléments de $D$.

Pour récapituler, nous avons donc considéré un modèle composé d'un perceptron d'entrée , combinés à deux autres perceptrons mis en parallèle et qui nourrissent un dernier perceptron. Une fois après avoir obtenu des candidats pour les paramètres optimaux, nous avons réussi à trouver une configuration de l'espace de départ du dernier perceptron ; obtenu en appliquant récursivement des transformations linéaires et non linéaires ; permettant de construire un hyperplan séparateur optimal et donc une séparation optimale en deux parties décrivant les deux classes de $D$.

Cependant, certains problèmes sont bien plus complexes et nécessitent des transformations bien plus complexes pour admettre, à terme, une séparabilité linéaire. Pour atteindre une telle complexité, nous pouvons chainer plus de perceptrons simples, mais aussi les chainer en parallèle. De manière générale, on parlera alors de perceptron multicouche.

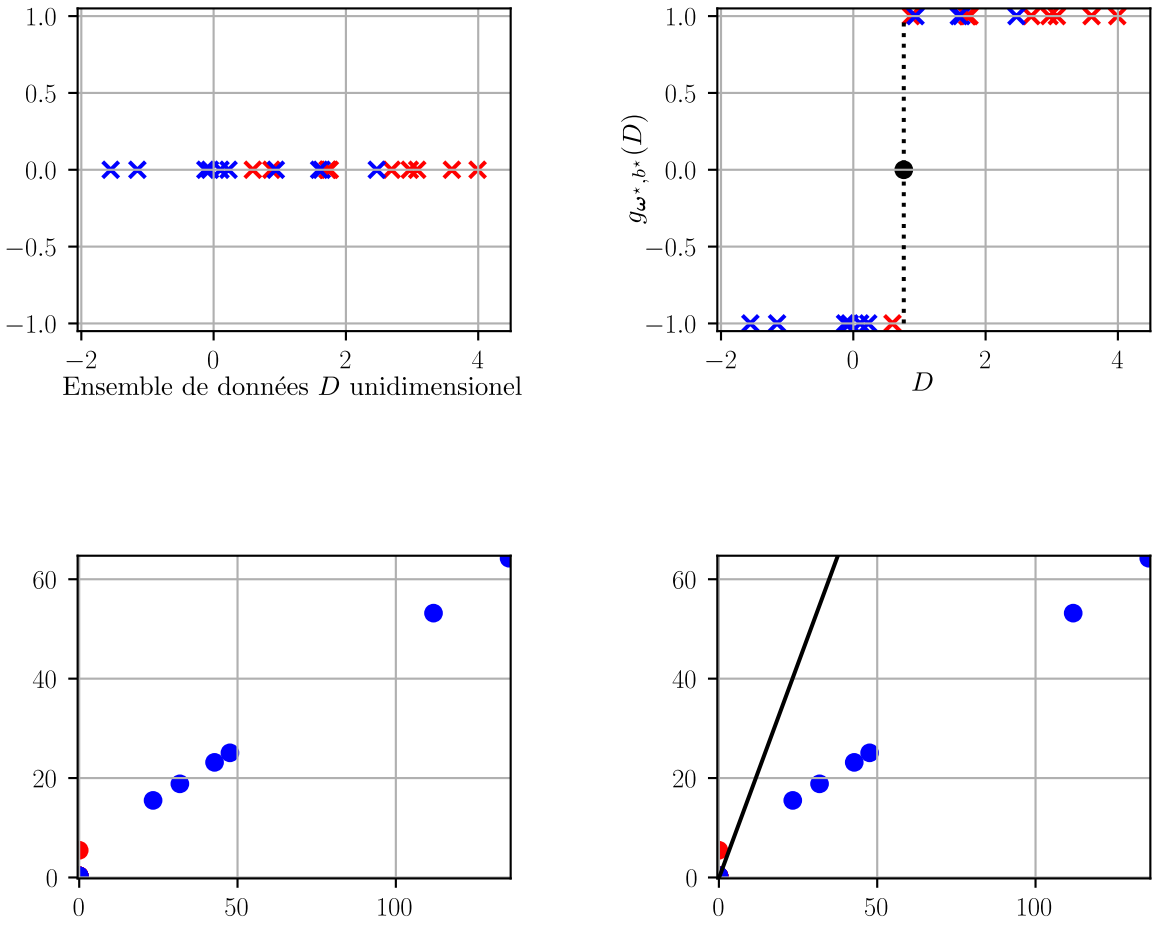



FIGURE 3.1 – a) Illustration de l'ensemble de données $D$ unidimensionnel - b) Classification de $D$ par le perceptron simple optimal $g_{\omega^\star,b^\star}$ - c) Projection de $D$ dans un espace linéairement séparable - d) Séparation du projeté de $D$ par un hyperplan optimal

### 3.1.3 Le perceptron multicouche

Le perceptron multicouche $f_{\boldsymbol{\theta}}$ [Rosenblatt, 1958] de profondeur $L \in \mathbb{N}$ a valeur dans $\mathbb{R}^d$ dans $\mathbb{R}$ est alors la donnée d'un $L$-uplets $(k_i)_{i\in\{1,\dots,L\}} \in \mathbb{N}^L$ tel que pour tout $\mathbf{x} \in \mathbb{R}^d$ on a :

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \begin{cases} \mathbf{x}^{(0)} = \mathbf{x} \text{ et } k_0 = d \\ \mathbf{x}^{(i)} = \boldsymbol{\sigma}_i(\boldsymbol{\Omega}_{(i)}^T \mathbf{x}^{(i-1)} + \mathbf{b}_{(i)}) \text{ pour } 1 \leq i \leq L+1 \text{ et } k_{L+1} = p \end{cases} \tag{3.1}$$

Où $\boldsymbol{\Omega}_{(i)} \in \mathbb{R}^{k_{i-1}\times k_i}$ est la matrice des poids synaptiques reliant la $i-1$-ème couche à la $i$-ème couche du modèle et où $\mathbf{b}_{(i)} \in \mathbb{R}^{k_i}$ et le vecteur des biais des $k_i$ perceptrons de la $i$-ème couche. La fonction d'activation $\boldsymbol{\sigma}_i$ vectorielle est une application définie par :

$$\begin{array}{rccc} \boldsymbol{\sigma}_i : & \mathbb{R}^{k_i} & \to & \mathbb{R}^{k_i} \\ & (\mathbf{x}_j)_{j\in\{1,\dots,k_i\}} & \mapsto & (\sigma_i(\mathbf{x}_j))_{j\in\{1,\dots,k_i\}} \end{array}$$

Et où $\sigma_i$ est l'une des fonctions d'activation réelle présentée précédemment, pour la $i$-ème couche. Et afin nous avons, $\boldsymbol{\theta} = ((\boldsymbol{\Omega}_{(i)})_i, (\mathbf{b}_{(i)})_i) \in \mathbb{R}^{\sum_{i=1}^{L} k_i(k_{i-1}+1)}$. Ce modèle permet donc d'obtenir une représentation récursivement transformée des données d'entrée de manière à obtenir à la couche $L$ une représentation qui est linéairement séparable. Il suffit alors de faire une classification linéaire dans cette espace pour résoudre la tâche de classification. D'un point de vue de l'espace de départ, ce chaînage de classifieur linéaire permet donc de construire des "loi" de classification linéaire par morceau, menant à une construction d'une "loi" complexe pour résoudre le problème selon une métrique d'évaluation. Le choix de $L$, de $(k_1, \dots, k_L)$ et de la famille $(\sigma_i)_i$ est ce que l'on appelle le choix de l'*architecture* de $f_{\boldsymbol{\theta}}$. Le choix a priori de l'architecture pour un problème donnée est un problème compliqué à résoudre donc on y répond souvent de manière empirique, a posteriori, et itérativement en pratique. Ce type de modèle, le perceptron multicouche est ce que l'on appelle aujourd'hui un réseau de neurones.

Supponsons maintenant que nous avons déterminé une bonne architecture pour $f_{\boldsymbol{\theta}}$, pour un problème donné. Comment pouvons-nous trouver $\boldsymbol{\theta}^\star$ , c'est à dire comment pouvons-nous optimiser la représentation de notre ensemble de données au travers de $f_{\boldsymbol{\theta}}$ pour trouver une configuration optimale permettant sa séparabilité

linéaire à la $L$-ième couche ? Utiliser la méthode de l'estimation du maximum de vraisemblance comme dans chapitre précédent peut paraitre compliqué dans le cas des perceptrons multicouches. En effet, dans le cas de la régression linéaire du chapitre précédent, la fonction à optimiser est convexe en les paramètres du modèle (c-à-d la vraisemblance du modèle). Cela permet d'avoir d'une part, unicité de la solution optimale, d'autre part on peut décrire cette solution par forme fermée et donc on peut la calculer. Dans le cas des perceptrons multicouches, nous n'avons pas un tel confort d'un point de vue de l'optimisation. Cependant notre fonction à optimiser partage aussi le faite d'être différentiable [3], en tant que composé de fonction différentiable. Ainsi au lieu de connaitre directement la solution optimale, nous allons utiliser un algorithme itératif qui navigue dans l'espace des paramètres de la fonction de cout, a tâtons, afin d'arriver à une solution convenable (c.-à-d, un paramétrage du modèle permettant la séparabilité de l'ensemble de données ou bien de manières équivalentes, d'obtenir une bonne représentation des données initiales permettant leurs séparabilités linéaires). Il s'agit de la méthode de la descente de gradient.

## 3.2 L'algorithme de la descente de gradient

La méthode de descente de gradient est une méthode d'optimisation du premier ordre. Pour une fonction $f : \mathbb{R}^d \to \mathbb{R}$ différentiable en $\mathbf{x}_0 \in \mathbb{R}^d$, la méthode permet d'atteindre un minimum local de $f$ au voisinage de $\mathbf{x}_0$.

L'idée de la méthode repose sur le principe suivant : du point $\mathbf{x}_0$, vers quelle direction $\mathbf{v} \in \mathbb{R}^d$ (on peut prendre $\mathbf{v}$ unitaire, c-à-d $\|\mathbf{v}\|_2 = 1$) devons-nous nous orienter pour atteindre une valeur de $f$ plus petite que $f(\mathbf{x}_0)$ ?

Soit $\lambda \in \mathbb{R}$ tel que

$$\begin{array}{rccc} g: & \mathbb{R} & \to & \mathbb{R} \\ & \lambda & \mapsto & f(\mathbf{x}_0 + \lambda \mathbf{v}) \end{array}$$

la fonction $g$ est donc la projection de l'image de $f$ selon l'axe décrit par $\mathbf{v}$ dans $\mathbb{R}^d$ depuis le point $\mathbf{x}_0$. On sait que $g$ est dérivable en 0 car $f$ différentiable en $\mathbf{x}_0$ par hypothèse. On a que

$$g'(\lambda) = \mathbf{v}^T \nabla_{\mathbf{x}_0 + \lambda \mathbf{v}} f(\mathbf{x}_0 + \lambda \mathbf{v}) \tag{3.2}$$

3. presque partout, on fera cette supposition dans toute la suite de ce document

en particulier,

$$g'(0) = \mathbf{v}^T \nabla_{\mathbf{x}_0} f(\mathbf{x}_0) \tag{3.3}$$

La quantité $g'(0)$ est donc une information du niveau de pente de $f$ au point $\mathbf{x}_0$ dans la direction $\mathbf{v}$. Mais quel est le $\mathbf{v}$ qui permet de descendre le plus vite possible la valeur de $f$ depuis $\mathbf{x}_0$, autrement dit pour quel vecteur $\mathbf{v}$, $g'(0)$ est-il minimal ?

Nous souhaitons donc résoudre :

$$\min_{\mathbf{v},\|\mathbf{v}\|_2=1} g'(0) = \min_{\mathbf{v},\|\mathbf{v}\|_2=1} \mathbf{v}^T \nabla_{\mathbf{x}_0} f(\mathbf{x}_0) \tag{3.4}$$

$$= \min_{\mathbf{v},\|\mathbf{v}\|_2=1} \|\mathbf{v}\|_2 \|\nabla_{\mathbf{x}_0} f(\mathbf{x}_0)\|_2 \cos(\alpha) \text{ où } \alpha \text{ est l'angle orienté entre } \mathbf{v} \text{ et } \nabla_{\mathbf{x}_0} f(\mathbf{x}_0) \tag{3.5}$$

$$= \min_{\mathbf{v}} \cos(\alpha) \tag{3.6}$$

$$\Leftrightarrow \alpha = \pi[2\pi] \tag{3.7}$$

Donc la direction à prendre pour minimiser le plus rapidement $f$ au voisinage de $\mathbf{x}_0$ est une direction colinéaire (positivement) à $-\nabla_{\mathbf{x}_0} f(\mathbf{x}_0)$. Une approche algorithmique pour minimiser localement une telle fonction $f$ au voisinage de $\mathbf{x}_0$ est donc de parcourir la trajectoire $(\mathbf{x}_k)_k$ dans $\mathbb{R}^d$ définie récursivement par :

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \epsilon_k \nabla_{\mathbf{x}_k} f(\mathbf{x}_k) \; \forall k \geq 0 \tag{3.8}$$

Où $\epsilon_k \in \mathbb{R}^+$ pour tout $k$, est un hyperparamètre de la méthode permettant de moduler la "longueur" du pas à faire dans $\mathbb{R}^d$ à l'étape $k$. Cet hyperparamètre est appelé le taux d'apprentissage. Une stratégie d'arrêt pour arrêter l'algorithme est de l'arrêter lorsque $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 < \delta \Leftrightarrow |\epsilon_k| \|\nabla_{\mathbf{x}_k} f(\mathbf{x}_k)\|_2 < \delta$, pour un $\delta \in \mathbb{R}$ petit. C'est-à-dire, si nous considérons $\epsilon_k$ constant pour tout $k \geq 0$, lorsque $\|\nabla_{\mathbf{x}_k} f(\mathbf{x}_k)\|_2$ est petit, c'est-à-dire lorsque nous sommes très proches d'un minimum local pour $f$. Il peut exister des configurations locales de $f$ tel que la suite $(\|\nabla_{\mathbf{x}_k} f(\mathbf{x}_k)\|_2)_k$ est stationnaire de niveau $\delta_0 \in \mathbb{R}^+$ à partir d'un certain rang $k_0 \in \mathbb{N}$ ou même que la suite soit strictement croissante à partir de $k_0$. Ainsi pour s'assurer une convergence et un arrêt il faut imposer une condition sur $(\epsilon_k)_k$ comme par exemple $\lim_{k\to\infty} e_k = 0$. Dans le cas où $f$ est convexe sur $\mathbb{R}^d$, alors avec les bonnes conditions sur la famille $(\epsilon_k)_k$, on peut montrer que la méthode de gradient converge vers le minimum global de $f$ [Cauchy, 1847] et cela peu importe les conditions d'initialisation $\mathbf{x}_0$.

Revenons sur les conditions d'initialisation de cet algorithme. Comme déja évoqué, lorsque nous supposons que nos données ont une structure sous-jacente linéaire, alors l'erreur quadratique moyenne est une fonction convexe en le paramètre du modèle. Dans le cas des réseaux de neurones, les transformations non-linéaire opérant dans le modèle rendent l'erreur quadratique moyenne non-convexe. Ainsi, la convergence vers le minimum global de la fonction $f$ n'est plus assurer et les conditions initiales sont importantes car elle peuvent fortement altérer le succès de l'algorithme. Il existe des approches permettant de mieux apprivoiser les structures locales de $f$, notamment les méthodes de second ordre faisant notamment appel au calcul de l'hessienne de $f$ par rapport aux paramètres à optimiser. Cependant si nous avons $\mathcal{O}(n)$ paramètre à optimiser alors pour la calculer il faut $\mathcal{O}(n^2)$ calculs ce qui peut être très long à calculer pour des modèles avec des millions de paramètres par exemple et c'est souvent le cas pour les modèles d'apprentissage machine moderne. Ainsi, bien que les méthodes apportent des informations plus fine sur les structures locales de la fonction a optimiser, et donc de mieux l'optimiser, elles sont mises de côté pour leur lourdeur calculatoire dans le cas des réseaux de neurones. Pour pouvoir utiliser malgré tout la méthode de descente de gradient dans le cas non convexe, il existe la méthode de l'inertie [Polyak, 1964], qui consiste à considérer que $\mathbf{x}_k$, par analogie avec la physique, est un corps ayant une masse non nulle, et que au cours de sa trajectoire il va accumuler de l'énergie cinétique en montant et descendant les structures localement convexe ou concave de $f$ de manière à se s'arrêter dans une localité à faible courbure assez étendue. Donc potentiellement, un minimum global. Une amélioration de cette approche a été proposée par [Nesterov, 1983, Neerincx et al., 2018]. Enfin, en conjonction avec l'introduction de l'inertie dans la méthode de descente de gradient, [Hinton, 2012] propose un moyen de correction automatique du taux d'apprentissage au cours de l'optimisation. Une version "normalisé" de cet algorithme a été proposée par [Kingma and Ba, 2017]. Cet algorithme est aujourd'hui le plus utilisé.

Par ailleurs, pour pouvoir utiliser la méthode de descente de gradient, il faut déja pouvoir calculer le gradient de la fonction à optimiser. Dans le cas des perceptrons multicouches, le calcul du gradient de la fonction de coût est facilité grâce à la méthode de la rétropropagation.

## 3.3 L'algorithme de rétropropagation

Nous avons vu que les perceptrons multicouches étaient une composition d'un certain nombre de couches de perceptrons. Autrement dit, les perceptrons multicouches sont des compositions de fonction vectorielle. En reprenant les notations de la partie précédente, on peut réécrire le modèle $f_{\boldsymbol{\theta}}$ du perceptron multicouche , comme :

$$f_{\boldsymbol{\theta}} = f_{L+1} \circ f_L \circ \cdots \circ f_0 \tag{3.9}$$

où pour tout $i \in \{0, \cdots, L+1\}$, $f_i$ est défini comme :

$$\begin{array}{rccc} f_i : & \mathbb{R}^{k_{i-1}} & \to & \mathbb{R}^{k_i} \\ & \mathbf{x} & \mapsto & \boldsymbol{\sigma}_i(\boldsymbol{\Omega}_{(i)}^T \mathbf{x} + \mathbf{b}_{(i)}) \end{array}$$

prenons par exemple $\mathbf{x}$ et son label $\mathbf{y}$ de notre ensemble de données $D$, nous pouvons alors calculer l'erreur de prédiction faite par $\boldsymbol{\theta}$ sur $\mathbf{x}$ par rapport à $\mathbf{y}$ et $\boldsymbol{\theta}$ et nous la définissons, pour l'illustration, par :

$$L(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{y} - f_{\boldsymbol{\theta}}(\mathbf{x})\|_2^2 \tag{3.10}$$

On suppose $L$ différentiable aux points d'intérêt par rapport à $\boldsymbol{\theta}$. Ainsi $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ à un sens sur son espace de définition.

La méthode de rétropropagation introduite dans [Rosenblatt, 1962, Rumelhart et al., 1986] permet de calculer exactement $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$. Nous définissons $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \in \mathbb{R}^{\sum_{i=0}^{L+1} k_i(k_{i-1}+1)}$ comme un vecteur colonne. Dans le contexte de $f_{\boldsymbol{\theta}}$, calculer $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ revient à calculer pour tout $i \in \{1, \cdots, L+1\}$, les quantités $\nabla_{\mathbf{b}_{(i)}} L(\boldsymbol{\theta})$ et $\nabla_{\boldsymbol{\Omega}_{(i)}} L(\boldsymbol{\theta})$.

Nous posons pour tout $i \in \{1, \cdots, L+1\}$ ,

$$\boldsymbol{\alpha}_i = \boldsymbol{\Omega}_{(i)}^T \mathbf{x}_{(i-1)} + \mathbf{b}_{(i)} \in \mathbb{R}^{k_i} \tag{3.11}$$

— Premièrement , calculons $\nabla_{\mathbf{b}_{(L+1)}} L(\boldsymbol{\theta})$ , c'est-à-dire , pour tout $j \in \{1, \cdots, k_{L+1}\}$,

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \mathbf{b}_j^{(L+1)}} \in \mathbb{R}$$

Nous avons :

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \mathbf{b}_j^{(L+1)}} = \frac{\partial L(\boldsymbol{\theta})}{\partial \mathbf{x}_j^{(L+1)}} \times \frac{\partial \mathbf{x}_j^{(L+1)}}{\partial \boldsymbol{\alpha}_j^{L+1}} \times \frac{\partial \boldsymbol{\alpha}_j^{L+1}}{\partial \mathbf{b}_j^{(L+1)}} \in \mathbb{R} \tag{3.12}$$

$$= (\mathbf{y}_j - f_{\boldsymbol{\theta}}(\mathbf{x})_j) \times \sigma'_{L+1}(\boldsymbol{\alpha}_j^{L+1}) \times 1 \tag{3.13}$$

$$= (\mathbf{y}_j - f_{\boldsymbol{\theta}}(\mathbf{x})_j) \times \sigma'_{L+1}(\boldsymbol{\alpha}_j^{L+1}) \tag{3.14}$$

Sous forme vectorielle cela revient à dire que :

$$\nabla_{\mathbf{b}_{(L+1)}} L(\boldsymbol{\theta}) = (\mathbf{y} - f_{\boldsymbol{\theta}}(\mathbf{x})) \circ \boldsymbol{\sigma}'_{L+1}(\boldsymbol{\alpha}_{L+1}) \tag{3.15}$$

$$= \nabla_{\mathbf{x}_{(L+1)}} L(\boldsymbol{\theta}) \circ \boldsymbol{\sigma}'_{L+1}(\boldsymbol{\alpha}_{L+1}) \in \mathbb{R}^{k_{L+1}} \tag{3.16}$$

Ici, le symbole $\circ$ représente le produit d'Hadamard.

— Deuxièmement, calculons $\nabla_{\boldsymbol{\Omega}_{(L+1)}} L(\boldsymbol{\theta})$, c'est-à-dire, pour tout $(j,i) \in \{1, \cdots, k_L\} \times \{1, \cdots, k_{L+1}\}$,

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\Omega}_{j,i}^{(L+1)}} \in \mathbb{R}$$

Nous avons :

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\Omega}_{j,i}^{(L+1)}} = \frac{\partial L(\boldsymbol{\theta})}{\partial \mathbf{x}_i^{(L+1)}} \times \frac{\partial \mathbf{x}_i^{(L+1)}}{\partial \boldsymbol{\alpha}_i^{L+1}} \times \frac{\partial \boldsymbol{\alpha}_i^{L+1}}{\partial \boldsymbol{\Omega}_{j,i}^{(L+1)}} \in \mathbb{R} \tag{3.17}$$

$$= \frac{\partial L(\boldsymbol{\theta})}{\partial \mathbf{x}_i^{(L+1)}} \times \frac{\partial \mathbf{x}_i^{(L+1)}}{\partial \boldsymbol{\alpha}_i^{L+1}} \times \mathbf{x}_j^{(L)} \tag{3.18}$$

Sous forme matricielle cela revient à dire que :

$$\nabla_{\boldsymbol{\Omega}_{(L+1)}} L(\boldsymbol{\theta}) = \mathbf{x}_{(L)} . \nabla_{\boldsymbol{\alpha}_{L+1}} L(\boldsymbol{\theta})^T \in \mathbb{R}^{k_L \times k_{L+1}}$$

Le symbole . représente le produit matriciel.

Ainsi nous avons calculé toutes les dérivés partielles de $L$ selon les poids de la dernières couche de $f_{\boldsymbol{\theta}}$.

Maintenant, calculons ces mêmes quantités pour la couche antérieur. C'est à dire calculons , d'une part $\nabla_{\mathbf{b}_{(L)}} L(\boldsymbol{\theta})$ et d'autre part $\nabla_{\boldsymbol{\Omega}_{(L)}} L(\boldsymbol{\theta})$.

L'utilisation de la loi de la chaine doit se faire avec précaution, en effet dans ce contexte, calculer $\nabla_{\mathbf{b}_{(L)}} L(\boldsymbol{\theta})$ nécéssite de calculer apriori $\nabla_{\mathbf{x}_{(L)}} L(\boldsymbol{\theta})$ car $\mathbf{x}_{(L)}$ dépend de $\mathbf{b}_{(L)}$. Or nous avons vu que pour tout $i \in \{1, \cdots, L+1\}$, $\boldsymbol{\alpha}_i^{L+1}$ dépend de $\mathbf{x}_{(L)}$ , ainsi il faut utiliser ici la formule de la dérivé totale. Donc, on a :

$$\nabla_{\mathbf{x}_{(L)}} L(\boldsymbol{\theta}) = [\nabla_{\mathbf{x}_{(L)}} \boldsymbol{\alpha}_{L+1}]^T . \nabla_{\boldsymbol{\alpha}_{L+1}} L(\boldsymbol{\theta}) \tag{3.19}$$

$$= \boldsymbol{\Omega}_{(L)}.\nabla_{\boldsymbol{\alpha}_{L+1}} L(\boldsymbol{\theta}) \in \mathbb{R}^{k_L} \tag{3.20}$$

Cette formule est un moyen de passage entre les différentes couches , au sein d'une couche le calcul des dérivées partielles se fait comme précédemment.

Ainsi on obtient :

$$\nabla_{\mathbf{b}_{(L)}} L(\boldsymbol{\theta}) = \nabla_{\mathbf{x}_{(L)}} L(\boldsymbol{\theta}) \circ \nabla_{\boldsymbol{\alpha}_L} \mathbf{x}_{(L)} \circ \nabla_{\mathbf{b}_{(L)}} \boldsymbol{\alpha}_L \tag{3.21}$$

$$= \nabla_{\mathbf{x}_{(L)}} L(\boldsymbol{\theta}) \circ \boldsymbol{\sigma}'_L(\boldsymbol{\alpha}_L) \tag{3.22}$$

$$= \nabla_{\boldsymbol{\alpha}_L} L(\boldsymbol{\theta}) \in \mathbb{R}^{k_L} \tag{3.23}$$

Et ont fait le raisonement analogue pour le calcul de $\nabla_{\boldsymbol{\Omega}_{(L)}} L(\boldsymbol{\theta})$, c'est-à-dire :

$$\nabla_{\boldsymbol{\Omega}_{(L)}} L(\boldsymbol{\theta}) = \nabla_{\mathbf{x}_{(L)}} L(\boldsymbol{\theta}) \circ \nabla_{\boldsymbol{\alpha}_L} \mathbf{x}_{(L)} \circ \nabla_{\boldsymbol{\Omega}_{(L)}} \boldsymbol{\alpha}_L \tag{3.24}$$

$$= \mathbf{x}_{(L-1)}.\nabla_{\boldsymbol{\alpha}_L} L(\boldsymbol{\theta})^T \in \mathbb{R}^{k_{L-1} \times k_L} \tag{3.25}$$

On appliquant ce procédé récursif, on peut calculer, pour tout $i \in \{1, \cdots, L+1\}$, les quantités $\nabla_{\mathbf{b}_{(i)}} L(\boldsymbol{\theta})$ et $\nabla_{\boldsymbol{\Omega}_{(i)}} L(\boldsymbol{\theta})$ , et par concaténation $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$.

De plus, d'un point de vue logiciel et matériel, le calcule de $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ a certaines bonnes propriétés :

— d'un point de vue logiciel : pour tout pour tout $i \in \{1, \cdots, L+1\}$, on a vu précédemment que

  — $\nabla_{\mathbf{b}_{(i)}} L(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\alpha}_i} L(\boldsymbol{\theta})$

  — $\nabla_{\boldsymbol{\Omega}_{(i)}} L(\boldsymbol{\theta}) = \mathbf{x}_{(i-1)}.\nabla_{\boldsymbol{\alpha}_i} L(\boldsymbol{\theta})^T$

  Donc pour chaque $i$, la quantité $\nabla_{\boldsymbol{\alpha}_i} L(\boldsymbol{\theta})$ n'a pas besoin d'être calculée deux fois (i.e pour $\nabla_{\mathbf{b}_{(i)}} L(\boldsymbol{\theta})$ et $\nabla_{\boldsymbol{\Omega}_{(i)}} L(\boldsymbol{\theta})$). Seulement un seul calcul de $\nabla_{\boldsymbol{\alpha}_i} L(\boldsymbol{\theta})$ est nécessaire, cette quantité peut ainsi être stockée en machine. Par exemple, si on a commencé par calculer $\nabla_{\boldsymbol{\Omega}_{(i)}} L(\boldsymbol{\theta})$ alors nous pouvons directement réutiliser $\nabla_{\boldsymbol{\alpha}_i} L(\boldsymbol{\theta})$ pour calculer $\nabla_{\boldsymbol{\Omega}_{(i)}} L(\boldsymbol{\theta})$. Cela permet donc de faire des économies de calcul.

— d'un point de vue matériel : pour tout pour tout $i \in \{1, \cdots, L+1\}$, le calcul de $\nabla_{\mathbf{b}_{(i)}} L(\boldsymbol{\theta})$ et $\nabla_{\boldsymbol{\Omega}_{(i)}} L(\boldsymbol{\theta})$ ne nécessite que de faire des produits matriciels. Un type d'opération particulièrement facile pour les unités de calcul type GPU qui ont été développées principalement pour ce type d'opérations.

Le calcul des différentes dérivées se fait imultané dans la plupart des implémentations logicielles de la méthode de la rétropropagation. Le calcul numérique des dérivées partielles pour un modèle quelconque $f_{\boldsymbol{\theta}}$ se fait en général par différence finie avant ou arrière.

Cet algorithme est simple et il est adéquat pour n'importe quelle fonction définie de manière récursive, admettant au moins une certaine régularité sur les espaces voulus, et donc en particulier pour les réseaux de neurones. Et il peut être utilisé pour calculer des différentielles d'ordre supérieur à 1. Cet algorithme est l'un des points clés du succès des réseaux de neurones. Mais il y a aussi d'autres aspects qui motivent l'utilisation des réseaux de neurones pour la conception de modèle produisant de d'intelligence artificielle.

## 3.4 L'approche stochastique de l'optimisation

Nous l'avons vu précédemment, en pratique trouver un bon modèle prédictif c'est résoudre un problème d'apprentissage machine et l'hypothèse selon laquelle l'ensemble de données que nous avons en pratique pour résoudre le problème est un échantillon i.i.d issu d'une variable aléatoire suivant la "vrai" loi des donnés, joue un rôle majeur. En effet, en supposant son existence, la vraie quantité à minimiser est l'espérance de la fonction de coût du problème prise sur la vraie loi. Cette quantité n'existe que de manière asymptotique, ainsi nous l'estimons via l'estimateur de la moyenne empirique, qui est sans biais, prise sur l'ensemble de données qui est un échantillon i.i.d par hypothèse. Ainsi la loi des grands nombres assure la convergence vers l'espérance de la fonction de coût précédemment évoqué. L'un des avantages de l'estimateur de la moyenne empirique est qu'il a une variance qui décroît en $\mathcal{O}(N^{-\frac{1}{2}})$, avec $N$ la taille de l'ensemble de données. Donc l'augmentation relativement faible de la précision de l'estimation a un coût important, sur la quantité de traitement à effectuer. De plus, malgré que les données soit rares dans certains cas, il existe des problèmes d'apprentissage machine où les tailles de l'ensemble de données sont de l'ordre du million. Ainsi tout traiter prendrait un temps trop important pour une faible valeur ajoutée pour l'estimation de la valeur moyenne de la fonction de coût (et donc son gradient). Ainsi, et pour les mêmes raisons, il est usuel d'estimer

ces quantités avec des petits paquets de données construient aléatoirement. Ces traitements étant indépendant entre eux , ils sont parallélisablent d'un point de vue calculatoire. Ainsi, les quantités relatives à la fonction de coût, comme le gradient, sont eux aussi des estimations. Et dans le cas des réseaux de neurones, le gradient étant un opérateur linéaire, nous pouvons utiliser l'algorithme de rétropropagation en parallèle pour chaque élément du paquet de données, donnant ainsi le gradient de la fonction de coût par rapport au paquet de données, par linéarité. Ainsi les problèmes d'apprentissage machine, impliquant des réseaux de neurones ou non, ont une nature nécessairement inférentielle, on parle aussi d'optimisation stochastique. Intéressons-nous maintenant aux méthodes de régularisation pour les réseaux de neurones.

## 3.5 Les méthodes de régularisation pour les perceptrons multicouches

Nous avons vu précédemment la nécessité des méthodes de régularisation pour les modèles de issu de l'apprentissage machine dans le chapitre précédent. Plusieurs approches ont été proposées, nous avons déjà évoqué la méthode de l'augmentation de données. Pour rappel, l'idée était, pour une donnée $\mathbf{x} \in D$ de crée de nouvelles instances dérivées de $\mathbf{x}$ en appliquant une transformation perturbatrice. L'idée de cette technique repose sur le faite que les données que nous avons dans notre ensemble de données nous sont pas une illustration précise du phénomène sous-jacent que nous souhaitons modéliser, mais quel sont en faite altérée par des transformations, qui peut se matérialiser du point de vue pratique comme des erreurs de mesures, des bruits de capteurs, etc. Or, nous voulons que notre modèle soit insensible à ces petites perturbations. Nous proposons alors un voisinage de données, plutôt qu'une donnée unitaire pour en représenter une. Cela permet donc au modèle d'être moins sensible à ces altérations et donc de mieux s'approcher du phénomène cible à modéliser.

À l'inverse, au lieu de considérer que les données sont bruitées, on peut considérer que le modèle fait des prédictions bruitées à cause du bruit des sous-jacent des données. Ainsi, au lieu de faire une augmentation de donnée, on fait une augmenta-

tion des prédictions pour chaque donnée $\mathbf{x}$. Nous obtenons alors plusieurs modèles dont la prédiction moyenne est plus proche que celle que le phénomène approximé pourrait produire.

Dans le cas des perceptrons multicouches, il existe une méthode facile à mettre en oeuvre : c'est la méthode du dropout [Srivastava et al., 2014]. Pour un perceptron multicouche, l'idée du dropout est "d'éteindre" avec une certaine probabilité $p$ chaque neurone qui compose $f_{\boldsymbol{\theta}}$. En tirant un échantillon pour l'extinction des neurones, ont construit alors plusieurs modèles, proches entre eux, proposant ainsi plusieurs prédictions.

Plus formellement, considérons une famille de $N = \sum_{i=1}^{L} k_i$ paramètres $(p_i)_i$ à valeurs dans $[0, 1]$ , tels que la variable aléatoire $\mathbf{m} \sim \mathcal{B}(1, (p_i)_i)$ telle que :

$$f_{i,\mathbf{m}_i} = \mathbf{m}_i \times f_i \; \forall i \in \{1, \cdots, N\} \tag{3.26}$$

Nous obtenons alors un nouveau modèle $f_{\boldsymbol{\theta},\mathbf{m}}$. Ainsi pour un élément $\mathbf{x} \in D$, nous obtenons une nouvelle prédiction égale à $\mathbb{E}_{\mathbf{m}}[f_{\boldsymbol{\theta},\mathbf{m}}(\mathbf{x})]$. Cette méthode de régularisation est très facile à implémenter dans le cas des réseaux de neurones.

Les méthodes de régularisation générique sont un des moyens pour réussir à mieux approximer les phénomènes, mais dans certains cas, la nature des données fait que un grand travail des méthodes de régularisation, peut être évité. Autrement dit, au lieu de travailler a postériori sur la robustesse du modèle avec des méthodes de régularisation, on peut déja effectuer, dans certains cas, une grande partie du travail, en travaillant sur la robustesse du modèle a priori.

En effet, aujourd'hui les problématiques de traitement automatique de l'information sont des problèmes de grandes dimensions. Mais les données traitées par ces modèles possèdent parfois une certaine structure, une certaine régularité. Dans ces cas là, la "vrai" loi des données possède surement un apriori elle aussi dans sa définition qu'il faut ensuite retranscrire dans la définition du modèle qui interagit avec ces données particulières. Savoir utiliser algorithmiquement les bons outils capables de formaliser ces aprioris est au coeur de la recherche en théorie de l'apprentissage des représentations. Nous allons lister quelqu'un de ces aprioris et nous allons voir que les perceptrons multicouches, une fois adaptés, permettent d'en intégrer quelqu'uns dans des cas particulier et contemporain. Avant de spécifier nos propos , énumérons brièvement ces a priori partager par un grand nombre des problèmes d'apprentissage

machine d'aujourd'hui.

## 3.6 La nature statistique des données contemporaines

La recherche en théorie de l'apprentissage des représentations porte sur la formalisation mathématique de la structure des données contemporaines et donc sur la formalisation de modèle capable d'interagir efficacement avec ces données. En voici une liste non exhaustive :

- **régularité de la vraie loi :** Comme évoqué dans le précédent chapitre, on suppose que deux causes proches engendrent des conséquences proches, et donc que le modèle modélisant la loi doit faire deux prédictions "proche" pour deux entrées "proche", voir [Bengio and Monperrus, 2004, Bengio et al., 2007, Bengio and LeCun, , Bengio, , Glorot and Bengio, , Glorot and Bengio, ].
- **dépendance linéaire des variables explicatives de la vraie loi :** a une certaine granularité, les interactions des variables produisant les descriptions des phénomènes, à l'échelle microscopique, sont assez simples. Si simple que ces interactions fines peuvent être représentées par un modèle linéaire, voir [Goodfellow et al., 2015].
- **la multitude des variables explicatives constituant la vraie loi :** les variables explicatives produisant la vraie loi sont relativement nombreuses. Les techniques de réduction de la dimension cherchent à les concentrer et permettent de les identifier. Un des objectifs est donc de les caractériser et une fois que celles-ci sont satisfaisantes, nous avons obtenu de bonnes caractérisations, nous pouvons plus aisément modéliser numériquement la vraie loi, voir [Bengio, , McCLELLAND et al., ].
- **la hiérarchie des variables explicatives et leurs interactions :** bien que multiples les variables explicatives sont en faite hiérarchisées entre elles. En effet, une variable explicative d'un certain niveau est en faite le produit d'une multitude de "sous-variables" explicatives. Ainsi la vraie loi est une combinaison causale de facteurs explicatifs multiéchelle qui interagissent de

manière pyramidale, voir [Hastad, 1986,Håstad and Goldmann, 1991,Bengio, ,Glorot and Bengio, ,Delalleau and Bengio, 2011].

— **la paramétrisation de la vraie loi est creuse :** pour chaque donnée d'intérêt, le processus qui la générer n'as utilisé qu'une partie des paramètres à sa disposition, voir [Olshausen and Field, 1996].

— **la distribution des données n'est pas uniforme :** en grandes dimensions, les espaces des données possiblement ajustables par le modèle sont très grands, cependant les données d'intérêt sont localisées et concentrées dans un petit sous-espace (c'est une des motivations pour l'utilisation d'opérateur d'aggrégation et pour la recherche sur la réduction de dimension) . Autrement dit, la vraie loi possède des masses singulières d'un point de vue spatial et elles ne sont pas distribuées uniformément dans celui-ci. Ce que nous entendons par données d'intérêts ce sont les données que nous manipulons en pratique, comme des photos de paysage ou bien des textes d'auteur.

  En effet, si nous considérons un tirage aléatoire uniforme d'un vecteur de taille $m \times n$ représentant une image de taille $m \times n$ en nuance de gris où chaque composante est i.i.d, alors il est très probable d'obtenir une image représentant simplement du bruit et il est improbable que nous obtenions une image d'un paysage ayant un sens sémantique. De la même manière pour un texte échantillonnée sur une espace des caractères admissibles assez grands, nous risquons d'obtenir une chaîne de caractère n'ayant aucune signification. Nos données d'intérêt se trouvent dans des localités de manière compacte dans l'espace que le modèle peu explorer. Le modèle doit donc être capable de capter ces masses de la vraie loi dans sa définition, voir [Goodfellow et al., 2015].

— **Les données ont une certaine cohérence spatiale :** nous avons évoqué précédemment un tirage i.i.d des composantes constituant nos données. Mais les composantes des données ont souvent des dépendances. En effet deux pixels proches d'une image, ont de grande chance de partager les mêmes valeurs. Un mot d'une langue répondue a un certain principe et il est courant, par exemple en français d'avoir une voyelle après une consonne. Pour des données temporelles, certaines valeurs à un instant de données peuvent dépendre

de valeurs d'étape précédente, on peut donc caractériser la dépendance spatiale du vecteur décrivant la série comme une dépendance temporelle des données . Nous avons donc la nécessité d'intégrer cette cohérence spatiale de la donnée dans la définition des modèles, voir [Becker and Hinton, 1992].

Il semblerait que les perceptrons multicouches aient les bonnes propriétés pour encoder la plupart de ces a priori statistiques des données contemporaines. Et donc que ces modèles soient particulièrement adaptés pour résoudre un large panel de problème d'apprentissage machine d'aujourd'hui. Nous présentons ici une liste non exhaustive des bonnes propriétés des perceptrons multicouches pour les problèmes d'apprentissage machine d'aujourd'hui.

## 3.7 Les bonnes propriétés des perceptrons multicouches pour l'apprentissage machine contemporain

Pour résoudre les tâches d'apprentissage machine contemporaines, il semblerait que les perceptrons multicouches soient particulièrement adaptés pour définir des modèles d'apprentissage machine moderne efficaces. Nous présentons ici une liste non exhaustive des bonnes propriétés des perceptrons multicouches pour les problèmes d'apprentissage machine d'aujourd'hui.

— **les arguments théoriques :**
  — la définition des perceptrons multicouche permet de satisfaire plusieurs a priori évoquées précédemment. En effet, les perceptrons multicouches sont une mise en cascade (c.-à-d. dépendance et hiérarchisation) d'une multitude de couches (c.-à-d. une multitude de facteurs expliquant à une échelle donnée) composée elle-même d'une multitude (c-à-d interaction de ces multiples facteurs à une échelle donnée) de classifieurs linéaires (c-à-d dépendance linéaire à une granularité donnée et caractère interpolant des modèles linéaires)
  — l'espace des perceptrons multicouches est dense dans l'espace des fonctions continues, voir [Hornik, 1991, Cybenko, 1989]. Autrement dit les

perceptrons multicouches peuvent approximer n'importe quelles fonctions continues [4].

- lorsque le modèle prédictif est un réseau de neurones, le gradient de la fonction de coût associé au modèle est facile à calculer grâce à la méthode de la rétropropagation, voir [Rumelhart et al., 1986].
- l'implémentation facile du principe de parcimonie avec la méthode du dropout, voir [Srivastava et al., 2014].
- le graphe acyclique orienté décrit par la structure des réseaux de neurones s'il est observé du point de vue probabiliste est un modèle graphique modélisant la vraie loi sous-jacente. Les modèles graphiques probabilistes sont un formalisme dédié à la modélisation d'interaction entre variables latentes en jeu dans une loi de probabilité. Les aprioris des données d'intérêt peuvent se résumer par le faite que, une donnée est le produit de variables représentatives latentes, qui sont distribués et hiérarchiques entre elles. La vaste littérature des modèles probabilistes graphiques est alors une précieuse aide pour améliorer la représentativité des réseaux de neurones.

— **les arguments pratiques :**

- les réseaux de neurones ont une définition naturellement parallélisable d'un point de la vue du calcul. En effet, pour chaque élément $\mathbf{x} \in D$, et pour chaque couche $i \in \{1, \cdots, L\}$ le calcul de $f_i(\mathbf{x})$ est parallélisable et nécessite principalement de calculer des produits scalaires. De la même manière, le calcul du gradient est parallélisable pour des raisons analogues et ne nécessite principalement que d'effectuer des calculs de produit scalaire. Les cartes graphiques sont des unités de calcul matérielles dont la conception matérielle a été spécifiquement conçue pour effectuer ces produits matrice-vecteur de manière parallèle. Ainsi nous avons un moyen matériel rapide pour obtenir les inférences des réseaux de neurones ainsi que pour optimiser leurs représentations.
- Nous avons aussi évoqué la nécessité statistique d'avoir une grande quantité de données pour avoir des modèles performants, en particulier pour

4. avec des hypothèses supplémentaires, souvent rencontrer en pratique

les réseaux de neurones. Avec la grande numérisation de notre monde, nous avons accès à un grand volume de donnée, permettant ainsi l'apprentissage en pratique des modèles d'apprentissage machine moderne et en particulier des réseaux de neurones.

Cependant, on ne voit pas comment les réseaux de neurones peuvent intégrer efficacement les propriétés de cohérence spatiale (ou temporelle) de certaines des données que l'on rencontre fréquemment. En effet les opérateurs pour l'instant définis ne semblent pas être particulièrement adaptés à ces données particulières. Dans ces cas où la cohérence spatiale ou temporelle est centrale à intégrer pour résoudre le problème statistique, il faut donc comprendre au mieux les données d'intérêt, leurs propriétés, et utiliser des opérateurs mathématiquement adaptés à ces propriétés.

Un type de donnée qui suppose ce genre d'apriori particulier de cohérence spatiale et qui nécessite une adaptation des opérateurs utilisés dans les perceptrons multicouches est l'image numérique. On considèrera qu'une image numérique est échantillonnée sur 1 octet, et possède les 3 canaux usuels, rouge, vert, bleu. Ainsi une image est une grille régulière d'une taille $m \times n$ et possède un signal qui est une fonction $I : \{1, \cdots, n\} \times \{1, \cdots, m\} \to \{0, \cdots, 255\}^3$.

Souvent on cherche à caractériser le contenu sémantique de l'image lorsque l'on cherche à les classifier. Par exemple, nous pouvons chercher à classifier deux types d'animaux différents comme le chien et le chat. Supposons que nous sommes assis dans un parc et que nous prenons une photo d'un chien qui marche au loin et que nous reprenons un cliché de ce même chien une seconde après le premier cliché. Nous demandons ensuite à un modèle d'apprentissage machine moderne d'inférer la classe des deux photos selon les étiquettes "chien" ou "chat" pour les deux photos. Nous nous attendons alors à ce que le modèle infère deux fois de suite la classe "chien". En effet, le contenu sémantique des deux clichés est le même. Cependant, ce contenu sémantique s'est déplacé spatialement puisque le chien a avancé entre les deux clichés. Une propriété alors souhaitable pour nos opérateurs est qu'il traite de la même manière un signa et son homologue translaté spatialement.

De la même manière, les séries temporelles ont une représentation usuelle sous forme de grille unidimensionnelle et d'un signal qui se propage sur cette grille. Si nous avons, par exemple, une série temporelle représentant des valeurs boursières d'un

actif et que nous souhaitons savoir s'il est bon de vendre ou d'acheter plus d'actifs. A priori cette décision ne dépend pas du temps de manière absolue, elle dépend simplement d'une fenêtre de temps autour du moment temporelle où l'on prend la décision. Autrement dit si nous sommes dans une situation favorable pour l'achat d'actifs a un instant $t$ est que nous sommes dans les mêmes conditions favorables à un instant $t + \Delta t$ alors nous devons aussi acheté des actifs.

Pour répondre à ce genre d'apriori spécifiques pour un perceptron multicouche, nous pouvons par exemple faire du partage de paramètre. C'est-à-dire que pour une couche $f_i$ avec les paramètres $\mathbf{\Omega}_{(i)} = (\omega_{i,j})_{(i,j)\in\{1,\cdots,k_i\}\times\{1,\cdots,k_{i-1}\}}$ au lieu de considérer que :

$$\omega_{i,j} = \omega_{k,l} \Leftrightarrow i = k \text{ et } j = l \tag{3.27}$$

nous considérons plutôt que nous avons un vecteur de paramètre $\boldsymbol{\omega} \in \mathbb{R}^{k_{i-1}}$ tels que :

$$\mathbf{\Omega}_{i,j} = \mathbf{\Omega}_{i+1[k_{i-1}],j+1[k_{i-1}]} = \boldsymbol{\omega}_{j-i+1[k_{i-1}]} \tag{3.28}$$

On dit alors que la matrice $\mathbf{\Omega}_{(i)}$ est circulante. Par exemple pour $k_{i-1} = 5$, on à :

$$\mathbf{\Omega}_{(i)} = \begin{pmatrix} \boldsymbol{\omega}_1 & \boldsymbol{\omega}_2 & \boldsymbol{\omega}_3 & \boldsymbol{\omega}_4 & \boldsymbol{\omega}_5 \\ \boldsymbol{\omega}_5 & \boldsymbol{\omega}_1 & \boldsymbol{\omega}_2 & \boldsymbol{\omega}_3 & \boldsymbol{\omega}_4 \\ \boldsymbol{\omega}_4 & \boldsymbol{\omega}_5 & \boldsymbol{\omega}_1 & \boldsymbol{\omega}_2 & \boldsymbol{\omega}_3 \\ \boldsymbol{\omega}_3 & \boldsymbol{\omega}_4 & \boldsymbol{\omega}_5 & \boldsymbol{\omega}_1 & \boldsymbol{\omega}_2 \\ \boldsymbol{\omega}_2 & \boldsymbol{\omega}_3 & \boldsymbol{\omega}_4 & \boldsymbol{\omega}_5 & \boldsymbol{\omega}_1 \end{pmatrix} \tag{3.29}$$

Donc pour un vecteur $\mathbf{x} \in \mathbb{R}^{k_{i-1}}$, cela produit la transformation suivante :

$$\mathbf{\Omega}_{(i)}\mathbf{x} = (\sum_{j=1}^{k_{i-1}} \boldsymbol{\omega}_{j-m+1\equiv[k_{i-1}]}\mathbf{x}_j)_m \tag{3.30}$$

D'une part, l'aspect circulant permet de rendre la transformation ci-dessus équivariante à la translation. Par exemple, pour une matrice de translation $\mathbf{T}$ décalant les coordonnées d'un vecteur de manière cyclique d'un pas 1 vers la droite :

$$\mathbf{T}\mathbf{x} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \mathbf{x} = (\mathbf{x}_{m-1\equiv[k_{i-1}]})_m \tag{3.31}$$

alors $\mathbf{T}$ commute avec $\mathbf{\Omega}_{(i)}$, c'est à dire :

$$\mathbf{T}\mathbf{\Omega}_{(i)}\mathbf{x} = \mathbf{\Omega}_{(i)}\mathbf{T}\mathbf{x}, \forall \mathbf{x} \in \mathbb{R}^{k_{i-1}} \tag{3.32}$$

En effet, dans notre cas :

$$(\mathbf{T}\mathbf{\Omega}_{(i)}\mathbf{x})_m = \sum_{j=1}^{k_{i-1}} \boldsymbol{\omega}_{j-(m-1)+1[k_{i-1}]}\mathbf{x}_j \tag{3.33}$$

$$= \sum_{j=1}^{k_{i-1}} \boldsymbol{\omega}_{(j+1)-m+1[k_{i-1}]}\mathbf{x}_j \tag{3.34}$$

$$= \sum_{j=2}^{k_{i-1}+1} \boldsymbol{\omega}_{j-m+1[k_{i-1}]}\mathbf{x}_{j-1} \tag{3.35}$$

$$= \sum_{j=1}^{k_{i-1}} \boldsymbol{\omega}_{j-m+1[k_{i-1}]}\mathbf{x}_{j-1} \tag{3.36}$$

or le dernier terme de la somme est $\boldsymbol{\omega}_{k_{i-1}+1-m+1[k_{i-1}]}\mathbf{x}_{k_{i-1}}$ or :

$$\boldsymbol{\omega}_{k_{i-1}+1-m+1[k_{i-1}]} = \boldsymbol{\omega}_{1-m+1[k_{i-1}]} \tag{3.37}$$

car $k_{i-1} + 1 \equiv 1[k_{i-1}]$, donc $\boldsymbol{\omega}_{k_{i-1}+1-m+1}\mathbf{x}_{k_{i-1}} = \boldsymbol{\omega}_{1-m+1}\mathbf{x}_{k_{i-1}}$, c'est-à-dire le cas où $j = 1$ dans la somme, quitte à considérer que $\mathbf{x}_0 = \mathbf{x}_{k_{i-1}}$ ou plus généralement en intégrant la cyclicité dans la notation, c'est-à-dire :

$$(\mathbf{T}\mathbf{\Omega}_{(i)}\mathbf{x})_m = \sum_{j=1}^{k_{i-1}} \boldsymbol{\omega}_{j-m+1[k_{i-1}]}\mathbf{x}_{j-1[k_{i-1}]} = (\mathbf{\Omega}_{(i)}\mathbf{T}\mathbf{x})_m \tag{3.38}$$

Par définition de $\mathbf{Tx}$ et cela pour tout $m \in \{1, \cdots, k_{i-1}\}$, d'où la commutation entre $\mathbf{T}$ et $\mathbf{\Omega}_{(i)}$.

Remarque : On peut aussi simplement dire que l'ensemble de matrices circulantes de taille $n$ , muni du produit matriciel à une structure de groupe abélien, ce qui conclut la démonstration directement, car $\mathbf{T}$ est elle-même un cas particulier de matrice circulante.

Ainsi le faite de partager les paramètres entre les neurones de la couche permet d'avoir cette propriété d'équivariance spatiale. D'un point de vue pratique, si $f_{\boldsymbol{\theta}}$ répond à ce principe d'équivariance alors il traitera de la même manière la reconnaissance d'un chien qui s'est déplacé entre deux photos successives ou bien deux opportunités financières dans le cas de la série temporelle [5]. Dans cette configuration, chaque composante de $\mathbf{x}_{i-1}$ interagit avec chacune des composantes de $f_i$. Or

5. on parle bien ici d'équivariance et non d'invariance

dans une image par exemple, seuls les pixels et ses voisins proches peuvent avoir une certaine cohérence en termes de valeurs. En tout cas ils ont probablement plus de cohérence dans un voisinage contigu à un pixel donné, qu'entre ce même pixel et un autre pixel qui lui est opposé spatialement. C'est un apriori spécifique aux images, mais il doit aussi rentrer en compte dans la conception de $f_{\boldsymbol{\theta}}$. Ainsi il y a deux idées centrales qui émergent :

— il faut conserver la propriété d'équivariance à la translation spatiale en partageant les paramètres

— introduire des fenêtres d'interactions locales, de taille plus petite que l'entrée, qui opère que sur les pixels contigus à un pixel donné, et cela sur tous les pixels

Nous avons vu que les matrices circulant étaient uniquement déterminées par le vecteur $\boldsymbol{\omega}$ dont les coordonnées sont sur la première ligne de la matrice. Pour satisfaire le dernier point, il faut enfaite contraindre à ce que $\boldsymbol{\omega}$ soit creux, c'est-à-dire qu'il existe $p \in \mathbb{N}$ (avec $p < k_{i-1}$) tel que :

$$\|\boldsymbol{\omega}\|_0 = k_{i-1} - p \tag{3.39}$$

Et en imposant que les coefficients nuls de $\boldsymbol{\omega}$ soient les $p$ dernières. Maintenant il faut ajouter une nouvelle contrainte pour ajouter l'aspect contigu de l'interaction locale, pour cela il suffit de considérer que les $p+1$ première lignes de $\boldsymbol{\Omega}_{(i)}$ et de mettre les $k_{i-1} - p - 1$ dernières lignes nulles. Ainsi de cette manière la $m$-ième coordonnée de $\boldsymbol{\Omega}_{(i)}\mathbf{x}$ est combinaison linéaire des éléments qui sont à une distance $k_{i-1} - p - 1$ à droite de la $m$-ième coordonnée de $\mathbf{x}$. On peut utiliser un raisonnement analogue pour des images sauf que la notion de "à droite" ou d'orientation n'est plus claire. Il faut donc adapter le raisonnement sur quelques points. Considérons que la matrice qui représente une image en nuance de gris peut être aplatie, c'est à dire où l'on a concaténé ses lignes les unes après les autres en un grand vecteur colonne. Les images nécessitent d'adapter le raisonnement ci-dessus mais en deux dimensions. Ainsi si notre image est de taille $n \times m$ alors nous obtenons un vecteur ligne $\mathbf{I}$ de taille $nm$. Alors nous devons maintenant considérer non seulement les interactions "horizontales", mais aussi "verticales". Autrement dit nous devons considérer maintenant que $\boldsymbol{\Omega}_{(i)}$ est une forme de matrice circulante par bloque où nous avons appliqué notre principe d'interaction locale contiguë, c'est-

à-dire mettre à zéros certaines lignes à zéros et adapter l'aspect creux en deux dimensions où on considère que nos éléments creux ne sont eux aussi tous nulle à partir d'un certain rang, mais pour les deux dimensions comme supposé dans le cas unidimensionnel, par exemple si nous voulons que la matrice soit creuse selon les lignes d'une quantité $p_1 < n$ et selon les colonnes d'une quantité $p_2 < m$ est en respectant notre paradigme, nous obtenons par exemple dans le cas où $n = 5, m = 4, p_1 = 2, p_2 = 2$ alors notre "vecteur bidimensionnel" $\boldsymbol{\omega}$ est :

$$\boldsymbol{\omega} = \begin{pmatrix} \omega_{1,1} & \omega_{1,2} & 0 & 0 \\ \omega_{2,1} & \omega_{2,2} & 0 & 0 \\ \omega_{3,1} & \omega_{3,2} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{3.40}$$

Par souci de pédagogie, nous avons aplati notre image pour pouvoir faire facilement l'analogie avec le cas unidimensionnel que nous avons détaillé. Nous aplatissons alors aussi notre "vecteur bidimensionnel" $\boldsymbol{\omega}$ et nous obtenons alors aussi un vecteur de taille $mn$ puis nous dérivons de ce vecteur sa matrice circulante comme précédemment pour obtenir, dans ce cas, $\boldsymbol{\Omega}_{(i)} \in \mathbb{R}^{mn \times mn}$, qui est alors défini par :

$$\boldsymbol{\Omega}_{(i)} = \begin{pmatrix}
\omega_{1,1} & \omega_{1,2} & 0 & 0 & \omega_{2,1} & \omega_{2,2} & 0 & 0 & \omega_{3,1} & \omega_{3,2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \omega_{1,1} & \omega_{1,2} & 0 & 0 & \omega_{2,1} & \omega_{2,2} & 0 & 0 & \omega_{3,1} & \omega_{3,2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \omega_{1,1} & \omega_{1,2} & 0 & 0 & \omega_{2,1} & \omega_{2,2} & 0 & 0 & \omega_{3,1} & \omega_{3,2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \omega_{1,1} & \omega_{1,2} & 0 & 0 & \omega_{2,1} & \omega_{2,2} & 0 & 0 & \omega_{3,1} & \omega_{3,2} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \omega_{1,1} & \omega_{1,2} & 0 & 0 & \omega_{2,1} & \omega_{2,2} & 0 & 0 & \omega_{3,1} & \omega_{3,2} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \omega_{1,1} & \omega_{1,2} & 0 & 0 & \omega_{2,1} & \omega_{2,2} & 0 & 0 & \omega_{3,1} & \omega_{3,2} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \omega_{1,1} & \omega_{1,2} & 0 & 0 & \omega_{2,1} & \omega_{2,2} & 0 & 0 & \omega_{3,1} & \omega_{3,2} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \omega_{1,1} & \omega_{1,2} & 0 & 0 & \omega_{2,1} & \omega_{2,2} & 0 & 0 & \omega_{3,1} & \omega_{3,2} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \omega_{1,1} & \omega_{1,2} & 0 & 0 & \omega_{2,1} & \omega_{2,2} & 0 & 0 & \omega_{3,1} & \omega_{3,2} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix} \tag{3.41}$$

Ainsi nous obtenons bien $(p_1 + 1) \times (p_2 + 1) = 9$ lignes d'intérêt et donc $nm - (p_1 + 1) \times (p_2 + 1) = 11$ lignes qui sont mises à zéro pour conserver la propriété de contiguïté. On observera que toutes les $m$ lignent on reboucle du début pour éviter de perdre la contiguïté locale. On observe par ailleurs que $\mathbf{\Omega}_{(i)}$ est doublement circulaire par bloque dans sa version non creuse. Cette propriété permet de conserver l'équivariance par translation. Une fois avoir calculé $\mathbf{\Omega}_{(I)}\mathbf{I}$ il suffit de lui redonner sa forme initiale en taille $n \times m$ est de retenir que les premières $(p_1 + 1) \times (p_2 + 1)$ composantes pour obtenir la transformation de $\mathbf{I}$. ainsi nous avons défini une matrice de poids $\mathbf{\Omega}_{(i)}$ permettant l'équivariance par translation et d'instaurer une propriété d'interactions locales contiguës, qui est un apriori de cohérence spatiale important à considéré dans la représentation de fonction traitant d'image numérique. On observe que l'introduction de cet apriori permet de toujours d'avoir une expression linéaire et donc comme un produit de matrice vecteur. Donc elle garde les propriétés concernant la facilité pour calculer le gradient et le produit en question avec les unités de calculs comme les cartes graphiques mentionées précédemment. L'opérateur linéaire que nous avons défini à l'équation 3.30 s'appelle la convolution et nous avons calculé $\mathbf{x} \star \boldsymbol{\omega}$ ou dans le cas des images $\mathbf{I} \star \boldsymbol{\omega}$. La convolution à une multitude de propriétés utiles en lien avec la transformé de Fourier, le faite que l'opérateur commute, etc.

La convolution permet donc de faire le traitement linéaire de la donnée. Mais par exemple dans le cas du cliché du chien, si nous tenons notre appareil de manière maladroite par exemple, en penchant légèrement notre appareil avant la prise de la photo. Cela entraîne une légère rotation du signal et on aimerait que $f_{\boldsymbol{\theta}}$ soit au moins localement invariant à ces petites transformations. C'est-à-dire que peu importe ces petites transformations locales, on aimerait que $f_{\boldsymbol{\theta}}$ construise la même loi locale. En effet, il est plus important de savoir si un point d'intérêt est présent que de savoir où il se trouve dans une localité pour discriminer des éléments. Malheureusement, la convolution ne permet pas cela. On introduit alors une autre transformation invariante à ces petites transformations, que l'on utilise après convolution. La classe de ce genre opérateurs [6] s'appelle les opérateurs mutualisant ou *pooling* dans la littérature. Ces opérateurs sont des fonctions prennant la valeur maximum ou une la somme de composantes dans une localité donnée, laissant invariant le signal local. L'utilisation

6. non paramétriques, eux

de ces éléments (c-à-d parti convolutif et mutualisant) est sont des moyens pour obtenir une représentation des données étant "stable" par rapport à une classe de transformation décrite par l'apriori que demande les données de type images numériques. Une fois avoir intégré ces aprioris, une fois avoir rendu les représentations des données insensibles à certaines transformations propres au problème, nous pouvons classifier ces nouveaux descripteurs robustes pour construire une loi la plus robuste possible elle aussi. La représentation de ces données étant elles aussi partiellement paramétriques, cela signifie que la "bonne" représentation de la donnée et sa bonne discrimination s'effectuent en même temps. Les réseaux de neurones incluant ces aprioris dans leurs définitions s'appellent les réseaux de neurones convolutifs. Nous l'avons vu ici, la formalisation de ces aprioris apporte un effet régularisant intrinsèque à notre modèle $f_{\boldsymbol{\theta}}$. Prenons maintenant un peu de hauteur concernant le raisonnement que nous venons de faire : nous avons spécifié notre raisonnement conditionnellement à la structure géométrique sous-jacente du domaine des images (c-à-d une grille régulière). L'utilisation de la régularité du domaine à enfaite toujours été implicite et nous a permit de parler informellement "d'horizontalité" ou de "verticalité", ou bien encore de la translation d'éléments. Cette translation est enfaite une transformation elle même "régulière" et sa régularité est dû à la régularité du domaine sous-jacent lui-même. En effet, si la composante $i$ allait à la place de celle en $i+1$ alors cela impliquait que celle en $i+1$ allait en $i+2$ et cela pour tout $i$ à congruences près. Cette spécification est enfaite régie par la nature du domaine et le faite que les composantes $i$ et $i+1$ soit adjacente et que celle en $i+1$ le soit avec celle en $i+2$.

On observe bien l'importance d'intégrer les propriétés du domaine des données contemporaines et leurs signaux associés pour mener à bien la formalisation d'outil mathématique répondant à des régularités décrit par les aprioris. De plus, les aprioris spécifient les régularités des données, en d'autres termes, ils spécifient ce à quoi les données ne sont pas sensibles ou ne change pas sous certaines perturbations. Comme nous l'avons vu, la formalisation de "stabilité" de ces perturbations doit se faire conditionnellement au domaine décrivant la structure géométrique des données. Cette étude de l'apprentissage machine sou l'angle des données contemporaines spécifiques comme les images ou les séries temporelles est l'objet d'étude des

chercheurs en apprentissage géométrique profond ou *geometric deep learning*. C'est l'objet de notre prochain chapitre.

# Chapitre 4

# L'apprentissage géométrique profond

## Sommaire

L'apprentissage géométrique profond [Bronstein et al., 2017] est une branche de recherche de l'apprentissage profond s'intéressant à la formalisation d'outils mathématiques permettant d'obtenir une bonne représentation pour des données admettant une structure géométrique particulière. En exploitant la particularité géométrique de la donnée, il est possible de dériver des opérateurs "insensibles" à une certaine transformation permettant d'apporter de la régularisation aux modèles profonds interagissant avec ce genre de données. Dans un premier temps, nous introduirons la notion de géométrie puis nous l'élargirons au domaine de l'apprentissage machine et particulièrement au domaine de l'apprentissage de la représentation.

## 4.1 La géométrie : de la mesure de notre monde à l'étude des symétries

Dans l'histoire des sciences, la géométrie était utilisée pour modéliser les lois physiques de notre monde. Dans notre monde, nous pouvons mesurer des distances et des angles, nous déplacer selon 3 axes, mesurer des aires ou des volumes, etc. Notre monde a donc une structure fondamentalement euclidienne et peut être formalisé mathématiquement comme un espace euclidien de dimension 3. La géométrie euclidienne a été une formalisation mathématique pour la métrologie de notre monde et a été pendant plusieurs millénaires, *la* géométrie. Elle a été principalement formalisée par Euclide et elle repose sur cinq axiomes décrits dans [Euclide, ]. La plupart des résultats en mécanique classique ont été dressés via l'utilisation de la géométrie euclidienne par exemple. D'autres géométries ont été créées comme la géométrie hyperbolique (Lobatchevski, Gauss, Bolyai) ou encore la géométrie elliptique (Riemann). Certaines de ces géométries ont été utilisées pour décrire des lois de la relativité générale par exemple. Toutes ces géométries reposent sur des bases axiomatiques distinctes, car utilisées dans des contextes distincts. Cependant pour chacune des géométries citées ci-dessus, nous pouvons adopter un autre angle de vue pour les décrire.

Par exemple, pour la géométrie euclidienne, certaines transformations laissent invariantes les mesures faites dans cette géométrie. Par exemple, si nous appliquons une rotation à un segment entre deux-points, alors la longueur de ce segment ne

change pas, c'est-à-dire que "l'écart" entre ces deux points n'a pas changé suite à la transformation. De plus, si nous considérons ce même segment et que nous translatons ce segment, alors il fait toujours la même longueur. Un autre exemple, si nous considérons un triangle et que nous lui appliquons une rotation, alors la valeur de chacun de ses angles est conservée et de même pour la longueur de ses côtés. On appelle ces transformations des isométries, car elles conservent les longueurs. Algébriquement, l'ensemble des isométries d'un espace euclidien à une structure de groupe, c'est le groupe euclidien. Pour les autres types de géométrie, y compris les non euclidiennes, on peut effectuer le même raisonnement et exhiber un groupe de transformation sur lequel certaines propriétés des objets sont invariantes. Et c'est ce que fait Félix Klein dans son programme [Klein, 1893].

Dans ce programme, Felix Klein propose d'approcher les géométries comme étant des espaces topologiques munis de groupes qui agissent transitivement sur ces espaces topologiques. De manière informelle, nous allons introduire ces notions et leurs donner une intuition dans le cas de l'apprentissage géométrique profond [1] :

— un espace topologique est un espace $A$ muni d'une topologie $T$ . Une topologie $T$ est un sous-ensemble des parties de $A$ tel que :
  — $A \in T, \emptyset \in T$
  — $T$ est stable par union finie ou infinie
  — $T$est stable par intersection finie ou infinie

  On peut toujours trouver une topologie pour un ensemble quelconque $A$, en effet $\{A, \emptyset\}$ est une topologie pour $A$. On appelle les éléments de $T$ les ouverts de $T$. Pour un ensemble discret $A = \{1, 2\}$, une topologie possible est $T = \{\emptyset, A, \{1\}, \{2\}\}$. Même si cela suppose de parler d'espace métrique, pour un élément $x$ de $A$, les ouverts sont des moyens de décrire le "voisinage" de $x$, c'est-à-dire une partie de $A$ "plus ou moins proche" de $x$. Par exemple pour décrire le voisinage de 1 on peut considérer tous les ouverts de $T$ qui contiennent 1 comme $\{1\}$ ou $\{1, 2\}$ et on sent alors que décrire 1 par $\{1\}$ est une description plus fine que $\{1, 2\}$. Et on peut parfaitement donner une approche numérique à la notion de voisinage. En effet, lorsque $A$ a une

1. Les concepts présentés dans ce chapitre ont des définitions formelles et ont chacun un cadre théorique très rigoureux, dont la portée est bien au-delà de ce document. La rédaction de ce chapitre a été fortement inspirée de [Bronstein et al., 2021].

structure d'espace métrique (c'est-à-dire muni d'une distance $d$), on peut alors décrire les ouverts de $A$ comme étant l'ensemble des boules ouvertes de $A$, où la boule ouverte centrée en $x \in A$, de rayon $r > 0$ est l'ensemble $B_r(x) = \{y \in A, d(x,y) < r\}$.

— Un groupe est un type de structure algébrique, qui est le couple d'un ensemble $A$ et une loi de composition interne $\circ$. Une loi de composition interne $\circ$ sur $A$ est une application telle que pour tout $(x,y) \in A^2$, alors $x \circ y \in A$. Un groupe est alors le couple $(A, \circ)$ tel que :
  — pour tout $(x,y,z) \in A^3, (x \circ y) \circ z = x \circ (y \circ z)$
  — il existe un unique $e \in A$, tel que pour tout $x \in A, x \circ e = e \circ x = x$
  — pour tout $x \in A$, il existe $y \in A, x \circ y = y \circ x = e$

— une action $\phi$ d'un groupe $(A, \circ)$ agissant sur un ensemble $B$ est une application définie par :

$$\begin{array}{rccc} \phi : & A \times B & \to & B \\ & (x,y) & \mapsto & \phi(x,y) \end{array}$$

tel que
  — pour tout $y \in B, \phi(e,y) = y$
  — pour tout $(x,x') \in A^2, \forall y \in B, \phi(x', \phi(x,y)) = \phi(x' \circ x, y)$

et elle est définie transitive si pour tout $(b,b') \in B^2$ , il existe $a \in A$ tel que :

$$b' = \phi(a,b) \tag{4.1}$$

— de plus, si $B$ à une structure de $\mathbb{R}$-espace vectoriel de dimension $n \in \mathbb{N}$ et que l'action $\phi$ est linéaire, alors on peut décrire les actions $\phi(a,\cdot)$ comme étant une application linéaire et donc par une matrice $\mathbf{\Phi}_a \in GL_n(\mathbb{R})$. L'application $\rho : A \to GL_n(\mathbb{R})$ est alors appelée représentation du groupe $A$ et cette application est compatible avec la loi de $A$ au sens où pour tout $x, x' \in A$ :

$$\rho_\phi(x) \cdot \rho_\phi(x') = \rho_\phi(x \circ x') \tag{4.2}$$

La géométrie de Klein est alors un espace topologique $B$ et un groupe $A$ qui agit via $\phi$ transitivement sur $B$ et de manière "symétrique" dans les termes de Klein. C'est-à-dire que l'action $\phi$ rend invariantes certaines propriétés des points de $B$. Dans son ouvrage [Klein, 1893], Klein y propose une vision "universelle" des géométries. Dans le cadre de l'apprentissage géométrique profond, ce caractère universel

de la description des géométries est donc un moyen pour définir efficacement des opérateurs ayant des propriétés intéressantes pour l'apprentissage sur des structures de données particulières, que nous allons évoquer ci-après.

## 4.2 Les aprioris géométriques

Nous avons vu que pour les données de grandes dimensions et en particulier les données contemporaines, nous avons vu qu'il était impossible d'utiliser un modèle prédictif générique et qu'il était donc nécessaire de spécifier et d'adapter les modèles à la tâche que l'on souhaite résoudre. Cependant, pour la classe de problème impliquant des données issues d'un processus de numérisation d'entités physiques réelles, nous pouvons étudier la géométrie de ces entités, décrire les invariants géométriques et construire des opérateurs numériques compatibles avec ces invariants. Permettant ainsi de spécifier nos modèles prédictifs et donc de les régulariser pour obtenir une bonne représentation des données, en accord avec la structure du problème considéré. Comme pour le programme d'Erlangen, il existe un cadre général pour décrire les contraintes auxquelles les modèles prédictifs opérant sur des données "physiques" doivent répondre. Ils sont décrits dans [Bronstein et al., 2021].

### 4.2.1 Signal et domaine

Dans la suite nous supposerons que l'entité physique qui a été numérisée (par exemple notre donnée contemporaine) peut être décrite conjointement selon deux composantes : un domaine $\mathcal{D}$ et un signal $\mathcal{X}(\mathcal{D}, C)$ selon la modalité $C$. Le domaine $\mathcal{D}$ est la structure physique sur laquelle le signal $\mathcal{X}(\mathcal{D}, C)$ se propage. Le signal $\mathcal{X}(\mathcal{D}, C)$ est une description de $\mathcal{D}$ selon la modalité $C$. Nous pouvons avoir une multitude de modalités pour un même domaine, c'est-à-dire une famille $(C_i)_i$ tel que nous avons une famille de modalité $(\mathcal{X}(\mathcal{D}, C_i))_i$ pour le même domaine $\mathcal{D}$. Ces modalités peuvent être indépendantes. Formellement, on définit $\mathcal{X}(\mathcal{D}, C)$ par l'ensemble des applications $s : \mathcal{D} \to C$ où $C$ à une structure d'espace de Hilbert et où $\mathcal{X}(\mathcal{D}, C)$ à une structure d'espace vectoriel.

Un exemple concret d'une entité physique numérisée est l'image numérique de taille $n \times m$ que nous avons décrite précédemment, tel que :

— Le domaine $\mathcal{D} = \{1, \cdots, n\} \times \{1, \cdots, m\}$ décrit sous forme d'une grille régulière de taille $n \times m$.

— Le signal est décrit selon trois modalités (le canal rouge, vert, bleu), des éléments de $\mathcal{X}(\mathcal{D}, \mathbb{R})$.

Dans la suite, nous considérons que les modalités sont indépendantes entre elles, donc nous notons simplement le signal se propageant sur $\mathcal{D}$ par $\mathcal{X}(\mathcal{D})$ et que les résultats prochains s'appliquent à chaque modalité, sauf mention contraire.

Dans le cas des signaux sur $\mathcal{D}$, comme l'espace $\mathcal{X}(\mathcal{D})$ admet une structure d'espace vectoriel réel, nous pouvons alors pour chaque élément $x$ d'un groupe $A$ agissant par $\phi$ sur $\mathcal{D}$ lui associer une représentation $\rho(x) \in GL_n(\mathbb{R})$ et définir une action agissant sur $\mathcal{X}(\mathcal{D})$, tel que pour tout signal $s \in \mathcal{X}(\mathcal{D})$ :

$$\rho(x)s(d) = s(\phi(x^{-1}, d)), \forall d \in \mathcal{D} \tag{4.3}$$

Cela est bien une représentation, car si nous prenons $x, x' \in A, s \in \mathcal{X}(\mathcal{D}), d \in \mathcal{D}$, nous avons :

$$\rho(x) \cdot (\rho(x')s(d)) = \rho(x) \cdot (s(\phi(x'^{-1}, d))) \text{ par définition 4.3} \tag{4.4}$$

$$= s(\phi(x'^{-1}, \phi(x^{-1}, d))) \text{ par définition 4.3} \tag{4.5}$$

$$= s(\phi(x'^{-1} \circ x^{-1}, d)) \text{ car } \phi \text{ est une action sur } A \text{ et d'après 4.2} \tag{4.6}$$

$$= s(\phi([x \circ x']^{-1}, d)) \text{ car le produit des inverses est l'inverse du produit commuté} \tag{4.7}$$

$$= \rho(x \circ x') \cdot s(d) \text{ par définition 4.3} \tag{4.8}$$

Donc pour tout $s \in \mathcal{X}(\mathcal{D})$, nous avons bien que

$$\rho(x) \cdot \rho(x') \cdot s(d) = \rho(x \circ x') \cdot s(d) \tag{4.9}$$

pour tout $d \in \mathcal{D}$. De plus cette représentation est bien linéaire sur $\mathcal{X}(\mathcal{D})$, car pour tout $s_1, s_2 \in \mathcal{X}(\mathcal{D})$, on sait que pour tout réel $\alpha, \beta \in \mathbb{R}, \alpha s_1 + \beta s_2$ à un sens, car $\mathcal{X}(\mathcal{D})$ à une structure d'espace vectoriel réel, ainsi pour tout $d \in \mathcal{D}, x \in A$ on a :

$$\rho(x)(\alpha s_1 + \beta s_2)(d) = (\alpha s_1 + \beta s_2)(\phi(x^{-1}, d)) \tag{4.10}$$

$$= \alpha s_1(\phi(x^{-1}, d)) + \beta s_2(\phi(x^{-1}, d)) \text{ car } \mathcal{X}(\mathcal{D}) \text{ est un } \mathbb{R} \text{ espace vectoriel} \tag{4.11}$$

$$= \alpha\rho(x) \cdot s_1(d) + \beta\rho(x) \cdot s_2(d) \text{ par définition 4.3} \tag{4.12}$$

L'application définie à l'équation 4.3 est alors une action de $GL_n(\mathbb{R})$ sur $\mathcal{X}(\mathcal{D})$ définie à partir d'une action de $A$ sur $\mathcal{D}$. Ainsi cette nouvelle action est un moyen de passage entre le domaine $\mathcal{D}$ et $\mathcal{X}(\mathcal{D})$. Autrement dit, c'est un moyen algébrique de décrire la géométrie de $\mathcal{D}$ sur $\mathcal{X}(\mathcal{D})$ et d'avoir un moyen d'effectuer des opérations sur $\mathcal{D}$ et d'en avoir les résultats sur $\mathcal{X}(\mathcal{D})$.

En effet, en pratique lorsque nous traitons ce type de donnée, nous avons plus souvent accès à une description du signal et nous n'avons qu'une description induite du domaine. Nous avons donc un moyen d'agir sur le domaine via le signal. Et cela est notamment intéressant pour définir les "insensibilités" (ou les "symétries" dans les termes de Klein) de notre modèle prédictif.

### 4.2.2 Equivariance et invariance

Il y a deux notions principales dans l'apprentissage géométrique profond, deux propriétés que l'on aimerait que notre modèle prédictif inclue dans sa définition pour construire des représentations robustes des données. Il s'agit de la propriété d'équivariance et d'invariance.

On dit qu'une fonction $f : \mathcal{X}(\mathcal{D}) \to \mathcal{Y}$ est $A$-invariante si pour tout $s \in \mathcal{X}(\mathcal{D}), x \in A$ :

$$f(\rho(x) \cdot s) = f(s) \tag{4.13}$$

et que $f$ est $A$-équivariante si pour tout $s \in \mathcal{X}(\mathcal{D}), x \in A$ :

$$f(\rho(x) \cdot s) = \rho(x) \cdot f(s) \tag{4.14}$$

Premièrement il faut noter que ces notions d'invariance et d'équivariance sont bien définies lorsque nous avons une définition exacte de $A$, de ses éléments et de $\mathcal{D}$.

Mais cela n'est pas évident dans certains cas pratiques : en effet, reprenons le cas de la photo du chien marchant dans le parc, si le temps d'obturation de l'appareil photo est trop grand, alors nous obtenons une photo floue et ainsi l'image que nous obtenons n'est plus une translation stricte de l'image obtenue lorsque le

chien marche. Nous obtenons quelque chose qui s'apparente de loin à une translation mais plus à une translation stricte. Ainsi il est nécessaire de généraliser la notion d'invariance et d'équivariance.

### 4.2.3 Stabilité géométrique et description locale

Comme évoqué, les transformations que nous observées en pratique ne sont parfois pas des transformations géométriques exactes au sens d'éléments d'un groupe de transformations. Ainsi, il faut adapter la notion d'équivariance et d'invariance à ce genre de cas, il faut alors définir une notion d'équivariance approximative et d'invariance approximative. L'idée est alors de considérer les transformations proches (par exemple, la translation floue) de la même manière que celles définies strictement au sens du groupe de transformation. Une chose que l'on constate cependant est que les transformations approximatives que nous observons en pratique sont des transformations souvent assez régulières et bijectives. On les appelle des difféomorphismes. L'ensemble des difféomorphismes pour des espaces bien choisis a une structure de groupe mais l'ensemble des "petits" difféomorphismes n'en est pas un. En effet, on peut trouver des cas où le composé de plusieurs petits difféomorphismes est un difféomorphisme entraînant une grande transformation. Nous n'avons donc pas la stabilité de la composition, ce qui empêche donc d'avoir la structure de groupe. Les transformations comme les translations, les rotations sont des difféomorphismes dans les espaces euclidiens. Mais si notre fonction $f$ est invariante aux translations par exemple, alors il existe bien une classe de difféomorphisme à laquelle notre fonction est insensible par définition. En effet, dans le cas d'espace euclidien, les translations forment un sous-groupe de l'ensemble des difféomorphismes des espaces considérés. L'idée est alors d'avoir une certaine tolérance pour les difféomorphismes "proches" de ceux appartenant au groupe auquel $f$ doit être insensible. Pour être plus précis sur la notion de "petit" ou "proche", nous devons munir $\mathcal{X}(\mathcal{D})$ d'une norme $\|\cdot\|^{(1)}$ et $f(\mathcal{X}(\mathcal{D}))$ d'une autre norme $\|\cdot\|^{(2)}$. Ainsi, nous redéfinissons la notion d'$A$-invariance d'une fonction $f$, par la notion $A$-invariance approximative de rapport modulé par $c$, si pour tout $x \in Diff(\mathcal{D})$ et tout $s \in \mathcal{X}(\mathcal{D})$ , il existe $C \in \mathbb{R}^+$ tel que :

$$\|f(\rho(x) \cdot s) - f(s)\|^{(2)} \leq Cc(x)\|s\|^{(1)} \tag{4.15}$$

Où $c : Diff(\mathcal{D}) \to \mathbb{R}^+$ est une fonction positive tel que $c(x) = 0 \Leftrightarrow x \in A$. Cette nouvelle définition est plus générale et compatible avec la définition de $A$-invariance. En effet si $x \in A \Rightarrow c(x) = 0 \Rightarrow \|f(\rho(x) \cdot s) - f(s)\|^{(2)} = 0 \Rightarrow f(\rho(x) \cdot s) = f(s)$

On peut de manière analogue définir la notion d'$A$-équivariance approximative d'une fonction $f$ de rapport modulé par $c$ si pour tout $x \in Diff(\mathcal{D})$ et tout $s \in \mathcal{X}(\mathcal{D})$ , il existe $C \in \mathbb{R}^+$ tel que :

$$\|f(\rho(x) \cdot s) - \rho(x) \cdot f(s)\|^{(2)} \leq Cc(x)\|s\|^{(1)} \tag{4.16}$$

Et là aussi, et pour les mêmes raisons, la notion d'$A$-équivariance approximatif est compatible avec celle de $A$-équivariance. Si $f$ respecte ces propriétés, alors elle est dite géométriquement stable.

Il y a un autre apriori que les données d'intérêt présupposent souvent dans leurs définitions. Comme évoqué précédemment, les entités physiques qui nous entourent admettent souvent une caractérisation multiéchelle, c'est-à-dire que ces entités sont la résultante d'une composition d'autres entités, mais locales, comme évoqué dans le chapitre précédent. Et ces entités locales peuvent admettre elles aussi, à leur niveau, des symétries. Ainsi, pour une bonne représentativité de la donnée, il est nécessaire de conserver l'information contenue dans chacune de ces entités afin de pouvoir reformer la structure multiéchelle sous-jacente. Cependant, si ces entités locales subissent une transformation, on aimerait que le modèle conserve l'information contenue dans ces entités locales par rapport au groupe de symétries considéré. Ainsi, nous souhaitons que la fonction traitant ces informations locales ait la propriété de $A$-équivariance (ou dans sa version approximative). Par ailleurs, non seulement il est nécessaire d'avoir ces descriptions locales, mais il faut aussi considérer comment elles interagissent entre elles, car nécessaires à la définition de l'entité globale. Il faut donc que notre fonction ait une structure récursive permettant la composition l'information des entités locales menant à la bonne représentation de la donnée. Dans le cadre de l'apprentissage géométrique profond, cet apriori s'appelle la séparation d'échelle. L'idée est d'alors de "concentrer" l'information contenue dans chaque entité locale et de considérer cette information concentrée comme une nouvelle entité, puis de réappliquer un traitement $A$-équivariant (ou approximativement $A$-équivariant) pour obtenir l'interactivité entre les entités locales précédentes. C'est

une idée analogue à celle du chaînage des descripteurs des perceptrons multicouche, mais pour la représentation de données ayant un apriori géométrique.

### 4.2.4 La séparation d'échelle et l'invariance finale

La structure physique des données présuppose d'une définition multiéchelles dépendantes entre elles. Ainsi la fonction qui modélise la représentation de la donnée doit admettre une définition récursive (c-à-d chaînage des opérateurs) et synthétique (c-à-d utilisation de la réduction de dimension : opérateur de mutualisation). En pratique cela se fait en composant des fonctions conservant l'information des descripteurs locaux (c-à-d application équivariante et application mutualisante) à une échelle donnée. Ainsi ce processus d'agrégation et de synthétisation permet de construire les dépendances, ou les interactions des différents descripteurs locaux qui composent le descripteur global ; car nécessaires à sa bonne définition. Une fois que nous avons construit une représentation de la donnée d'intérêt comprenant assez d'informations issues des descripteurs locaux et de leurs interactions, nous pouvons appliquer une opération $A$-invariante de manière à rendre la représentation de la donnée invariante à certaines transformations comme présenté initialement dans ce chapitre. L'ensemble de ces aprioris forme un ensemble de conditions nécessaires, mais pas suffisantes pour la définition de fonction capable de produire une "bonne" représentation de donnée admettant des régularités géométriques.

### 4.2.5 Les modèles d'apprentissage géométrique profond

De manière plus formelle, un modèle d'apprentissage géométrique profond de profondeur $L \in \mathbb{N}$ traitant une donnée décrite par :

— un domaine $\mathcal{D}_1$ admettant des symétries par rapport au groupe $A$

— un signal $s \in \mathcal{X}(\mathcal{D}_1, C_1)$ où $C_1$ est de dimension $d \in \mathbb{N}$

est la donnée d'un $L$-uplet $(k_1, \cdots, k_L) \in \mathbb{N}^L$ et d'une fonction $g : \mathcal{X}(\mathcal{D}_1, C_1) \to \mathcal{Y}$ où $Y$ est un espace de Hilbert de dimension $p \in \mathbb{N}$ telle que :

$$g(s) = s^{(L)} = I(\sigma_L(E_L(s_{(L-2)}))) \tag{4.17}$$

tel que :

$$s_{(L-2)} = \begin{cases} s^{(0)} = s \\ s^{(j)} = P_j(\sigma_j(E_j(s_{(j-1)}))) \text{ pour } 1 \leq j \leq L-2 \end{cases} \tag{4.18}$$

où pour tout $j$ on a :

- $E_j$ est une application approximativement $A$-équivariante tel que $E_j : \mathcal{X}(\mathcal{D}_j, C_j) \to \mathcal{X}(\mathcal{D}_{j+1}, C_{j+1})$ où $C_j$ est de dimension $k_j$ et $C_{j+1}$ de dimension $k_{j+1}$.
- $\sigma_j$ est une application non linéaire $\sigma_j : C_{j+1} \to C_{j+1}$.
- $P_j$ est une application de projection définie par $P_j : \mathcal{X}(\mathcal{D}_{j+1}, C_{j+1}) \to \mathcal{X}(\mathcal{D}_{j+2}, C_{j+1})$ tel qu'informellement $\mathcal{D}_{j+2} \subset \mathcal{D}_{j+1}$
- $I$ est une application approximativement $A$-invariante tel que $I : \mathcal{X}(\mathcal{D}_L, C_L) \to \mathcal{Y}$
- d'un point de la vue de la terminologie, l'application qui à $s^{(j-1)}$ associe $s^{(j)}$ est appelée "couche" de rang $j$.

Notons que $E_j$ peut être paramétrique pour certains $j \in \{1, \cdots, L\}$ si la représentation de donnée dépend d'un paramètre $\boldsymbol{\theta}_j$ , ce qui est le cas en pratique dans le cadre de l'apprentissage géométrique profond où l'on cherche justement à construire une bonne représentation de la donnée en optimisant ce paramètre. Si le processus d'optimisation se fait via une méthode du premier ordre alors on suppose $E_j$ et $\sigma_j$ différentiable sur les espaces d'intérêts. Et $g$ est alors paramétrique par construction de paramètre $\boldsymbol{\theta} = (\boldsymbol{\theta}_j)_{j \in \{1,\cdots,L\}}$. Illustrons maintenant quelques structures de données usuelles, leurs géométries et leur symétrie respective.

## 4.3 Quelques données usuelles avec apriori géométrique

Dans les types de données que nous rencontrons dans les problèmes d'apprentissage statistiques contemporains avec des aprioris géométriques, il y a les images ou les séries temporelles pour ne citer que les plus usuelles. Ces deux types de données ont un domaine décrit sous forme de grille régulière unidimensionnelle pour les séries temporelles, et bidimensionnelles (ou tridimensionnelles pour les images IRM par exemple) pour les images numériques. Les signaux évoluant sur ces domaines

sont souvent en pratique multidimensionnels et à valeurs réelles. En imagerie médicale, où encore en informatique graphique, les domaines des données peuvent être décrits par des variétés différentielles qui ont été numérisées en maillage. Tous ces types de données dans leur version numérisée reposent sur une structure commune : la structure de graphe.

Un graphe est une structure de donnée utilisée dans de nombreux domaines pour modéliser des informations relationnelles entre des entités. Formellement un graphe $G$ est un couple $(V, E)$ où $V$ est un ensemble appelé l'ensemble des noeuds de $G$, et $E \subset V^2$ est l'ensemble des arrêtes modélisant les liens entre les éléments de $V$. Définir un graphe c'est définir un ensemble d'éléments et la manière dont ils sont reliés entre eux. Bien entendu, $V$ ne présupposent pas d'une structure particulière et la seule contrainte que $E$ doit respecter est d'être un sous-ensemble de $V^2$. Remarque : on considère en pratique l'énumération de $V$, qui est une partie de $\mathbb{N}$ plutôt que $V$ elle-même pour représenter les noeuds du graphe $G$.

Cette définition est la définition d'un graphe homogène. Il existe des graphes non homogènes, dit hétérogènes ou encore hypergraphes lorsque $E \subset V^n$ pour un entier naturel $n \geq 3$. Dans la suite, les graphes seront tous considérés comme homogènes sauf mention contraire. Il existe aussi des graphes où $V$ est un multiensemble, c'est-à-dire que deux noeuds $u$ et $v$ peuvent avoir plusieurs liens distincts. Nous ne considérerons pas non plus ce genre de graphe dans la suite.

Dans le contexte de l'apprentissage géométrique profond, on peut munir un signal à $V$ à valeurs dans $\mathbb{R}^d$ et un signal à $E$ a valeurs dans $\mathbb{R}^p$ où $d, p \in \mathbb{N}$. Ainsi on définit un signal sur $V$ par l'application $s_V : V \to \mathbb{R}^d$ et celle sur $s_E : V^2 \to \mathbb{R}^p$. On associe aussi à $V^2$ une autre application $a : V^2 \to \mathbb{R}$, décrivant la connectivité du graphe.

L'application $a$ mesure la "force" de connexion entre deux éléments $(v, u) \in V^2$, tel que par exemple pour tout $(u, v) \in V^2$ :

$$a(u, v) = w_{u,v} \in \mathbb{R} \tag{4.19}$$

tel que

$$w_{u,v} = 0 \Leftrightarrow (u, v) \notin E \tag{4.20}$$

Et que la valeur de $|w_{u,v}|$ est proportionnelle à la force de connexion entre $u$ et $v$. Dans ce cas, on dit que $G$ est un graphe pondéré. De plus si $a$ est symétrique sur $V^2$ alors

on dit que le graphe est orienté. Dans la suite du document, nous considérerons que les graphes ne sont pas orientés. Pour modéliser les relations d'interactivité entre les noeuds (dans le cas non orienté), on peut aussi simplement utiliser la caractérisation décrite dans l'équation 4.20, tel que pour tout $u, v \in V^2$ :

$$a(u,v) = w_{u,v} = \mathbf{1}_{(u,v)\in E}(u,v) \in \{0,1\} \tag{4.21}$$

L'application $s_E$ quant à elle, peut servir à donner des informations caractérisant non seulement la présence de connectivité, mais aussi la connectivité en elle-même. On peut penser par exemple à un graphe qui représente une molécule chimique dont les atomes qui la constituent forment l'ensemble des noeuds de $G$ et les liaisons covalentes définissent l'ensemble des arrêtés de $G$ et tel que pour chaque arrêté on ait des mesures de grandeur physique comme la polarité, l'électronégativité, etc. Ce cas-là peut être aussi vu sous la forme d'un hypergraphe où chaque type de grandeur physique est associé à un graphe homogène pondéré. On représente algébriquement l'information de connexion d'un graphe homogène par une matrice $\mathbf{A} \in \mathbb{M}_{|V|}(\mathbb{R})$ dite d'adjacence, où les entrées de $\mathbf{A}$ sont

$$\mathbf{A}_{u,v} = a(u,v) = w_{u,v} \in \mathbb{R}^+ \tag{4.22}$$

pour tout $(u,v) \in \{1, \cdots, V\}^2$.

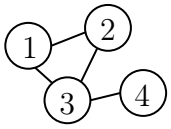
Dans le cas d'un graphe pondéré, cette représentation caractérise complètement le graphe et elle est complètement équivalente à la définition ensembliste du graphe. Dans la géométrie de Klein, on peut trouver une topologie pour chacun des graphes. Ainsi chaque graphe décrit un espace topologique. Par exemple, si $G$ est connexe et que son adjacence est définie de manière binaire, tel que pour tout $u, v \in V$ :

$$d(u,v) = \min\{l(p) | p \in P_{u,v}\} \tag{4.23}$$

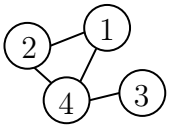
Où $P_{u,v}$ est l'ensemble des chemins entre $u$ et $v$ et/ou pour tout $p \in P_{u,v}, l'p) \in \mathbb{N}$ est le nombre d'arrêtés que contient $p$. Alors $(G,d)$ définie un espace métrique et la topologie est l'ensemble des boules décrit par $d$ sur $G$.

Pour un graphe à $n$ noeuds, vus comme espace topologique, le groupe des symétries de cet espace est le groupe des permutations $\Sigma_n$. En effet, dans le cas où $|V|$ est finie, disons de taille $n \in \mathbb{N}$, l'énumération des éléments de $V$ est parfaitement arbitraire, de telle manière que si on permute l'énumération, alors on

décrit toujours le même graphe, seule l'information relationnelle entre les noeuds de $G$ compte. Considérons, par exemple, un graphe $G_1$ tel que $V = \{1, 2, 3, 4\}$ et $E = \{(1, 2), (2, 3), (3, 1), (3, 4)\}$ et un graphe $G_2$ où nous permettons l'indice 1 avec l'indice 2 et l'indice 3 avec l'indice 4, alors nous avons :




(a) $G_1$ - Choix d'énumération de $V$ n°1




(b) $G_2$ - Choix d'énumération de $V$ n°2

Malgré le choix d'énumération n°1 ou n°2, qui sont parfaitement arbitraires, nous représentons le même graphe, les mêmes connectivités entre les entités décrites par $V$. On dit que ces deux graphes sont isomorphes. Dans notre cas, nous avons utilisé la permutation de $\pi \in \Sigma_4$ :

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix} \tag{4.24}$$

Ainsi, si nous décrivons $V$ comme un vecteur $\begin{pmatrix} 1 & 2 & 3 & 4 \end{pmatrix}^T$ est que nous appliquons la représentation de $\pi$ , alors on a :

$$\rho(\pi) \cdot \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 4 \\ 3 \end{pmatrix} \tag{4.25}$$

Et nous obtenons bien le choix d'énumération n°2 en appliquant $\pi$ aux choix d'énumération n°1. Cependant, cela a aussi une incidence sur les descriptions matricielles entre $G_1$ et $G_2$, en effet la matrice d'adjacence de $G_1$ n'est pas égale à la matrice d'adjacence de $G_2$. Cependant si nous notons $\mathbf{A}_{G_1}$ la matrice d'adjacence de $G_1$ et par $\mathbf{A}_{G_2}$ de $G_2$, alors elles sont reliées par la relation suivante :

$$\mathbf{A}_{G_2} = \rho(\pi) \cdot \mathbf{A}_{G_1} \cdot \rho(\pi)^T \tag{4.26}$$

Donc dans le cas des graphes, effectuer une opération sur $G = (V, E)$ implique de faire une opération conjointement sur $V$ et $E$. Les fonctions opérant sur des graphes doivent donc opérer conjointement sur les deux ensembles. Ainsi de manière générale,

pour un graphe $G = (V, E)$ à $n$ noeuds ayant une matrice d'adjacence $\mathbf{A}$, un signal $s_V$ évoluant sur les noeuds, une fonction $f$ $\Sigma_n$-invariante sur $G$ si pour tout $\pi \in \Sigma_n$ :

$$f(\rho(\pi) \cdot s_V, \rho(\pi) \cdot \mathbf{A} \cdot \rho(\pi)^T) = f(s_V, \mathbf{A}) \tag{4.27}$$

De manière analogue, pour ce même graphe $G$, une fonction $f$ $\Sigma_n$-équivariante sur $G$ si pour tout $\pi \in \Sigma_n$ :

$$f(\rho(\pi) \cdot s_V, \rho(\pi) \cdot \mathbf{A} \cdot \rho(\pi)^T) = \rho(\pi) \cdot f(s_V, \mathbf{A}) \tag{4.28}$$

Par extension, toutes fonctions traitantes des signaux $s_V \in \mathbb{R}^{n \times d}$ définie sur des domaines décrit par des graphes à $n$ noeuds doivent être $\Sigma_n$-invariante.

Nous avons vu que pour obtenir une bonne représentation de la donnée, il était nécessaire d'enchevêtrer des représentations locales et de les assembler pour produire la représentation finale. Dans le contexte des graphes, nous définissons la notion de localité d'un élément $v \in V$ comme la notion de voisinage de $x$. C'est-à-dire , pour tout $v \in V$, on définit le voisinage de rang $k \in \mathbb{N}$ (ou $k$-voisinage) de $v$ par $N_k(v)$ tel que :

$$N_k(v) = \{u | l(p) = k, p \in P_{u,v}\} \tag{4.29}$$

Le traitement du signal $s_V|_{N_k(i)}$ au $k$-voisinage du $i$-ème noeud est donc un moyen explicite de construire une représentation locale à la granularité $k$ du noeud $i$. Une description à un niveau de granularité $k = 1$ est la description la plus fine que l'on puisse obtenir donc on la privilégiera de manière à pouvoir construire une représentation de la donnée la plus hiérarchisée. Pour un signal $s_V \in \mathbb{R}^{n \times p}$ définie sur $V$ où pour chaque noeud $i$, est associé un vecteur $\mathbf{x}_i \in \mathbb{R}^p$ qui est constitué des $p$ descripteurs du noeud $i$ ; on peut construire une fonction $F$ $\Sigma_n$-équivariante en appliquant en parallèle pour tout noeud $i$ la même fonction $\phi$ $\Sigma_{|N_1(i)|}$-invariante traitant le signal $s_V|_{N_1(i)}$, c'est-à-dire :

$$F(s_V, \mathbf{A}) = \begin{bmatrix} \phi(s_V|_{N_1(1)}) \\ \phi(s_V|_{N_1(2)}) \\ \phi(s_V|_{N_1(3)}) \\ \vdots \\ \phi(s_V|_{N_1(n)}) \end{bmatrix} \tag{4.30}$$

Le fait que, pour tout $i$, la valeur de $\phi(s_V|_{N_1(i)})$ soit indépendante de l'ordre des noeuds dans $N_1(i)$ (car $\phi$ est $\Sigma_{|N_1(i)|}$-invariante part construction) et que l'on partage cette même fonction pour chaque noeud $i$, permet de rendre $F$ $\Sigma_n$-équivariante. En respectant ensuite le schéma des modèles d'apprentissage géométrique profond décrit à l'équation 4.17, on peut alors construire une fonction $g$ adaptée pour la représentation de données sur graphe. Ces modèles sont souvent dénommés réseaux de neurones sur graphe ou Graph Neural Network (GNN) dans la littérature.

## 4.4 Les réseaux de neurones sur graphes

Les réseaux de neurones sur graphes sont des réseaux de neurones répondant au paradigme de l'apprentissage géométrique profond pour les données de type graphes. Le groupe de symétrie étant ici le groupe des permutations. Leur introduction est antérieure à la formalisation de l'apprentissage géométrique profond tout comme l'ont été la géométrie euclienne et celle de Klein. Ainsi la couche définie à l'équation 4.30 est globalement ce qui caractérise la définition d'un GNN. Il existe une multitude de définitions pour une couche d'un GNN qui respecte la stabilité géométrique, mais la plupart des variantes sont dérivées d'un modèle général introduit dans [Gilmer et al., 2017a, Battaglia et al., 2018a]

L'implémentation générale d'une couche au rang $j$ d'un GNN repose sur le principe du "passage de message", où en pratique où voit un graphe décrit au rang $j$ comme un réseau de communications entre ses noeuds, et où on considère que deux noeuds $i$ et $i'$ "conversent" entre eux via des messages. Ces messages sont leurs descripteurs vectoriels $\mathbf{x}_i$ et $\mathbf{x}'_i$ au rang $j$ et la conversation est matérialisée par $\psi(\mathbf{x}_i, \mathbf{x}'_i)$ où $\psi$ est une fonction qui souvent dépend du paramètre $\boldsymbol{\theta}_{j,i,i'}$ et ayant de bonnes propriétés de régularité. Ainsi pour obtenir la nouvelle représentation de la donnée du noeud $i$, au niveau $j$ on effectue le calcul suivant :

$$\mathbf{x}_i = \phi\left(\mathbf{x}_i, \bigoplus_{i' \in N_1(i)} \psi(\mathbf{x}_i, \mathbf{x}'_i)\right) \tag{4.31}$$

Où $\bigoplus$ représente un opérateur non paramétrique invariant aux permutations comme une somme, une moyenne ou un maximum/minimum, etc.

Nous allons maintenant détailler les principales variantes de ces modèles et donner des formulations explicites de fonction $\Sigma_n$ équi/in-variante.

Depuis ce modèle "communicatifs" ont été dérivés deux modèles principaux :

— l'un où

$$\psi(\mathbf{x}_i, \mathbf{x}'_i) = \tilde{a}_{i,i'}\psi'(\mathbf{x}'_i) \tag{4.32}$$

Où $\tilde{a}_{i,i'} \in \mathbb{R}$ est uniquement dépendant de l'adjacence du graphe, en pratique égale à l'inverse de la racine carrée des produits des degrés du noeud $i$ et $i'$ et où $\psi'$ est une fonction paramétrique ne dépendant que du vecteur de $\mathbf{x}'_i$. Les GNN utilisant cette formulation sont qualifiés de réseau convolutionnel sur graphe ou encore de *graph convolutional network* dans la littérature, plus largement décrit dans [Kipf and Welling, 2016a, Defferrard et al., 2016, Wu et al., 2019a]

— l'autre où.

$$\psi(\mathbf{x}_i, \mathbf{x}'_i) = b(\mathbf{x}_i, \mathbf{x}'_i)\psi'(\mathbf{x}'_i) \tag{4.33}$$

Où $b$ est une fonction paramétrique implémentant le mécanisme "d'attention", représentant "l'importance" du lien entre le noeud $i$ et $i'$. Ce modèle est qualifié de réseau de neurones attentionnels sur graphe ou de *graph attention networks*. Ces modèles sont plus largement décrits dans [Veličković et al., 2018a, Monti et al., 2016, Zhang et al., 2018b].

— GraphSage [Hamilton et al., 2017], Graph Isomorphism Network (GIN) [Xu* et al., 2018] sont aussi des fonctions équivariantes largement utilisées dans la communauté.

Il existe différents types de tâches pour les réseaux de neurones sur graphe, les plus communes sont les tâches de classification de noeuds, de graphes et de la prédiction de lien entre des noeuds. De plus il existe d'autres types de modèles pour définir les couches d'un GNN, et les GNN sont utilisés pour résoudre diverses tâches, en voici une vue globale.

## 4.5 État de l'art relatif aux réseaux de neurones sur graphes

Il existe d'autres types de couches pour les réseaux de neurones sur graphes. Nous allons les évoquer dans cette partie, les principales contributions relatives aux réseaux de neurones sur graphes. Notez que ces contributions ne sont pas toutes en lien avec nos travaux de manière directe et donc non pas toutes reçues une étude approfondie. Cependant, elles sont majeures dans la littérature des réseaux de neurones sur graphes. Nous listons les principales dans l'Annexe B.2.2.

Les réseaux de neurones convolutionnels ont montré leur suprématie dans le domaine du traitement de l'image. Cependant, il existe des comportements de ces modèles qui sont difficiles à comprendre comme par exemple [Su et al., 2019] ou encore [Szegedy et al., 2014]. Bien que les chercheurs aient tenté de caractériser ces comportements interpellant, il reste délicat de pleinement expliquer le comportement des réseaux de neurones, principalement dû à leur sur-paramétrisation. Ainsi, l'utilisation de ce genre de modèles, bien que performants, pour résoudre des tâches sensibles comme dans le domaine du médical, de la navigation autonome doivent être controlés. Il y a d'autres interrogations, mais une partie des chercheurs en apprentissage machine se sont intéressés à ces question d'explicabilité des réseaux de neurones. C'est l'objet du prochain chapitre.

# Chapitre 5

# Explicabilité des modèles issus de l'apprentissage machine

## Sommaire

Du fait de la jeunesse de l'utilisation massive des réseaux de neurones pour résoudre des problèmes d'apprentissage machine contemporain, la recherche en explicabilité de ces modèles est a fortiori encore très jeune. Dans ce chapitre, nous allons mettre la lumière sur les travaux réalisés pour l'explicabilité de ce genre de modèles mais aussi pour les modèles non-profonds [1].

## 5.1 Explicabilité, interprétabilité, paradigmes, propriétés et mesures

Lorsque nous terminons un raisonnement prédictif, nous obtenons une prédiction. Puis nous avons deux issues possibles : soit nous nous intéressons à pourquoi nous obtenons cette prédiction, c'est à dire quels sont les éléments dans notre raisonnement qui nous ont menés à la valeur de la prédiction en question, ou bien nous ne nous posons pas cette question du pourquoi et nous contentons de la valeur de la prédiction. Dans le premier cas, et c'est celui-ci qui va nous intéresser dans cette partie, il peut exister une multitude de motivations pour se poser cette question du pourquoi. Lorsque nous répondons à cette question du pourquoi, nous construisons une explication. Ainsi l'objet de la recherche en explicabilité revient à savoir si on peut formuler ces explications, sous quelles contraintes, dans quel contexte, quelles sont les propriétés qui doivent satisfaire les explications, etc. L'explication devient alors un objet d'étude et on cherche alors à le caractériser le mieux possible. Cependant, la notion d'explicabilité n'est pas bien définie et encore moins formellement [Miller, 2019a, Kim et al., 2016a, Murdoch et al., 2019]. Ainsi l'explicabilité n'est pas quelque chose qui se calcule ou s'optimise directement. Cependant, la méthode dite "explicative" doit satisfaire quelques propriétés établies par la communauté. Dans un premier temps, nous allons évoquer les motivations derrière l'explicabilité et cela dans le contexte de l'apprentissage machine.

1. Les concepts et arguments évoqués dans ce chapitre sont issus de [Molnar et al., 2020,Murphy, ] ou de notre propre réflexion

### 5.1.1 Motivations pour l'explicabilité

Dans le contexte de la l'apprentissage machine, ce sont les modèles prédictifs qui fournissent les prédictions. Et pour un modèle de l'apprentissage machine donné, il existe une multitude de motivations pour se poser la question du pourquoi ce modèle a effectué une certaine prédiction. Nous présentons ici une liste de quelques unes de ces motivations dans le contexte de l'apprentissage machine :

— **découvrir** : nous en avons quelques exemples aujourd'hui, certains modèles de l'apprentissage machine surpassent l'humain pour certaines tâches. Et dans ce cas, ce n'est plus l'humain qui fournit de la connaissance au modèle, mais c'est ce dernier qui en devient le fournisseur. Ainsi être capable de comprendre les raisons qui l'ont amené à faire les prédictions optimales , que nous humains, ne sommes pas capables de faire en moyenne, est donc un moyen de faire des découvertes. [Badgeley et al., 2019, Zech et al., 2018, Gururangan et al., 2018]. Cela peut permettre aussi d'améliorer certains modèles ou de les "débugger" [Staff, 2020, Zerilli et al., 2019, Goodman and Flaxman, 2017, Gen, , Bechtel and Abrahamsen, 2005, Hempel and Oppenheim, 1948, Glennan, 2002, Fiedler and Juslin, 2005]

— **prévenir** : l'étendue applicative de certains modèles de l'apprentissage machine en particulier les plus complexes, à certains domaines où l'utilisation de la prédiction peut avoir de lourdes conséquences sur la société (médicale, voiture autonome, etc.). Ainsi, il est bon de se prémunir de certaines mauvaises prédictions et donc d'avoir des explications quant à leurs occurrences [Amodei et al., 2016]. De plus, si nous avions un modèle prédictif qui ne se trompe presque jamais, alors on pourrait se dire qu'il n'est pas nécessaire de l'inspecter car les cas où il se trompe sont quasi-inexistants. Malheureusement, rappelons que la définition de nos modèles de l'apprentissage machine est nécessairement sous-contrainte et donc attribuer des caractéristiques de manière absolue à un modèle n'est pas à faire puisque nous n'avons qu'une mesure partielle de ses performances [Doshi-Velez and Kim, 2017a, Keil, 2006].

À l'inverse, lorsque nous connaissons bien le phénomène que nous modélisons par un modèle prédictif, c'est-à-dire lorsque il n'y a plus de découverte à faire le concernant, alors il n'est plus utile de répondre à la question du pourquoi. De la même manière,

lorsque les prédictions d'un modèle n'ont pas d'impacts sur la société, alors il n'y a plus d'intérêt à formaliser des explications. Cependant, dans le cas où l'intérêt d'avoir des explications est là, comment pouvons-nous les obtenir et comment pouvons-nous les qualifier ?

### 5.1.2 La formulation d'explication, le facteur contextuel et intentionnel

La notion d'explicabilité, qui pourrait signifier "ce qui peut être expliqué" est mal définie. Qu'est-ce qu'une explication formellement ? Est-ce un concept que l'on peut décrire par une équation ? Il existe des formulations sur des cas particuliers, mais il n'existe pour l'instant pas de consensus pour définir ces notions. En effet, pour un modèle de l'apprentissage machine moderne, il existe bien un calcul, parfaitement défini, qui a produit la prédiction d'une donnée. Mais l'interprétation crue de calcul est probablement difficilement représentable intellectuellement. Est-ce qu'il est si informatif que ça ? En effet, "interpréter" signifie "présenter en des termes compréhensibles pour un humain" [Def, 2023]. La notion d'interprétation a donc une composante intrinsèquement "humaine" et le fait de "comprendre" ou de rendre "quelque chose de compréhensible" signifie que le concept/schéma mental associé au "quelque chose" a été intériorisé alors qu'il était intellectuellement "étranger", avant compréhension, à notre esprit. On peut alors voir la compréhension d'un "quelque chose" comme un alignement intellectuel entre un concept intellectuel externe à notre esprit et le concept intellectuel interne de ce "quelque chose". Cette représentation interne est dépendante de plusieurs facteurs comme le contexte sémantique définitionnel du "quelque chose" et aussi l'objectif final de ce quelque chose, et la personne qui se confronte à la tâche de compréhension. Ces derniers points font sens dans le contexte de la l'apprentissage machine. En effet, la compréhension d'un modèle de l'apprentissage machine se fait au travers du prisme du contexte sémantique de la tâche, du but pour lequel le modèle a été construit et des concepts internes de celui qui souhaite comprendre le modèle ; et spécifiquement à ce contexte, la formulation externe du concept (c.-à-d. l'explication fournie par la méthode explicative) dépend du matériel et du logiciel utilisé par la plateforme qui formule l'explication [Miller, 2019b]. On dit alors que l'interprétation est intrinsè-

quement dépendante du contexte socio-technique [Selbst et al., 2019]. Dans la forme calculatoire de l'interprétation, il y a d'autres modalités par lesquelles les interprétations sont formulées ou caractérisées, souvent même, conjointement à plusieurs de ces modalités simultanément, les voici :

— **contexte sémantique de la tâche d'apprentissage :** le contexte sémantique de la tâche est indissociable du principe d'interprétation.
— **la tâche :** que doit résoudre la méthode, son but final, est un moyen pivot pour orienter la formulation des interprétations.
— **le processus de construction de la base de données pour l'optimisation du modèle prédictif** [Crawford, ,Mitchell et al., 2019].
— **la méthode explicative fournit une explication par rapport à un modèle, mais à plusieurs échelles :** en considérant l'entièreté du modèle ou bien que sur certain de ses composants (p. ex., une couche dans le cas des méthodes de l'apprentissage machine moderne) pour formuler ses explications. De plus, pour un même modèle prédictif, ces méthodes explicatives peuvent donner des explications différentes. Le choix de la méthode explicative est alors dépendant du contexte applicatif et du but final.
— **la qualité de l'explication fournie est évaluée par des critères numériques :** mais la définition de ces critères doit se faire en accord avec le contexte sémantique et le but final du modèle expliqué. En effet, pour qualifier des modèles de l'apprentissage machine, nous utilisons parfois la précision sur l'ensemble de tests (par exemple pour un problème de classification, ou bien l'indice de Jacquard pour qualifier la pertinence d'un masque de segmentation avec la vérité terrain). Ainsi chaque problème (et donc à chaque contexte et but final) a sa propre métrique pour qualifier un modèle prédictif, ces métriques étant différentes d'un problème à l'autre.
— **de la même manière :** les explications doivent aussi satisfaire des propriétés (qui certes sont souvent elles aussi évaluées numériquement) et l'utilité de ces propriétés varie selon le contexte de la tâche ainsi que le but final.

Pour un modèle prédictif donné, la combinaison d'une méthode explicative, de métrique d'évaluation, de propriétés par rapport à un contexte et un but peut être définie comme étant un écosystème explicatif du modèle expliqué. Ainsi le choix

d'un écosystème explicatif doit être fait relativement au contexte et au but du modèle expliqué [Van Fraassen, 1980]. De manière analogue, on peut voir le choix de cet écosystème comme étant le choix d'un langage de programmation et de son paradigme pour résoudre une tâche selon un contexte précis (par exemple, pour de l'embarqué ou pour un serveur avec beaucoup de ressources, etc.) avec un but précis (par exemple déterminer si le patient est atteint de cancer ou non, etc.). Nous allons donner dans la suite les différents "paradigmes" usuels pour les méthodes explicatives, les métriques d'évaluation et les propriétés.

### 5.1.3 Type des méthodes explicatives

En fonction du contexte et du but du modèle prédictif, il existe plusieurs types de méthode explicative :

— **intrinsèque ou non :** il existe des modèles de l'apprentissage machine qui sont directement interprétables. Ils sont dits directement interprétables s'il n'est pas nécessaire d'effectuer des calculs supplémentaires pour les interpréter dans leur contexte et par rapport à leur but [Dash et al., 2015]. C'est le cas des modèles linéaires [Ustun et al., 2014], des arbres de décisions [Breiman et al., 1984, Rivest, 1987, Wang et al., a, Letham et al., 2015a, Angelino et al., ], voir [Com, b] pour une étude de l'état de l'art sur ce point. Puis il y a les modèles qui sont plus complexes et qui nécessitent de faire des calculs supplémentaires pour les interpréter.

— **post-hoc ou non :** avant d'atteindre un régime prédictif intéressant, les modèles de l'apprentissage machine paramétriques doivent être optimisés. Ainsi, on qualifie de "post-hoc" les méthodes explicatives qui sont utilisées après la phase d'optimisation, ou en tout cas, lorsqu'on estime que le modèle prédictif a atteint un certain niveau de justesse. C'est le cas de la plupart des méthodes de l'état de l'art qui supposent dans leur définition d'avoir atteint un niveau de justesse optimal pour être utilisable. Cependant rien n'empêche de développer des méthodes explicatives qui prennent en compte la phase d'optimisation dans la construction des explications.

— **global ou local :** on peut formuler des explications à plusieurs niveaux. Par exemple, pour une prédiction en particulier d'un modèle. On peut aussi les

effectuer de diverses manières : en estimant l'importance de certains descripteurs dans la donnée d'entrée selon le modèle [Sundararajan et al., 2017a, Smilkov et al., 2017a, Zeiler and Fergus, 2014a, Selvaraju et al., 2017, Bengio, 2009, Springenberg et al., 2015, Shrikumar et al., 2017] ; pour les modèles non directement interprétables, on peut approximer le comportement du modèle complexe au voisinage de la donnée d'entrée par un un modèle directement interprétable [Ribeiro et al., 2016a] et ainsi le caractériser facilement ; ou bien encore, on identifie les données voisines les plus influentes pour effectuer la prédiction de celle considérée [Koh and Liang, 2020a, Mittelstadt et al., 2019, Karimi et al., 2020]. Les méthodes explicatives définies pour des explications globales s'intéressent elles, à expliquer le modèle pour plusieurs données d'un coup. On s'intéresse donc à caractériser l'entité modèle prédictive en tant que telle. Par exemple, dans le cas d'un modèle non directement interprétable, on approxime le comportement global du modèle complexe sachant un échantillon avec un modèle directement interprétable qui permet donc l'interprétation en approximant le modèle [Hinton et al., 2015], ou bien on caractérise statistiquement les descripteurs des données qui sont influentes pour le modèle complexe [Kim et al., 2018a] ; ou bien encore en construisant des représentations de donnée issue de la base de données qui sont représentatives du comportement global du modèle [Yeh et al., 2018, Zhu et al., 2018, Amir and Amir, ].

— **spécifique ou agnostique :** pour construire les explications, les méthodes explicatives peuvent considérer les modèles prédictifs comme "boite noire", c'est-à-dire qu'elles ne considèrent que les entrées et sorties du modèle pour construire son explication, on dit alors que la méthode explicative est agnostique au modèle. Sinon , la méthode peut utiliser les résultats des calculs intermédiaires effectués par le modèle pour construire ces explications, on qualifie alors la méthode explicative de spécifique au modèle. Cependant, il faut noter que les méthodes agnostiques au modèle peuvent être utilisées par une plus grande classe de modèles que les méthodes spécifiques, car tout modèle a des entrées et sorties alors que les méthodes spécifiques au modèle peuvent dépendre de certains calculs qui ne sont pas forcément définis pour

tous les modèles.

— **unitaire ou modulaire :** pour les méthodes spécifiques, certaines méthodes explicatives produisent leurs explications en se basant uniquement sur une partie du modèle (par exemple une couche d'un modèle en particulier), elles sont dites modulaires. Celle qui considère les modèles dans leur entièreté est alors dite unitaire.

Ces types d'approches peuvent être combinés ensemble ?par exemple usuel dans la littérature de rencontrer des méthodes explicatives locales, post-hoc et agnostiques. Là aussi, certaines approches sont parfois plus adaptées au contexte et au but du modèle prédictif que d'autres. Cependant, l'ensemble de ces méthodes, indépendamment de leurs paradigmes, doivent satisfaire un certain nombre de propriétés dans le but de données de "bonnes" explications. Nous présentons ici les plus récurrentes dans la littérature :

### 5.1.4 Propriétés des méthodes explicatives

Nous présentons ici une vue partielle des propriétés que doivent satisfaire les méthodes explicatives dans certains contextes, voir [Robnik-Šikonja and Bohanec, 2018].

— **fidélité :** cette propriété porte sur comment la méthode explicative conserve les propriétés de précision du modèle prédictif. Par exemple, dans le cas d'une méthode explicative globale, comme l'utilisation d'un modèle directement interprétable interpolant le comportement du modèle prédictif complexe, la fidélité peut être la précision relative du modèle interprétable par rapport au modèle complexe. Dans le cas d'une méthode explicative locale, la fidélité se mesure souvent en mesurant la perte de précision d'un modèle lorsque les descripteurs importants, relevés par la méthode explicative, sont occlus. Ces notions sont décrites dans [Jacovi and Goldberg, 2020a,Jacovi and Goldberg, 2021a]

— **consistance :** la consistance d'une méthode explicative est sa capacité à produire des explications similaires par rapport à un ensemble de méthodes prédictives entraîné pour la même tâche.

— **stabilité :** sachant un modèle prédictif, la propriété de stabilité est une

mesure du phénomène suivant : si deux données sont "similaires" alors leurs explications sont "similaires" aussi, voir [Alvarez-Melis and Jaakkola, 2018a, Miller, 2019b]

— **compréhensibilité :** sachant un modèle et un utilisateur final de la méthode explicative, est-ce que les explications fournies par la méthode explicative est exprimée selon les concepts mentaux de l'utilisateur final ? Autrement dit, est-ce que l'explication issue de la méthode explicative est intellectuellement compréhensible par l'utilisateur final ? Ou bien, est-ce cette explication illustre (numériquement) des connaissances relatives au contexte ou au but, identifiées à priori par l'utilisateur final ? Si oui, alors la méthode est dite "compréhensible", voir [Kim et al., 2018b, Clough et al., 2019b].

— **certitude :** nous avons vu que les prédictions issues de modèle prédictif ont une nature intrinsèquement probabiliste. Cependant, quantifier des intervalles de confiance sur un modèle prédictif, quelconque n'est pas trivial. La quantification de ce genre d'information est pourtant importante dans le contexte de l'explicabilité de modèle prédictif, car elle permettrait de savoir si l'on peut prendre appui sur les prédictions du modèle pour une utilisation de ce dernier dans un contexte sensible par exemple (santé, etc.). Si une méthode explicative est capable de fournir ce genre d'informations dans ses explications, alors elle satisfait la propriété de certitude.

— **nouveauté :** une méthode explicative a la propriété de nouveauté si les explications qu'elle fournit apportent de nouvelles connaissances, mais il se peut aussi que l'on produise des explications "nouvelles" du fait que le modèle prédictif n'est pas sûr de sa prédiction, ce genre de propriété est donc compliquée à mesurer en pratique.

— **représentativité :** cette propriété caractérise la capacité d'une méthode explicative à produire un petit nombre d'explications pertinentes et différentes pour un grand nombre de prédictions différentes.

— **compacité :** cette propriété mesure en quoi les explications d'une méthode explicative sont assez "petites", c'est-à-dire en pratique calculables, mais aussi, et c'est lier, si la taille des explications est au moins aussi petite que les prédictions du modèle prédictif évalué. Par exemple, seul un petit nombre

de descripteurs sont considérés par la méthode explicative comme étant pertinents devant le nombre de descripteurs utilisables par la méthode explicative. Dans ce cas, les explications sont dites parcimonieuses ou compactes, voire [Myt, ,Murdoch et al., 2019].

— **complétude :** cette propriété caractérise la capacité d'une méthode explicative à intégrer tous les éléments pertinents relativement au contexte et au but du modèle prédictif dans ses explications, voir [Yeh et al., ,Kulesza et al., 2013]

— **adaptativité :** cette propriété caractérise la capacité d'une méthode explicative à filtrer ses explications a posteriori en fonction des attentes des divers utilisateurs finaux, [Karimi et al., 2021,Poyiadzi et al., 2020a,Tomsett et al., ].

— **partionnabilité :** cette propriété caractérise la capacité d'une méthode explicative à fournir des explications partionnables, au sens ou chaque partie de l'explication répond à une partie du raisonnement fait par le modèle prédictif pour une prédiction donnée, voir [Myt, ,Murdoch et al., 2019]

— **interactivité :** cette propriété caractérise, pour une méthode explicative, la possibilité d'action affectable a posteriori par l'utilisateur final. C'est à dire qu'une fois une explication est obtenue pour une prédiction donnée, est-ce que l'on peut interagir avec l'explication, pour la préciser sur certains points, avoir une vue plus générale, etc. Voir [Tenney et al., 2020,Zeiler and Fergus, 2014b,Olah et al., 2017,Inc, 2015,Nguyen et al., 2016,Hohman et al., 2020]

— **transfluence :** expliquer un modèle prédictif permet notamment d'identifier les limites du modèle prédictif qu'elle évalue, mais la méthode explicative peut elle-même avoir des limites, sont-elle clairement considérées et évoquées dans la définition de cette dernière ? Si oui, alors la méthode explicative est qualifiée de transfluente, voir [Sokol and Flach, 2020,Liao et al., ].

— **simulabilité :** nous avons vu en introduction que les prédictions des modèles de la l'apprentissage machine sont calculables, en est-il de même pour les méthodes explicatives ? Autrement dit, les explications d'une méthode explicatives sont elles calculables avec une complexité temporelle et mémoire raisonnable ? Si oui, alors la méthode explicative est qualifiée de cumulable [Myt,

, Murdoch et al., 2019]

— **contrastive :** pour deux prédictions issues de deux phénomènes distincts, alors les explications doivent être, elles aussi, distinctes. On dit alors que la méthode explicative fournit des explications contrastives [Miller, 2019a].

Pour un contexte et un but donnés, toutes les propriétés évoquées ci-dessus ne sont pas toutes forcément adaptées et/ou même mesurables numériquement. L'utilisation d'une méthode explicative satisfaisant un large panel de ces propriétés est donc à privilégier. Pour déterminer si une méthode explicative est "meilleure" qu'une autre conditionnellement à un contexte et un but final, il faut pourvoir les évaluer. Nous présentons ici les protocoles usuels pour évaluer les méthodes explicatives.

### 5.1.5 Évaluation des méthodes explicatives

Pour un modèle prédictif, un contexte et un but final, il existe a priori un ensemble de méthodes explicatives non vides pour expliquer ledit modèle. Cependant, dans cet ensemble, certaines méthodes sont plus propices que d'autres. Pour les hiérarchiser, il est nécessaire de les évaluer individuellement. Pour mener ces évaluations, il existe deux approches principales [Doshi-Velez and Kim, 2017a]. La première approche est dit "objective" puisqu'elle évalue numériquement les méthodes explicatives. L'autre approche est une approbation humaine de la méthode, par conséquent "subjective".

— **évaluation objective :** pour évaluer objectivement une méthode explicative, cela se fait par la mesure numérique de différentes propriétés parmi celles évoquées avant, en accord avec le contexte et le but final du modèle prédictif évalué. Une fois les propriétés pertinentes pour l'évaluation de la méthode explicative retenues, il faut les définir numériquement si tenté qu'elles admettent une telle définition. Bien entendu, on peut définir numériquement une propriété de différentes façons. Par exemple, une notion de parcimonie peut être définie par la norme $p$ de l'explication (si elle a une représentation vectorielle) pour tout $p \in ]0, 1[$, [Hurley and Rickard, 2009]. Il y a donc une infinité de façons de définir la propriété de parcimonie et donc un choix à faire sur la définition numérique. Cela peut introduire des biais lors de l'évaluation de la méthode explicative, voir [Tomsett et al., 2020, Sayres et al., 2019]. Ceci

étant dit, voici les principales approches objectives :

- **preuve théorique :** dans certains cas, avec des hypothèses sur l'ensemble des composantes de l'évaluation, il est possible de prouver mathématiquement qu'une méthode satisfait une certaine propriété. Sinon, on adopte une approche empirique, et donc on mesure des tendances statistiques relativement à la propriété considérée.
- **test de vérification :** dans le cas empirique, nous pouvons dresser des hypothèses auxquelles nous voulons que notre méthode explicative ne réponde pas. Ainsi, une approche est de créer une situation dans laquelle l'hypothèse pourrait être satisfaite et observer empiriquement que la méthode explicative n'y répond pas, voir [Adebayo et al., 2020, Kindermans et al., 2017, Ghorbani et al., 2019a].
- **ensembles de données synthétiques :** dans cette approche, l'idée est de constituer des données comprenant des descripteurs définis artificiellement comme importants et les descripteurs complémentaires comme non importants. De cette manière, on obtient artificiellement une vérité terrain d'explicabilité, au sens où on sait où se trouve la partie "importante" de la donnée, c'est-à-dire la partie discriminante pour sa prédiction par le modèle prédictif, voir [Yang and Kim, 2019, Sanchez-Lengeling et al., 2020, Ghandeharioun et al., 2022, Plumb et al., 2019, Kim et al., 2022, Yeh et al., , Kim et al., 2018a, Ying et al., 2019c]
- **ensembles de données non synthétiques :** dans le cas où l'ensemble de données est réel il n'y a, a priori, pas de vérité terrain associée. Dans ce cas-là, il est usuel de prendre les descripteurs les plus importants des données et de masquer cette donnée avec l'explication fournie puis d'effectuer la mesure de la propriété souhaitée. Cette approche est souvent utilisée, voire [Ghorbani et al., 2019c, Fong and Vedaldi, 2017, Ribeiro et al., 2016b, Dabkowski and Gal, 2017]

— **validation experte :** une autre approche pour évaluer une méthode explicative est de valider les explications d'une méthode par un individu expert par rapport au contexte et au but final du modèle prédictif. Dans ce cas, il faut apporter une attention particulière à la représentation de l'explication pour

l'expert qui n'est pas forcément expert capable d'exploiter directement la représentation numérique de l'explication. Cette représentation peut aider à mener une étude quantitative ou qualitative et donc conclure statistiquement sur la pertinence de la méthode explicative. Ce genre d'étude statistique peut se faire soit sur des données réelles ou synthétiques, le processus de construction de ce dernier ou aussi comment le modèle est optimisé pour ces tâches. Cependant, il faut noter que l'évaluation humaine est nécessairement subjective, avec ce que cela peut entraîner (biais, variance des évaluations), ce qui peut être limitant pour conclure sur la pertinence absolue d'une méthode explicative.

De plus, comme nous l'avons déjà évoqué, il a plusieurs façons de formaliser une même propriété, ce qui par conséquent peut biaiser la conclusion de la pertinence d'une méthode explicative [Tomsett et al., 2020,Sayres et al., 2019]. Mais il n'y a pas que dans la formalisation des propriétés que cette variabilité dans les conclusions peut subvenir. En voici d'autres :

### 5.1.6 Représentation et calcul des explications

Dans le contexte de l'explicabilité des méthodes prédictives (c-à-d des modèles de l'apprentissage machine modernes), les méthodes explicatives sont, elles aussi, décrites en pratique algorithmiquement. Ainsi les explications, indépendamment du paradigme choisi, sont décrites numériquement. Cependant il y a une multitude de façons de représenter numériquement un même objet et ce choix peut avoir un impact sur les propriétés de l'explication (comme dans [Lundberg and Lee, 2017]) et donc dans certains cas, sur son évaluation. Par exemple dans ce corpus de contributions [Simonyan et al., 2014a, Dabkowski and Gal, 2017, Fong and Vedaldi, 2017, Datta et al., 2016, Adler et al., 2016, Bach et al., 2015a], chacune souhaite mesurer l'impact d'une région de pixel ou encore l'importance de ces régions. Elles cherchent pourtant à expliquer le même phénomène ici : la classification d'images. Cependant, la variance de la définition de l'importance sur les études entraîne une variation sur les conclusions de celles-ci, ce qui est problématique. Ce phénomène est dû principalement au fait que la notion d'explication, d'explicabilité n'a pas de définition formelle consensuelle. Ainsi chacun introduit et formalise ses propres no-

tions , entraînant cette variabilité. On observe aussi le phénomène dans cet autre corpus de contributions [Sundararajan et al., 2017b, Smilkov et al., 2017a, Selvaraju et al., 2020, Bengio, 2009, Shrikumar et al., 2017] Par ailleurs, il y a aussi la notion de similarité entre explications qui accuse d'une variabilité dans sa définition, comme ce corpus de contributions [Wachter et al., 2017, Lucic et al., 2020, Kusner et al., 2018, Karimi et al., 2020] Pour les raisons citées ci-dessus, on constate que l'évaluation de méthodes explicatives, même objective, est dépendante d'un choix de représentation et de quantification qui par définition est subjective. Pour des motivations purement académiques, c'est problématique pour comparer les méthodes, pour la reproductibilité non absolue des résultats, en bref pour valider quelle méthode est objectivement meilleure de manière absolue que les autres dans un contexte et un but donné. Et d'un point de vue sociétal, cette variabilité peut donner l'illusion que nous avons gagné en connaissance par rapport au modèle prédictif étudié alors qu'il n'en est ne peut être rien. Ce point n'est qu'une première limite de l'approche explicative des modèles de l'apprentissage machine moderne, en voici d'autres.

### 5.1.7 L'explicabilité des modèles de l'apprentissage machine modernes, variabilité, partialité

L'essor des méthodes de l'apprentissage machine modernes (c-à-d méthodes profondes) pour résoudre des problèmes académiques ou industriels est relativement récent. Ainsi les motivations pour l'interprétabilité de ces modèles sont, elles aussi, très récentes. C'est la principale cause du manque de consensualité sur les définitions d'interprétabilité, d'explicabilité, d'explications ou encore les différentes propriétés ou bien celles concernant les protocoles d'évaluation de ces méthodes. Cependant, est-ce que la recherche en explicabilité des méthodes profondes peut atteindre un stade absolument consensuel dans la définition des objets de recherche de ce domaine ? Et de manière générale, quelles sont les limites actuelles des modèles explicatifs ? Nous présentons ici quelques éléments de réponse :

— **un formalisme de l'écosystème explicatif nécessairement relatif :** L'objectif des études scientifiques est de, rigoureusement, gagner des connaissances relativement à un objet d'étude. Comme dans toute approche utilisant la méthode scientifique (c-à-d l'approche logico-déductive), la définition des

objets d'études et comment ils interagissent sont fondamentales pour caractériser les problématiques. Dans le cas de l'explicabilité, la définition des objets et leurs interactions se font en fonction du contexte et du but final du modèle prédictif, mais aussi par rapport à la personne qui reçoit les explications, la modalité la plus difficile à définir, à ce jour au moins, étant "la personne". Il va sans dire que définir "la personne" rigoureusement et logiquement (pour utiliser à bien le principe logico-déductif), au niveau absolu le plus bas est aujourd'hui infaisable. Il n'y a que des définitions approximatives et donc variables. Or, la définition de l'écosystème explicatif est dépendant de ce concept "personne", et donc subit nécessairement la variabilité de la définition de la personne. Cet écosystème a lui aussi une définition nécessairement variable. Pour illustrer cela, prenons une personne experte d'un phénomène donné et une personne non experte par rapport à ce même phénomène, alors, pour une méthode explicative donnée, si cette dernière fait gagner en connaissance l'experte (en mettant en lumière des arguments uniquement traitables par des personnes de niveau experte), il y a de fortes chances pour que ce ne soit pas le cas pour la personne non experte (car par définition elle n'est pas experte, donc elle ne peut pas traiter ce type d'information) et inversement, une explication "trop simple" n'apportera de valeur ajoutée en termes de connaissance qu'à la personne ignorante, il n'y aura pas de "gain" pour l'experte. Une illustration pratique ce phénomène [Kleinberg et al., 2016]. Ainsi chercher un cadre formel pour l'explicabilité ne peut se faire que de manière relative et non de manière absolue, du fait de sa dépendance à la "variabilité intrinsèque humaine". Par extension pour les définitions formelles, une définition formelle et universelle n'existe pas, elle est nécessairement relative et donc variable. Ce cas de la nécessité de la non-universalité de l'écosystème explicatif peut être vu comme une application du théorème No Free Lunch [Wolpert and Macready, 1997].

— **une explication nécessairement partielle :** nous l'avons vu en introduction de ce document, nous ne pouvons décrire le "code source" d'un modèle de l'apprentissage machine moderne alors que pourtant l'accès à ce code serait un moyen explicatif et scientifique (au sens du principe logico-déductif) du

comportement du modèle (si tenté que ce code soit humainement compréhensible). Le but premier de l'explicabilité de gagner en connaissance vis-à-vis d'un modèle n'est pas forcément d'atteindre la borne supérieure de la connaissance relative à ce modèle car elle est humainement inatteignable. Nous pouvons au mieux faire un petit pas vers cette borne supérieure. Chaque explication fournie par une méthode explicative, qui permet d'effectuer ce "pas", est nécessairement incomplète, l'explication absolue explicative étant celle produite par la méthode explicative ayant atteint "la borne supérieure de l'explicabilité". Ainsi les explications que nous avons en pratique ne sont que des explications ne caractérisant le modèle prédictif que de manière marginale et n'offrent nécessairement qu'une vision partielle du modèle prédictif.

— **la partialité peut entraîner la non-unicité :** si une méthode explicative atteint la borne supérieure de l'explicabilité, alors pour un écosystème explicatif donné, il n'existe qu'une seule bonne explication possible du modèle prédictif, car nous avons une vue absolue du modèle en question, par unicité de la borne supérieure (en admettant que la notion d'explicabilité soit représentable numériquement, ce que l'on supposera ici). Or, en pratique, les méthodes explicatives courantes n'atteignent pas cette borne, c'est-à-dire qu'elles ont une vue partielle du modèle prédictif. Mais il peut y avoir des variables latentes par rapport à la méthode explicative qui peuvent altérer le comportement de la méthode explicative elle-même , et donc engendrer une instabilité et une fragilité des explications fournies. On l'observe en pratique, souvent avec des modèles prédictifs surparamétrisés comme les réseaux de neurones , où par exemple une méthode explicative indique qu'un descripteur dans une donnée est non important alors que le réseau de neurones l'utilise fortement pour effectuer ses prédictions [Alvarez-Melis and Jaakkola, 2018b]. On observe aussi que les méthodes explicatives sont sujettes aux attaques. [Yeh et al., 2019a, Ghorbani et al., 2019b, Dombrowski et al., 2019, Slack et al., 2020].

— **la sur-paramétrisation pour un but donné n'est pas toujours nécessaire :** l'utilisation de modèle sur-parémétré plus sujet aux phénomènes d'instabilité des explications n'est pas toujours justifiée à postériori. [Rudin,

2019a]. Cependant, plus la qualité de représentation des données est importante, plus a priori cette représentation est "simple", ou "concise" au sens où, en effet, du point de vue algorithmique, la représentation de la donnée n'est plus à effectuer par l'algorithme (c-à-d le modèle prédictif), il peut donc être plus simple (à contexte et but égaux), et dans notre contexte, plus simple signifie moins paramétré. Ainsi, l'étape de ce que l'on peut grossièrement qualifier de "phase de prétraitement" de la donnée est essentielle, car si elle est bien effectuée, elle permet de résoudre la même tâche dans le même contexte, avec un modèle prédictif moins complexe donc plus explicable. En voici quelques illustrations concrètes [Wang et al., a, Caruana et al., 2015, Letham et al., 2015b, Ustun and Rudin, 2016, Futoma et al., , Kim et al., ].

— **L'existence d'une explication n'est pas une condition suffisante pour l'utilisation aveugle des modèles de l'apprentissage machine modernes ou pour émettre une conclusion hâtive sur leur comportement :**
  — L'existence d'une explication n'implique pas forcément la pertinence de celle-ci (globalement ou localement) par rapport au contexte et but du modèle prédictif, au sens des propriétés comme évoquées, mais aussi au sens de la valeur ajoutée quant au gain de connaissance relativement au modèle prédictif. Toutes les explications n'aident pas forcément.
  — Si nous, humains, avons des croyances concernant un modèle prédictif qui ne sont pas des vérités absolues et/ou fausses et qu'une explication illustre ces croyances, alors que notre biais de confirmation pourrait nous mener à conclure sur le comportement du modèle de manière erronée. [Smilkov et al., 2017b, Sundararajan et al., 2019, Kaur et al., 2020, Leavitt and Morcos, 2020]. Rappel : les explications sont nécessairement non absolues.
  — L'existence d'explication illustrant des représentations haut niveau de la donnée que nous humains pouvons faire en conscientisant la donnée peut induire que le modèle prédictif a alors lui aussi une forme de conscience et on donne une au modèle prédictif une forme anthropomorphe. Non, le modèle prédictif ne "pense" pas il a simplement un but numérique à atteindre avec des arguments numériques eux aussi.

— L'existence d'explication n'est pas une condition suffisante pour donner une confiance aveugle aux modèles prédictifs, l'explicabilité de ces modèles n'est qu'un moyen, marginal, pour caractériser les modèles prédictifs, a fortiori lorsque cela est complexe. En effet pour les modèles prédictifs de la l'apprentissage machine moderne, il a d'autres domaines de recherche s'intéressant à caractériser d'autres aspects opaques ou sur la régulation de ces modèles vis-à-vis de leur impact grandissant sur les sociétés humaines.

— **L'explicabilité est en perpétuel mouvement avec le progrès scientifique :** de manière générale, l'explication permet d'éclaircir quelque chose qui est flou à un moment donné. On utilise les modèles prédicitfs pour associer des entités d'entrée à des entités de sortie. Il semble y avoir un lien entre la compréhension du phénomène physique sous-jacent à modéliser et l'utilisation de modèle prédictif simple (c-à-d, non profond, et donc parfois interprétable) ou complexe (c-à-d, profond). En effet, si la connaissance scientifique courante a permis de bien comprendre le phénomène physique, alors nous avons une bonne représentation et donc des descripteurs peuvent être obtenue de manière complexe mais tout en ayant une relation simple entre eux. Dans ce cas, nous pouvons alors utiliser un modèle prédictif plus simple. Nous avons en quelque sorte effectué en amont la tâche de réduction de dimension du problème. A l'inverse, si le phénomène physique n'est pas bien compris vis à vis de la connaissance scientifique courante, ou en tout cas que l'écart entre les meilleurs descripteurs (entité d'entrée) du phénomène que nous sommes capables de formuler nécessite d'être amplement transformé pour produire les descripteurs souhaités (entités de sortie) est a priori "grand", "long" et que, comme nous n'avons pas une bonne compréhension du phénomène sous jacent, nous ne sommes pas capable de formuler clairement le lien entre ces entités d'entrées et de sorties, alors nous utilisons un modèle qui construit ce lien de manière automatique (c-à-d, via des modèles profonds). Ainsi l'utilisation de modèles profonds est due en partie au manque de connaissances de la communauté scientifique relativement au phénomène physique. Et en particulier, la nécessité de l'explicabilité pour ces derniers est donc aussi due au manque de connaissances scientifiques par rapport à ce même phénomène physique.

> Autrement dit, le jour où la connaissance scientifique sera bien plus avancée, des descripteurs plus complexes des mêmes phénomènes seront formulables. Nous pourrons alors utiliser des modèles plus simples, mieux compris, voire directement interprétables, rendant vaine l'utilisation de modèles profonds et donc la nécessité de leur explicabilité. Cependant, de nouvelles questions scientifiques apparaissent tous les jours, ainsi la connaissance scientifique ne devrait pas se stabiliser puisque nous n'atteindrons jamais la connaissance absolue de notre univers. Ainsi, la recherche en explicabilité des modèles profonds devrait donc être en perpétuel mouvement, coordonnée avec le progrès scientifique.

Nous présentons ici l'état de l'art des méthodes explicatives pour les modèles de l'apprentissage machine, pour les modèles directement interprétables puis non directement interprétables.

## 5.2 Les méthodes explicatives pour les modèles de l'apprentissage machine

Nous présentons ici une vue globale de la recherche en explicabilité des méthodes profondes ou non utilisées pour résoudre des problématiques d'apprentissage machine. Notez que, certaines contributions citées n'ont pas reçu de notre part une étude approfondie car bien que connexes et en lien avec les problématiques d'explicabilité, elles restent parfois loin de notre domaine de recherche avec les réseaux de neurones sur graphes et de ses contraintes. Cependant, ce sont des contributions majeures et reconnues dans la communauté des chercheurs s'intéressant à l'explicabilité des modèles de l'apprentissage machine, c'est pourquoi nous les avons citées dans ce manuscrit.

Nous listons dans B.2.1 ces principales contributions.

## 5.3 Les méthodes explicatives pour les réseaux de neurones sur graphes

Nous présentons ici maintenant une vue globale de la recherche en explicabilité spécifiquement développée pour les réseaux de neurones sur graphes et quelques contributions connexes.

Cependant les méthodes principales sont :

— GNNExplainer [Ying et al., 2019a] : une méthode local et post-hoc reposant sur la maximisation de l'information mutuelle entre un sous graphe et le graphe initial conservant les propriétés classifiantes du classifieur évalué.

— PGExplainer [Luo et al., 2020] : une approche similaire à GNNExplainer ne considérant pas les descripteurs de noeuds dans leurs globalités et utilisant un perceptron multi-couches pour le calcul des explications.

— SubgraphX [Yuan et al., 2021] : est une approche post-hoc locale qui décompose le graphe d'entrée via un arbre de recherche de Monte Carlo et considère l'importance de chaque découpe avec la valeur de Shapley.

— XGNN [Yuan et al., 2020] : est une méthode explicative globale qui génére des graphes expliquant via un processus d'apprentissage par renforcement.

— GNN-LRP [Schnake et al., 2022] : est une approche de décomposition qui propage la sortie du modèle classifiant vers l'entrée. C'est une méthode dérivée de l'approche LRP utilisée en vision par ordinateur.

— PGM-Explainer [Vu and Thai, 2020] : cette méthode utilise les modèles probabiliste graphiques pour construire les explications. Chaque noeuds représente une variable aléatoire et la méthode permet d'identifier ainsi les noeuds importants.

— GradCAM GNN [Baldassarre and Azizpour, 2019a] : est l'extension aux réseaux de neurones sur graphes de la méthode GradCAM pour les réseaux de neurones convolutionnels [Selvaraju et al., 2020].

Pour rédiger cette partie, nous nous sommes inspirés des résumés des contributions citées. Certaines contributions ne sont pas issues de revues reconnues et/ou ne fournissent pas le code nécéssaire pour tester la véracité et la reproductibilité des travaux. Aussi, certaines n'ont pas été reconnus par la commaunté. Ces contri-

butions sont quand même citées ici mais n'ont, par conséquent, pas reçu d'étude approfondie de notre part. Voici les autres méthodes de l'état de l'art : voir Annexe B.2.3.

# Chapitre 6

# Explicabilité dans le contexte de l'imagerie IRM fonctionnelle pour la classification du genre

## Sommaire

## 6.1 Contexte global de la contribution et motivations

La volonté d'intégrer des méthodes d'apprentissage machine contemporaines pour résoudre des problématiques liées à la santé est grandissante depuis ces dernières années. La santé est un exemple de milieu applicatif pour lequel il y a un réel besoin de contrôle de la fiabilité des méthodes profondes si ces dernières contribuent au diagnostic pour la patientèle. Cette première contribution est une illustration empirique du potentiel des méthodes d'apprentissage machine moderne pour l'aide au diagnostic en neurologie. En voici maintenant les détails [1]. La neurologie est une spécialité de la médecine s'intéressant aux maladies du système nerveux et en particulier du cerveau. Ces objets d'étude ne sont pas directement observables par les médecins, car une mesure directe (c-à-d opération) sur ces objets d'étude à de lourdes conséquences sur la santé du patient. L'Imagerie par Résonance Magnétique (IRM) est un moyen indirect (c-à-d non invasif pour le patient) de mesure des grandeurs en jeu dans le système nerveux (c-à-d composition chimique). L'IRM repose sur le principe de la Résonance Magnétique Nucléaire (RMN) utilisé en analyse chimique. L'analyse par RMN d'une molécule évalue l'énergie relaxée par les atomes composant la molécule suite à une excitation de ces derniers par un rayonnement électromagnétique. L'énergie relâchée est alors quantifiable du point de vue vibratoire et est la signature de ces éléments. En exploitant ce principe, l'IRM permet alors d'établir la composition chimique de la zone évaluée, et ce de manière non-invasive pour le patient. Pour la neurologie, il existe un cas applicatif de la méthode par l'IRM très important. Lorsqu'il y a une activité neuronale, la quantité d'oxygène contenue dans le sang varie au voisinage de ce neurone. Cette variation locale est alors caractérisée, via l'IRM, et permet alors de mesurer l'activité neuronale locale, c'est le signal Blood-Oxygen-Level dependent (BOLD). En effectuant une mesure globale de ce signal, nous obtenons alors une mesure de l'activité cérébrale du patient. On parle alors de l'Imagerie par Résonance Magnétique fonctionnelle (IRMf).

L'IRMf est devenue un outil central pour les neurologues et a permis de nom-

1. La rédaction de ce chapitre, présentant une de nos contributions, a fortement été inspiré du document originel rédigé en anglais

breuses avancées en neurologie. Une des questions qui a été posée par la communauté des médecins est la suivante : nous savons qu'hommes et femmes se différencient sur de nombreux aspects physiologiques et physiques, mais y a t-il une différence au niveau de l'activité cérébrale entre hommes et femmes, au repos ? Dans la littérature neurologique, des résultats en faveur d'une différence existent, comme évoqués dans [Gong et al., 2009, Xu et al., 2015, Goldstone et al., 2016, Weis et al., 2020, E et al., 2020]

Nous nous sommes posés la même question, mais selon le point de vue de l'apprentissage machine moderne.

## 6.2 Motivations

Dans cette contribution [2], nous avons voulu, à l'aide de méthodes d'apprentissage machine moderne, mener une étude discriminatoire du genre par rapport aux activités cérébrales des individus d'une population saine d'hommes et de femmes, puis identifier les points de différences moyens entre les deux groupes. La littérature étant déjà consensuelle sur cette question, la concordance entre les points de différences illustrés par la littérature neurologique et ceux identifiés par les méthodes d'apprentissage machine moderne serait alors, pour cette expérience au moins, une validation pour l'utilisation de ce type d'approches pour cette tâche. De plus, dans ce contexte les enjeux sociétaux, au premier degré en tout cas, l'utilisation de ce type de méthode ne semble pas impactant. Nous présentons ici une formalisation des hypothèses pour résoudre notre problématique.

## 6.3 Hypothèses

L'enregistrement de l'activité cérébrale grâce aux méthodes d'IRMf est obtenu par la mise en séquence d'enregistrements d'instantanés représentant numériquement le signal BOLD dans des localités. Grâce au signal BOLD, les instantanés sont donc des enregistrements (c-à-d échantillonnage) spatiaux de l'activité cérébrale à un instant donné. La mise en séquence de ces clichés permet alors de caractériser

2. Notre implémentation est disponible : `github.com/araison12/genderxaifmri`

temporellement et spatialement l'activité cérébrale pour une résolution spatiale et temporelle donnée. Typiquement pour un IRM de champ magnétique 3 Tesla que l'on utilise souvent pour les études cliniques, la résolution temporelle est de $0.4Hz$ et la résolution spatiale est de $0.29 \times 0.29 \times 0.8mm^3$. Sous cette forme, on parle de représentation de voxels. Par ailleurs, la littérature neurologique a établi que le cerveau humain est organisé sous forme d'une multitude de réseaux indépendants, chacun étant lié au moins à une fonctionnalité cognitive (parole, vision, etc.) [Hagmann, 2005], on parle alors de connexion et numériquement, de représentation connectomique. Dans notre contexte et comme nous le verrons dans la suite, la représentation connectomique de l'activité cérébrale a certains avantages numériques. De plus, il est possible d'obtenir une représentation connectomique à partir d'une représentation de voxels (c-à-d celle issue de l'acquisition IRMf). En effet, si nous considérons que nous avons une représentation de voxels $\mathbf{I} \in \mathbb{R}^{T \times M \times N \times P}$ où $T \in \mathbb{N}$ représente la dimension temporelle, et $M, N, P \in \mathbb{N}$ représente les trois dimensions spatiales, alors la représentation connectomique de $\mathbf{I}$ avec $Q$ réseaux fonctionnels est donnée par $C(\mathbf{I}) \in \mathbb{R}^{Q \times M \times N \times P}$ tel que :

$$\mathbf{I} = \mathbf{W} \cdot C(\mathbf{I}) \tag{6.1}$$

Où $\mathbf{W} \in \mathbb{R}^{T \times Q}$ est la matrice de mélange de l'information spatiale des représentants numériques des réseaux fonctionnels et leur impact temporel dans $\mathbf{I}$. En effet, on peut bien décomposer numériquement chaque composante de $\mathbf{I}$ comme étant la résultante de la combinaison linéaire des composantes de $Q$ sous signaux. De plus, la littérature neurologique valide l'existence d'une telle décomposition, de l'aspect linéaire du mélange et de l'indépendance statistique (c-à-d indépendance mutuelle des composantes de $C(\mathbf{I})$) des $Q$ sous signaux dus à l'indépendance mutuelle fonctionnelle (c-à-d deux à deux, les réseaux jouent des rôles fonctionnels distincts). Il existe une solution algorithmique pour solutionner 6.1, il s'agit de l'analyse en composantes indépendantes ( [Hyvarinen, 1999]) et notamment son implémentation rapide [Hyvarinen, 1999], qui permet, notamment par le procédé d'orthonormalisation de Gram-Schmidt, d'obtenir les représentants fonctionnels ayant les propriétés statistiques désirés. Bien entendu, une des conditions sur les dimensions vectorielles du problème est que $T \geq Q$. Là où ce procédé algorithmique est intéressant, c'est lorsqu'on choisit $Q$ tel que $T >> Q$. En effet, dans ce cas, les données issues d'un

processus d'acquisition par IRMf, et donc représentées sous forme de voxels peuvent être traduites sous leur forme connectomique impliquant ainsi une réduction de la dimension du problème si $Q$ est bien choisi devant $T, M, N$ et $P$ et en accord avec la littérature neurologique. Comme évoqué, cette réduction est souhaitable lors de l'utilisation de modèles paramétriques comme ceux utilisés dans le contexte de l'apprentissage machine (moderne ou non). Nous présentons ici la description de notre problématique.

## 6.4 Problématique, état de l'art et protocole

Considérons un ensemble de données $\mathcal{D}$ composé de $N \in \mathbb{N}$ enregistrements issus d'acquisitions via IRMf de $N$ individus sains et au repos (c-à-d ne subissant aucun stimulus externe), représentés sous forme de voxels. Ces données ont été converties dans leurs représentations connectomiques, via analyse par composante indépendante, pour les raisons numériques décrites ci-dessus. En effet, nous allons partitionner $\mathcal{D}$ via un modèle d'apprentissage machine moderne $f_{\boldsymbol{\theta}}$ par rapport au genre (c-à-d homme ou femme) des $N$ individus considérés. Cependant, pour des raisons que nous évoquerons dans le paragraphe suivant, ce n'est pas directement la représentation connectomique des acquisitions qui va nous intéresser, mais plutôt leurs matrices de mélange associées. En effet, la représentation connectomique est relativement stable pour une population saine et au repos, car, grossièrement dit, le fonctionnement du cerveau ne varie que peu pour des personnes saines et au repos au sein de cette population [Beckmann and Smith, 2005, Calhoun et al., 2001, Esposito et al., 2005, Guo and Pagnoni, 2008, Vj and Sk, 2004, M et al., 2002]. Ainsi, pour tout $i \in \{1, \cdots, N\}$, chaque individu est représenté par une matrice de mélanges $\mathbf{X}_i \in \mathcal{D}_{\mathbf{X}}$ de longueurs temporelles $T \in \mathbb{N}$ évoluant sur $Q \in \mathbb{N}$ réseaux fonctionnels, on a donc, numériquement, $\mathbf{X}_i \in \mathbb{R}^{p \times q}$. Puis on définit la quantité $y_i \in \{0, 1\}$ comme représentant le genre du $i$-ème individu. Nous avons donc $\mathcal{D} = \mathcal{D}_{\mathbf{X}} \times \{0, 1\}$. La fonction $f_{\boldsymbol{\theta}}$ va alors de $\mathcal{D}_{\mathbf{X}}$ dans $\{0, 1\}$. Elle est paramétrée par $\boldsymbol{\theta}$ par un vecteur réel de taille définie par la structure de $f_{\boldsymbol{\theta}}$. Dans notre cas, $f_{\boldsymbol{\theta}}$ sera un réseau de neurones convolutif, car nos données répondent à la propriété d'insensibilité de la détermination du genre à la translation temporelle des acquisitions des activités

cérébrales et donc, par constance de la représentation sous forme de concepts sur $\mathcal{D}$, des mélanges. Nos reviendrons sur l'ensemble des choix et hypothèses énoncés ici dans la conclusion de ce chapitre. Nous présentons ici le protocole numérique expérimental :

— Nous cherchons dans un premier temps à résoudre le problème d'optimisation suivant :

$$\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} L(f_{\boldsymbol{\theta}}(\mathbf{X}_i), y_i) \tag{6.2}$$

— puis, pour tout $i \in \{1, \cdots, N\}$, nous souhaitons obtenir une explication $\psi(\mathbf{X}_i, f_{\boldsymbol{\theta}^\star}) \in \mathbb{R}^{T\times Q}$ calculée par la méthode explicative $\psi : \mathbb{R}^{T\times C} \times \{f | f : \mathcal{D}_{\mathbf{X}} \to \{0,1\}\} \to \mathbb{R}^{T\times C}$ local et post-hoc, telle que :

$$\min_{\psi} \sum_{i=1}^{N} \|\mathbf{R}(\psi(\mathbf{X}_i, f_{\boldsymbol{\theta}^\star}) \circ \mathbf{X}_i) - \mathbf{R}(\mathbf{X}_i)\|_2 \tag{6.3}$$

où pour tout $\mathbf{X} \in \mathbb{R}^{T\times Q}$,

$$\mathbf{R}(\mathbf{X}) = \left( \frac{Cov(\mathbf{X}_{:,q_1}, \mathbf{X}_{:,q_2})}{Var(\mathbf{X}_{:,q_1}) Var(\mathbf{X}_{:,q_2})} \right)_{(q_1,q_2)\in\{1,\cdots,Q\}^2} \tag{6.4}$$

Est la matrice de corrélation (temporelle, sur $T$ étapes temporelles) des signaux d'activités entre les $Q$ réseaux fonctionnels où $\mathbf{X}_{:,q_1}$ représente les valeurs temporelles de la région $q_1$ pour $q_1 \in \{1, \cdots, Q\}$ pour le mélange du $i$-ème individu de $\mathcal{D}$.

Les motivations pour satisfaire la formulation 6.3 sont les suivantes :

— pour une méthode explicative $\psi$, et pour tout $i \in \{1, \cdots, N\}$, l'explication $\psi(\mathbf{X}_i, f_{\boldsymbol{\theta}^\star})$ est issue d'une méthode explicative
  — Adaptée aux réseaux de neurones convolutifs : car c'est le choix architectural que nous avons utilisé pour mener notre expérience et nous avons justifiez ce choix plus haut.
  — locale : car nous souhaitons mener une analyse patient par patient afin de mesurer une statistique sur $\mathcal{D}$, et le paradigme local n'est a priori pas le seul pour obtenir de telles statistiques, mais est probablement le plus naturel pour ça.
  — Post-hoc : cette caractérisation ne doit se faire qu'une fois que le modèle prédictif a atteint un régime stable et précis (selon les critères pratiques usuels).

— Le procédé de masquage via le produit d'Hadamard est usuel dans ce contexte et dans la littérature relative à celui-ci. Dans ce contexte, la matrice $\psi(\mathbf{X}_i, f_{\boldsymbol{\theta}^\star})$ est assimilée à une carte d'importance où chacune des composantes a une valeur réelle proportionnelle à l'importance du facteur présent aux mêmes coordonnées dans $\mathbf{X}_i$. Autrement dit, on pondère $\mathbf{X}_i$ avec l'importance des descripteurs qui la composent de manière à caractériser la "connaissance numérique" de $f_{\boldsymbol{\theta}^\star}$ par rapport à $\psi$ et $\mathcal{D}$. Notez que $\psi(\mathbf{X}_i, f_{\boldsymbol{\theta}^\star})$ a subi préalablement une normalisation pour des raisons de comparabilité entre les explications fournies par $\psi$.

— La caractérisation d'interaction des différents réseaux fonctionnels joue le rôle de signature du genre ici. Dans la littérature neurologique, cette caractérisation est considérée comme linéaire. Ainsi l'outil statistique le plus adéquat sur ces hypothèses est l'étude de la corrélation de l'activité de ces réseaux. Dans ce contexte, ces études statistiques ont été menées , voir [Gong et al., 2009,Xu et al., 2015,E et al., 2020,Weis et al., 2020,Goldstone et al., 2016, Allen et al., 2011]. Ainsi la matrice $\mathbf{R}(\mathbf{X}_i)$ joue le rôle de signature et de vérité terrain pour la caractérisation du genre en fonction de l'activité cérébrale. Ainsi la formulation 6.3 permet, sur un panel de plusieurs méthodes explicatives, de sélectionner celle qui caractérise le mieux la connaissance neurologique déjà établie dans la littérature via des arguments numériques construits par $f_{\boldsymbol{\theta}^\star}$.

Nous présentons maintenant les applications numériques.

## 6.5 Résultats

Dans cette partie, nous précisons les données que nous avons utilisées pour mener cette étude, ainsi que l'architecture du modèle prédictif $f_{\boldsymbol{\theta}}$, ses performances par rapport aux métriques usuelles, les différentes méthodes explicatives utilisées et les résultats obtenus en suivant le protocole ci-dessus.

### 6.5.1 Données

Les données que nous utilisons dans cette étude sont issues de S1200 WU-Minn Human Connectome Project (publié en juin 2017 nommé HCP1200 Parcellation+Timeseries+Netmats) [3]. L'ensemble de données est constitué d'enregistrements d'activités cérébrales de $N = 812$ patients sains au repos, représenté sous forme connectomique. Les données consistent en $T = 1200$ volumes pour chaque enregistrement pour un total de 4 800 volumes pour chaque sujet sur les quatre enregistrements de 15 minutes chacun. Chaque série de IRMf de chaque sujet a été prétraitée par le consortium HCP [Sm et al., 2013]. Les données ont été prétraitées de manière minimale [Mf et al., 2013] et les artefacts ont été supprimés à l'aide d'ICA+FIX, [L et al., 2014] et [Salimi-Khorshidi et al., 2014]. Nous avons choisi de mener l'étude avec un nombre faible de composantes indépendantes ($C = 25$). En effet, des études antérieures ont démontré que dans le cas du processus de prétraitement utilisé ci-dessus, le choix de cette résolution permet de produire des composantes qui correspondent à des segmentations anatomiques et fonctionnelles connues selon [A et al., 2010, Kiviniemi et al., 2009, Sm et al., 2013, M et al., 2010] . Les caractéristiques démographiques des sujets de l'étude sont présentées dans le Tableau 6.1 ci-dessus.

| | 22-25 ans | 26-30 ans | 31-35 ans | 36+ ans | Total |
|---|---|---|---|---|---|
| Femmes | 70 | 181 | 153 | 6 | 410 |
| Hommes | 125 | 176 | 99 | 2 | 402 |
| Total | 195 | 357 | 252 | 8 | 812 |

TABLE 6.1 – Distribution des sujets de l'ensemble des données

Pour la résolution $C$, nous avons construit la correspondance entre les réseaux anatomofonctionnels connus et nos $C$ composantes. Cette cartographie de base a été élaborée à la main par un expert. La cartographie de référence est résumée dans le Tableau 6.2.

3. Les données ont été fournies [en partie] par le Human Connectome Project, WU-Minn Consortium (chercheurs principaux David Van Essen et Kamil Ugurbil ; 1U54MH091657) financé par les 16 instituts et centres du NIH qui soutiennent le NIH Blueprint for Neuroscience Research ; et par le McDonnell Center for Systems Neuroscience de l'Université de Washington.

| **Composante** | **Nom du réseau** |
|---|---|
| 1 | Cortex Visuel Bilatéral (BVC) |
| 2 | Cortex Visuel Interne (IVC) |
| 3 | Réseau d'Attention Dorsale (DAN) |
| 4 | Réseau Mode Défaut (DMN) |
| 5 | Réseau Executif Central Droite (RCEN) |
| 6 | Réseau de Saillance (SN) |
| 7 | Réseau Executif Central Gauche (LCEN) |
| 8 | Precuneus / Réseau Mode Défaut (P/DMN) |
| 9 | Réseau Visuel Postérieur (PVN) |
| 10 | Réseau Exécutif Central Bilatéral (BCEN) |
| 11 | Réseau temporelle postérieur (TPN) |
| 12 | Neocerebellum (N) |
| 13 | Cortex Moteur (MC) |
| 17 | Réseau Frontal Singulier (FSN) |
| 18 | Cerebellum Gauche (LC) |
| 19 | Precuneus (P) |
| 20 | Réseau d'Attention Ventral (VAN) |
| 23 | Réseaux Fronto-perculaire (FPN) |
| 25 | Striatum Ventral (VS) |

TABLE 6.2 – Tableau de correspondance entre les ICs et les réseaux anatomofonctionnels. Les composantes n°14,15,16,21,22,24 sont assimilées à des composantes bruitées.

Nous présentons ici la définition du modèle numérique et ses paramètres.

### 6.5.2 Optimisation du paramètre du modèle prédictif

L'analyse en composantes indépendantes génère des cartes spatiales statistiquement indépendantes les unes des autres, dont les composantes représentent les fluctuations BOLD spontanées synchronisées pendant l'état de repos du cerveau. Ces cartes représentent non seulement des réseaux intrinsèquement connectés, mais aussi du bruit dû à la respiration, aux mouvements des yeux ou de la tête et aux pulsations du liquide céphalo-rachidien, facilement identifiable par inspection visuelle ou par des algorithmes entraînés. Les artefacts sont identifiés et éliminés par [L et al., 2014]. Pour classer avec précision les sujets en fonction de leur sexe et tenir compte des différences entre les sexes dans l'organisation fonctionnelle du cerveau d'une manière spatio-temporelle, nous utilisons un réseau de neurones convolutif sur 4 séries temporelles moyennées. Le réseau est composé de 3 blocs convolutifs suivis de 3 blocs linéaires. Chaque bloc convolutif est suivi d'une fonction d'activation relue et chaque bloc linéaire est muni de dropout. Ensuite, nous appliquons la transformation sigmoïde pour obtenir une valeur dans l'intervalle [0,1]. Nous utilisons l'entropie croisée binaire comme fonction de perte. Comme montré dans la Figure 6.1 notre modèle formé atteint une erreur de test négligeable autour de la valeur 0,1 pour l'entropie croisée binaire (BCE) après 22 époques. Nous notons que le processus d'apprentissage est relativement stable à partir d'un certain moment .

### 6.5.3 Méthodes explicatives utilisées et analyses

Dans cette étude, nous comparons la pertinence de plusieurs méthodes explicatives de la littérature répondant aux pré-requis précédemment décrits. Voici celles qui ont été impliquées :

— Saliency (SA) [Simonyan et al., 2014b] est une approche simple pour calculer l'attribution de l'entrée, renvoyant le gradient de la sortie par rapport à l'entrée. Cette approche peut être comprise comme une expansion de Taylor du premier ordre du réseau à l'entrée, et les gradients sont simplement les coefficients de chaque caractéristique dans la représentation linéaire du modèle. La valeur absolue de ces coefficients peut être considérée comme représentant

l'importance de la caractéristique.

— Input × Gradient (IXG) est une extension de l'approche Saliency, qui prend les gradients de la sortie par rapport à l'entrée et les multiplie par les valeurs des caractéristiques d'entrée. Les gradients sont simplement les coefficients de chaque entrée, et le produit de l'entrée par un coefficient correspond à la contribution totale de la caractéristique à la sortie du modèle linéaire.

— Feature Ablation (FA) [Molnar, 2022] est une approche basée sur la perturbation pour calculer l'attribution, impliquant le remplacement de chaque caractéristique d'entrée par une valeur de base / référence donnée (par exemple 0), et le calcul de la différence dans la sortie. Les caractéristiques d'entrée peuvent également être regroupées et éliminées ensemble plutôt qu'individuellement.

— Feature Permutation (FP) [Molnar, 2022] est une approche basée sur la perturbation qui prend chaque caractéristique individuellement, permute aléatoirement les valeurs des caractéristiques au sein d'un lot et calcule le changement dans la perte résultant de cette modification.

— Layer Wise Relevant Propagation (LRP) [Bach et al., 2015b] est un système de rétropropagation de la pertinence des caractéristiques distribuées de manière égale. La méthode LRP a été largement utilisée par la communauté de l'apprentissage profond, en particulier dans le domaine de la vision par ordinateur.

— Deconvolution (DEC) [Zeiler and Fergus, 2013] calcule le gradient de la sortie cible par rapport à l'entrée, mais la rétropropagation des fonctions relue est neutralisée de sorte que seuls les gradients non négatifs sont rétropropagés. Dans la déconvolution, la fonction relue est appliquée aux gradients de sortie et directement rétropropagée.

— ScoreCAM (SCAM) [Wang et al., 2020a] est une méthode d'explication visuelle post-hoc basée sur la cartographie d'activation de classe (CAM) [Zhou et al., 2015], Score-CAM se débarrasse de la dépendance sur les gradients en obtenant le poids de chaque carte d'activation par son score de passage vers l'avant sur la classe cible. Le résultat final est obtenu par une combinaison linéaire de ces poids et des cartes d'activation suréchantillonnées de la dernière couche convolutive.

Selon notre protocole, nous observons dans la Figure 6.1 que les méthodes DEC, FA, FP, IXG et SA sont regroupées dans un voisinage relativement éloigné du seuil de vérité terrain par rapport à la méthode SCAM. Cela signifie que ces méthodes ne caractérisent pas assez les interactions entre les $C$ réseaux. Le manque de pertinence des méthodes évoquées a été étudié dans [Adebayo et al., ]. Par ailleurs, les méthodes basées sur la perturbation des descripteurs ne permettent pas d'obtenir de meilleurs résultats pour la récupération des activités temporelles des circuits intégrés. La méthode SCAM permet, elle, d'obtenir la plus grande fidélité par rapport à la vérité terrain. Parmi les méthodes sélectionnées, la méthode SCAM est celle qui a donc été retenue pour conclure sur les résultats de notre étude.

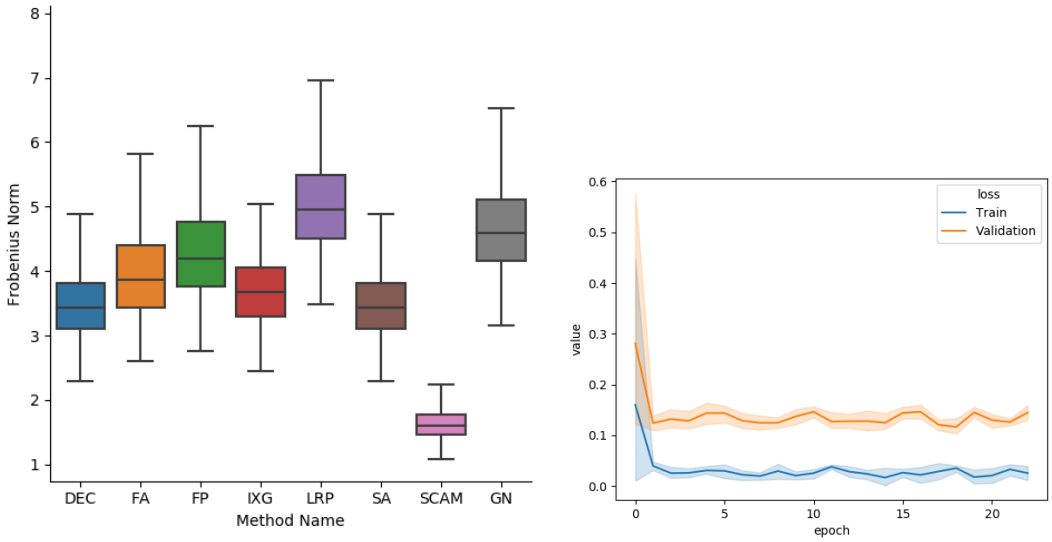



FIGURE 6.1 – *Gauche* - Dispersion des méthodes selon notre pipeline. La boîte montre les quartiles des distributions de la norme de Frobenius dérivées de notre évaluation. *Droite* - la ligne indique la perte moyenne pour chaque époque et la zone colorée montre l'intervalle de confiance de 95% de la moyenne estimée sur 5 entraînements indépendants.

## 6.6 Conclusion

La littérature a toujours besoin d'annotations d'experts pour identifier les réseaux anatomofonctionnels dans les cerveaux. Grâce à notre score de classification élevé basé sur notre modèle d'apprentissage machine moderne et à l'utilisation d'une méthode explicative adaptée, nous avons émulé l'expertise d'un expert pour extraire

et récupérer les réseaux fonctionnels discriminant le genre. Les études [Gong et al., 2009], [Xu et al., 2015], [Goldstone et al., 2016], [Weis et al., 2020], [E et al., 2020] indiquent que les réseaux discriminant le genre le plus fort sont le Default Mode Network (IC n°4), le réseau de saillance (IC n°6), le réseau exécutif central gauche (IC n°7) ainsi que le réseau exécutif central droit (IC n°5). Pour mettre en évidence les différences intrinsèques entre les femmes et les hommes, nous évaluons l'amplitude des différences de corrélation entre ces deux sexes (voir la Figure 6.2). Sans surprise, nous constatons qu'il existe des différences entre les femmes et les hommes principalement pour le réseau du mode par défaut, le réseau de la saillance et les réseaux exécutifs centraux gauches/droites. Étant donné que la méthode LRP n'est pas pertinente pour notre tâche, elle n'entraîne pas de différences significatives entre les deux sexes. À l'inverse, pour la méthode SCAM nous retrouvons les tendances statistiques déjà établies dans la littérature.

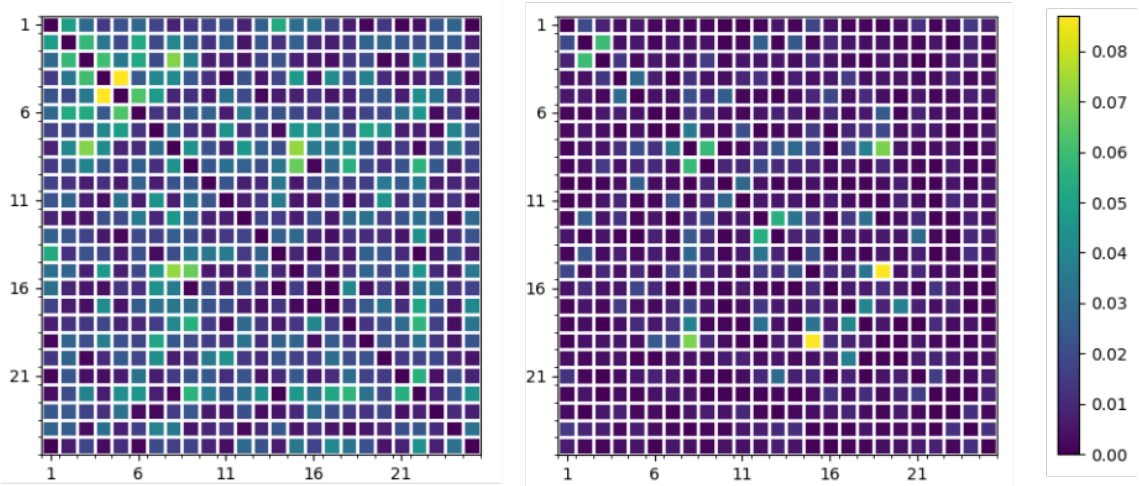



FIGURE 6.2 – Amplitude des différences de corrélation entre les hommes et les femmes : Matrice des corrélations de Pearson : (Gauche) par SCAM (Droite) par LRP

De nombreux algorithmes classiques pour la classification des sexes basée sur les activités des réseaux cérébraux au repos ont été proposés par l'état de l'art. Leur grande facilité d'interprétation est souvent en contradiction avec leurs performances absolues. À des fins de diagnostic, un modèle interprétable est plus approprié qu'un modèle opaque. Pour améliorer les diagnostics médicaux, nous utilisons la puissance des approches profondes et nous appliquons le modèle SCAM pour saisir les règles

de décision de classification de notre modèle prédictif. Nos résultats montrent qu'un tel modèle suit les mêmes règles de décision qu'un modèle classique, mais avec une performance de classification accrue tout en récupérant les résultats de la littérature neurologique.

## 6.7 Discussions, étude épistémologique, limites et ouvertures

L'expérience présentée ci-dessus peut être vue comme une prémisse pour établir un argument favorable pour l'utilisation de méthodes d'apprentissage machine moderne dans des contextes médicaux connexes à celui de l'expérience. Cette expérience est une étude des références déjà existantes. Cependant les degrés de libertés de l'étude, et les choix faits peuvent être discutables. Voici quelques uns d'entre eux :

— L'apriori géométrique des données : nous avons utilisé un réseau de neurones convolutif en tant que modèle prédictif pour notre étude. Cependant la représentation connectomique des données d'activité cérébrale ne répond pas globalement au principe d'invariance par translation. En effet, cela n'est vrai que pour la dimension temporelle des données, mais pas pour la dimension positionnelle. La dimension positionnelle, elle, répond au principe d'invariance par permutation, qui est un cas plus général que pour les translations. En effet, comme déjà évoqué, on peut voir une translation comme une permutation cyclique qui est un cas particulier de permutation. Cette invariance à la permutation n'est pas prise en compte dans la conception du modèle prédictif et est donc une des limites de ces travaux.

— L'apriori géométrique du modèle prédictif : la conception de notre modèle prédictif comme un réseau de neurones convolutif répond à la propriété d'équivariance, mais pas à celle d'invariance. En effet, entre la partie descriptive et classifiante du réseau se trouve un opérateur d'aplatissement vectoriel. Ce dernier ne répond pas au principe d'invariance et cette propriété n'est pas possédée par la partie classifiante du réseau qui est simplement un perceptron multicouche. Ainsi, notre modèle ne répond pas strictement au paradigme de l'apprentissage géométrique profond. Cette conception des réseaux de neu-

rones convolutifs est cependant la plus répandue dans la littérature.

— Optimalité du modèle prédictif : comme décrit dans l'équation 6.2, nous supposons que le paramètre du modèle prédictif est optimal, au sens où la fonction de coût associée a atteint son minimum global, ce qui en pratique est faux, nous n'avons qu'un candidat permettant d'avoir des performances satisfaisantes. En utilisant Adam comme algorithme d'optimisation, nous savons qu'il n'existe pas de moyen théorique pour la convergence vers un minimum global de la fonction de coût et donc de pouvoir de représentation et de classification optimale pour le modèle prédictif. Ainsi, la non-optimalité du modèle prédictif et l'utilisation de méthodes explicatives ayant elles-mêmes leurs propres limites peut altérer les conclusions de ces travaux. Notez que cette limitation est inhérente à tout problème d'apprentissage machine et est inéluctable. Cela reste cependant une limite en soi.

— Représentativité de l'ensemble de données : la définition du modèle prédictif ou l'optimisation de sa représentativité n'est pas la seule limite pour obtenir un comportement optimal de celui-ci. En effet, ce dernier est dépendant de la qualité de représentativité statistique de l'ensemble de données et ce point est particulièrement important dans le domaine médical. Les données numériques issues du monde médical sont peu volumineuses, ce qui est dû aux coûts des acquisitions (humainement, financièrement, temporellement et légalement). De plus, le prétraitement de ces acquisitions engendre le même genre de coût. Or, les tâches de traitement de l'information dans le domaine médical sont lourdes et complexes. La limite du volume de données entraîne donc nécessairement une sous-représentativité de l'ensemble de données devant celle nécessaire pour résoudre statistiquement la tâche. Ainsi la pertinence du modèle prédictif peut être facilement remise en question et nécessite donc une forte régularisation qui n'est peut-être pas forcément aussi importante en pratique que nécessaire.

— Choix de l'ensemble de méthodes explicatives : pour mener notre étude comparative, nous avons choisi les méthodes utilisées et reconnues par la communauté au moment de l'étude, et celles répondant aux paradigmes que nous avions choisis (local et post-hoc). Cependant, ces méthodes sont nécessaires

et incomplètes dans le calcul de leurs explications (par exemple biais, définitionnel, sujet à des attaques, etc.). Par conséquent, le choix de l'ensemble des méthodes explicatives est lui-même nécessairement sous représentatif en termes d'explicabilité et ne peut apporter qu'une vision partielle de l'explication. Ainsi, le choix de cet ensemble de méthodes impacte fortement les conclusions de cette étude et est limitant.

— Confiance absolue dans les méthodes explicatives : dans la formulation 6.3, le critère de minimaliste ne porte ni sur les données, ni sur le calcul des méthodes explicatives, ni sur le traitement statistique des données et des explications, mais bien sur les méthodes explicatives en elles-mêmes. C'est-à-dire que la pertinence de tous ces facteurs, pourtant cruciaux, est occultée dans l'étude et donc dans ses conclusions. Dans l'étude, ces facteurs sont justement considérés, implicitement, comme optimaux, et ne sont pas rediscutés. En particulier, les explications fournies sont toutes considérées comme absolument pertinentes, ce qui n'est en pratique jamais le cas et ne peut pas l'être comme évoqué précédemment. Elles ne sont mises de côté que lorsqu'elles ne recouvrent pas la vérité terrain alors que cette dernière est déterminée par de multiples facteurs dont une partie est au moins discutable et qu'elles ne font pas forcément l'unanimité dans la littérature.

— Dépendance de la conclusion à un expert humain : une autre considération concernant les conclusions de cette étude est que la validation des résultats est traitée par un neurologue, dont les connaissances et les raisonnements ne sont pas fondamentalement à remettre en question, mais, si nous avions collaboré avec un autre neurologue, il est possible que les conclusions soient différentes (mais probablement pas diamétralement opposées). La conclusion comporte donc une part de subjectivité et donc peuvent être objectivement être remise en question.

— Significativité statistique des résultats : les amplitudes des corrélations observées entre les réseaux considérés comme saillants par les méthodes explicatives sont relativement faibles comparativement à d'autres études statistiques que l'on trouve dans la littérature. Ainsi, les conclusions sur les interactivités entre composantes sont-elles pertinentes ? Cette question est d'autant plus

légitime lorsque l'échantillon n'est pas le plus fourni, comme cela est le cas dans cette expérience. Par extension, les conclusions sont-elles valides ?

— Représentativité statistique du protocole : Les statistiques utilisées dans ce protocole sont celles utilisées communément dans la littérature neurologique. Cependant il est probable que les phénomènes d'interactions entre les réseaux aient des composantes non linéaires. En effet, le cerveau est un objet d'étude très complexe. Ainsi, quelle est la pertinence de l'utilisation de mesure de corrélations entre ces réseaux ?

— Choix du protocole et communauté : l'ensemble de tous ces constats montrent que les conclusions de cette étude peuvent être fragiles pour des raisons qui sont, soit inévitables de par la nature du problème, soit en lien avec un manque de connaissances scientifiques sur les objets en jeu dans l'étude. Il peut exister aussi d'autres constats non évoqués ici. Dans les parties de l'étude où les connaissances scientifiques manquent, les choix faits dans l'établissement du protocole ont été faits en accord avec la littérature existante, bien que parfois pas forcément consensuelle sur ces points.

Ainsi, cette étude a permis, via des arguments numériques nouveaux, de retrouver des résultats de la littérature neurologique. C'est la contribution de cette étude. Pour son caractère original, elle a fait l'objet d'une publication dans une conférence internationale à comité de lecture :

A. Raison, P. Bourdon, C. Habas and D. Helbert, *"Explicability in resting-state fMRI for gender classification"*, 2021 Sixth International Conference on Advances in Biomedical Engineering (ICABME), Werdanyeh, Lebanon, 2021, pp. 5-8, doi : 10.1109/ICABME53305.2021.9604842.

## 6.8 Inscription de la contribution avec le projet de recherche et la littérature

Nous présentons ici les points d'articulation de cette contribution dans l'ensemble des travaux de cette thèse et l'état de l'art :

— **Inscription de la contribution dans la littérature existante :** comme évoqué précédemment, les apports de cette contribution sont techniques. Elle

propose une nouvelle vision, formulation et solution d'une problématique déjà résolue avec des arguments contemporains.

— **Inscription de la contribution dans la problématique de recherche :** les modèles prédictifs utilisés répondent au paradigme de l'apprentissage géométrique profond. Nous avons dans cette contribution mis en lumière différentes méthodes explicatives adaptées à ces modèles et déjà existantes dans la littérature. Cependant l'interprétation des calculs issus de ces méthodes et donc leurs valorisations pour la communauté, est nouvelle. Nous avons donc proposé une analyse et une interprétation des méthodes d'apprentissage machine contemporaines appliquées au domaine médical.

L'ensemble des points soulevés précédemment ont nourri les réflexions pour mener les travaux suivants. Bien entendu, en réponse à la multitude de ces points, les degrés de liberté accessibles pour mener les futurs travaux sont nombreux. Nous avons donc fait des choix arbitraires en accord avec nos sensibilités et les questions encore en suspens dans la littérature. Nous les détaillons dans la partie suivante. [4]

4. Ces travaux avaient été expertisé par Pr. Christophe Habas, PU-PH à l'hopital du Quinze-Vingt en neuro-imagerie. Ces travaux matérialisaient le début de nos collaborations pour mener ce projet de recherche dans le contexte du médicale. Malheureusement, Pr. Christophe Habas est décédé 27 mai 2022. Nous avons donc dû réorienter le cadre applicatifs de nos recherches.

# Chapitre 7

# ScoreCAM GNN : une méthode explicative locale et post-hoc pour l'apprentissage géométrique profond, définition et application

## Sommaire

## 7.1 Contexte global de la contribution et motivations

La majorité des cas d'usages des GNN portants sur des problèmes de classification de noeuds ou de graphes. Dans ce genre de problématiques, les méthodes explicatives, locales, post-hoc attributives sont le plus largement utilisées dans le monde académique ou industriel, car, empiriquement,elles sont les plus à même de fournir de la connaissance directement. Par directement, nous entendons que sous ce paradigme, l'analyse des explications ne demande pas plus de prérequis que celui pour analyser l'écosystème courant. De plus, le traitement statistique des explications est souvent proche de celui utilisé pour l'écosystème. Pour les GNN, il existe des méthodes explicatives locales post-hoc et attributives comme nous l'avons vu. Dans l'étude précédente, nous n'avons utilisé que des méthodes locales, post-hoc et attributives pour l'obtention de nos résultats. Empiriquement, nous avons mesuré que la méthode ScoreCAM [Wang et al., 2020a] avait de bien meilleurs résultats devant le reste de la population des méthodes explicatives considérées dans l'étude. Nous présentons ici les motivations pour ces travaux [1].

## 7.2 Motivations

Dans l'étude précédente, nous n'avons utilisé que des méthodes explicatives locales, post-hoc et attributives. De plus, nous avons mesuré empiriquement les qualités de la méthode ScoreCAM comparativement à la cohorte de méthodes utilisées dans l'étude précédente. Or, la méthode ScoreCAM est utilisable pour n'importe quel type de modèle prédictif basé sur les Convolutional Neural Network (CNN). Comme évoqué, les GNN sont une généralisation sur plusieurs points des CNN. Des méthodes explicatives locales, post-hoc et attributives pour les GNN on été proposées dans la littérature notamment celles basées sur l'approche Class Activation Mapping (CAM) [Baldassarre and Azizpour, 2019b] dont des généralisations aux GNN des méthodes pour CNN utilisées dans l'étude précédente. Ainsi, notre ques-

1. La rédaction de ce chapitre, présentant une de nos contributions, a fortement été inspiré du document originel rédigé en anglais

tion a été la suivante : La méthode ScoreCAM peut-elle aussi se généraliser aux GNN et conserve t-elle ses qualités dans le contexte des GNN ? Quelles sont ses potentielles limites ? Nous présentons ici nos éléments de réponses [2].

## 7.3 Hypothèses

Par la suite, nous allons considérer que le modèle prédictif $f_{\boldsymbol{\theta}}$ sous-jacent dans cette étude répond au paradigme des GNN.

## 7.4 Problématique, état de l'art et protocole

Dans le cadre du paradigme local, post-hoc et attributif, les méthodes explicatives fournissent des explications en construisant un masque d'importance des descripteurs. Cela se fait principalement en cherchant une partie pertinente dans la donnée d'entrée qui préserve le mieux le comportement du modèle prédictif par rapport à cette donnée. La recherche de cette partie pertinente peut être formulée au premier abord comme un problème d'optimisation. Cependant, dans le contexte de l'apprentissage géométrique profond, ce problème d'optimisation s'exprime conjointement selon les deux modalités suivantes : le domaine et le signal. Dans le cas des GNN, la complexité du dit problème d'optimisation entre les deux modalités est inégale. En effet :

— Selon la modalité du domaine : la nature discrète du domaine rend le problème d'optimisation de sélection de sous structures (c-à-d sous graphes) combinatoire, ainsi la complexité de toute heuristique naïve est nécessairement exponentielle en la taille du graphe d'entrée.

— Selon la modalité du signal : la nature topologique de l'espace sur lequel le signal prend ses valeurs (c-à-d souvent un espace de Hilbert réel) permet d'utiliser des stratégies connues et efficaces pour extraire des sous-signaux (c-à-d des processus de filtrage). Souvent, ce sont des problèmes d'optimisation difficiles, mais où il existe des approches ayant un coût modéré.

Ainsi, une approche commune, dans la littérature, pour développer des méthodes explicatives locales, post-hoc et attributives adaptées au GNN, est d'occulter ex-

2. Notre implémentation est disponible : `github.com/araison12/scgnn`

plicitement la modalité de domaine dans le processus d'explication et d'exploiter la structure de l'espace des signaux pour les construire [Baldassarre and Azizpour, 2019b, Pope et al., 2019a]. Les méthodes locales et attributives fournissent leurs explications dans le même format que les données d'entrée, et donc en attribuant à chaque descripteur de la donnée d'entrée une valeur d'importance du descripteur selon l'écosystème (donnée d'entrée, modèle prédictif, etc.). En cueillant ces valeurs, qui s'expriment sur le domaine entier de la donnée d'entrée, on obtient alors un moyen d'extraire une partie du domaine pertinente (en binarisant le domaine selon le seuil), ce qui permet alors de réunir les deux modalités de départ pour formuler l'explication. La méthode ScoreCAM [Wang et al., 2020a] repose sur ce principe. Nous présentons ici son fonctionnement dans le cas des CNN et sa généralisation au GNN.

### 7.4.1 ScoreCAM CNN

ScoreCAM [Wang et al., 2020a] est une méthode explicative locale, post-hoc et attributive développée initialement pour les modèles prédictifs types CNN. Elle est dérivée de la méthode GradCAM [Selvaraju et al., 2020]. GradCAM a répondu aux principaux défauts de la méthode CAM, à savoir la dépendance à l'égard de l'architecture du modèle prédictif, la saturation du gradient, l'évanouissement du gradient ou les problèmes de fausses attributions. GradCAM pour GNN a été introduite dans [Pope et al., 2019a]. ScoreCAM a adressée certains problèmes de la méthode GradCAM, parmi eux : l'augmentation de la confiance, le lissage avec normalisation et l'utilisation de méthodes de suréchantillonnage, voir [Wang et al., 2020a] . Mis à part l'utilisation de méthodes de suréchantillonage, ces problèmes ne sont pas en fonction de la structure géométrique de la donnée. Ainsi, la méthode GradCAM subit les mêmes défauts dans le cas des CNN ou des GNN. Cependant, ScoreCAM pour CNN utilise des aprioris spécifiques à la structure géométrique des images (c-à-d domaine type grille) dans sa définition. Ainsi, la généralisation de ScoreCAM aux GNN nécessite de traiter quelques points, que voici :

### 7.4.2 Points clés pour la généralisation

L'implémentation originale de ScoreCAM a introduit des concepts intéressants, nous généralisons ici chacun de ces points clés afin de pouvoir fournir des explications sûres GNN.

#### Informations de haut niveau et normalisation

l'implémentation originale de la ScoreCAM utilise des informations de haut niveau et une normalisation des descripteurs pour calculer ses explications. Étant donné que CNN et GNN suivent le modèle théorique du Geometric Deep Learning (GDL), en particulier ils utilisent le principe de hiérarchisation ou de représentation des données. Avec la profondeur du modèle prédictif, nous avons accès à des représentations de plus en plus haut niveau des données et l'utilisation de la dernière couche, celle des descripteurs de plus haut niveau, reste pertinente pour construire l'explication, car c'est le niveau de représentation de la donnée le plus intelligible humainement. Comme pour les CNN, les signaux évoluant sur les GNN restent de la même nature du point de vue algébrique et analytique. Nous définissons l'opérateur de normalisation $S$ de vecteur réel ou pour tout vecteur réel $\mathbf{x} = (x_i)_i$, tel que $\mathbf{x} \neq \lambda \mathbf{1}_K$ pour tout $\lambda \in \mathbb{R}$,

$$S(\mathbf{x}) = \frac{\mathbf{x} - \min_i \mathbf{x}_i}{\max_i \mathbf{x}_i - \min_i \mathbf{x}_i} \tag{7.1}$$

Ainsi, la notion de normalisation a toujours un sens numériquement. L'agrégation d'informations de haut niveau et la normalisation sont des propriétés compatibles avec la définition théorique des GNN et sont donc des arguments en faveur de la possibilité de la généralisation de la méthode ScoreCAM aux GNN.

#### Reconsidération de la pertinence de l'échantillonnage

En outre, la méthode ScoreCAM originale utilise une approche de suréchantillonnage des descripteurs de haut niveau afin de s'adapter à toute architecture de CNN. Le suréchantillonnage à la dimension d'entrée originale permet d'effectuer un masquage de l'entrée pour calculer le score des descripteurs masqué. Dans le contexte de GNN, en raison d'aprioris beaucoup plus faibles sur la structure géométrique

du domaine par rapport au domaine type grille, la définition de l'opérateur de suréchantillonnage n'est pas claire. Bien qu'il en existe théoriquement pour les GNN comme vu précédemment, ils ne font pas forcément sens sémantiquement dans le contexte de l'explicabilité. En effet, le suréchantillonnage étend un signal sur un domaine "étendu" par rapport au domaine initial. Avec les régularités de la structure géométrique de domaines de type grille, il est facile d'effectuer cette extension en étendant, selon les régularités, les bords du domaine. Par exemple, prendre des photos d'une scène du monde réel n'est qu'une procédure de discrétisation d'une scène physique continue (c'est-à-dire que la lumière est diffusée grâce aux photons) sur une structure de grille finie. La quantité d'informations contenues dans le phénomène physique capturé ne varie pas en fonction de la taille de la grille, même dans un petit voisinage du domaine. Cette indépendance est due à la régularité du domaine type grille. Pour les données de type graphe, qui ne satisfait plus ce genre d'hypothèses, nous ne pouvons pas définir un tel étirement du domaine qui préserve la quantité d'informations incorporées. Il y a nécessairement un ajout d'information et la légitimité de cet ajout est questionnable : cela peut avoir un sens mathématiquement parlant, mais peut ne pas en avoir sémantiquement et donc dans le contexte de l'explicabilité. En effet, quelle est la signification de l'introduction d'un nouveau noeud dans un graphe, pour un voisinage donné, en particulier lorsque le graphe concerné représente des molécules chimiques ou un réseau social par exemple ? Par conséquent, nous renonçons à l'utilisation de procédures de suréchantillonnage pour notre implémentation de ScoreCAM pour GNN.

#### Invariance du score de base

L'implémentation initiale de ScoreCAM s'appuie sur le score relatif (par rapport à une référence) pour calculer le score d'un ensemble de descripteurs donné. Sur la base de notre définition, nous montrons que ce score est indépendant de toute référence donnée. Par conséquent, le calcul du score, relativement à un point de référence ou non, n'a pas d'impact sur le calcul de l'explication. Ainsi, nous utiliserons le score de manière non relative. Nous présentons ici l'implémentation de ScoreCAM GNN.

## 7.5 Formalisation de ScoreCAM GNN

En utilisant les prérequis détaillés ci-dessus, nous définissons ScoreCAM GNN pour un graphe d'entrée $G = (\mathbf{X}, \mathbf{A})$ appartenant à la classe $c \in \mathbb{N}$ composée de $N \in \mathbb{N}$ noeuds, chacun représenté par un vecteur à valeur réelle et un modèle prédictif optimisé $f_{\boldsymbol{\theta}^\star}$ avec une profondeur $L \in \mathbb{N}$ pour la partie descripteur de graphe de telle sorte que la dernière couche est $\mathbf{X}^L \in \mathbb{R}^{N \times F_L}$, nous avons :

$$ScoreCAM(G, f_{\boldsymbol{\theta}^\star}) = ReLU[\sum_{k=1}^{F_L} g_k^c \mathbf{I}_k] \tag{7.2}$$

où

$$s_k^c = [f_{\boldsymbol{\theta}}^\star(G_{S(\mathbf{X}_k^L)})]_c \text{ and } g_k^c = \frac{e^{s_k^c}}{\sum_{k=1}^{F_L} e^{s_k^c}} \tag{7.3}$$

$S_k^c$ est la probabilité que le graphe d'entrée $G$ masqué par $S(\mathbf{X}_k^L)$ appartienne à la classe de $G$, c'est-à-dire la $c$-ième classe [3]. Notez que notre implémentation est

**Algorithme 2 :** Algorithme de ScoreCAM GNN

**Entrées :** Un graphe $G = (\mathbf{X}, \mathbf{A})$ de classe $c$, un réseau de neurones sur graphe $f_{\boldsymbol{\theta}^\star}$

**Output :** ScoreCAM($G, f_{\boldsymbol{\theta}^\star}$)

$\mathbf{X}^L \leftarrow$ the lastest activated feature map of $f_{\boldsymbol{\theta}^\star}$ through $G$;
$F_L \leftarrow$ the feature dimension of $\mathbf{X}^L$;
**pour chaque** $k \in \{1, \ldots, F_L\}$ **faire**
  $\hat{\mathbf{X}_k^L} \leftarrow S(\mathbf{X}_k^L)$;
  $\mathbf{I}_k \leftarrow \mathbf{X} \circ \hat{\mathbf{X}_k^L}$;
  $G_k \leftarrow G_{\hat{\mathbf{X}_k^L}}$ procédure de masquage du signal du graphe;
  $s_k^c \leftarrow [f_{\boldsymbol{\theta}}^\star(G_k)]_c$;
**pour chaque** $k \in \{1, \ldots, F\}$ **faire**
  $g_k^c \leftarrow \frac{e^{s_k^c}}{\sum_{k=1}^{F_L} e^{s_k^c}}$;
$\mathbf{X}_{exp} \leftarrow ReLU[\sum_{k=1}^{F_L} g_k^c \mathbf{I}_k]$;
**retourner** graphe expliqué $(\mathbf{X}_{exp}, \mathbf{A})$

adaptable aux problèmes de classification de noeuds. Nous présentons ici les résultats

3. Nous dénotons ici par souci de clarté $\mathbf{I}_k = \mathbf{X} \circ S(\mathbf{X}_k^L)$

de notre implémentation de ScoreCAM pour GNN selon les protocoles d'évaluation rencontrés usuellement dans la littérature.

## 7.6 Résultats

Nous avons comparé ScoreCAM GNN avec quatre méthodes explicatives locales et post-hoc de l'état de l'art :

- Deux méthodes spécifiques au modèle prédictif : DeepLIFT [Baldassarre and Azizpour, 2019a] et GradCAM [Pope et al., 2019a]
- deux méthodes agnostiques au modèle prédictif : GNNExplainer [Ying et al., 2019c] et GraphSVX [Duval and Malliaros, 2021].

Pour mettre en évidence la pertinence de ScoreCAM GNN, nous avons comparé notre méthode selon trois points de vue :

- Les avantages théoriques des approches spécifiques aux modèles prédictifs
- Evaluation quantitative des performances de la méthode selon des propriétés usuellement rencontrées dans l'état de l'art
- Evaluation qualitative sur des ensembles de données de l'état de l'art tel que *MNISTSuperpixel* et *REDDIT-BINARY* ne nécessitant pas de connaissance experte préalable. Nous n'avons pas choisi les ensembles de données comme *BBBP, BACE, PROTEINS*, utilisés dans la littérature, qui eux nécessitent des connaissances expertes préalables.

### 7.6.1 Les apports des méthodes spécifiques par rapport aux méthodes agnostiques dans le contexte de l'explicabilité des réseaux de neurones sur graphes

Dans le paradigme explicatif local et post-hoc, les deux principaux choix de paradigmes d'explication restants sont ceux dites agnostiques ou non au modèle prédictif. Ce choix entre les deux paradigmes est important. En effet, dans le cas des GNN, ce choix peut avoir un fort impact sur le coût de calcul des explications. Par exemple, pour GNNExplainer [Ying et al., 2019c] ou GraphSVX [Duval and Malliaros, 2021], qui sont des méthodes agnostiques au modèle prédictif, construisent leurs explications , via inférences du modèle prédictif, en explorant le domaine. Et

comme évoqué, dans ce cas, le nombre de combinaisons est exponentiel en la taille du graphe traité, ce qui peut avoir rapidement un coût de calcul important dans la majorité des cas. De plus, lorsque ces derniers optimisent des grandeurs pour construire leurs explications, de par la dépendance des objectifs aux modèles prédictifs et leurs complexités, les fonctions objectives ont une faible régularité, et donc peuvent être difficiles à optimiser, avoir des solutions instables, variables, dépendantes de conditions d'initialisations, etc.,ce qui n'est pas forcément le cas des méthodes explicatives spécifiques. En effet, la complexité du calcul par une méthode explicative spécifique est souvent linéaire en la taille de la donnée et ne nécessite souvent pas de phase d'optimisation. Les propriétés numériques des calculs des explications ne sont pas les seules à être prises en compte, les propriétés sémantiques sont aussi à prendre en compte pour qualifier la qualité d'une méthode explicative.

### 7.6.2 Protocole expérimental

#### Ensemble de données

Nous réalisons nos expériences avec cinq ensembles de données réelles pour les tâches de classification de graphes. Nous les avons utilisés parce que ces ensembles de données ont été largement utilisés dans [Baldassarre and Azizpour, 2019a, Duval and Malliaros, 2021, Huang et al., 2022b, Lucic et al., 2022, Pope et al., 2019a, Schnake et al., 2022, Ying et al., 2019c, Yuan et al., 2020, Yuan et al., 2021] pour illustrer les méthodes d'explication de GNN. *REDDIT-BINARY* [Yanardag and Vishwanathan, 2015] un jeu de données composé de 2000 graphes. Chacun d'entre eux représente un fil de discussion de Reddit basé sur des questions/réponses, à savoir *r/IAMA* et *r/AskReddit*. Dans ces graphes, les noeuds représentent un utilisateur et il existe un lien entre deux utilisateurs si l'un a répondu à l'autre ; *BBBP* pour pénétration barrière sang-cerveau (*BBBP*) [Wu et al., 2017] provient d'une étude récente sur la modélisation et la prédiction de la perméabilité de la barrière cellulaire ; *PROTEINS* [Borgwardt et al., 2005] est un ensemble de données de protéines classées en tant qu'enzymes ou non-enzymes. Les noeuds représentent les acides aminés et deux noeuds sont reliés par une arête s'ils sont à une distance inférieure à une courte distance prédéfinie ; *BACE* fournit des résultats quantitatifs (IC50) [Wu et al., 2017] et qualitatifs de liaison pour un ensemble d'inhibiteurs de la *BACE*-1. Toutes les

| | BBBP | BACE | REDDIT-BINARY | PROTEINS | MNISTSuperpixel |
|---|---|---|---|---|---|
| Nombre de graphes | 2039 | 1513 | 2000 | 1113 | 70000 |
| Dimension des descripteurs ($F$) | 9 | 9 | 1 | 3 | 1 |
| Nombre de classes ($C$) | 2 | 2 | 2 | 2 | 10 |

TABLE 7.1 – Caractéristiques des ensembles de données

données sont des valeurs expérimentales rapportées dans la littérature au cours de la dernière décennie, certaines avec des structures cristallines détaillées disponibles ; *MNISTSuperpixel* a été introduit par [Bronstein et al., 2017] est une version superpixel de l'ensemble de données MNIST de chiffres manuscrits bien connus. L'utilisation de cet ensemble de données nous permet d'utiliser en parallèle l'ensemble de données MNIST original (version image en nuance de gris), car celui-ci est interprétable directement par des personnes non expertes. Chaque graphe de l'instance *MNISTSuperpixel* est un graphe qui représente l'instance MNIST originale de $28 \times 28$ pixels. Les graphes partagent la même étiquette. Ils sont composés de 75 noeuds (superpixels), ces noeuds étant liés en fonction de leurs relations spatiales. L'ensemble de données MNIST (et, par extension, l'ensemble de données *MNISTSuperpixel*) a été conçu pour classer les chiffres écrits à la main. Distinguer un chiffre d'un autre tourne autour de la recherche de modèles discriminants ou de singularités qui identifient un chiffre (en termes d'étiquettes de vérité de terrain) comme une possibilité unique.

Dans le Tableau 7.1, nous avons repris les caractéristiques de chaque ensemble de données.

### Les procédures d'apprentissage

Nous avons utilisé deux configurations principales de GNN pour classer les éléments de nos ensembles de données. Cette architecture est dérivée de celles utilisées dans [Pope et al., 2019a] qui fonctionnent aussi bien GNN explications. Une configuration est basée sur Graph Convolutional Network (GCN) un module que nous appelons `GCNModel`. L'autre est basée sur le module Graph Attention Network (GAT) et sera appelée `GATModel`.

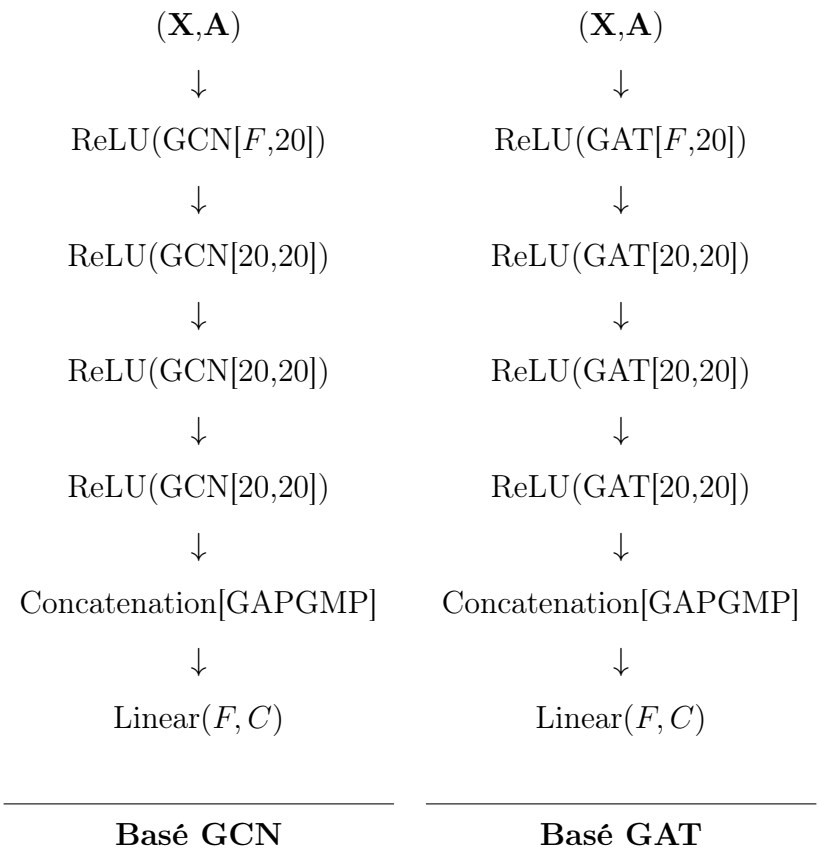
Remarquons que les notations pour la dimension des espaces des descripteurs $F$

et le nombre de classes $C$ sont introduits dans le Tableau 7.1. Pour les ensembles de données *BBBP, BACE, PROTEINS*, les classifieurs étaient construits de la manière suivante : [4] :

| **Basé GCN** | **Basé GAT** |
|---|---|
| (**X**,**A**) | (**X**,**A**) |
| ↓ | ↓ |
| ReLU(GCN[$F$,20]) | ReLU(GAT[$F$,20]) |
| ↓ | ↓ |
| ReLU(GCN[20,20]) | ReLU(GAT[20,20]) |
| ↓ | ↓ |
| GAP | GAP |
| ↓ | ↓ |
| Linear($F, C$) | Linear($F, C$) |

et pour les ensembles de données *MNISTSuperpixel* et *REDDIT-BINARY*, les classifieurs étaient construits comme suit :

4. Global Average Pooling (GAP) est l'opérateur de mutualisation moyennant les descripteurs de chaque dimension et Global Maximum Pooling (GMP) est l'opérateur de mutualisation retournant le maximum des descripteurs de chaque dimension.




Pour chaque ensemble de données, nous avons procédé à plusieurs ajustements de la taille des lots et du taux d'apprentissage. Nous désignons par `GCNModel` le modèle prédictif ayant une architecture donnée (selon l'ensemble de données) et obtenant les meilleurs résultats sur l'ensemble de tests (avec un ratio de division de 80/20). Par analogie, nous désignons par `GATModel` le modèle basé sur GAT répondant aux mêmes exigences et obtenant également les meilleurs résultats.

### Standardisation des explications

Pour comparer les différentes méthodes explicatives, nous avons mis en place un protocole standard. Nous avons besoin d'uniformiser les explications afin de produire une quantification comparable de l'importance des noeuds et ainsi effectuer une évaluation qualitative et quantitative. Ainsi, étant donné un graphe avec des descripteurs de noeuds $\mathbf{X} \in \mathbb{R}^{N \times F}$, si l'explication $\hat{\mathbf{X}}$ est un vecteur à valeur $\mathbb{R}^F$, nous considérons $\mathbf{X} \circ \hat{\mathbf{X}} \in \mathbb{R}^{N \times F}$. Si l'explication $\hat{\mathbf{X}}$ est un vecteur à valeurs $\mathbb{R}^{N \times F}$, nous avons considéré $\mathbf{X} \circ \hat{\mathbf{X}} \in \mathbb{R}^{N \times F}$. Ensuite, nous considérons $S(\mathbf{X} \circ \hat{\mathbf{X}})$ qui est ensuite projeté sur l'espace des noeuds du graphe. Pour comparer objectivement notre méthode aux autres, nous évaluons, pour un modèle prédictif $\phi$, une méthode

explicative $\psi$ et un ensemble de données $\mathcal{D}$, l'infidélité moyenne et la parcimonie moyenne des explications appartenant à $\mathcal{D}_\psi$.

### Propriétés souhaitables pour une méthode explicative

Pour justifier de la pertinence des méthodes explicatives, la littérature a défini des propriétés souhaitables auxquelles les méthodes doivent répondre. Ces propriétés n'ont pas encore de fondements théoriques et numériques, mais elles ont un sens commun et significatif pour de nombreuses applications explicatives. Plus la méthode vérifie les propriétés, plus elle est pertinente. Nous avons listé ci-dessous les propriétés les plus utilisées et nous avons mené une étude comparative de la réalisation statistique des propriétés entre la méthode ScoreCAM GNN et les autres méthodes.

— *stabilité* [Alvarez-Melis and Jaakkola, 2018b] met l'accent sur la capacité de la méthode à fournir des explications similaires pour des entrées similaires.

— *contrastivité* [Miller, 2019a] met l'accent sur la capacité à fournir des explications basées sur une approche relative plutôt qu'absolue. En effet, en tant qu'être humain, nous préférons expliquer un phénomène P par rapport à un autre phénomène Q, et non P par lui-même.

— *complétude* [Kulesza et al., 2013] est la capacité de fournir des explications en encodant toutes les informations explicatives de la vérité fondamentale.

— *justesse* [Kulesza et al., 2013] est la capacité d'une méthode d'explication à fournir des explications uniquement en ce qui concerne les informations sémantiques nécessaires.

— *parcimonie* [Lipton, 2018, Murdoch et al., 2019] est la capacité d'une méthode d'explication à fournir des explications concises (par exemple, peu de caractéristiques, peu de paramètres, etc.)

— *simplicité* [Miller, 2019a] est la capacité à fournir des explications conviviales plutôt que des explications inutiles et compliquées. Cette propriété est étroitement liée à *compacité* et *justesse*.

— *fidélité* [Jacovi and Goldberg, 2020b, Jacovi and Goldberg, 2021b] est la capacité à retrouver le comportement initial du modèle expliqué lorsque les instances sont masquées par leurs propres explications.

— *consistance* [Miller, 2019a] est la capacité à fournir des explications cohérentes concernant une instance unique exprimée dans le même contexte, à expliquer.

Certains ont proposé des formulations numériques de ces propriétés. À titre d'exemple, [Pope et al., 2019a, Duval and Malliaros, 2021, Yuan et al., 2021] pour la *contrastivité* , [Pope et al., 2019a, Yuan et al., 2021, Duval and Malliaros, 2021] pour la *fidélité* [Pope et al., 2019a, Yuan et al., 2021, Duval and Malliaros, 2021] pour la *compacité* ou [Pope et al., 2019a, Yuan et al., 2021, Duval and Malliaros, 2021] pour la *fidélité*. Ces formulations numériques dépendent fortement du contexte de l'étude (par exemple, le contexte binaire dans [Pope et al., 2019a]) et aucune généralisation numérique n'a encore été proposée. Mais une notion analogue de *fidélité* a été proposée numériquement : l'infidélité [Yeh et al., 2019b].

### Métriques d'évaluation objectives

Pour quantifier de la pertinence des méthodes explicatives, la littérature a établi des métriques mesurant certaines propriétés relatives aux explications. Une évaluation par un expert est aussi possible, mais plus difficile à mettre en place et à valoriser. Les auteurs de [Yeh et al., 2019b] proposent une métrique objective, indépendante des experts, pour évaluer la qualité des cartes d'explication fournies par une méthode XAI.

— *Infidélité* : [Yeh et al., 2019b] quantifie la manière dont les explications fournies par une méthode explicative $\phi$ des prédictions faites par un modèle prédictif optimisé $f_{\boldsymbol{\theta}^\star}$ changent lorsqu'une entrée $\mathbf{X}$ est perturbée par une variable aléatoire $\mathbf{I}$ suivant une densité de perturbation $P$. Il est défini par :

$$\texttt{Infd}(\phi, f_{\boldsymbol{\theta}^\star}, \mathbf{X}) = \mathop{\mathbb{E}}_{\mathbf{I}\sim P}[(\mathbf{I}^T\phi(\mathbf{X}, f_{\boldsymbol{\theta}^\star}) - (f_{\boldsymbol{\theta}^\star}(\mathbf{X}) - f_{\boldsymbol{\theta}^\star}(\mathbf{X}-\mathbf{I})))^2] \tag{7.4}$$

— *Parcimonie spatiale* : D'une manière générale, les explications concises sont préférées aux explications diffuses sur le domaine, car elles noient les informations pertinentes. Cette affirmation ne dépend pas du contexte des explications, il s'agit donc d'une affirmation objective. Une façon naturelle de mesurer cette parcimonie est la norme $l_0$ [Pope et al., 2019a]. Il s'agit donc d'un objectif approprié que nous cherchons à atteindre. Formellement, étant donné un vecteur $\mathbf{X}$, un modèle prédictif optimisé $f_{\boldsymbol{\theta}^\star}$ et une méthode d'ex-

plication $\phi$, nous avons :

$$\texttt{Spar}(\lambda, f_{\boldsymbol{\theta}^\star}, \phi, \mathbf{X}) = 1 - \frac{\|\phi(\mathbf{X}, f_{\boldsymbol{\theta}^\star}) < \lambda\|_0}{dim(\mathbf{X})} \tag{7.5}$$

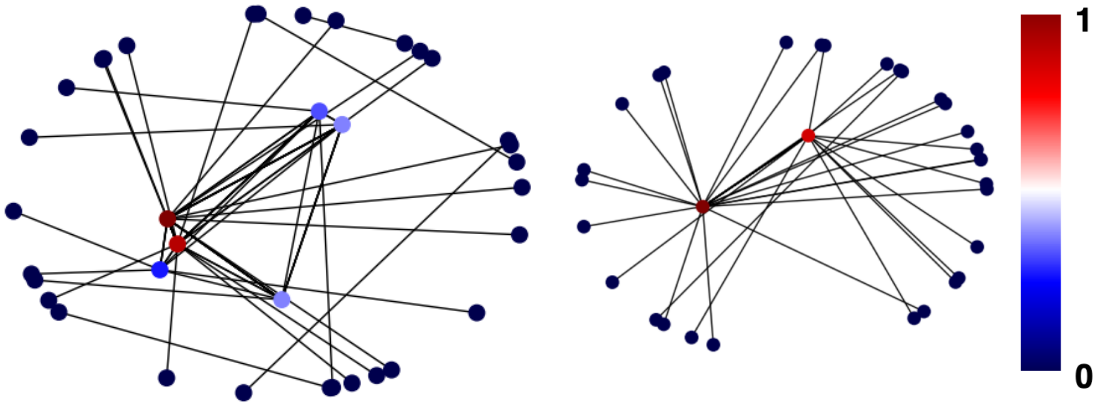
### 7.6.3 Résultats

Dans cette partie, nous fournissons les résultats numériques et non numériques concernant la pertinence objective et subjective de la méthode ScoreCAM GNN par rapport aux méthodes de la littérature.

#### Évaluation de la qualité des explications : applications sur des données réelles

Dans un premier temps, nous avons construit des modèles prédictifs précis sur les ensembles de données *MNISTSuperpixel* et *REDDIT-BINARY*. Ici, nous présentons les résultats obtenus, en termes d'explication, sur *REDDIT-BINARY* afin de montrer qu'avec peu de contexte et de connaissances expertes, les capacités de ScoreCAM GNN à localiser des informations significatives sur des ensembles de données du monde réel. Pour les résultats sur *MNISTSuperpixel*, nous utilisons des caractérisations visuelles de bons sens pour juger de la pertinence des explications fournies par ScoreCAM GNN. En effet, pour autant que nous le sachions, cet ensemble de données n'a pas été très utilisé dans l'état de l'art de l'explicabilité des GNN, alors que, pourtant, largement étudié dans le domaine de la vision par ordinateur. De par son accessibilité, nous avons pris l'initiative d'utiliser cet ensemble de données pour mener notre étude.

— **REDDIT-BINARY :** Cet ensemble de données contient deux, trois ou quatre utilisateurs experts par discussion en moyenne [Yanardag and Vishwanathan, 2015], les noeuds pertinents sont les utilisateurs experts puisqu'ils apportent les informations pertinentes à la discussion considérée. Ici, nous présentons deux discussions dans lesquelles deux experts discutent du sujet de la discussion.




(a) Première discussion (b) Seconde discussion

FIGURE 7.1 – Illustration des résultats de ScoreCAM GNN sur l'ensemble de données REDDIT-BINARY

Dans la Figure 7.1a, les experts (noeuds rouges) sont clairement identifiés par ScoreCAM GNN tandis que les personnes qui posent des questions (noeuds bleu foncé) ne sont pas pertinentes pour faire la lumière puisqu'elles n'apportent aucune valeur ajoutée au processus de question/réponse. Les noeuds bleu clair sont des personnes ayant des connaissances intermédiaires ou discutant avec des experts afin d'acquérir des connaissances. Dans la Figure 7.1b, seuls deux experts sont présents dans le fil de discussion, correctement mis en évidence par ScoreCAM GNN. Nous observons que ScoreCAM GNN met l'accent sur les noeuds à haut degré, ce qui est conforme à la mise en évidence des utilisateurs experts. Comme indiqué dans [Ying et al., 2019c], les utilisateurs experts (noeuds de haut degré) sont au coeur de la discussion. Ils jouent un rôle central dans l'orientation de la discussion, car ils peuvent interagir avec n'importe quel autre utilisateur en diffusant leurs connaissances d'expert sur lui.[5]. Notre méthode est donc conforme aux explications de la vérité de terrain.

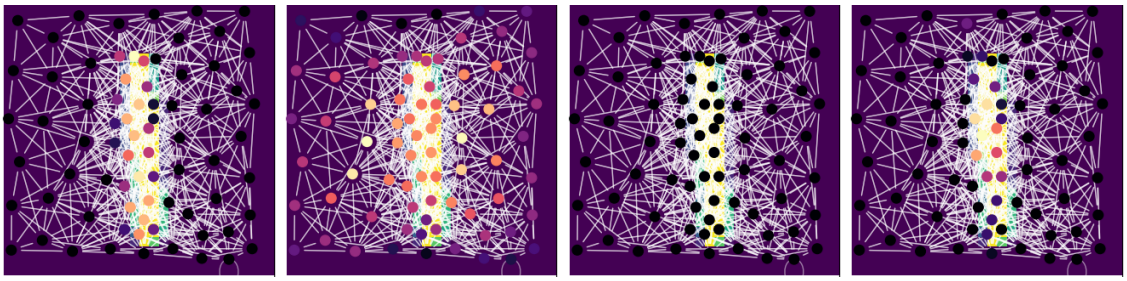
— **MNISTSuperpixel :** En ce qui concerne *MNISTSuperpixel*, nous expliquons six instances choisies au hasard et nous surchargeons les instances MNIST

5. Nous ne comparons pas nos résultats avec ceux fournis dans [Duval and Malliaros, 2021, Ying et al., 2019c, Pope et al., 2019a, Baldassarre and Azizpour, 2019a] puisqu'ils ont montré qu'ils produisaient des résultats similaires sur cette tâche.

originales associées, considérées comme une vérité de terrain [6]. MNIST et sa version superpixel traitent des chiffres écrits à la main. La distinction des chiffres est essentiellement une tâche de séparation multimodale entre la courbure des chiffres, les symétries axiales. Dans le contexte, une bonne explication doit encoder et inclure les priorités géométriques sous-jacentes. Par conséquent, nous devons retrouver ces symétries dans les explications fournies. Nous avons également ajouté une étude de la redondance spatiale afin d'établir un lien avec la mesure de la parité, étudiée en profondeur dans la partie quantitative.

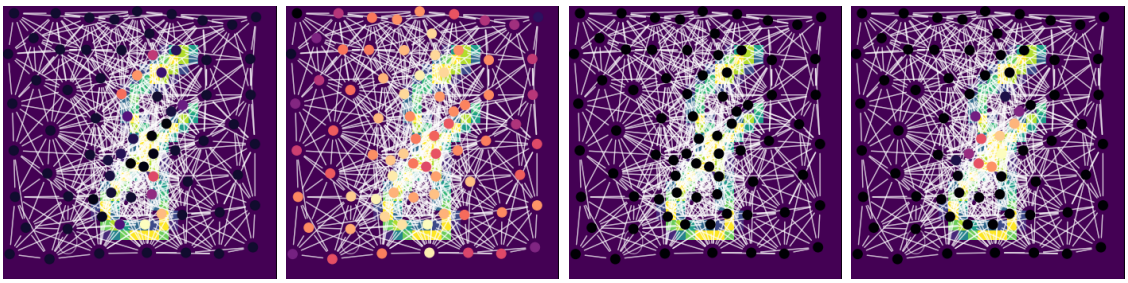
— *L'exploitation des symétries* : Par exemple $n°1193$, le chiffre est un "1". Il induit donc une symétrie axiale avec une courbure faible. Les sorties de DeepLIFT (Figure 7.2a) et de GradCAM (Figure 7.2b) n'exploitent pas ces symétries et se concentrent sur des noeuds non pertinents et diffus sur la majeure partie du domaine, ce qui n'est pas pertinent. ScoreCAM GNN (Figure 7.2d) exploite les caractéristiques géométriques intrinsèques du chiffre et distribue principalement son explication sur le barycentre du chiffre d'une manière compacte spatialement. De manière analogue, le chiffre "8" (instance $n°444$) est le seul chiffre qui a un autocroisé lorsque qu'il est écrit : c'est l'une des caractéristiques les plus pertinentes qui permet de caractériser le chiffre huit. Notre méthode cible entièrement cette signature symétrique (c'est-à-dire avec des symétries multiaxiales) (Figure 7.3d) alors que d'autres méthodes ne parviennent pas à caractériser cette singularité.

6. Notez que, comme GraphSVX ne fournit pas une densité suffisamment élevée, nous avons choisi de l'éliminer de l'évaluation subjective



(a) DeepLIFT (b) GradCAM (c) GNNExplainer (d) ScoreCAM GNN

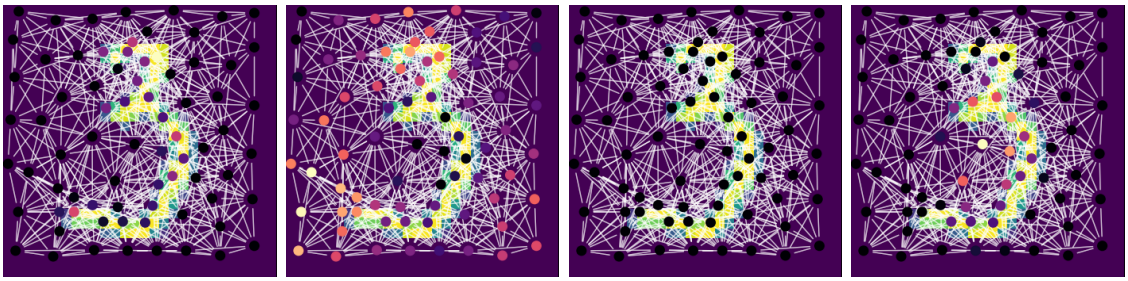
FIGURE 7.2 – Exemple $n°$ 1193 de *MNISTSuperpixel*



(a) DeepLIFT (b) GradCAM (c) GNNExplainer (d) ScoreCAM GNN

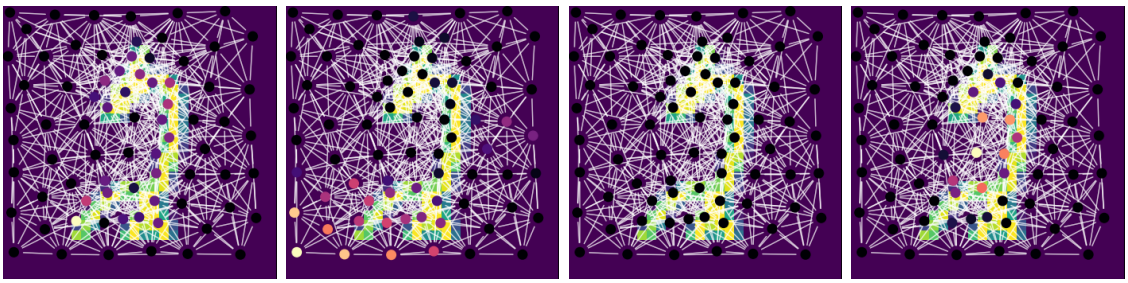
FIGURE 7.3 – Exemple $n°$ 444 de *MNISTSuperpixel*

— *Caractérisation par la courbure des chiffres* : Les instances $n°969$ et $n°1036$ sont respectivement les chiffres "3" et "2" qui sont particulièrement courbés dans leurs représentations. Attribuer une importance aux noeuds en fonction de la partie la plus courbée de ces chiffres n'est pas suffisant. En effet, ScoreCAM GNN ne se concentre pas seulement sur les points saillants qui caractérisent la structure curviligne, mais aussi sur le noeud qui est voisin de la plupart de ces noeuds caractérisant la courbure. Dans ces exemples, nous observons également que les explications fournies par ScoreCAM GNN ont une plus grande densité spatiale que les autres méthodes. La localité des explications fournies est également adaptée à la description manuscrite des chiffres, ce qui n'est pas le cas des autres méthodes.



(a) DeepLIFT (b) GradCAM (c) GNNExplainer (d) ScoreCAM GNN

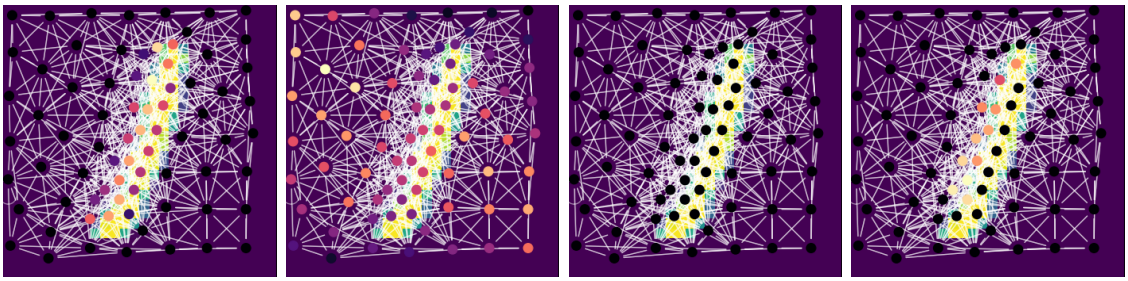
FIGURE 7.4 – Exemple $n°$ 969 de *MNISTSuperpixel*



(a) DeepLIFT (b) GradCAM (c) GNNExplainer (d) ScoreCAM GNN

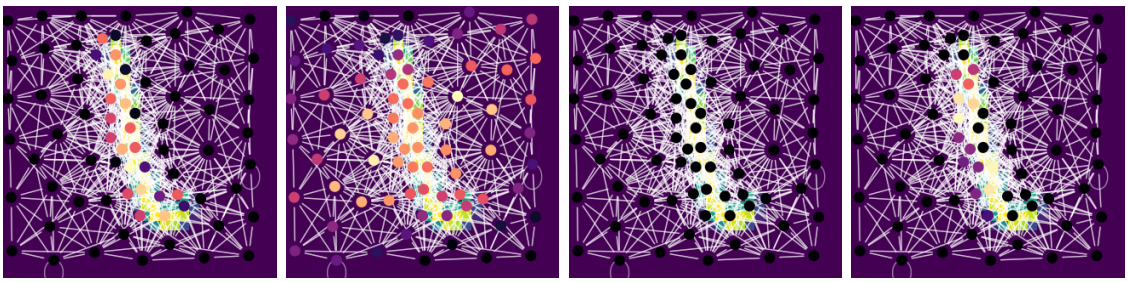
FIGURE 7.5 – Exemple $n°$ 1036 de *MNISTSuperpixel*

— *Non redondance* : Nous avons observé que ScoreCAM GNN produit des explications plus compactes dans l'espace. Mais nous ne devons pas négliger la modalité d'amplitude de l'importance des descripteurs. En effet, dans la Figure 7.6a en écartant les symétries, la distribution de l'intensité est presque uniforme sur les explications. Cela ne permet pas d'ordonner efficacement quelle partie de l'explication est plus importante qu'une autre spatialement. Comme le montrent les figures 7.7b et 7.6b, cette uniformité de l'explication apporte des informations floues et redondantes à l'utilisateur, dans les deux modalités : dans l'espace et en intensité. Dans la Figure 7.6d, l'importance des noeuds est principalement répartie sur les noeuds qui correspondent à la représentation manuscrite des chiffres, en ce qui concerne la modalité d'intensité, ScoreCAM GNN concentre la masse explicative sur cette représentation, ce qui est un comportement attendu.



(a) DeepLIFT (b) GradCAM (c) GNNExplainer (d) ScoreCAM GNN

FIGURE 7.6 – Exemple $n^\circ$ 346 de *MNISTSuperpixel*



(a) DeepLIFT (b) GradCAM (c) GNNExplainer (d) ScoreCAM GNN

FIGURE 7.7 – Exemple $n^\circ$ 478 de *MNISTSuperpixel*

**Évaluation quantitative : mesures objectives**

Les métriques d'infidélité et de parcimonie sont largement utilisées pour évaluer quantitativement et objectivement la pertinence des méthodes explicatives [Pope et al., 2019a, Baldassarre and Azizpour, 2019a, Yuan et al., 2020]. Elles sont un moyen de quantifier l'optimalité des méthodes explicatives. Nous évaluons d'abord la méthode proposée en termes objectifs en utilisant ces métriques avec $\lambda = 0.2$ et $P = \mathcal{N}(\mathbf{0}, \mathbf{1})$.

— Concernant `GCNModel`, ScoreCAM GNN est respectivement 88%, 74%, 99%, 89% moins infidèle que le score d'infidélité moyen des méthodes d'état de l'art pour les ensembles de données *BBBP*, *PROTEINS*, *REDDIT-BINARY* , *MNISTSuperpixel* (voir les Figures 7.2, 7.3, 7.4, 7.6). Tout en atteignant la limite inférieure du score d'infidélité sur l'ensemble de l'expérience, ScoreCAM GNN atteint également de manière significative la limite inférieure de l'écart-type moyen du score d'infidélité. En outre, en ce qui concerne l'ensemble de données *BACE* (voir le Tableau 7.5), le ScoreCAM GNN a un

score d'infidélité 20% plus élevé que le score d'infidélité moyen des méthodes de l'état de l'art.

— Concernant `GATModel`, ScoreCAM GNN est respectivement 94%, 97%, 99%, 94%, 93% moins infidèle que le score d'infidélité moyen des méthodes d'état de l'art pour le jeu de données *BBBP*, *PROTEINS*, *REDDIT-BINARY*, *BACE* et *MNISTSuperpixel* (voir les Figures 7.2, 7.3, 7.4, 7.5, 7.6). Tout en atteignant la limite inférieure du score d'infidélité sur l'ensemble de l'expérience, ScoreCAM GNN atteint également de manière significative la limite inférieure de l'écart-type du score d'infidélité moyen.

— Concernant `GCNModel`, ScoreCAM GNN est respectivement 3.9, 1.4, 2.5, 3. 3 fois plus parcimonieuse que le score moyen de parcimonie des méthodes d'état de l'art pour *BBBP*, *REDDIT-BINARY*, *BACE* et *MNISTSuperpixel* (voir les Figures 7.2, 7.4, 7.5 7.6). Tout en atteignant la limite inférieure du score de parcimonie sur l'ensemble de l'expérience, ScoreCAM GNN atteint également de manière significative la limite inférieure de l'écart-type moyen du score de parcimonie. En outre, en ce qui concerne l'ensemble de données *PROTEINS* (voir le Tableau 7.3), le ScoreCAM GNN a un score de parcimonie inférieur de 6% par rapport au score d'infidélité moyen des méthodes de l'état de l'art.

— Concernant `GATModel`, ScoreCAM GNN est respectivement 3,9, 2,6, 1,4, 6,4, 2. 2 fois plus parcimonieuse que le score de parcimonie moyen des méthodes d'état de l'art pour *BBBP*, *PROTEINS*, *REDDIT-BINARY*, *BACE* et *MNISTSuperpixel* (voir les Figures 7.2, 7.3, 7.4, 7.5, 7.6). Tout en atteignant la limite inférieure du score de parcimonie sur l'ensemble de l'expérience, ScoreCAM GNN atteint également de manière significative la limite inférieure de l'écart-type moyen du score de parcimonie. Nous devons noter que puisque `GATModel` permet de représenter une classe plus large de fonctions, ils peuvent potentiellement être plus optimaux d'un point de vue de la classification de graphes. Comme ScoreCAM GNN est une méthode de spécifique, elle tire parti de cette qualité de représentation accrue par rapport aux modèles `GCNModel` pour expliquer les instances, ce qui peut entraîner une diminution d'infidélité par rapport aux autres

méthodes de l'état de l'art.

| **Méthode explicative** | **Type de modèle prédictif** | **Infidélité moyenne** (↓) | **Parcimonie moyenne** (↑) |
|---|---|---|---|
| DeepLIFT | GAT | 8.31 (± 26.35) | 0.06 (± 0.01) |
| GradCAM | GAT | 5.85 (± 6.83) | 0.77 (± 0.01) |
| GNNExplainer | GAT | 6.70 (± 29.19) | 0.11 (± 0.01) |
| GraphSVX | GAT | 7.62 (± 29.94) | 0.05 (± 0.01) |
| **ScoreCAM GNN** | **GAT** | **0.43(± 0.30)** | **0.98 (± 0.01)** |
| DeepLIFT | GCN | 5.84 (± 11.43) | 0.08 (± 0.01) |
| GradCAM | GCN | 5.32 (± 8.14) | 0.77 (± 0.01) |
| GNNExplainer | GCN | 7.89 (± 32.37) | 0.10 (± 0.01) |
| GraphSVX | GCN | 7.20 (± 29.21) | 0.05 (± 0.01) |
| **ScoreCAM GNN** | **GCN** | **0.25 (± 0.18)** | **0.98 (± 0.01)** |

TABLE 7.2 – Ensemble de données BBBP - (↑) le plus haut le mieux (↓) le plus bas le mieux

| **Méthode explicative** | **Type de modèle prédictif** | **Infidélité moyenne** (↓) | **Parcimonie moyenne** (↑) |
|---|---|---|---|
| DeepLIFT | GAT | 20.83 (± 729.67) | 0.07 (± 0.02) |
| GradCAM | GAT | 6.67 (± 49.26) | 0.85 (± 0.02) |
| GNNExplainer | GAT | 21.11 (± 624.80) | 0.49 (± 0.06) |
| GraphSVX | GAT | 6.09 (± 46.62) | 0.05 (± 0.01) |
| **ScoreCAM GNN** | **GAT** | **0.44 (± 1.38)** | **0.97 (± 0.01)** |
| DeepLIFT | GCN | 12.75 (± 246.14) | 0.08 (± 0.01) |
| GradCAM | GCN | 6.25 (± 41.60) | **0.89 (± 0.02)** |
| GNNExplainer | GCN | 22.97 (± 1093.25) | 0.48 (± 0.05) |
| GraphSVX | GCN | 13.60 (± 516.29) | 0.05 (± 0.01) |
| **ScoreCAM GNN** | **GCN** | **3.60 (± 21.68)** | 0.73 (± 0.08) |

TABLE 7.3 – Ensemble de données PROTEINS - (↑) le plus haut le mieux (↓) le plus bas le mieux

| **Méthode explicative** | **Type de modèle prédictif** | **Infidélité moyenne** (↓) | **Parcimonie moyenne** (↑) |
|---|---|---|---|
| DeepLIFT | GAT | 9.56 (± 1.36) | 1.0 (± 0.01) |
| GradCAM | GAT | 6.95 (± 0.97) | 0.86 (± 0.02) |
| GNNExplainer | GAT | 6.34 (± 3.62) | 0.71 (± 0.54) |
| GraphSVX | GAT | 7.32 (± 12.42) | 0.04 (± 1.34) |
| **ScoreCAM GNN** | **GAT** | **0.01 (± 0.1)** | **0.97 (± 0.03)** |
| DeepLIFT | GCN | 5.46 (± 9.56) | 0.76 (± 0.06) |
| GradCAM | GCN | 6.98 (± 1.01) | 0.86 (± 0.02) |
| GNNExplainer | GCN | 4.36 (± 8.05) | 1.0 (± 0.01) |
| GraphSVX | GCN | 7.34 (± 11.75) | 0.04 (± 0.01) |
| **ScoreCAM GNN** | **GCN** | **0.01 (± 1.2)** | **0.97 (± 0.02)** |

TABLE 7.4 – Ensemble de données REDDIT-BINARY - (↑) le plus haut le mieux (↓) le plus bas le mieux

| **Méthode explicative** | **Type de modèle prédictif** | **Infidélité moyenne** (↓) | **Parcimonie moyenne** (↑) |
|---|---|---|---|
| DeepLIFT | GAT | 6.23 (± 8.22) | 0.04 (± 0.01) |
| GradCAM | GAT | 13.67 (± 24.03) | 0.36 (± 0.11) |
| GNNExplainer | GAT | 6.62 (± 29.33) | 0.06 (± 0.01) |
| GraphSVX | GAT | 5.08 (± 17.48) | 0.04 (± 0.01) |
| **ScoreCAM GNN** | **GAT** | **0.51 (± 1.81)** | **0.97 (± 0.01)** |
| DeepLIFT | GCN | 7.61 (± 31.34) | 0.06 (± 0.01) |
| GradCAM | GCN | 9.12 (± 16.1) | 0.35 (± 0.11) |
| **GNNExplainer** | **GCN** | **3.39 (± 7.67)** | **0.16 (± 0.04)** |
| GraphSVX | GCN | 5.29 (± 16.43) | 0.04 (± 0.01) |
| ScoreCAM GNN | GCN | 7.58 (± 19.8) | 0.40 (± 0.02) |

TABLE 7.5 – Ensemble de données BACE - (↑) le plus haut le mieux (↓) le plus bas le mieux

| **Méthode explicative** | **Type de modèle prédictif** | **Infidélité moyenne** (↓) | **Parcimonie moyenne** (↑) |
|---|---|---|---|
| DeepLIFT | GAT | 307.08 (± 469.43) | 0.10 (± 0.05) |
| GradCAM | GAT | 85.84 (± 62.94) | 0.48 (± 0.15) |
| GNNExplainer | GAT | 201.67 (± 106.18) | 0.51 (± 0.38) |
| GraphSVX | GAT | 851.77 (± 132.90) | 0.01 (± 0.34) |
| **ScoreCAM GNN** | **GAT** | **23.89 (± 0.37)** | **0.62 (± 0.08)** |
| DeepLIFT | GCN | 862.51 (± 26.31) | 0.14 (± 0.06) |
| GradCAM | GCN | 340.88 (± 92.17) | 0.46 (± 0.15) |
| GNNExplainer | GCN | 592.14 (± 31.33) | 0.27 (± 0.19) |
| GraphSVX | GCN | 583.58 (± 23.15) | 0.01 (± 0.03) |
| **ScoreCAM GNN** | **GCN** | **68.67 (± 7.01)** | **0.74 (± 0.04)** |

TABLE 7.6 – Ensemble de données MNISTSuperpixels - (↑) le plus haut le mieux (↓) le plus bas le mieux

### Évaluation des propriétés non numériques

Nous résumons ici les propriétés désirables que les méthodes explicatives doivent satisfaire. Ces propriétés n'ont souvent pas de formalisation numérique. Nous faisons alors une étude subjective pour chacune des méthodes explicatives ainsi que pour ScoreCAM GNN.

— Pour *stabilité*, les méthodes considérées satisfassent toutes la propriété. Pour les instances, DeepLIFT explique le chiffre un en se basant sur les mêmes arguments (voir Figures 7.2a, 7.6a, 7.7a) pour caractériser le chiffre un. Nous observons le même comportement pour GradCAM (voir Figures 7.2b, 7.6b, 7.7b) ainsi que pour GNNExplainer (voir Figures 7.2c, 7.6c, 7.7c) et pour ScoreCAM GNN (voir Figures 7.2d, 7.6d, 7.7d).

— Pour le *contrastivité*, seules DeepLIFT et ScoreCAM GNN remplissent cette propriété. En effet, ces deux méthodes font une différence importante entre les chiffres un et huit (voir Figure 7.2a et Figure 7.3a pour DeepLIFT, Figure 7.2d et Figure 7.3d pour ScoreCAM GNN) par exemple, alors que ce n'est pas le cas pour GNNExplainer et GradCAM.

— Pour la *complétude*, il n'est pas clair si l'une des explications fournies exploite l'entièreté des descripteurs, nous ne sommes donc pas en mesure de conclure ici.

— Pour la *justesse*, seuls DeepLIFT et ScoreCAM GNN ont des explications

valables, car ils se concentrent sur le chiffre lui-même alors que GNNExplainer ou GradCAM se concentrent souvent sur des caractéristiques non pertinentes pour fournir leurs explications (voir la Figure 7.4b pour GradCAM et la Figure 7.3c pour GNNExplainer).

— Pour *compacité*, grâce à la mesure de la parcimonie spatiale, nous pouvons objectivement conclure que seul ScoreCAM GNN remplit la propriété.

— Pour *simplicité*, ScoreCAM GNN est la seule méthode qui fournit des informations claires, légères et facilement interprétables par l'homme.

— Pour la *fidélité*, grâce à la mesure de l'infidélité, nous pouvons objectivement conclure que seul ScoreCAM GNN remplit la propriété.

— Pour la *consistance*, seuls DeepLIFT, GradCAM et ScoreCAM GNN remplissent cette propriété. En effet, GraphSVX ou GNNExplainer impliquent soit des méthodes d'échantillonnage, soit des procédures d'optimisation complexes au comportement variable ; ces deux méthodes sont dépendantes des graines d'initialisation, ce qui conduit à différentes explications possibles de la même instance exprimée dans le même contexte. Donc pas de résultats consistants a priori pour ces méthodes.

Nous observons que la méthode ScoreCAM GNN remplit en majorité les propriétés désirables alors que les autres méthodes explicatives de l'étude ne remplissent pas souvent ces propriétés.

| | DeepLIFT | GradCAM | GraphSVX | GNNExplainer | ScoreCAM GNN |
|---|---|---|---|---|---|
| *stabilité* | ✓ | ✓ | ✓ | ✓ | ✓ |
| *constrastivité* | ✓ | ✓ | ✓ | ✓ | ✓ |
| *complétude* | ∼ | ∼ | ∼ | ∼ | ∼ |
| *justesse* | ✓ | x | x | x | ✓ |
| *compacité* | ✓ | x | x | x | ✓ |
| *simplicité* | ✓ | x | x | x | ✓ |
| *fidélité* | x | x | x | x | ✓ |
| *consistance* | ✓ | ✓ | x | ✓ | ✓ |

TABLE 7.7 – Résumé des propriétés satisfaites

## 7.7 Conclusion

Dans cette contribution, nous avons proposé une extension des méthodes Score-CAM CNN qui a montré sa pertinence dans le domaine de la vision par ordinateur. Or, les CNN étant un cas particulier de GNN, nous avons généralisé cette méthode explicative aux modèles GNN. Nous avons conçu cette généralisation conformément à la théorie de l'apprentissage géométrique profond. Nous avons montré empiriquement que la méthode proposée possède des propriétés numériques et sémantiques souhaitables. Notre méthode tire parti de la complexité linéaire des méthodes explicatives spécifique, alors que les méthodes agnostiques ont souvent recours à des approches combinatoires dont les coûts de calculs sont bien plus importants pour une valeur ajoutée relative. Il apparaît également que d'autres méthodes explicatives spécifiques ont montré des résultats faibles en termes de d'évaluation qualitative, même sur des expériences simples. Mais aussi quantitativement en ce qui concerne les métriques d'évaluation objectives qui sont aussi généralement utilisées pour mesurer la qualité des méthodes d'explication. Notre méthode s'est révélée efficace dans ces trois contextes : théorique, qualitatif et quantitatif.

## 7.8 Discussions, étude épistémologique, limites et ouvertures

Notre généralisation de la méthode ScoreCAM aux GNN et notre protocole d'évaluation ont montré que notre implémentation permettait d'avoir des résultats pertinents vis-à-vis des méthodes proposées dans la littérature. Cependant, notre étude dans sa globalité admet quelques limites, en voici quelques-unes :

— Notre généralisation s'est concentrée sur les composantes de la méthode ScoreCAM qui ont des dépendances aux structures type image dans leur définition et nous avons strictement réutilisés les composantes qui n'en dépendaient pas dans le but de ne pas perdre la nature de la méthode. Cependant, on peut contester l'impact numérique de certaines de ces composantes :
  — Occultation de l'aspect domaine de la donnée dans le calcul de l'explication : le calcul des explications, comme pour toute méthode CAM, ne

dépend pas directement de la structure du domaine de la donnée. C'est pourtant une information majeure dans le contexte des GNN. La sous-estimation de cette modalité dans le calcul des explications peut donc mener à des conclusions plus superficielles.

— Pour la procédure de masquage normalisé : rappelons que la quantité

$$\mathbf{I}_k = \mathbf{X} \circ S(\mathbf{X}_k^L), k \in \{1, \cdots, F_L\} \tag{7.6}$$

. Ainsi, pour $(i,j) \in \{1, \cdots, N\} \times \{1, \cdots, F_L\}$, tel que

$$(i,j) = \arg\min_{(i,j)} (\mathbf{X}_k^L)_{i,j} \tag{7.7}$$

alors

$$S(\mathbf{X}_k^L)_{i,j} = 0 \tag{7.8}$$

d'après l'équation 7.1, ce qui implique que

$$(\mathbf{I}_k)_{i,j} = 0 \tag{7.9}$$

par construction. Donc, cela signifie que l'importance accordée à la quantité $(\mathbf{X}_k^L)_{i,j}$ est nulle ou bien en d'autres termes, que ce descripteur n'est pas important pour caractériser le comportement du modèle prédictif, ce qui ne fait pas sens. En effet, la quantité $(\mathbf{X}_k^L)_{i,j}$ n'est que le minimum de $\mathbf{X}_k^L$. Et dans le cas d'un calcul de représentation, être le minimum ne veut pas dire être inutile pour obtenir un modèle prédictif performant (c-à-d obtenir une bonne représentation). Autrement dit, cela ne veut pas dire que ce descripteur n'a pas un rôle déterminant dans le comportement du modèle prédictif, c'est à dire, que ce descripteur n'est pas important pour expliquer le comportement du modèle prédictif. Par exemple, le choix d'une normalisation $l_1$ permet d'éviter l'aspect absorbant du 0 et le descripteur normalisé en norme $l_1$ est égal à 0 si et seulement s'il est nul initialement. La normalisation $l_1$ permet donc d'éviter ce cas de figure et donc éviter de mener à de fausses conclusions.

— Pour la procédure de pondération des descripteurs : si le descripteur $\mathbf{I}_k$ a une masse de pertinence $s_k^c = 0$ , autrement dit que le descripteur $\mathbf{I}_k$ a un poids nul dans l'explication, alors il se trouve qu'en pratique il

aurait l'importance $g_k^c > 0$ dans le calcul final de l'explication, ce qui peut aussi fausser les conclusions. Pour certaines architectures (avec une couche de softmax en sortie du modèle prédictif notamment), avoir $s_k^c = 0$ est impossible, mais en toute généralité cela est parfaitement envisageable pour des architectures quelconques.

Ainsi les processus de normalisation en jeu dans la définition de la méthode ScoreCAM originelle ont introduit d'autres effets numériques qui peuvent mener à des explications qui ne font pas sens. Dans notre implémentation, nous les avons aussi utilisés et donc la qualité de nos conclusions peut être altérée pour les mêmes raisons.

— La prévalence effective de ScoreCAM par rapport aux autres méthodes CAM : les présents travaux n'ont pas seulement une visée généralisatrice de la méthode ScoreCAM originelle, mais cette étude montre que, avec peu d'adaptation, n'importe quelle méthode CAM, développée pour les réseaux de neurones convolutifs est facilement généralisable aux GNN. En effet, les points centraux des méthodes CAM sont de reposer sur l'hypothèse suivante : les données de haut niveau nécessaire au calcul de l'explication se trouvent dans la dernière couche des descripteurs du modèle prédictif (basé sur les réseaux de neurones convolutionels). Or, n'importe quel modèle prédictif répondant au paradigme de l'apprentissage profond géométrique est construit en accord avec le prérequis des méthodes CAM. Les variantes des méthodes CAM ne sont que des variantes calculatoires corrigeant des points particuliers des méthodes antérieures. Ainsi, notre étude s'est portée sur ScoreCAM, car la méthode s'était illustrée dans l'étude précédant celle-ci, mais parmi toutes les méthodes CAM de la littérature, il n'est pas certain que ScoreCAM soit la plus performante une fois que celle-ci est généralisée aux GNN.

— Valeurs ajoutées relatives de ScoreCAM dans la littérature : les motivations originelles pour le développement de ScoreCAM étaient construites en réponse à la méthode GradCAM. Cependant les méthodes CAM ne sont, en majeure partie, que développées en réponse aux défauts des méthodes CAM qui leurs sont antérieures, construisant ainsi une sorte de littérature sous forme d'arbre dont la première feuille est la méthode CAM. Ainsi leurs ap-

ports ne sont majoritairement que valorisables dans le paradigme CAM et en dehors de cette valorisation relative, elles n'apportent pas réellement de valeur ajoutée absolue, au sens où elles ne proposent pas des avancées majeures pour la communauté. Il en va de même pour notre implémentation et pour notre état de l'art. Ainsi, bien que nous ayons utilisé le protocole usuel de la littérature pour évaluer la ScoreCAM GNN, les conclusions de notre étude sur la supériorité de ScoreCAM GNN n'est que relative aux :

- Méthodes explicatives comparées : nous avons fait un choix sur les méthodes, un choix arbitraire, qui conditionne l'issue de l'étude. C'est donc limitant.
- Au protocole d'évaluation : qui est analogue à celui que l'on trouve dans la littérature. Ce dernier est bien défini structurellement (c-à-d enchaînement de mesure qualitative et quantitative), mais il existe une multitude de formulations pour une même métrique d'évaluation, ou pour un procédé de mesure qualitative. De plus, on peut en considérer certain(e)s et d'autres non, de manière parfaitement arbitraire aussi. Ainsi, ce protocole admet une grande variabilité et on l'observe en pratique en lisant les papiers de la littérature.

Pour cela, il faudrait mener une étude théorique, mais cette dernière serait limitée par le manque de fondement théorique de l'explicabilité à ce jour. Aujourd'hui nous n'avons principalement que des protocoles similaires entre eux, mais variables dans leurs instances pratiques, empiriques et subjectives, dans leurs réalisations et leurs conclusions. Ainsi, il n'est pas clair que ScoreCAM GNN soit une méthode fondamentalement meilleure que celles de la littérature, CAM ou non.

- Par ailleurs , nous considérons que le modèle prédictif expliqué à chaque fois est celui qui est optimal pour l'ensemble de données, ce qui n'est pas le cas en pratique. Cela biase donc nos conclusions.
- Aussi, nous n'avons pas mesuré la sensibilité de la méthode ScoreCAM GNN aux perturbations des poids du modèle prédictif ni des données issues de l'ensemble de données, comme conseillé dans [Adebayo et al., ]. Les conclusions auront pu être assez variables suite à ces perturbations, donc à ce titre, nos

conclusions peuvent être fragiles.

— ScoreCAM GNN pour la classification de noeud : bien que l'adaptation de ScoreCAM GNN aux problèmes de classification de noeud soit très légère, nous l'avons simplement évoqué et n'avons pas à illustrer la méthode sous cet angle. Cela aurait pu corroborer nos conclusions actuelles.

— ScoreCAM GNN sans réduction de dimension : dans le cas de ScoreCAM original, du fait de l'opération de convolution, les dimensionnalités des descripteurs réduisent (sans opération de rembourrage) avec la profondeur du modèle prédictif. L'utilisation de sur-échantillonnage dans la définition de ScoreCAM était d'ailleurs à l'origine de cette réduction de dimension, pour des besoins calculatoires évidents. Pour les raisons évoquées dans l'étude, nous n'utilisons pas de sur-échantillonnage dans la définition de ScoreCAM GNN. De manière analogue, cela signifie que nous nous interdisons d'utiliser toute opération entraînant une réduction de dimension (vis-à-vis du domaine) dans nos modèles prédictifs. Or, la réduction de dimension et la construction de sémantique numériquement sont intimement liées. Cette sémantique joue un rôle central dans le domaine de l'explicabilité des modèles prédictifs profonds. Ce point limite donc probablement le pouvoir explicatif de ScoreCAM GNN.

— Absence de vérité terrain pour l'évaluation de la méthode : dans le choix de notre protocole, nous n'avons utilisé que des données issues du monde réel n'ayant pas de vérité terrain pour l'explicabilité (si tenté qu'elle existe) , empêchant ainsi une évaluation plus objective de notre méthode.

— Biais de confirmation dans les études subjectives : comme dans toute étude subjective, elles sont sujettes au biais de confirmation. Le manque de fondements théoriques et de consensus en recherche sur l'explicabilité des modèles prédictifs profonds laisse alors place à ce genre d'approche, biaisant alors les conclusions des études, y compris la nôtre.

Ainsi, cette étude a permis, via la généralisation d'une méthode de référence à l'entièreté des modèles prédictifs répondant au paradigme de l'apprentissage géométrique profond, d'établir de nouveaux résultats et de proposer une méthode pertinente à la communauté. C'est la contribution de cette étude. Cette étude ainsi que

son implémentation est disponible publiquement en ligne. Les premiers résultats ont fait l'objet d'une publication dans un congrès national avec comité de lecture :

A. Raison, P. Bourdon and D. Helbert, *"ScoreCAM GNN : une explication optimale des réseaux profonds sur graphes "*, XXVIIIe Colloque GRETSI - Traitement du Signal et des Images, Sep 2022, Nancy, France

La publication complète est donnée par :

A. Raison, P.Bourdon and D.Helbert, *"ScoreCAM GNN : a generalization of an optimal local post-hoc explaining method to any geometric deep learning models"* 2022. ⟨hal-03808171v1⟩

## 7.9 Inscription de la contribution avec le projet de recherche et la littérature

Nous présentons ici les points d'articulation de cette contribution dans l'ensemble des travaux de cette thèse et l'état de l'art :

- **Inscription de la contribution dans la littérature existante :** Comme évoqué précédemment, les apports de cette contribution sont théoriques. Elle propose une extension théorique d'une méthode explicative pertinente à des modèles qui leurs sont plus généraux. Elle apporte de meilleurs résultats que les méthodes de l'état de l'art, sur le plan qualitatif et quantitatif, en utilisant un protocole d'évaluation analogue à ceux utilisés dans la littérature.
- **Inscription de la contribution dans la problématique de recherche :** La méthode explicative développée est adaptée au paradigme de l'apprentissage géométrique profond. Elle n'est pas illustrée dans le cas de données médicales, mais peut l'être. Cette contribution s'inscrit donc dans notre problématique globale de recherche.

Ainsi la méthode ScoreCAM GNN et son évaluation sont perfectibles sur plusieurs aspects. En nous affranchissant de tout paradigme, et en incluant les deux modalités descriptives des données répondant à l'apriori géométrique , nous avons défini une nouvelle méthode explicative adaptée aux réseaux de neurones sur graphe :

EiX-GNN (Eigencentrality eXplainer for Graph Neural Networks), que nous détaillerons dans un prochain chapitre de ce document. Cependant, bien que notre méthode ScoreCAM GNN ait des défauts, nous avons souhaité l'illustrer sur une tâche de reconnaissance d'émotions faciales. Le protocole et les résultats de cette expérience sont détaillés dans le prochain chapitre.

# Chapitre 8

# Cas d'utilisation : Compréhension des expressions faciales par une caractérisation géométrique et profonde des visages humains

## Sommaire

## 8.1 Contexte global de la contribution et motivations

Les émotions sont l'un des principaux canaux de communication de l'Homme. Elles sont le résultat de nombreux processus internes humains complexes qui ont eux-mêmes été déclenchés par des stimuli externes. La compréhension du fonctionnement des émotions humaines est un sujet de recherche académique et industrielle. Leurs applications permettent par exemple l'amélioration de la compréhension de la psychologie humaine ou de faire du ciblage publicitaire. Dans cette étude [1], nous proposons un nouveau cadre permettant de comprendre visuellement les émotions faciales ; de quelle manière elles se distinguent les unes des autres ; et tout cela, numériquement. Notre approche repose sur l'encodage des visages humains sur graphes dont les noeuds représentent des points de repère sémantiques et positionnels du visage humain. La distribution spatiale de ces points est étroitement liée aux processus de révélation des émotions faciales. Nous commençons par classer ces graphes à l'aide d'un modèle prédictif GNN. Nous présentons ensuite une étude qualitative qui explique le processus décisionnel interne pris par le modèle prédictif, ce qui, dans ce contexte, est une approximation de la définition numérique du processus de révélation des émotions faciales. La caractérisation des émotions est avant tout un problème dépendant de la reconnaissance des émotions. Les problèmes de reconnaissance des émotions ont fait l'objet d'études approfondies sous différents angles. Par exemple, via l'analyse du traitement du langage naturel, [Lian et al., 2020]. D'autres études ont été menées avec des signaux EEG [Bos, ] qui nous permettent d'avoir une notion approfondie de ce qu'est l'émotion d'un point de vue neurologique. D'autres études se sont intéressées à la caractérisation des émotions humaines, à savoir [Neerincx et al., 2018, Pham et al., 2021] et une étude basée sur les graphes a été réalisée par [Guerdan et al., 2021]. La compréhension des émotions faciales d'un point de vue visuel peut se faire via un traitement de l'alignement facial. L'alignement facial est au coeur de l'intérêt de la communauté des chercheurs s'intéressant à la vision par ordinateur. Il consiste à spécifier la position des repères faciaux en fonction d'une sémantique spécifique. Certaines solutions pertinentes ont été proposées par cette

1. Notre implémentation est disponible : `github.com/araison12/emotixaignn`

communauté, ce qui a conduit à un large déploiement dans les applications industrielles, bien que cela reste un sujet de recherche toujours actif. Dans cette étude, nous proposons un procédé entier qui, tout d'abord, encode les visages humains présents dans les images couleur sous forme de graphes ; ensuite, nous expliquons la classification, via GNN, de la représentation des visages via une représentation sous forme de graphe, ce qui donne une définition numérique de ce que sont les émotions faciales humaines [2]..

## 8.2 Problématique

L'émotion est une réponse humaine multi-factorielle à l'interaction environnementale. Dans la communication interhumaine, les émotions aident à spécifier les messages échangés. L'expression des émotions peut se faire par différents moyens. L'un d'entre eux consiste à placer le visage dans une configuration spatiale spécifique qui est socialement reconnaissable. À partir de là, la révélation d'une émotion est un processus de transfert entre une forme initiale du visage et une autre forme spécifique. La pose qui en résulte peut être définie comme étant l'émotion exprimée. Par exemple, dans une conversation à deux, l'expression de la joie se fait la plupart du temps par le sourire et le soulèvement des joues de l'une des personnes impliquées. Le message émotionnel est donc reçu si les deux individus partagent le même processus de révélation de l'émotion faciale. Dans le cadre d'une approche assistée par ordinateur, la compréhension et la caractérisation des émotions du visage humain peuvent être abordées en comprenant de manière numérique la géométrie faciale conditionnée sous-jacente et surtout en tenant compte de ces configurations précises. Nous présentons ici l'approche que nous proposons et qui fournit des résultats pertinents, même dans un large éventail de contextes sociaux. La déformation de la forme du visage est due à l'activation conjointe des muscles faciaux. Lors de l'expression d'émotion, certains motifs spatiaux apparaissent. L'identification de ces motifs et leur caractérisation numérique permettent alors la compréhension numérique des émotions faciales. Le codage de ces motifs nécessite de les définir physiquement et de les transcrire numériquement. Les visages sont des structures biologiques complexes,

2. La rédaction de ce chapitre, présentant une de nos contributions, a fortement été inspiré du document originel rédigé en anglais

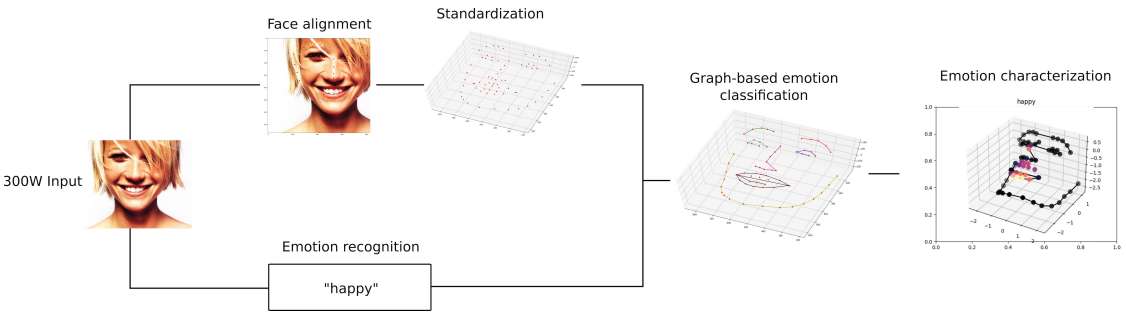



FIGURE 8.1 – Notre méthode repose sur trois composants : un module d'étiquetage qui est utilisé pour construire nos ensembles de données d'émotions faciales sous forme de graphe ; une fois cet assemblage effectué, un modèle profond va classifier nos graphes en fonction des positions spatiales normalisées des sommets du graphe ; une conversion est ensuite opérée pour mettre cette représentation complexe dans un format plus intelligible qui permet de caractériser les émotions faciales humaines grâce à une méthode explicative adaptée aux GNN.

mais nous pouvons représenter les groupes de muscles faciaux comme des repères positionnels sémantiques. Pour adresser cette tâche, nous utilisons une méthode d'alignement des visages de l'état de l'art. La variation de la spatialité des repères est due aux muscles faciaux adjacents qui se contractent ou s'étirent conjointement dans certaines parties du visage. Grâce au modèle prédictif , nous allons pouvoir encoder numériquement les interactions intra-groupes et inter-groupes musculaires et ainsi encoder numériquement les émotions. Dans la suite, nous présentons notre procédé de conversion de la représentation basée en image des visages, en graphe. Puis nous présenterons le cadre de la classification des émotions basées sur ces représentations sous forme de graphes et leurs caractérisations via une méthode explicative adaptée au modèle prédictif. Ces images sont les entrées de notre étape de prétraitement : la détection de repères faciaux et la reconnaissance d'émotions basées sur l'image. Par souci de clarté, nous désignons $D$ comme étant notre ensemble de données d'images. La taille de l'ensemble de données est notée $|D| \in \mathbb{N}$ et nous considérons une image de taille $k, l \in \mathbb{N}$ comme un triplet $\mathbf{X} \in \mathbb{M}_{k,l}([0, 255] \cap \mathbb{N})^3$.

### 8.2.1 Inférence de repères faciaux

A partir d'images couleur, nous déterminons des caractéristiques graphiques à apprendre grâce à une méthode d'alignement des visages. En supposant que nous disposions d'un détecteur de repères faciaux optimal basé sur l'image $\mathtt{fa}^\star$ qui détermine soigneusement $p \in \mathbb{N}$ repères faciaux tridimensionnels, c'est-à-dire :

$$\begin{array}{rccc} \mathtt{fa}^\star : & D & \to & \mathbb{R}^{3\times p} \\ & \mathbf{X} & \mapsto & (\mathbf{x}_i)_{i\in\{1,\cdots,p\}} \end{array}$$

Étant donné que chaque instance $\mathbf{X}$ de $D$ a son propre contexte d'enregistrement, ainsi l'inférence $\mathtt{fa}^\star(\mathbf{X})$ peut prendre ses valeurs dans des sous-espaces de $\mathbb{R}^3$ très variables. Sans procédure de normalisation, cela conduira à un apprentissage machine non pertinent, car le signal d'émotion intrinsèque n'est qu'une interaction entre les points de repère de manière relative et donc non absolue. Pour obtenir une plus grande stabilité statistique, nous appliquons un processus de normalisation reposant sur la recherche de transformations affines rigides optimales par rapport à une configuration de référence. Nous considérerons ici les transformations d'échelle, de rotation et de translation, car ce sont ces trois transformations affines, ou degré de liberté auquels l'acquisition est sensible et auquels le contexte est adapté. Cette normalisation place $\mathtt{fa}^\star(D)$ dans une espace commun, permettant la stabilité statistique et donc une meilleure classification de l'ensemble des données. Ce processus de normalisation est fait relativement à un visage moyen. Ce visage moyen est sans émotion, synthétique, et a été conçu en faisant la moyenne spatiale de $p$ points de repère de visages neutres alignés, déterminée en amont. Cela signifie également que l'émotion neutre sera notre référence pour caractériser les autres émotions faciales. Ce visage moyen est noté $(\mathbf{m}_i)_{i\in\{1,\cdots,p\}} \in \mathbb{R}^{3\times p}$. Le processus de normalisation est formalisé comme suit ; pour tout $(\mathbf{x}_i)_i \in \mathtt{fa}^\star(D)$ :

$$(s^\star_{\mathbf{X}}, \mathbf{R}^\star_{\mathbf{X}}, \mathbf{t}^\star_{\mathbf{X}}) = \underset{(s,\mathbf{R},\mathbf{t}^T)\in\Omega}{\arg\min} \sum_{i=1}^{p} \|\mathbf{m}_i^T - s\mathbf{R}\mathbf{x}_i^T - \mathbf{t}^T\|_2 \tag{8.1}$$

With $\Omega = \mathbb{R}^+ \times SO(3) \times \mathbb{R}^3$. De plus, les transformations géométriques considérées ne doivent pas altérer la structure biologique globale des visages humains qui est, sans considérations abusives, supposée constante entre tous les individus. Et comme les transformations affines préservent notamment la colinéarité, le parallélisme, le

rapport des longueurs et des barycentres, elles conservent l'aspect physique global des visages et n'introduisent pas d'altérations sémantiques ou positionnelles. Ce problème (8.1) a une solution sous forme fermée dont la preuve est donnée dans [Horn, 1987]. En raison de la variation globale de la forme du visage entre les individus dans $D$, la variation non linéaire (c'est-à-dire la singularité individuelle) n'est pas prise en compte, mais peut être considérée comme une caractéristique bruyante qui peut ne pas affecter de manière significative la phase d'apprentissage. Nous désignons l'ensemble de données normalisé par $D^\star$, ce qui signifie :

$$D^\star = \{s^\star_{\mathbf{X}}\mathbf{R}^\star_{\mathbf{X}}\mathbf{X}^T - \mathbf{t}^{\star T}_{\mathbf{X}} | \mathbf{X} \in \mathtt{fa}^\star(D)\} \tag{8.2}$$

Nous intégrons ensuite cette nouvelle représentation normalisée des données sous la forme d'un graphe selon le schéma suivant. D'après la définition originale, les graphes sont des couples $(V, E)$ où $V$ est l'ensemble des noeuds et $E \subset V \times V$ est l'ensemble des arêtes. L'ensemble $E$ est la représentation de l'adjacence du graphe considéré. Il existe une formulation matricielle pour la représentation des arêtes qui est totalement équivalente et qui permet d'effectuer des opérations algébriques. La matrice d'adjacence $\mathbf{A} \in \mathbb{M}_{|V|}(\{0, 1\})$ ayant comme entrée $a_{i,j}$ est définit par $a_{i,j} = 1$ si $(i, j) \in E$, 0 sinon. Dans le contexte de la représentation profonde des signaux évoluant sur le graphe, les graphes sont plutôt considérés comme $(\mathbf{L}, \mathbf{A})$ de sorte que $\mathbf{L}$ est un vecteur colonne de taille $|V|$ évalué dans $\mathcal{H}$, un espace de Hilbert de dimension $d$ ($d \in \mathbb{N}$). Cela signifie donc que chaque noeud de $V$ est un vecteur ligne de taille $d$ à valeurs dans $\mathcal{H}$ représentant le signal évoluant sur ce noeud. Nous transformons de manière bijective $D^\star$ vers

$$D^\star_G = \{G_{\hat{\mathbf{X}}} | \hat{\mathbf{X}} \in D^\star\} \tag{8.3}$$

où pour chaque $\hat{\mathbf{X}} \in D^\star$

$$G_{\hat{\mathbf{X}}} = (\hat{\mathbf{X}}, \mathbf{A}) \in D^\star \times \{\mathbf{A}\} \tag{8.4}$$

. Notons que $\mathbf{A}$ ne dépend pas de $\hat{\mathbf{X}}$, car la topologie des graphes induits (c'est-à-dire l'adjacence des interactions entre les muscles du visage) est une caractéristique commune à tous les êtres humains. En effet, comme nous l'avons déjà mentionné, lorsqu'un contexte social et une émotion sont donnés, les personnes l'expriment selon les mêmes schémas d'activation des muscles faciaux. La conception de $\mathbf{A}$ a été

| **Descripteur sémantiques** | **Indices des noeuds** |
|---|---|
| Sourcil gauche | 17,18,19,20,21 |
| Sourcil droite | 22,23,24,25,26 |
| Oeil gauche | 36,37,38,39,40,41 |
| Oeil droite | 42,43,44,45,46,47 |
| Nez | 27,28,29,30,31,32,33,34,35 |
| Menton et joues | 0,1,2,3,4,5,6,7,8,9,10,11,...,16 |
| Bouche (supérieur) | 48,49,50,51,52,53,54,64 |
| Bouche (inférieur) | 48,60,59,58,57,56,55,54,64,65,66,67 |

TABLE 8.1 – Lien entre sémantique concernant les parties du visage et les indices des noeuds

guidée par l'hypothèse de connexité locale de l'activation musculaire. L'hypothèse de connexité locale de l'activation musculaire est basée sur le fait que dans toutes les localités du visage, les muscles physiquement connectés localement agissent ensemble en tant que groupe musculaire (c'est-à-dire qu'un muscle donné, lorsqu'il s'active, les muscles adjacents sont plus susceptibles d'être entraînés par cette activation et d'agir à leur tour). Nous supposons donc que cette hypothèse s'applique à la révélation des émotions faciales.

### Inférence de reconnaissance des émotions

Tout comme la méthode de détection des repères faciaux, la reconnaissance des émotions à partir d'images en couleur a fait l'objet de nombreuses recherches dans le domaine de la vision par ordinateur. De nombreuses études ont été menées et les problèmes de reconnaissance des émotions sont souvent formulés comme des problèmes de classification supervisée, car de nos jours, de nombreuses données annotées sont disponibles publiquement. En considérant $C \in \mathbb{N}$ différentes émotions, un système de reconnaissance des émotions est simplement une fonction optimale $\mathtt{er}^\star$ telle que :

$$\begin{array}{rccl} \mathtt{er}^\star : & D & \rightarrow & \{1, \cdots, C\} \\ & \mathbf{X} & \mapsto & c_{\mathbf{X}} \end{array}$$

Nous déduisons ensuite $\texttt{er}^\star(D)$ pour obtenir des étiquettes qui seront ensuite utilisées pour classer les émotions par rapport à $D_G^\star \times \texttt{er}^\star(D)$, c'est-à-dire d'une manière basée sur les graphes plutôt que dans un cadre basé sur les images.

*Note :* Les déductions faites à l'aide de $\texttt{fa}^\star$ et $\texttt{er}^\star$ peuvent ne pas être absolument exactes en ce qui concerne, respectivement, les émotions et les repères faciaux. La qualité de $D_G^\star \times \texttt{er}^\star(D)$ dépend fortement de la précision de $\texttt{fa}^\star$ et de $\texttt{er}^\star$. Il convient de noter qu'une mauvaise affectation peut conduire à des résultats non pertinents et fausser la conclusion de cette étude.

### Classification des caractéristiques

Afin d'acquérir une compréhension approfondie des émotions faciales et d'encoder efficacement la distribution de la position relative des points de repère pour une émotion donnée, nous tirons parti des puissantes capacités de représentation des données qu'ont les modèles prédictifs profonds. Par conséquent, nous avons utilisé un modèle profond pour classer les émotions adaptées aux données représentées sous forme de graphes, les GNN. Les GNN sont des modèles profonds introduits dans [Scarselli et al., 2009] qui sont capables de traiter des données représentées sous forme de graphe efficacement. Des variantes de GNN ont été proposées [Kipf and Welling, 2017, Veličković et al., 2018a]. Nous avons utilisé ces modèles pour résoudre notre tâche de classification.

### Compréhension des caractéristiques

Après avoir déterminé $\theta^\star$ le paramètre optimal pour le modèle prédictif $f_{\boldsymbol{\theta}^\star}$, nous pouvons analyser comment les processus de décision internes de $f_{\boldsymbol{\theta}^\star}$ ont été conçus. Le modèle prédictif $f_{\theta_\star}$ combine de manière non linéaire de nombreuses échelles différentes de représentations d'entrée. Nous utilisons ensuite une méthode explicative qui cherche à mettre en évidence de manière intelligible et précise ce processus décisionnel.

## 8.3 Protocole

Dans cette section, nous présentons l'ensemble de données que nous avons utilisé pour étiqueter les données, ainsi que notre configuration pour $f_{\theta^\star}$ et les résultats de nos expériences.

### 8.3.1 Ensemble de données

Par soucis de reproductibilité de notre expérience, nous utilisons des images RVB disponible publiquement provenant de célèbres ensembles de données de vision par ordinateur.

#### 300W

L'ensemble de données [Sagonas et al., 2016, Sagonas et al., 2013, Sagonas et al., 2013] est un ensemble de données de vision par ordinateur largement utilisé pour les problèmes de repères faciaux. Cet ensemble de données est intéressant, car il fournit un très grand nombre d'exemples contextuels comprenant une large gamme d'émotions. Cette variété de données couvre la majorité des émotions humaines qu'une personne utilisant notre outil peut rencontrer dans la vie quotidienne. Cet ensemble de données est composé respectivement de 300 instances contextualisées en intérieur et de 300 instances contextualisées en extérieur. C'est à partir de cet ensemble de données que nous avons inféré des repères tridimensionnels et des émotions.

#### Outils d'alignement facial

Nous avons utilisé [Bulat and Tzimiropoulos, 2017] comme détecteur de repères tridimensionnels `fa`$^\star$. Ce modèle est l'un des meilleurs détecteurs de points de repère 2D à 3D et il détecte $p = 68$ points de repère facial. Il s'appuie sur le réseau d'alignement des visages. Plusieurs méthodes (par exemple [Zhu et al., ]) ont été conçues pour trouver directement des points de repère tridimensionnels à partir d'images brutes. Mais pour des raisons de précision, la régression des modèles dans l'espace bidimensionnel est plus facile du point de vue de l'optimisation que dans l'espace tridimensionnel, car l'espace de recherche est dans la dimension inférieure. Ainsi, un modèle 2D-3D qui étend simplement la dernière dimension sur la base

d'une position 2D déjà précise est préférable à un modèle de régression 3D qui est optimisé à partir de zéro.

**Outils de reconnaissance des émotions**

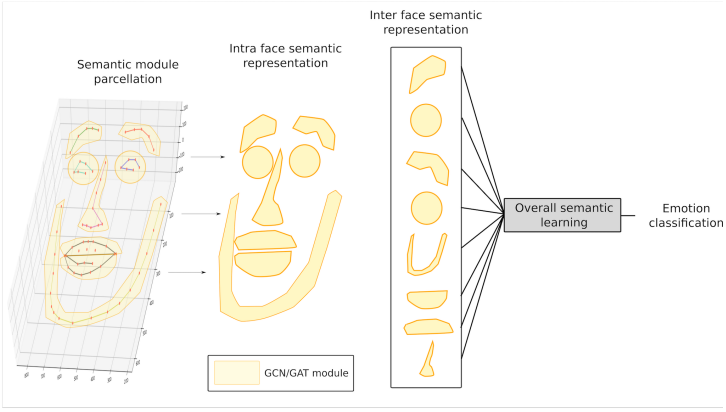
Comme outil de reconnaissance des émotions `er`$^\star$, nous avons choisi le modèle d'état de l'art [Pham et al., 2021]. Il atteint une précision de 76,82% sur l'ensemble de données FER2013, qui est l'ensemble de données d'état de l'art pour ce type de tâche de classification. Elle a été entraînée à détecter 7 types d'émotions : la colère, la peur, le dégoût, la joie, la tristesse, la surprise et la neutralité. Cette méthode utilise un réseau de segmentation pour affiner les cartes de caractéristiques qui sont ensuite intégrées à un réseau résiduel profond et à une architecture basée sur un U-Net. Pour autant que nous le sachions, ces deux outils ont obtenu de bons résultats par rapport à leurs états de l'art respectifs. C'est pourquoi nous les avons choisis pour mener cette étude. Nous présentons maintenant notre propre architecture de modèle prédictif en fonction de ces $C = 7$ types d'émotions.

### 8.3.2 Architecture du modèle prédictif et hyperparamètres

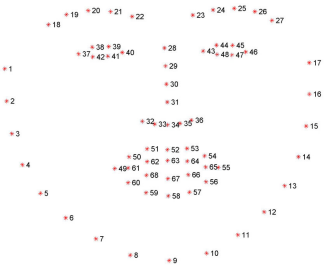
L'architecture de $f_{\boldsymbol{\theta}}$ (Figure 8.2a) est conçue avec un GNN sur chacune des composantes connexes du visage mises en parallèle, l'information partielle est alors fusionnée dans un dernier GNN global et est classifiée par un perceptron multicouche à une couche de 512 neurones permettant la classification des émotions. Dans cette section, nous présentons tout d'abord les performances que nous avons atteintes pour la classification des émotions dans le cadre de notre approche basée sur les graphes. Ensuite, nous illustrerons nos résultats concernant la singularisation des émotions obtenue par la méthode explicative ScoreCAM GNN, qui s'est avérée être la plus pertinente en ce qui concerne les métriques d'évaluation objectives comme montré dans le Tableau 8.2.

### 8.3.3 Performances de classification

Pour la tâche de classification, la distribution des étiquettes doit être équilibrée pour une meilleure représentativité statistique. Il s'avère que les émotions ne sont pas uniformément réparties sur les instances de 300W. Par conséquent, nous nous




(a) Structure de classificateur basée sur des graphiques : nous avons conçu le schéma de préparation du signal en concentrant les localités sémantiques entre elles. Chaque partie sémantique est considérée comme une composante convexe du graphe évalué. Tant que chaque composant code une information partielle sur l'émotion, nous rassemblons chacun d'eux dans un graphe sémantique avec huit composants connectés de manière non contrainte, cela signifie qu'il s'agit d'un graphe complet. La distribution de données intégrées obtenue est ensuite transmise à un module GCN et à une partie linéaire qui aide à la classification.



(b) Indexation initiale des nœuds sémantiques que nous avons utilisée pour notre expérimentation. Cette indexation est largement utilisée pour résoudre les problèmes d'attribution de repères de visage.

sommes concentrés uniquement sur les classes qui apparaissent au moins à 10% parmi les instances de 300W. Ces émotions sont la joie, la neutralité, la tristesse et la peur, ce qui signifie que $C = 4$. Nous notons également quelques erreurs d'attribution d'émotions par $\texttt{er}^{\star}$. Dans notre configuration expérimentale, nous atteignons 66% de précision sur la variante graphique de l'ensemble de données 300W. Ce résultat reste acceptable compte tenu du problème de l'attribution erronée et de la précision du modèle de classification général, qui est de l'ordre de 80%. Sur la base de ces résultats, nous fournissons maintenant quelques mesures qualitatives de caractérisation des émotions.

### 8.3.4 Étude comparative des méthodes explicatives de l'état de l'art

Afin de fournir des résultats aussi pertinents que possible, nous avons mené une étude statistique comparative entre les méthodes d'explication de l'état de l'art. Dans cette étude, nous retenons GNNExplainer, la méthode d'état de l'art, LRP-GNN et ScoreCAM GNN qui fournissent des explications dans des délais réalistes par rapport à XGNN ou SubgraphX. Comme le montre le Tableau 8.2, nous avons mesuré l'évaluation objective de certaines méthodes d'explication (la fidélité et l'éparpillement) et ScoreCAM GNN semble être la méthode d'explication la plus pertinente. Par conséquent, nous ne prendrons en compte que les explications ScoreCAM GNN pour mener notre étude sur la compréhension émotionnelle. ScoreCAM GNN [Raison, 2022] est une extension au domaine non euclidien du ScoreCAM, initialement introduit par [Wang et al., 2020a]. ScoreCAM combine linéairement le plus haut niveau de représentation des données et pondère chacun d'entre eux par son propre poids de contribution relativement à la tâche de classification. Le résultat de ScoreCAM GNN est, étant donné un modèle prédictif et une instance, une distribution normalisée de ces contributions à la classification de l'instance, de chaque élément du domaine sur lequel l'instance s'appuie (c.-à-d. les noeuds). Dans notre contexte, elle fournit l'impact de chaque noeud sur la classification du graphe considéré selon $f_{\boldsymbol{\theta}^{\star}}$. En d'autres termes, elle fournit l'importance de chaque point de repère qui a une correspondance physique presque bijective (voisinage des muscles faciaux) impliquée dans la description des émotions, de sorte qu'une compréhension approfondie

des émotions faciales est possible.

| | Infidélité | Parcimonie |
|---|---|---|
| GNNExplainer | 1.67 | 0.23 |
| LRP-GNN | 0.79 | 0.14 |
| **ScoreCAM GNN** | **0.21** | **0.89** |

TABLE 8.2 – Étude statistique comparative des méthodes d'explication : la colonne de gauche représente l'infidélité moyenne sur l'ensemble de données 300W pour chaque méthode d'explication évaluée. L'infidélité mesure l'infidélité d'une méthode d'explication par rapport à une instance et à un modèle profond. Plus l'infidélité est faible, mieux c'est. Nous avons également mesuré l'espacement moyen des explications fournies par rapport à chaque méthode d'explication. La parité est une autre métrique objective qui mesure la concision de chaque explication fournie. Plus l'espacement est important, mieux c'est. Selon les deux mesures objectives, ScoreCAM GNN a montré empiriquement de meilleurs résultats.

## 8.4 Résultats

La compréhension des comportements des modèles profonds peut dépendre de la précision du modèle, puisque celle-ci reflète la compréhension numérique des tâches d'apprentissage du modèle. Compte tenu de ce qui précède, certains résultats peuvent être quelque peu hors de propos, mais il apparaît expérimentalement que l'explication des comportements de $f_{\boldsymbol{\theta}^\star}$ est conforme à la référence sociale de ce qu'est une émotion par rapport à d'autres émotions, et donc à la caractérisation visuelle des émotions faciales. Comme le montre la Figure 8.3, nous avons affiché la caractérisation des quatre émotions avec leur représentation d'image initiale. Pour la tâche de classification, une méthode d'explication à instance unique fournit toujours son explication en incluant les informations pertinentes (compréhension de la classe actuelle) et celles qui ne le sont pas (classes restantes) puisque la classification est en permanence un problème de partition de l'espace. Nous avons tout d'abord décrit ce qu'est l'émotion neutre, car elle peut être considérée comme une base de référence pour comprendre et caractériser les autres émotions. En effet, nous pouvons consi-

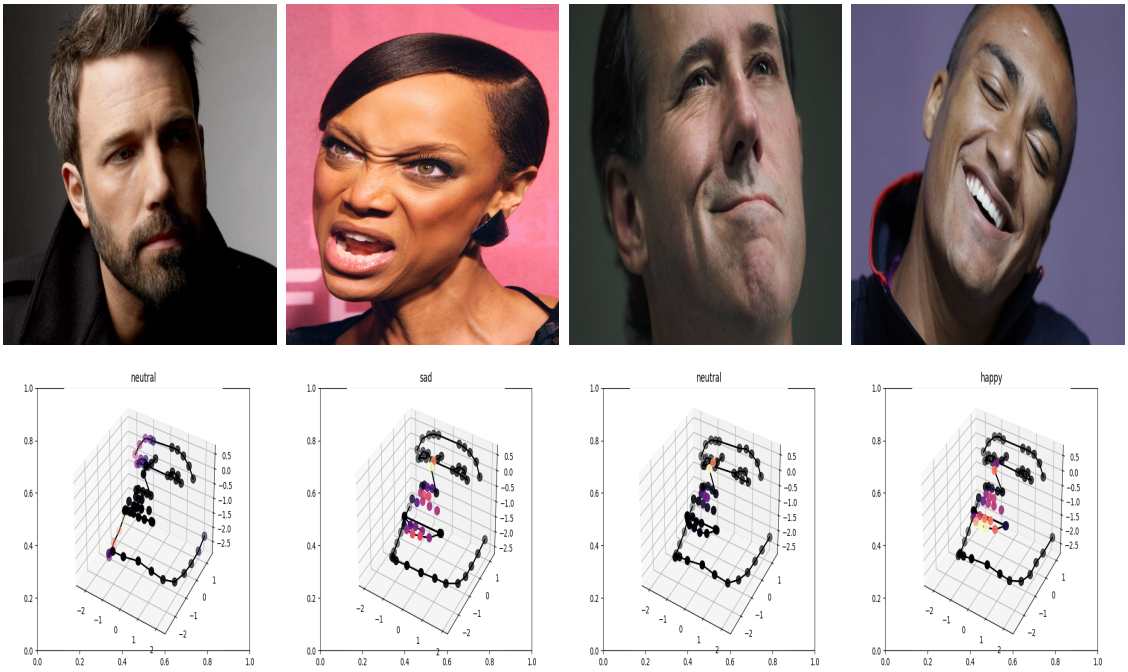



FIGURE 8.3 – Caractérisation des émotions par ScoreCAM GNN : Dans ces quatre représentations doubles de visages humains, nous mettons en évidence la définition numérique des émotions faciales. L'émotion neutre est l'émotion de base pour caractériser les autres. Elle montre que le bonheur et la tristesse sont tous deux fortement intégrés dans les mouvements de la bouche et du nez, qui sont socialement pertinents.

dérer l'émotion neutre comme une émotion non émotionnelle. La caractérisation des autres émotions sera ensuite fournie de manière contrastive, avec comme base de référence, l'émotion neutre qui est géométriquement non ambiguë. Pour l'émotion de bonheur, ce que nous avons remarqué et ce que ScoreCAM GNN a révélé, c'est que la bouche supérieure, la bouche inférieure et le nez sont conjointement impliqués dans la description brute de ce qu'est le bonheur sur un visage humain. De même, l'émotion de tristesse peut être considérée comme le symétrique du bonheur et nous avons découvert que la reconnaissance de la tristesse d'un point de vue géométrique se fait également par une configuration spécifique des muscles de la bouche et du nez qui est très différente de la configuration du bonheur. Ces résultats sont en accord social avec ce que nous pouvons attendre de la réponse de ce qui caractérise les émotions faciales.

## 8.5 Conclusion

Dans cette étude, nous proposons une approche originale pour caractériser visuellement les émotions du visage humain. En utilisant des méthodes d'état de l'art pour récupérer des informations de haut niveau, telles que la reconnaissance des émotions basée sur l'image ou le positionnement sémantique des repères faciaux, nous encodons les visages humains sous forme de graphe pour renforcer et exploiter les aspects relationnels des schémas d'activation musculaire impliqués dans les processus de révélation des émotions humaines. Même avec des informations incomplètes dues principalement à la faiblesse des méthodes d'étiquetage, qui se distinguent comme étant des méthodes d'état de l'art dans leur propre domaine d'application, nous obtenons les résultats sociaux escomptés, à savoir quelles sont les émotions et quels sont les groupes musculaires impliqués dans leur description visuelle. D'autres travaux pourraient inclure des méthodes d'étiquetage plus précises afin d'augmenter la précision globale du modèle prédictif basé sur les graphes. D'autres méthodes d'explication peuvent être utilisées pour fournir une compréhension des émotions.

## 8.6 Discussions, étude épistémologique, limites et ouvertures

Concernant notre application à la caractérisation numérique des émotions faciales, notre étude admet aussi quelques limites, en voici quelques unes :

- Ensemble de données restreint : l'ensemble de données que nous avons retenu pour cette étude à un volume très restreint et notre modèle prédictif probablement pas assez régularisé. L'optimalité du modèle prédictif peut alors être largement remise en question. De plus, la qualité des annotations des données qui le constitue est largement discutable aussi :
  - Annotation des repères faciaux : l'annotation des repères faciaux est faite via l'inférence d'un modèle prédictif profond ayant lui aussi ces sensibilités et n'est objectivement pas optimale, donc nos annotations peuvent être fausses sur certaines données.
  - Annotation des émotions : de manière analogue au cas de l'annotation des repères faciaux, l'annotation des émotions admet les mêmes défauts.
- Normalisation des données incomplète : nos données ont subi une première normalisation permettant une première régularisation, mais leurs représentations sont encore sous forme de repère spatial tridimensionnel. Ainsi, on peut imaginer aisément que le modèle prédictif peut considérer un paramètre optimal, mais qui est encore sensible à des rotations, ou des effets d'échelle, alors que l'attribution d'émotions est invariante sous ces transformations. La représentation spatiale a alors ces limites. De plus les variations positionnelles entre émotions peuvent être légères devant celle de nos repères dans l'espace tridimensionnel. Cela peut alors être considéré comme une sorte de bruit, alors qu'il s'agit d'information pertinente. En d'autres termes, nos données répondent aux aprioris géométriques des graphes, mais aussi à d'autres aprioris que nous avons que partiellement pris en compte. Pour une meilleure représentativité de l'ensemble de données, il est donc fortement probable qu'un encodage de donnée, moins naïf que celui employé, soit à adopter.
- Adjacence arbitraire des groupes de données : pour construire nos données, l'adjacence des noeuds est constante au long de l'ensemble de données. Elle a

été déterminée arbitrairement, et il est possible que notre étude ne considère pas certaines interactions entre groupes musculaires pertinentes pour l'expression d'émotions faciales. Cela limite donc le pouvoir expressif de notre modèle prédictif.

— Absence d'étude statistique : bien que nous connaissions les liens entre sémantiques, positionnement et indexation des noeuds, nous n'avons pas mené d'étude statistique, notamment de caractérisation "moyenne" des émotions sur un visage moyen. Ainsi nous ne connaissons pas la robustesse de nos conclusions qui peuvent donc admettre une certaine variance.

— Absence de vérité terrain émotionnelle : bien que nous ayons des informations positionnelles sémantiques, nous n'avons pas de vérité terrain sur ce que sont les émotions, et encore moins une version numérique de cette vérité terrain. Cela empêche donc d'apporter un aspect plus objectif à notre étude.

— Utilisation de ScoreCAM et conclusions : nous avons voulu illustrer notre méthode explicative ScoreCAM, GNN mais cette dernière a des limites dont certaines sont évoquées précédemment. Ce point peut donc aussi fortement altérer les conclusions de notre étude.

Ainsi, cette étude a permis, via des arguments numériques nouveaux, de retrouver des attentes sociales bien établies. C'est la contribution de cette étude. Pour son caractère original, elle a fait l'objet d'une publication dans une conférence internationale à comité de lecture :

A. Raison, T.Biardeau, P. Bourdon, and D. Halbert, *"Face expressions understanding by geometrical characterization of deep human faces representation"*, Electronic Imaging (EI), San Francisco, United States, 2023, doi : http ://dx.doi.org/10.2139/ssrn.4170493

## 8.7 Inscription de la contribution avec le projet de recherche et la littérature

Nous présentons ici les points d'articulation de cette contribution dans l'ensemble des travaux de cette thèse et l'état de l'art :

— **Inscription de la contribution dans la littérature existante :** Comme évoqué précédemment, les apports de cette contribution sont techniques. Elle propose une nouvelle vision, formulation des représentations sociales avec des arguments contemporains.

— **Inscription de la contribution dans la problématique de recherche :** Les modèles prédictifs utilisés répondent au paradigme de l'apprentissage géométrique profond. Nous avons dans cette contribution établi un protocole permettant une caractérisation numérique des émotions faciales. La mise en place du protocole et l'interprétation des résultats sont nouvelles pour la communauté.

# Chapitre 9

# EiX-GNN : une approche spectrale et sociale pour l'explicabilité pour réseaux de neurones sur graphes

## Sommaire

## 9.1 Contexte global de la contribution et motivations

Les méthodes explicatives doivent fournir des informations pertinentes et intelligibles pour les humains concernant les aspects internes des modèles prédictifs profonds en expliquant le comportement de ces modèles. Expliquer, comprendre ou interpréter, bien qu'il s'agisse de notions différentes, dépend intrinsèquement de l'homme et du contexte. Il s'agit donc de notions sociales. L'une de ces notions sociales est qu'un explicateur doit adapter la formulation de son explication en fonction du contexte relativement à celui qui reçoit l'explication en ce qui concerne le phénomène à expliquer. [Clough et al., 2019a, Kim et al., 2018b]. Notamment, un explicateur doit partager ses connaissances sur un phénomène avec celui qui reçoit les explications d'une manière compréhensible pour ce dernier, afin que le processus d'explication soit efficace et profitable pour celui qui reçoit les explications [Bechtel and Abrahamsen, 2005, Glennan, 2002, Chater and Oaksford, 2005, Keil, 2006]. Plusieurs méthodes explicatives ont été proposées pour expliquer les modèles prédictifs utilisant les réseaux de neurones sur graphes, mais ces méthodes, souvent, ne tiennent pas compte de la dépendance sociale du processus explicatif lorsqu'elles fournissent leurs explications et se concentrent uniquement sur le côté purement numérique des modèles profonds pour fournir des informations sur les aspects internes des modèles profonds pour graphes. Dans cette contribution, nous proposons une méthode explicative tenant compte de la dimension sociale qui exploite la variabilité des connaissances de celui qui reçoit une explication, tout en conservant un score élevé par rapport aux mesures d'évaluation objectives les plus récentes de l'état de l'art. Tout d'abord, nous définirons le contexte social dont dépend le processus d'explication. Ensuite, nous présenterons notre approche [1] conformément à la formulation numérique du contexte social. Ensuite, nous montrerons la pertinence de notre méthode par rapport aux méthodes comparées dans une étude qualitative objective dans des cas réels, et une étude quantitative objective par rapport aux métriques objectives utilisées dans la littérature. Dans cette étude, nous fournissons tout d'abord une méthode pertinente, qui permet d'obtenir des résultats d'expli-

1. Notre implémentation est disponible : `github.com/araison12/eixgnn`

cation plus pertinents, de manière absolue, que les méthodes de pointe, mais aussi relativement à la dépendance entre celui qui reçoit l'explication et plus largement le contexte social dont dépend tout processus d'explication [2]..

## 9.2 Problématiques

L'explication aide les experts humains à inspecter les modèles prédictifs profonds, afin de mettre en évidence les problèmes divers et pour empêcher ces modèles de nuire potentiellement à la société. Pour expliquer les modèles prédictifs profonds, le contexte social [Selbst et al., 2019] et le facteur humain [Murdoch et al., 2019, Miller, 2019a] sont des éléments essentiels. En effet, ce processus d'explication implique un alignement des modèles mentaux. Cet alignement se fait entre ce que fait le modèle prédictif profond et ce que l'utilisateur pense que le modèle fait. En termes d'ordre, pour réaliser cet échange d'informations, il faut un ensemble d'arguments dont la machine et l'utilisateur sont conjointement conscients et qu'ils sont capables de traiter. L'explication d'un modèle prédictif profond est donc un processus centré sur l'homme et, afin de fournir à l'utilisateur un aperçu significatif du comportement du modèle, le processus d'explication doit être adapté [Kim et al., 2018b, Clough et al., 2019a]. Il convient de noter qu'il est plus facile de façonner la représentation des résultats d'une machine que d'obliger les humains à penser d'une manière très différente de celle à laquelle ils sont habitués. Par conséquent, c'est la machine qui doit être adaptée à l'utilisateur. Mais lorsqu'un utilisateur souhaite obtenir des explications pertinentes, celles-ci doivent être exprimées selon une granularité spécifique à celui qui reçoit les explications. En effet, un utilisateur ayant un niveau de connaissance élevé a des attentes différentes de celles d'un utilisateur moins averti en ce qui concerne le modèle prédictif profond concerné. De manière plus formelle, il peut être reformulé comme suit : l'explication est un processus de transfert de connaissances humaines impliquant un explicateur (par exemple, une machine) $E^{\star}$ et un destinataire $\tilde{E}$ (par exemple, un utilisateur, un ingénieur) concernant un phénomène $P$ (par exemple, un modèle prédictif profond). Afin d'avoir une conversation fructueuse (par exemple, fournir l'explication de $P$ à partir de $E^{\star}$ à $\tilde{E}$), les deux

2. La rédaction de ce chapitre, présentant une de nos contributions, a fortement été inspiré du document originel rédigé en anglais

personnes impliquées doivent partager un vocabulaire commun. Cela signifie que les idées partagées doivent être exprimées sur un ensemble de concepts partagés par les deux individus. Cela permet à la conversation d'être profitable pour eux. Pour les besoins de l'explication, le terme profitable signifie augmenter la quantité de connaissances relativement à $P$ de $\tilde{E}$ grâce à l'explication de $E^\star$. Pour l'explication, ces concepts sont présentés comme des éléments atomiques qui, lorsqu'ils sont soigneusement mélangés, permettent à l'explicateur de fournir une explication de $P$ à celui qui reçoit l'explication $\tilde{E}$. Cependant, ces briques élémentaires sont choisies en fonction de la quantité de connaissances de $E^\star$ et de $\tilde{E}$ qui dépendent également de $P$. En effet, si celui qui reçoit l'explication $\tilde{E}$ possède déjà un solide bagage concernant $P$, les connaissances de base permettant une compréhension superficielle de $P$ sont déjà acquises par $\tilde{E}$. Seuls des détails plus fins doivent être fournis par l'explicateur $E^\star$ à $\tilde{E}$ pour qu'il ait une compréhension totale de $P$. À l'inverse, si $\tilde{E}$ vient de commencer à s'intéresser à $P$, il doit assimiler les concepts grossiers relatifs à $P$ avant d'atteindre ceux plus fins, $E^\star$ devant adapter la complexité de son vocabulaire afin d'être compréhensible par $\tilde{E}$. Nous présentons ici la formulation de notre méthode explicative intégrant cette dimension sociale dans le calcul numérique de ces explications.

## 9.3 Protocole

**EiX-GNN** (**ei**gencentrality e**x**plainer for **g**raph **n**eural **n**etwork) est une méthode explicative locale post-hoc adaptée à tous les GNN adapté aux problèmes de classification de graphes. Dans notre terminologie, elle fournit ses explications en fonction d'un ensemble de concepts atomiques. Ces concepts sont pour les processus d'explication, ce que les pièces sont pour les échanges d'argent, c'est-à-dire qu'ils sont les parties élémentaires du processus d'explication que les explicateurs utilisent pour construire leurs explications. Ces concepts doivent être soigneusement choisis par l'explicateur afin de correspondre au niveau de connaissance a priori de celui qui reçoit l'explication sur le phénomène expliqué. En supposant que l'explicateur a une connaissance optimale d'un phénomène $P$, la sélection des concepts dépend des connaissances a priori de celui qui reçoit l'explication et du phénomène $P$ à

expliquer. EiX-GNN a été conçu pour intégrer cette dépendance à la connaissance a priori du celui qui reçoit l'explication en fonction d'un phénomène à expliquer. Formellement, nous présentons l'ensemble des concepts admissibles pour celui qui reçoit les explications comme un espace de probabilité $\mathcal{C}_p$ où les concepts sont des variables aléatoires à valeur dans $\mathcal{C}_p$. Le paramètre $p$ est la contrainte d'assimilation du concept de celui qui reçoit l'explication. C'est un paramètre réel compris tel que $p \in [0, 1]$ et est proportionnel à l'assimilabilité du concept de celui qui reçoit l'explication sachant $P$.

Dans ce qui suit, sauf mention contraire, nous considérons le phénomène $P = (f_{\boldsymbol{\theta}^\star}, G, Y)$ où $f_{\boldsymbol{\theta}^\star}$ est un modèle prédictif profond basé sur les GNN de profondeur $D \in \mathbb{N}$ qui a été entraîné sur un ensemble de données $\mathcal{D}$ auquel appartient $(G, Y)$. Conformément à la formulation des graphes que nous avons utilisée précédemment, nous considérons dans ce qui suit un graphe $G = (\mathbf{X}, \mathbf{A})$ composé de $N \in \mathbb{N}$ noeuds et $M \in \mathbb{N}$ arêtes. Nous supposons également que celui qui reçoit les explications a une contrainte d'assimilation du concept pour celui qui reçoit les explications de valeur $p \in [0, 1]$.

EiX-GNN fournit son explication sur la base d'un ordre local et global des concepts adaptés selon $p$. Nous présentons tout d'abord la procédure de génération de concepts, puis le processus d'ordonnancement de concept global qui est le coeur de notre processus d'explication. Enfin, nous présentons la procédure d'ordonnancement des concepts locaux, qui est une procédure d'affinage de l'explication qui permet de préciser au niveau des noeuds, l'explication.

### 9.3.1 Procédure de génération des concepts

Comme mentionné ci-dessus, les concepts sont des éléments atomiques qui permettent à l'explicateur de fournir son explication. Étant donné l'assimilabilité $p$ du concept pour celui qui reçoit l'explication, le concept $C_p$ est une variable aléatoire à valeur $\mathcal{C}_p$, le concept $C_p$ est une variable aléatoire à valeur $\mathcal{C}_p$. Cette variable est un sous-graphe de $G$ tel que $|C_p| = \lfloor |G| \times p \rfloor$. Notre motivation du point de vue du signal est de décrire une sous-partie du signal évoluant sur un sous-domaine de $G$. Nous présentons ici les motivations pour l'introduction du paramètre $p$ dans le cas de l'apprentissage profond géométrique. Dans de nombreuses représentations profondes

de données faites par des modèles répondant au paradigme de l'apprentissage profond géométrique, de nombreuses informations de bas niveau (par exemple, les fréquences élevées de l'image) sont rassemblées le long de la profondeur du modèle pour produire une information unifiée de haut niveau (par exemple, une distribution de probabilité des classes auxquelles cette image appartient). La probabilité des classes est compréhensible par toute personne intéressée par les approches d'apprentissage profond, tandis que la compréhension des hautes fréquences n'est possible que pour les personnes ayant des connaissances spécialisées dans le traitement des images. Nous avons conçu notre processus de génération de concepts en tenant compte de cette hiérarchisation de la représentation des données, allant d'une représentation détaillée compréhensible par les experts à une représentation plus communément compréhensible.

Après avoir déterminé la granularité d'explication souhaitée grâce à $p$, nous devons échantillonner ces concepts à partir du graphe initial $G$. Nous avons sélectionné des approches d'échantillonnage qui dépendent ou non d'une distribution préalable. L'échantillonnage de concepts est donc un processus d'échantillonnage de sous-graphes qui présente un aspect combinatoire inhérent à tout problème d'échantillonnage de sous-graphes naïf. Les concepts sont les éléments clés de notre approche et sont soigneusement sélectionnés puisqu'ils nous fournissent la matière première pour concevoir nos explications. De tous les $\binom{|G|}{|C_p|}$ sous-graphes possibles que nous pouvons dériver de $G$, certains sont mieux adaptés que d'autres pour l'explication de $P$ sachant $p$. L'hypothèse d'une distribution uniforme de la pertinence de l'explication de $P$ parmi tous ces sous-graphes possible n'est pas adaptée. Nous considérons plutôt une approche d'échantillonnage préférentiel qui quantifie la pertinence des noeuds conditionnellement à $P$.

Pour construire cette distribution, nous appliquons une approche quantifiant l'importance des noeuds par ablations de ces derniers ; elle évalue l'importance des noeuds au sein de leur voisinage par rapport à $P$. Formellement, pour un noeud voisin $v_j \in \mathcal{N}_i = \{v_j | (v_i, v_j) \in E\} \subset V$ de $v_i$. Pour quantifier l'importance de l'ablation d'un noeud, nous définissons la quantité $s : V^2 \to \mathbb{R}^+$ qui mesure l'effet de perturbation ablative relative entre deux noeuds par rapport à $P$ (par exemple, l'impact relatif de $f_{\boldsymbol{\theta}^\star}$ sur l'altération de la performance lors de la suppression de

$v_j$ dans $\mathcal{N}_i$). En supposant une distribution uniforme de la pertinence $\mathbb{U}_{|\mathcal{N}_i|}$ [3], où les noeuds composant $\mathcal{N}_i$, nous définissons la distribution de pertinence a priori $\alpha_P$ du noeud $v_i$ conditionnellement à $P$ par :

$$\alpha_P(v_i) = \mathop{\mathbb{E}}_{v_j \sim \mathbb{U}_{|\mathcal{N}_i|}} [s(v_i, v_j)|v_i, P] \tag{9.1}$$

Avec une constante de normalisation $F \in \mathbb{R}^*$ telle que $F^{-1} \sum_{v_i \in V} \alpha_P(v_i) = 1$, nous obtenons une distribution de probabilité de l'importance des noeuds a priori qui permet un processus d'échantillonnage plus efficace pour déterminer les concepts pertinents par rapport à $P$. Une fois cette distribution préalable déterminée, nous échantillonnons de manière i.I d. $L \in \mathbb{N}$ réalisations de $C_p$ que nous désignons par $(C_i)_{i \in \{1,\dots,L\}}$ où chaque noeud composant le sous-graphe $C_i$ a été échantillonné grâce à la distribution préalable des noeuds. Ensuite, nous présenterons la procédure de hiérarchisation de ces $L$ concepts par rapport à $P$.

### 9.3.2 Procédure d'ordonnancement globale des concepts

Une fois les concepts échantillonnés, nous devons trouver une relation d'ordre afin de classer leur pertinence selon $P$. Grâce à l'approche de l'échantillonnage de l'importance des noeuds, nous avons déjà établi une telle hiérarchisation, mais parmi tous les sous-graphes possibles de $G$ de taille $|C_p|$ ce qui réduit considérablement le périmètre de périmètre de recherche de la sous-structure optimale qui expliquera $P$. Au lieu de cela, nous présentons une méthode d'ordonnancement qui hiérarchise deux à deux les concepts parmi les $L$ concepts échantillonnés.

En considérant ces $L$ concepts, nous construisons un arbre de recherche avec $G$ comme racine et ces $L$ concepts comme feuilles. En l'absence d'autres travaux, nous ne savons pas encore si un concept $C_i$ est plus pertinent qu'un autre concept $C_j$ pour expliquer $P$. afin de construire cet ordre, nous dérivons de l'échantillon, un graphe complet $K_L$ où chaque noeud représente un concept et chaque poids d'arête de $K_L$ représente l'ordre relatif des concepts.

3. Notez que nous pouvons étendre l'utilisation de notre méthode explicative aux tâches de classification de noeuds. Pour expliquer le noeud $n$, nous pouvons considérer que l'échantillon est plutôt donné $\mathbb{U}_{|\mathcal{N}_i \cap Hop(n)_L|}$ où $Hop(n)_D$ est le saut $D$ du noeud $n$.

**Similarité relative des domaines des concepts**

Nous définissons la similarité de domaine entre deux concepts $C_i, C_j \in (C_l)_{l\in\{1,\cdots,L\}}$ comme la densité relative des arêtes entre $C_i$ et $C_j$. La densité des arêtes d'un concept $C_i$, notée $d(C_i)$, est le rapport entre les arêtes réelles composant $C_i$ sur le nombre total d'arêtes possibles dont $C_i$ peut être composé. Pour un graphe $G = (\mathbf{X}, \mathbf{A})$ avec $N$ noeuds et $M$ arêtes, il est défini comme suit :

$$d(G) = \frac{(2 \times \mathbf{1}_{\{\mathbf{A}=\mathbf{A}^T\}} + \mathbf{1}_{\{\mathbf{A}\neq\mathbf{A}^T\}})M}{N(N-1)} \tag{9.2}$$

Elle mesure comment $C_i$ tend à être un graphe complet. Nous choisissons cette mesure en raison de l'opération d'agrégation locale définie dans les GNN. Empiriquement les sous-graphes complets sont plus représentatifs du signal local que les sous-graphes plus épars. Certes, l'agrégation de nombreux signaux voisins n'implique pas d'agréger plus d'informations pertinentes que dans des structures plus éparses. Néanmoins, cela produira statistiquement une estimation plus juste de la pertinence de l'information locale sachant $P$ que cela peut être fait sur des localités plus éparses.

**Similarité relative des signaux des concepts**

La similarité entre les signaux des concepts quantifie le degré de similarité entre les comportements de $f_{\boldsymbol{\theta}^\star}$ entre les concepts considérés relativement à $P$. Nous considérions deux concepts $C_i$ et $C_j$, le cas où $C_i$ est similaire à $C_j$ étant donné $P$ signifie que $f_{\boldsymbol{\theta}^\star}$ voit de manière équivalente $C_i$ et $C_j$. Ainsi, considérer $C_i$ n'apporte pas de valeur ajoutée par rapport à ne considérer uniquement $C_j$ lui-même, par rapport à $P$. En tant que métrique de similarité entre deux concepts $C_i$ et $C_j$, nous utilisons la divergence de Kullbach-Liebler entre les deux distributions de probabilité déduites de $C_i$ et $C_j$ selon $f_{\boldsymbol{\theta}^\star}$. Formellement, nous définissons $s_{f_{\boldsymbol{\theta}^\star}}(C_i, C_j) : \mathcal{C}_p \to \mathbb{R}^+$ pour $f_{\boldsymbol{\theta}^\star}$ comme étant la mesure de similarité selon les signaux de $C_i$ et $C_j$ par :

$$s_{f_{\boldsymbol{\theta}^\star}}(C_i, C_j) = \frac{1}{2}[D_{KL}(f_{\boldsymbol{\theta}^\star}(C_i)||f_{\boldsymbol{\theta}^\star}(C_j)) + D_{KL}(f_{\boldsymbol{\theta}^\star}(C_j)||f_{\boldsymbol{\theta}^\star}(C_i))] \tag{9.3}$$

Où $D_{KL}(\cdot\|\cdot)$ désigne la divergence de Kullbach-Liebler. On remarquera que $s$ est bien une application symétrique. Cette métrique est largement utilisée dans les problèmes d'apprentissage automatique. Elle a fait l'objet d'études approfondies dans diverses applications, en particulier pour les problèmes de classification en

profondeur. Dans ces problèmes, les données sont représentées sous forme de lois de probabilité, et la divergence de Kullbach-Liebler est utilisée dans ce contexte pour quantifier la similarité entre les lois de probabilité inférées et les lois de probabilité issue de la vérité terrain.

**Unification des similarités entre signaux et domaines**

Nous présentons ici notre processus d'unification de ces deux modalités afin d'obtenir un classement global des concepts. Pour chaque modalité, nous obtenons les valeurs quantifiant la pertinence marginale relative de chacun des concepts. Étant donné un concept, nous obtenons la pertinence jointe relative (c'est-à-dire la pertinence globale relative du concept) en multipliant chaque valeur marginale entre elles(c'est-à-dire la valeur relative de la pertinence du domaine du concept et la similarité relative du signal du concept). Plus formellement, le calcul de la pertinence relative du concept d'ordre global entre le concept $C_i$ et $C_j$ est la valeur réelle $a_{i,j}$ définis par :

$$a_{i,j} = \frac{d(C_i)}{d(C_j)} \times s_{f_{\boldsymbol{\theta}^\star}}(C_i, C_j), \forall i, j \in \{1, \cdots, L\}^2 \tag{9.4}$$

Désormais, au lieu de considérer $L \times L$ valeurs d'ordre local relatif, nous voulons hiérarchiser globalement ces $L$ concepts avec une stratégie d'ordonnancement des concepts globale. Ainsi, nous considérons le graphe $K_L$ composé de $L$ noeud qui représente des concepts avec une matrice d'adjacence $\mathbf{A}(K_L)$ dont les composantes sont données par la famille $(a_{i,j})_{i,j}$. Cependant, bien que nous ayons obtenu la plupart des $L$ concepts pertinents grâce aux approches précédentes, nous recherchons, à l'échelle globale, l'ordonnancement par paire des concepts les plus dissemblables selon notre mesure de similarité globale. En effet, du point de vue de l'explication, deux concepts très similaires $C_i$ et $C_j$ (c'est-à-dire que $a_{i,j}$ est faible) apportent des informations similaires concernant l'explication. En d'autres termes, ils produisent des informations redondantes qui ne sont pas nécessaires et qui peuvent perturber et altérer la compréhension du phénomène $P$ par l'utilisateur. Les valeurs les plus élevées de $\mathbf{A}(K_L)$ correspondent aux concepts les moins redondants (relativement) en ce qui concerne l'explication de $P$. Mais quel est le concept qui est à la fois pertinent et moins redondant parmi les concepts candidats ? Cette question peut être reformulée dans le cadre de la théorie des graphes en déterminant quel noeud

a la centralité normalisée la plus élevée. Une bonne approche consiste à calculer le PageRank [Brin and Page, 1998] de chaque noeud de $K_L$. Une fois obtenu, il donne une relation d'ordre totale entre les noeuds de $K_L$ (c'est-à-dire les concepts). Formellement, nous considérons :

$$\widehat{\mathbf{A}(K_L)} = \mathbf{\Lambda}(\mathbf{A}(K_L)\mathbf{e})^{-1}\mathbf{A}(K_L) \tag{9.5}$$

Qui est la version normalisée de $\mathbf{A}(K_L)$ où $\mathbf{e}$ le vecteur unitaire de taille $L$ et, pour tout vecteur $\mathbf{x}$ de taille $L$, $\mathbf{\Lambda}(\mathbf{x})$ désigne la matrice diagonale contenant $\mathbf{x}$ dans sa diagonale. Dans ce contexte, ce vecteur propre $\mathbf{r}$ définit une loi de probabilité et ses composantes sont les valeurs *PageRank* de chaque noeud. En ce qui concerne le processus d'explication, les mesures de centralité *PageRank* produisent un schéma d'ordonnancement global des concepts dans lequel le candidat concept ayant la valeur *PageRank* la plus élevée est le représentant explicatif qui propose le moins de redondance d'informations tout en étant pertinent dans les deux modalités. Cette procédure d'ordonnancement global des concepts permet de réduire considérablement l'espace de recherche pour trouver des sous-graphes pertinents, mais elle reste grossière. Une fois ce raffinement effectué, nous proposons une approche plus fine qui évalue la pertinence à l'échelle des noeuds de $G$ que nous définissons comme la procédure d'ordonnancement de concepts locaux.

### 9.3.3 Procédure d'ordonnancement locale des concepts

Le fait de ne considérer la pertinence des noeuds qu'au niveau des sous-graphes (c-à-d les noeuds d'un sous graphe ont tous la même importance) peut conduire à une formulation incomplète et moins fine des explications. En effet, bien que de manière sous-jacente, la pertinence des noeuds soit déjà partiellement encodée dans les processus d'échantillonnage des concepts, les noeuds composant ces sous-graphes peuvent eux-mêmes avoir leur propre rôle sur les résultats globaux de l'ordonnancement des concepts, c'est-à-dire qu'étant donné un concept, le noeud n'a pas forcément une contribution uniforme. De plus, dans de nombreuses applications réelles, les noeuds peuvent représenter des atomes pour les représentations de molécules ou des villes sur une carte routière pour les prévisions de trafic et ont donc leurs propres enchâssements sémantiques qui peuvent ne pas être rendus par une focali-

sation au niveau d'un seul sous-graphe. C'est pourquoi ces données au niveau des noeuds doivent être incluses dans notre processus de conception d'explications. Pour quantifier soigneusement la contribution de chaque noeud au sein d'un concept de candidat $C_i$, nous avons exploité la théorie des jeux. Il s'agit de calculer la valeur *Shapley* [Shapley, 1951] de chaque noeud $i$ composant $C_i$. La valeur de Shapley est issue de la théorie des jeux coopératifs qui quantifie l'importance du rôle marginal d'un joueur dans l'issue d'un jeu. Si l'on considère une coalition de $K \in \mathbb{N}$ joueurs indexés dans $Q = \{1, \cdots, K\}$ jouant un jeu avec un gain de jeu $v : \mathcal{P}(Q) \rightarrow \mathbb{R}$ où $\mathcal{P}(Q)$ désigne tous les sous-ensembles possibles de $Q$. La valeur de Shapley d'un joueur $i \in Q$ est définie par :

$$\gamma_Q(i) = K \mathop{\mathbb{E}}_{j \sim \mathbb{U}_K} \left[ \mathop{\mathbb{E}}_{S \subset Q \setminus \{i\}} [v(S \cup \{i\}) - v(S)) \mid |S| = j] \right] \tag{9.6}$$

Nous notons par $\gamma_j(i)$ la valeur *Shapley* du noeud $i$ du concept $C_j$.

### 9.3.4 Procédure de fusion de l'information globale et locale

Dans notre contexte, étant donné un noeud $i$ qui appartient à un concept $C_j$, le calcul de la valeur *Shapley* de $i$ nécessite de considérer tous les sous-graphes possibles de $C_j$ et de calculer, en fonction d'eux, les effets perturbateurs (c-à-d ablation) de $i$ concernant $f_{\boldsymbol{\theta}^\star}$ à l'échelle de $C_j$. Numériquement, $\gamma_j(i)$ fournit une valeur précise de la pertinence du concept du noeud $i$ appartenant à $C_j$ par rapport à $P$ ( [Duval and Malliaros, 2021, Yuan et al., 2021]). Notons que cette valeur $\gamma_j(i)$ reste dépendante de $C_j$. D'un point de vue informatique, l'évaluation nécessite $\mathcal{O}(2^{\lfloor |G| \times p \rfloor})$ inférences de $f_{\boldsymbol{\theta}^\star}$, ce qui peut être intensif, voire intraitable dans la pratique, et dépend par construction de la contrainte d'assimilabilité $p$ de celui qui reçoit l'explication. Pour surmonter ce problème, nous pouvons estimer chaque $\gamma_j(i)$ par une stratégie d'estimation type Monte-Carlo . Ces calculs produisent un ensemble de $L$ évaluation explicative au niveau des noeuds $(\boldsymbol{\gamma}_j)_j \subset \mathbb{R}^{\lfloor |G| \times p \rfloor}$ où chaque $\boldsymbol{\gamma}_j$ est normalisé par sa norme $l_1$. Nous étendons ensuite chaque $\boldsymbol{\gamma}_j$ à $\boldsymbol{\gamma}_j^{ext} \in \mathbb{R}^N$, de sorte que pour le noeud $i$, $\boldsymbol{\gamma}_j^{ext}[i] = \boldsymbol{\gamma}_j[i]$ si $i \in C_j$, 0 sinon. Et nous concaténons les colonnes de chaque $\boldsymbol{\gamma}_j^{ext}$ pour obtenir $\boldsymbol{\Gamma}_p \in \mathbb{R}^{N \times L}$. Enfin, notre carte d'explication $\texttt{EiX-GNN}_{L,p}(P)$ du phénomène $P$ avec une contrainte d'assimilabilité $p$ est définie comme suit :

$$\texttt{EiX-GNN}_{L,p}(P) = \boldsymbol{\Gamma}_p \boldsymbol{\Lambda}(\mathbf{r}) \mathbf{e}^T \tag{9.7}$$

La carte d'explication $\texttt{EiX-GNN}_{L,p}(P) \in \mathbb{R}^N$ décrit la pertinence de chaque descripteur décrivant le phénomène $P$ par rapport à $p$. Dans le contexte de la classification de graphes , elle décrit la pertinence normalisée de chaque noeud composant $G$ en ce qui concerne $f_{\boldsymbol{\theta}^\star}$ avec une contrainte d'assimilabilité $p$. Nous menons à présent une étude quantitative et qualitative sur des applications réelles et fournissons une étude d'impact concernant le concept d'assimilabilité sur les explications produit par notre méthode explicative.

## 9.4 Résultats

### 9.4.1 Protocole expérimental

#### Ensembles de données

Pour évaluer notre méthode, nous avons utilisé quatre ensembles de données du monde réel qui sont constitués de caractéristiques intelligibles par l'homme : MNIST-Superpixels [Monti et al., 2017], PROTEINS [Borgwardt et al., 2005], MSRC [Shotton et al., 2009,Winn et al., 2005], REDDIT-BINARY [Yanardag and Vishwanathan, 2015]. Ces ensembles de données sont largement utilisés dans la littérature pour illustrer GNN et explainer. Chacun des ensembles de données suivants est adapté aux problèmes de classification de graphes. Il convient de noter que REDDIT-BINARY ne nécessite aucune connaissance préalable pour évaluer la pertinence d'une explication et que la vérité terrain est facile déterminer, comme indiqué dans [Ying et al., 2019c], contrairement à PROTEINS qui nécessite des connaissances préalables en chimie. Dans ce qui suit, nous trouvons des détails concernant les ensembles de données utilisés :

**MNISTSuperpixel** [Monti et al., 2017] est un ensemble de données composé de 60000 graphiques, qui représentent chacun une version superpixel du célèbre ensemble de données MNIST [Lecun et al., 1998]. Chaque instance de MNISTSuperpixels est une représentation sous forme de graphe de l'instance d'image MNIST originale. Deux sommets sont liés en fonction de leur proximité spatiale.

**PROTEINS** [Borgwardt et al., 2005] est un ensemble de données comprenant 1113 graphes étiquetés. Chaque graphique représente une protéine classée comme enzyme ou non-enzyme. Les noeuds représentent les acides aminés et deux noeuds sont connectés s'ils partagent la même localité spatiale.

**MSRC** [Shotton et al., 2009, Winn et al., 2005] Les ensembles de données sont utilisés dans les problèmes de segmentation sémantique des images. Chaque image est convertie en une version sémantique de superpixels. Dans MSRC-9, qui est composé de 221 graphes étiquetés, les étiquettes sémantiques sont réparties entre 8 étiquettes sémantiques. Dans la version MSRC-21, composée de 563 graphes étiquetés, le nombre d'étiquettes sémantiques possibles est porté à 21.

**REDDIT-BINARY** [Yanardag and Vishwanathan, 2015] est un ensemble de données composé de 2000 graphes où chacun d'entre eux représente un fil de discussion de Reddit basé sur des questions/réponses, à savoir *r/IAMA* et *r/AskReddit*. Dans ces graphes, les noeuds représentent les utilisateurs et il existe un lien entre deux utilisateurs si l'un a répondu à l'autre.

### Protocole d'optimisation du modèle prédictif

nous avons utilisé deux configurations GNN principales pour classer nos instances. Qu'elles soient basées sur les modules GCN ou GAT, elles produisent étonnamment toutes deux des résultats similaires en termes de précision des tests. Mais le modèle basé sur GCN est moins paramétré que celui basé sur GAT, c'est pourquoi nous avons choisi les modèles GCN. Nos modèles prédictifs sont construits de la façon suivante : nous avons enchaîné deux modules GCN avec une couche de mutualisation moyenne globale. Un perceptron multicouche est ensuite utilisé pour la partie classifiant du modèle. Entre les couches, nous utilisons la fonction relue comme fonction d'activation. Pour l'ensemble de données MNISTSuperpixels, nous utilisons quatre modules de descripteurs enchaînés et tanh comme fonction d'activation. Toutes ces différentes implémentations utilisent l'optimiseur ADAM [Kingma and Ba, 2017] avec le même paramètre d'apprentissage égal à $10^{-4}$. Nos entraînements ont duré 100 époques. Dans ces conditions, nous avons obtenu pour chaque jeu de données un classificateur aussi précis que ceux utilisés dans les études comparatives de l'état

de l'art.

#### Mesure objective

L'évaluation de la qualité ou de la pertinence de l'explication d'un phénomène nécessite souvent l'approbation d'un spécialiste. Des méthodes objectives et ne nécessitant pas de contexte ont été proposées pour quantifier la pertinence des méthodes explicatives. Deux d'entre elles ont été largement utilisées dans la littérature [Duval and Malliaros, 2021, Pope et al., 2019a, Yuan et al., 2021], à savoir l'infidélité [Yeh et al., 2019b] et la rareté spatiale [Duval and Malliaros, 2021,Pope et al., 2019a,Yuan et al., 2021]. Nous les avons utilisés pour mener notre étude quantitative.

### 9.4.2 L'évaluation qualitative : cas applicatifs dans le monde réel

Pour illustrer notre méthode, nous avons choisi une configuration omnisciente pour notre méthode, c'est à dire : $L = 70$ permettant de tirer des explications complexes avec une large base d'argumentation et $p = 0.05$ pour se concentrer sur les détails les plus fins des données. Nous discutons ensuite de l'impact marginal de chacun de ces paramètres par rapport à la configuration omnisciente comme base de référence. Chaque instance de *REDDIT-BINARY* est une discussion impliquant des utilisateurs ayant des connaissances variées sur le sujet de la discussion. Certains utilisateurs ont une connaissance approfondie du sujet et peuvent être considérés comme des experts. L'explication de ces discussions, en termes d'interactivité avec les utilisateurs, consiste à rechercher ces experts, comme indiqué dans [Ying et al., 2019c].

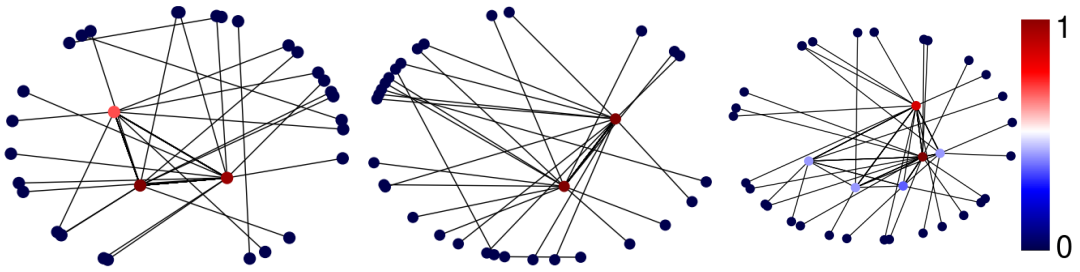
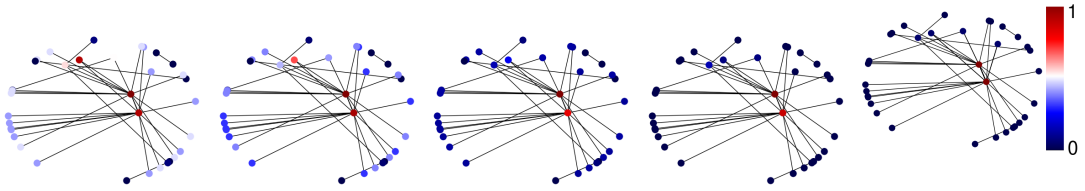



FIGURE 9.1 – Illustration de EiXGNN pour certaines des discussions de REDDIT-BINARY

Il s'avère que les experts sont en fait les utilisateurs qui répondent le plus à tous les autres utilisateurs. Dans la terminologie de la théorie des graphes, ces experts sont représentés par le noeud ayant le degré relatif le plus élevé. Dans le cadre de la configuration omnisciente, EiX-GNN met en évidence ces utilisateurs experts et les place dans un graphique avec une faible attribution aux utilisateurs ayant une faible interactivité (faible connaissance) et une forte attribution aux utilisateurs ayant une forte interactivité (forte connaissance), c'est-à-dire les experts (Figure 9.1). Ces résultats sont conformes à ceux obtenus dans [Ying et al., 2019c]. Nous mesurons maintenant l'impact marginal de chacun des $p$ et $L$ sur certaines explications de fils de discussion et nous comparons ces explications avec la ligne de base omnisciente, qui est considérée comme la limite supérieure pratique des explications de qualité qui nécessitent à la fois une grande compréhension et une grande connaissance (c'est-à-dire la récupération des informations les plus pertinentes, la localisation minutieuse des experts des fils de discussion). D'un point de vue social, nous cherchons à qualifier l'impact marginal de ces deux paramètres sur l'expressivité sociale des explications EiX-GNN. Ce que nous attendons, c'est une explication incomplète avec une base de concepts incomplète (c'est-à-dire faible $L$) et une explication grossière avec une contrainte d'assimilation des concepts de celui qui reçoit les explications importantes $p$, le tout au regard d'une explication complète et fine à travers la ligne de base omnisciente.

**Variation de $p$ : étude d'impact**




(a) $p = 0.7$ (b) $p = 0.3$ (c) $p = 0.2$ (d) $p = 0.1$ (e) Vue omnisciente

FIGURE 9.2 – Expressivité social de EiXGNN : variation de la contrainte d'assimilabilité des concepts selon celui qui reçoit les explications

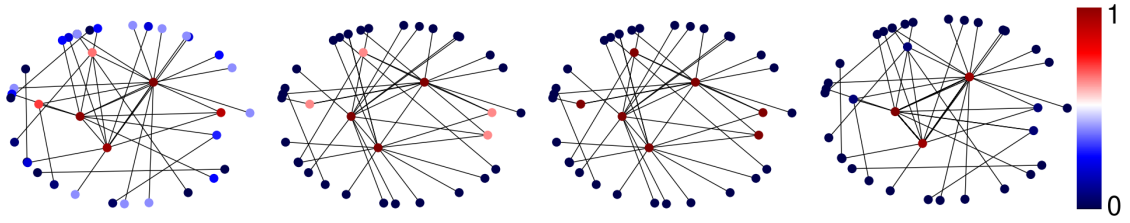
Dans cette séquence d'explications du même fil, nous avons fait varier la valeur de $p$ et nous avons fixé la valeur de $L$. Une faible valeur de $p$ correspond à une faible

contrainte d'assimilation des concepts de celui qui reçoit les explications, ce qui signifie que celui qui reçoit les explications est capable d'atteindre une compréhension plus fine du phénomène. En l'occurrence, il s'agit d'être en mesure de reconnaître précisément l'expert du fil de discussion dans les données de *REDDIT-BINARY*. Le schéma inverse apparaît avec une valeur élevée de la contrainte d'assimilabilité du concept pour celui qui reçoit 'explication. Nous observons que tant que nous abaissons la valeur de $p$, nous obtenons des informations (Figures 9.2c, 9.2d) concernant la précision de l'explication et nous augmentons progressivement la quantité de connaissances jusqu'à atteindre le régime omniscient (Figure 9.2e).

— $p = 0.7$ : l'explication est basée sur de grands concepts fournissant une connaissance grossière, les utilisateurs à interactivité unique ont un rôle assez important dans l'explication et les experts ont une valeur explicative plus élevée. Cette explication n'est pas la plus fine, mais permet à celui qui reçoit les explications d'avoir une vision imprécise mais globale du sujet. Cette explication est adaptée aux besoins de connaissances limitées (Figure 9.2a).
— $p = 0.3$ : nous observons ici que les informations de niveau spécialisé sont beaucoup plus mises en avant qu'auparavant. Les experts sont reconnus et les informations moins pertinentes sont largement écartées (utilisateurs à interaction unique qui ne sont pas des spécialistes) (Figure 9.2b).
— $p = 0.2$ : la tendance précédente a été accélérée, les connaissances spécialisées sont beaucoup plus mentionnées que les connaissances médiocres (Figure 9.2c).
— $p = 0.1$ : nous avons presque atteint le régime omniscient et nous avons acquis une compréhension comparable à celle des spécialistes (Figure 9.2d).

Globalement, nous observons que la valeur de $p$ se comporte comme un accordeur social vis-à-vis de l'explication. Une contrainte élevée fournit des informations de tendance générale, ces informations sont générales, mais imprécises. Elle donne une idée globale du phénomène sous-jacent. Elle permet donc aux non-spécialistes d'appréhender les explications du phénomène. Tant que la contrainte est élevée, nous tendons vers une compréhension experte du phénomène expliqué en n'incluant que les détails les plus fins et en écartant les entités qui ne sont capables que de fournir des généralités qui ne sont destinées qu'aux non-spécialistes.

**Variation de $L$ : étude d'impact**




(a) $L = 10$ (b) $L = 30$ (c) $L = 50$ (d) Vue omnisciente

FIGURE 9.3 – Expressivité social de EiXGNN : variation du nombre d'éléments dans la base de concepts

Nous observons que tant que le nombre de concepts *la* passe d'une faible valeur (Figure 9.3a) à une valeur élevée (Figure 9.3c), notre méthode explicative est en mesure de fournir des explications de plus en plus précises. Ainsi, plus la largeur de la base conceptuelle augmente, plus nous nous rapprochons d'une vision experte (Figure 9.3d). En fait, ce comportement est prévisible, car si la méthode est capable de fournir des explications sur une base large d'arguments, nous obtenons des explications précises et significatives.

### 9.4.3 L'évaluation quantitative : cas applicatifs dans le monde réel

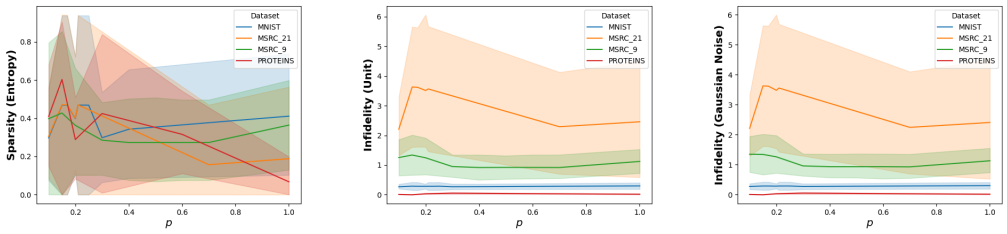
**Étude comparative numérique des mesures objectives**

Pour EiX-GNN, nous avons utilisé la configuration omnisciente présentée ci-dessus. Nous constatons que notre méthode est moins infidèle avec au moins un facteur $10^2$ sur MNISTSuperpixel et REDDIT-BINARY et MSRC-21 et un facteur 10 sur PROTEINS et MSRC-9. En outre, notre méthode surpasse les autres méthodes comparées en ce qui concerne la parcimonie des explications d'au moins un facteur $10^3$ sur MNISTSuperpixel et REDDIT-BINARY, d'un facteur $10^4$ sur MSRC-9, d'un facteur $10^2$ sur MSRC-21 et d'un facteur 10 sur PROTEINS. Nous présentons ici un tableau récapitulatif (Tableau 9.1) avec une version détaillée de ces mesures.

| Ensemble de données | Méthode explicative | Parcimonie (↓) | Infidélité (père. gaussienne) (↓) | Infidélité (père. unitaire) (↓) |
|---|---|---|---|---|
| MNISTSuperpixels | EiX-GNN | **9.41E-01** | **5.69E+00** | **5.69E+00** |
| | GNNExplainer | 1.30E+03 | 2.43E+05 | 2.44E+05 |
| | PGExplainer | 1.21E+03 | 1.80E+04 | 1.80E+04 |
| | SubgraphX | 1.70E+02 | 1.31E+03 | 1.31E+04 |
| PROTEINS | EiX-GNN | **9.37E-01** | **2.38E-01** | **2.78E-01** |
| | GNNExplainer | 5.21E+01 | 3.36E+02 | 3.47E+02 |
| | PGExplainer | 7.02E+01 | 8.23E+00 | 8.21E+00 |
| | SubgraphX | 1.40E+01 | 4.56E+01 | 4.56E+01 |
| MSRC-9 | EiX-GNN | **9.02E-01** | **2.29E-05** | **9.68E-05** |
| | GNNExplainer | 7.84E+01 | 1.14E+03 | 1.12E+03 |
| | PGExplainer | 8.69E+01 | 2.31E+03 | 2.29E+03 |
| | SubgraphX | 4.45E+01 | 2.69E+03 | 2.70E+03 |
| REDDIT-BINARY | EiX-GNN | **4.02E-01** | **2.64E-02** | **4.63E-01** |
| | GNNExplainer | 6.71E+01 | 5.17E+03 | 5.14E+03 |
| | PGExplainer | 5.61E+01 | 2.35E+02 | 2.36E+02 |
| | SubgraphX | 3.95E+01 | 1.11E+03 | 1.110E+03 |
| MSRC-21 | EiX-GNN | **8.54E-01** | **2.02E+00** | **2.05E+00** |
| | GNNExplainer | 2.79E+02 | 1.03E+04 | 1.04E+04 |
| | PGExplainer | 1.90E+05 | 8.74E+02 | 8.74E+02 |
| | SubgraphX | 4.82E+03 | 3.67E+04 | 3.67E+04 |

TABLE 9.1 – Tableau de comparaison entre EiX-GNN et les méthodes de l'état de l'art retenues par rapport aux métriques objectives sur les ensembles de données retenus pour l'étude

### Variation de $p$ : étude d'impact



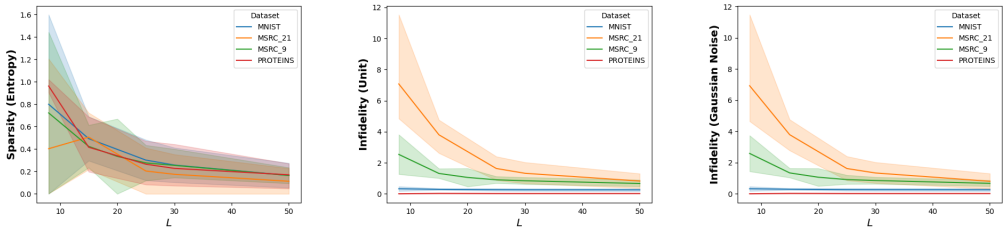

(a) Entropie de Von Neumann (b) Infidélité : perturbation unitaire (c) Infidélité : perturbation gaussienne

FIGURE 9.4 – Impact de $p$ par rapport aux métriques objectives

En ce qui concerne la valeur $p$, nous constatons qu'en moyenne, elle n'a pas d'impact sur l'infidélité de l'explication par rapport au classificateur, comme le montre

la Figure 9.4. En outre, l'explication au niveau du spécialiste est plus concise et donc intrinsèquement plus éparse, comme le montrent les Figures 9.2d et 9.3c. Cela signifie que la valeur de $p$ n'a pas d'impact sur la qualité de l'explication fournie par EiX-GNN et que EiX-GNN fournit toujours une explication pertinente relativement à la connaissance de celui qui reçoit l'explication.

**Variation de $L$ : étude d'impact**




(a) Entropie de Von Neumann (b) Infidélité : perturbation unitaire (c) Infidélité : perturbation gaussienne

FIGURE 9.5 – Impact de $L$ par rapport aux métriques objectives

En ce qui concerne la largeur de la base conceptuelle, nous retrouvons numériquement qu'une base argumentaire large favorise la production d'explications pertinentes qui sont susceptibles d'être, d'une part, moins infidèles et, d'autre part, plus concises, comme le montre la Figure 9.5. Dans cette étude, nous abordons ce problème avec EiX-GNN, une nouvelle approche qui intègre pleinement cette dépendance sociale à la connaissance a priori de celui qui reçoit l'explication. La notion d'assimilabilité permettant d'adapter le processus d'explication à ce dernier. Nous avons mené une étude qualitative concernant cet aspect social sur des données réelles et nous avons comparé, par rapport à des métriques objectives utilisées dans la littérature, les propriétés de fidélité et de compacité par rapport à des méthodes de la littérature. Dans les deux contextes, nous fournissons des résultats significatifs.

## 9.5 Conclusion

Il est courant de rencontrer des modèles d'apprentissage profond et en particulier des GNN pour aborder des problèmes académiques et industriels notamment dans

des contextes sensibles tels que la santé, la conduite autonome, etc. Leur puissance est souvent au détriment de l'inintelligibilité humaine des processus de décision que ces modèles conçoivent. Cela soulève de sérieuses questions quant à la sécurité du déploiement de ces modèles dans notre société. Ainsi, nous devons expliquer leur fonctionnement interne afin de mieux les comprendre et de les rendre dignes de confiance. Néanmoins, les explications ne sont rentables si et seulement si les explications sont adaptées à la personne qui les reçoit. Les méthodes explicatives de pointe fournissent souvent des explications absolues sans tenir compte des antécédents ou des attentes de la personne qui reçoit les explications. Cet aspect n'est quasiment jamais pris en compte numériquement dans notre littérature. Dans cette étude, nous abordons ce problème avec EiX-GNN, une nouvelle approche qui intègre pleinement cette dépendance à celui qui reçoit les explications, via la notion de contrainte d'assimilabilité de celui qui reçoit l'explication. Nous avons mené une étude qualitative concernant cet aspect social des explications, sur des données réelles et nous avons comparé, par rapport à des métriques objectives utilisées dans la littérature, les propriétés de fidélité et de parcimonie par rapport à des méthodes reconnues dans la littérature. Dans les deux contextes, nous fournissons des résultats significatifs du point de vue social et du point de vue purement numérique selon des mesures objectives utilisées dans la littérature.

## 9.6 Discussions, critiques, limites et ouvertures

notre méthode explicative a montré des propriétés intéressantes dans le cas de notre protocole d'évaluation, cependant dans sa globalité, notre étude admet elle aussi quelques limites, en voici quelques unes :

— La représentation de la contrainte d'assimilabilité : la représentation d'une quantité si haut niveau, complexe à définir, par un simple réel peut paraître assez légère. Cela peut être limitant dans l'optique de concevoir une méthode innovante prenant pleinement en compte ce facteur social dans sa définition. Cette méthode est simplement une première ébauche pour l'intégration de ces variables sociales, primordiales et trop souvent omises par l'état de l'art, dans le calcul des explications des modèles prédictifs profonds. Par ailleurs, nous

avons lié cette contrainte à la taille des concepts, c'est un choix naturel dans le contexte de l'apprentissage géométrique profond, mais d'autres choix auraient pu être faits avec des calculs plus complexes, intégrant certains aprioris socionumériques par exemple dans ces derniers.

— Forte dépendance au procédé d'échantillonnage des concepts : ce processus joue un rôle crucial dans le calcul des explications par notre méthode, car il fournit la matière première pour calculer ces explications. Le processus ablatif mis en place peut sembler lui aussi trop léger pour rendre compte de tous les effets ablatifs et leurs impacts. Bien que des calculs plus précis permettraient un meilleur échantillonnage , ils seraient sûrement plus coûteux (par exemple valeur de Shapley). C'est un point limitant pour notre méthode.

— Unification des similitudes domaine et signal : nous avons souhaité unifier les deux modalités à pied égal, souvent elles aussi omises dans la littérature, dans le calcul des explications pour les GNN. Cependant, le calcul effectué pour l'unification peut avoir certaines limites :

  — Décroissance en $x \mapsto x^{-1}$ : la comparaison relative des densités des concepts a une décroissance non linéaire (à numérateur fixé). Cela introduit un déséquilibre entre l'importance des densités si la densité d'un concept se trouve au numérateur ou au dénominateur, alors que ce comportement n'a a priori pas de sens d'un point de vue explicatif. Là aussi, cela peut sembler naturel dans la construction, mais c'est une réelle limite pour le calcul des explications.

  — produit des quantités : faire le produit entre les deux quantités, qui ont des amplitudes a priori différentes peut là aussi favoriser une modalité plus qu'une autre alors que d'un point de vue explicatif, cela ne semble pas justifié. Là aussi, cela peut sembler naturel dans la construction, mais c'est une réelle limite pour le calcul des explications.

— Coût de calcul : on considérant qu'une inférence du modèle prédictif évalué est en temps constant, le coût de calcul de la méthode varie en fonction de la contrainte d'assimilabilité de celui qui reçoit les explications et du nombre de concepts souhaités :

  — Des valeurs de centralité : pour un nombre de concepts $L \in \mathbb{N}$, le cal-

cul de centralité nécessite la diagonalisation d'une matrice de taille $L^2$, or l'algorithme le plus rapide pour diagonaliser une matrice à ce jour est l'algorithme QR qui est en $\mathcal{O}(L^3)$ ce qui peut devenir coûteux. Pour les calculs de centralité, des algorithmes d'approximation peuvent être utilisés, ce qui était d'ailleurs le cas de l'algorithme PageRank initialement. En pratique on observe que c'est peu limitant, mais dans certaines conditions, cela peut l'être, en effet il s'agit tout de même d'une complexité cubique.

  - Des valeurs de Shapley : de la même manière pour un concept à $N \in \mathbb{N}$ noeuds, le calcul de la valeur exacte de la valeur de Shapley nécessite $\mathcal{O}(2^N)$ calcul. Cette quantité est donc très vite incalculable à partir d'un certain rang pour $N$. De plus, ce calcul coûteux est répété pour chaque noeud de chaque concept, donc dépendant globalement de la contrainte d'assimilabilité et du nombre de concepts. Une approximation de cette valeur est donc nécessaire, c'est l'objet de l'approximation par méthode type Monte-Carlo, cela reste une approximation et peut donc être limitant pour le pouvoir explicatif de notre méthode.
- Mélange linéaire final : un mélange linéaire des calculs intermédiaire (c-à-d valeur de centralité et valeur de Shapley des noeuds des concepts) peut paraître assez pauvre pour rendre compte de toutes les interactions entre les objets. C'est un probablement un point limitant.
- Protocole d'évaluation : comme le protocole d'évaluation est proche de celui de l'étude précédente, celui utilisé dans cette étude subit donc les mêmes défauts :
  - supposition sur l'optimalité du modèle prédictif
  - absence de vérité terrain pour l'évaluation de la méthode
  - biais de confirmation dans les évaluations subjectives
- Pas d'étude de la méthode pour des problèmes de classification de noeuds : pour couvrir la plupart des cas applicatifs des GNN, il aurait fallu mener une étude de l'adaptation de la méthode, que nous avons détaillée, pour expliquer les cas relatifs à la classification de noeuds. Cela aurait pu corroborer nos actuelles conclusions.

Ainsi, cette étude propose une méthode pertinente pour l'explicabilité des réseaux de neurones sur graphes à la communauté via des arguments quantitatifs et qualitatifs. Cette étude, ainsi que son implémentation, est disponible publiquement en ligne.

La publication est donnée par :

A. Raison, P.Bourdon and D.Helbert, *"EiX-GNN : Concept-level eigencentrality explainer for graph neural networks"* 2022, arXiv preprint arXiv :2206.03491

## 9.7 Inscription de la contribution avec le projet de recherche et la littérature

Nous présentons ici les points d'articulation de cette contribution dans l'ensemble des travaux de cette thèse et l'état de l'art :

- **Inscription de la contribution dans la littérature existante :** Comme évoqué précédemment, les apports de cette contribution sont théoriques avec une forte dominante numérico-sociale dans sa définition, un aspect trop souvent omis par la littérature. C'est là toute l'originalité de cette méthode. De plus, dans sa définition actuelle et par rapport au protocole usuel, notre méthode montre de sérieux résultats qualitativement et quantitativement.
- **Inscription de la contribution dans la problématique de recherche :** La méthode explicative développée est adaptée au paradigme de l'apprentissage géométrique profond. Elle n'est pas illustrée dans le cas de données médicales, mais peut l'être. Cette contribution s'inscrit donc dans notre problématique globale de recherche sur l'explicabilité en introduisant l'aspect social des explications au centre de sa définition.

# Chapitre 10

# Conclusion

Les modèles d'intelligence artificielle, aujourd'hui principalement représentés par les réseaux de neurones, jouent un rôle déterminant dans le traitement de l'information et donc dans le développement de nos sociétés. Leurs utilisations de plus en plus massives, dans des domaines parfois sensibles où leurs impacts peuvent être importants sur la vie humaine et la pérennité de la société. Pour limiter ces potentiels impacts négatifs, des méthodes ont été mises en place. Elles permettent d'expliquer les comportements des modèles d'intelligence artificielle et donc de juger de la pertinence de leurs utilisations dans certains contextes.

Dans les premiers travaux, nous avons illustré empiriquement que, selon la méthode explicative ScoreCAM, les réseaux de neurones convolutifs peuvent être utilisés dans des problématiques de neurologie.

Nos seconds travaux, nous avons généralisé la méthode ScoreCAM aux réseaux de neurones sur graphes. Cette généralisation a été permise par le fait que les réseaux de neurones convolutifs sont un cas particulier des réseaux de neurones sur graphes d'après la théorie de l'apprentissage géométrique profond. Nous avons défini et implémenté cette nouvelle méthode et nous avons illustré sa pertinence par rapport à un protocole d'évaluation fréquemment utilisé dans la littérature.

Dans nos troisièmes travaux, nous avons illustré à nouveau la pertinence de la généralisation de la méthode ScoreCAM dans un contexte de caractérisation des émotions faciales par des réseaux de neurones sur graphes.

Les quatrièmes travaux, nous avons proposé une nouvelle méthode explicative pour réseaux de neurones sur graphes. Cette méthode a deux avantages : elle est

utilisable par une large gamme de réseaux de neurones sur graphes et inclue dans sa définition un paramètre proposant l'ajustement du niveau de granularité utilisé pour exprimer une explication. Ainsi, cette méthode permet d'expliquer un seul et même modèle à un large panel d'utilisateur ayant des connaissances apriori sur la tâche d'apprentissage du modèle expliqué ou non.

L'explicabilité et donc la pertinence des méthodes proposées dans la littérature ont une forte dépendance sociale et humaine qui est trop souvent omise dû faite notamment de la forte et complexe variabilité des notions dans ces domaines.

Nous avons proposé d'intégrer cette dépendance humaine dans la définition de la méthode explicative EiXGNN, et c'est l'une des perspectives à court terme pour nos travaux : l'intégration profonde des variables humaines dans les formulations numériques des méthodes explicatives. En parallèle de ces travaux, nous nous sommes intéressés au lien entre les méthodes spectrales sur graphes et les méthodes explicatives pour réseaux de neurones sur graphes. En effet, nous avons constaté que sous certaines conditions, les approches spectrales et les méthodes explicatives avaient des comportements analogues. Ainsi le formalisme déjà existant pour les approches spectrales pourrait compenser le manque de formalisme des méthodes explicatives dû à leurs dépendances aux aspects sociaux et ainsi permettre de développer de nouvelles méthodes explicatives pour les réseaux de neurones sur graphes.

À plus long terme, les efforts dans la recherche en explicabilité devraient être redistribués dans la compréhension a priori des phénomènes physiques que les modèles profonds modélisent aujourd'hui. En effet, le premier rempart pour éviter de se confronter au manque d'explicabilité des modèles profonds est de mieux comprendre les phénomènes physiques à modéliser et ainsi ne pas laisser un algorithme construire la connaissance à la place de l'humain. Cependant, les modèles profonds sont vus comme des réponses universelles à une multitude de problématiques, pour diverses raisons économiques ou sociétales. Leurs utilisations intensives en font des objets d'étude prioritaire pour une grande partie des scientifiques, mais c'est plutôt leurs conditions d'utilisation qui devraient être plus largement investiguées ainsi que le rôle qu'ils jouent aujourd'hui dans la société. En effet, s'ils sont utilisés à mauvais escient ou de la mauvaise façon alors ils risquent fortement de mal se comporter. Nous avons aujourd'hui quelques illustrations concrètes concernant ces mauvaises

utilisations. Or ces modèles profonds ne seront pas la cause des effets de bords qu'ils peuvent engendrer, ce sera leur mauvaise utilisation qui sera en cause. Ainsi, pour protéger la société, il est alors probablement plus efficace de mener des réflexions sur les contextes d'utilisation de ces modèles plutôt que sur ces modèles eux-mêmes.

# Bibliographie

[Com, a] The Compendious Book on Calculation by Completion and Balancing. https ://www.loc.gov/item/2021666184/.

[Com, b] Comprehensible classification models : A position paper : ACM SIGKDD Explorations Newsletter : Vol 15, No 1. *ACM SIGKDD Explorations Newsletter*.

[Dua, ] Dual Graph Convolutional Networks for Graph-Based Semi-Supervised Classification | Proceedings of the 2018 World Wide Web Conference. https ://dl.acm.org/doi/10.1145/3178876.3186116.

[Gen, ] General Data Protection Regulation (GDPR) – Official Legal Text. https ://gdpr-info.eu/.

[Gra, ] Graph Convolutional Networks with EigenPooling | Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. https ://dl.acm.org/doi/10.1145/3292500.3330982.

[Int, ] Interpretable Chirality-Aware Graph Neural Network for Quantitative Structure Activity Relationship Modeling.

[Myt, ] The Mythos of Model Interpretability : In machine learning, the concept of interpretability is both important and slippery. : Queue : Vol 16, No 3. *Queue*.

[Inc, 2015] (2015). Inceptionism : Going Deeper into Neural Networks. https ://blog.research.google/2015/06/inceptionism-going-deeper-into-neural.html.

[Def, 2023] (2023). Definition of INTERPRET. https ://www.merriam-webster.com/dictionary/interpret.

[A et al., 2010] A, A.-E., T, S., J, R., J, N., O, T., and V, K. (2010). The effect of model order selection in group PICA. *Human brain mapping*, 31(8).

[Abadal et al., 2021] Abadal, S., Jain, A., Guirado, R., López-Alonso, J., and Alarcón, E. (2021). Computing Graph Neural Networks : A Survey from Algorithms to Accelerators.

[Abadal et al., 2022] Abadal, S., Jain, A., Guirado, R., López-Alonso, J., and Alarcón, E. (2022). Computing Graph Neural Networks : A Survey from Algorithms to Accelerators. *ACM Computing Surveys*, 54(9) :1–38.

[Abu-El-Haija et al., 2019] Abu-El-Haija, S., Perozzi, B., Kapoor, A., Alipourfard, N., Lerman, K., Harutyunyan, H., Steeg, G. V., and Galstyan, A. (2019). MixHop : Higher-Order Graph Convolutional Architectures via Sparsified Neighborhood Mixing. In *Proceedings of the 36th International Conference on Machine Learning*, pages 21–29. PMLR.

[Adadi and Berrada, 2018] Adadi, A. and Berrada, M. (2018). Peeking Inside the Black-Box : A Survey on Explainable Artificial Intelligence (XAI). *IEEE access : practical innovations, open solutions*, 6 :52138–52160.

[Adamic and Glance, 2005] Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 U.S. election : Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, pages 36–43, Chicago Illinois. ACM.

[Adebayo et al., ] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity Checks for Saliency Maps.

[Adebayo et al., 2020] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2020). Sanity Checks for Saliency Maps. Technical Report arXiv :1810.03292, arXiv.

[Adler et al., 2016] Adler, P., Falk, C., Friedler, S. A., Rybeck, G., Scheidegger, C., Smith, B., and Venkatasubramanian, S. (2016). Auditing Black-box Models for Indirect Influence.

[Aflalo et al., 2022] Aflalo, E., Du, M., Tseng, S.-Y., Liu, Y., Wu, C., Duan, N., and Lal, V. (2022). VL-InterpreT : An Interactive Visualization Tool for Interpreting Vision-Language Transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21374–21383, New Orleans, LA, USA. IEEE.

[Agarwal et al., 2022] Agarwal, A., Tan, Y. S., Ronen, O., Singh, C., and Yu, B. (2022). Hierarchical Shrinkage : Improving the accuracy and interpretability of tree-based methods. Technical Report arXiv :2202.00858, arXiv.

[Agarwal et al., 2023a] Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., and Lakkaraju, H. (2023a). OpenXAI : Towards a Transparent Evaluation of Model Explanations. Technical Report arXiv :2206.11104, arXiv.

[Agarwal et al., 2023b] Agarwal, C., Queen, O., Lakkaraju, H., and Zitnik, M. (2023b). Evaluating Explainability for Graph Neural Networks. Technical Report arXiv :2208.09339, arXiv.

[Agarwal et al., 2021] Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., and Hinton, G. (2021). Neural Additive Models : Interpretable Machine Learning with Neural Nets. Technical Report arXiv :2004.13912, arXiv.

[Aggarwal et al., 2010] Aggarwal, C. C., Chen, C., and Han, J. (2010). The Inverse Classification Problem. *Journal of Computer Science and Technology*, 25(3) :458–468.

[Aïvodji et al., 2020] Aïvodji, U., Bolot, A., and Gambs, S. (2020). Model extraction from counterfactual explanations. Technical Report arXiv :2009.01884, arXiv.

[Akula et al., 2020] Akula, A., Wang, S., and Zhu, S.-C. (2020). CoCoX : Generating Conceptual and Counterfactual Explanations via Fault-Lines. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03) :2594–2601.

[Albini et al., 2020] Albini, E., Rago, A., Baroni, P., and Toni, F. (2020). Relation-Based Counterfactual Explanations for Bayesian Network Classifiers. volume 1, pages 451–457.

[Ali et al., 2022] Ali, A., Schnake, T., Eberle, O., Montavon, G., Müller, K.-R., and Wolf, L. (2022). XAI for Transformers : Better Explanations through Conservative Propagation. Technical Report arXiv :2202.07304, arXiv.

[Ali et al., 2023a] Ali, G., Al-Obeidat, F., Tubaishat, A., Zia, T., Ilyas, M., and Rocha, A. (2023a). Counterfactual explanation of Bayesian model uncertainty. *Neural Computing and Applications*, 35(11) :8027–8034.

[Ali et al., 2023b] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., and Herrera, F. (2023b). Explainable Artificial Intelligence (XAI) : What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99 :101805.

[Allen et al., 2011] Allen, E., Erhardt, E., Damaraju, E., Gruner, W., Segall, J., Silva, R., Havlicek, M., Rachakonda, S., Fries, J., Kalyanam, R., Michael, A., Caprihan, A., Turner, J., Eichele, T., Adelsheim, S., Bryan, A., Bustillo, J., Clark, V., Feldstein Ewing, S., Filbey, F., Ford, C., Hutchison, K., Jung, R., Kiehl, K., Kodituwakku, P., Komesu, Y., Mayer, A., Pearlson, G., Phillips, J., Sadek, J., Stevens, M., Teuscher, U., Thoma, R., and Calhoun, V. (2011). A Baseline for the Multivariate Comparison of Resting-State Networks. *Frontiers in Systems Neuroscience*, 5.

[Alon and Yahav, 2020] Alon, U. and Yahav, E. (2020). On the Bottleneck of Graph Neural Networks and its Practical Implications.

[Altmann et al., 2010] Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance : A corrected feature importance measure. *Bioinformatics (Oxford, England)*, 26(10) :1340–1347.

[Alvarez Melis and Jaakkola, 2018] Alvarez Melis, D. and Jaakkola, T. (2018). Towards Robust Interpretability with Self-Explaining Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

[Alvarez-Melis and Jaakkola, 2018a] Alvarez-Melis, D. and Jaakkola, T. S. (2018a). On the Robustness of Interpretability Methods.

[Alvarez-Melis and Jaakkola, 2018b] Alvarez-Melis, D. and Jaakkola, T. S. (2018b). On the Robustness of Interpretability Methods. Technical Report arXiv :1806.08049, arXiv.

[Amir and Amir, ] Amir, D. and Amir, O. HIGHLIGHTS : Summarizing Agent Behavior to People.

[Amodei et al., 2016] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete Problems in AI Safety.

[Amoukou and Brunel, ] Amoukou, S. I. and Brunel, N. J.-B. Consistent Sufficient Explanations and Minimal Local Rules for explaining any classifier or regressor.

[Amoukou et al., ] Amoukou, S. I., Salaun, T., and Brunel, N. J. B. Accurate Shapley Values for explaining tree-based models.

[Anandkumar et al., 2020] Anandkumar, A., Azizzadenesheli, K., Bhattacharya, K., Kovachki, N., Li, Z., Liu, B., and Stuart, A. (2020). Neural Operator : Graph Kernel Network for Partial Differential Equations.

[Ancona et al., 2019] Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2019). Gradient-Based Attribution Methods. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R., editors, *Explainable AI : Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pages 169–191. Springer International Publishing, Cham.

[Ancona et al., ] Ancona, M., Öztireli, C., and Gross, M. Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Values Approximation.

[Angelino et al., ] Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., and Rudin, C. Learning Certifiably Optimal Rule Lists for Categorical Data.

[Apley and Zhu, 2019] Apley, D. W. and Zhu, J. (2019). Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. Technical Report arXiv :1612.08468, arXiv.

[Arenas et al., ] Arenas, M., Barceló, P., Romero, M., and Subercaseaux, B. On Computing Probabilistic Explanations for Decision Trees.

[Arik and Pfister, 2020] Arik, S. O. and Pfister, T. (2020). TabNet : Attentive Interpretable Tabular Learning. Technical Report arXiv :1908.07442, arXiv.

[Armeni et al., 2016] Armeni, I., Sener, O., Zamir, A. R., Jiang, H., Brilakis, I., Fischer, M., and Savarese, S. (2016). 3D Semantic Parsing of Large-Scale Indoor Spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1534–1543, Las Vegas, NV, USA. IEEE.

[Arora et al., 2022] Arora, S., Pruthi, D., Sadeh, N., Cohen, W. W., Lipton, Z. C., and Neubig, G. (2022). Explain, Edit, and Understand : Rethinking User Study Design for Evaluating Model Explanations. Technical Report arXiv :2112.09669, arXiv.

[Arras et al., 2019] Arras, L., Arjona-Medina, J., Widrich, M., Montavon, G., Gillhofer, M., Müller, K.-R., Hochreiter, S., and Samek, W. (2019). Explaining and Interpreting LSTMs. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R., editors, *Explainable AI : Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pages 211–238. Springer International Publishing, Cham.

[Artelt and Hammer, 2020a] Artelt, A. and Hammer, B. (2020a). Convex Density Constraints for Computing Plausible Counterfactual Explanations. Technical Report arXiv :2002.04862, arXiv.

[Artelt and Hammer, 2020b] Artelt, A. and Hammer, B. (2020b). Efficient computation of counterfactual explanations of LVQ models. Technical Report arXiv :1908.00735, arXiv.

[Artelt et al., 2021] Artelt, A., Vaquet, V., Velioglu, R., Hinder, F., Brinkrolf, J., Schilling, M., and Hammer, B. (2021). Evaluating Robustness of Counterfactual Explanations. Technical Report arXiv :2103.02354, arXiv.

[Asher et al., 2022] Asher, N., De Lara, L., Paul, S., and Russell, C. (2022). Counterfactual Models for Fair and Adequate Explanations. *Machine Learning and Knowledge Extraction*, 4(2) :316–349.

[Atwood and Towsley, 2016] Atwood, J. and Towsley, D. (2016). Diffusion-Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

[Augustin et al., ] Augustin, M., Boreiko, V., Croce, F., and Hein, M. Diffusion Visual Counterfactual Explanations.

[Auten et al., 2020] Auten, A., Tomei, M., and Kumar, R. (2020). Hardware Acceleration of Graph Neural Networks. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6, San Francisco, CA, USA. IEEE.

[Azabou et al., 2023] Azabou, M., Ganesh, V., Thakoor, S., Lin, C.-H., Sathidevi, L., Liu, R., Valko, M., Veličković, P., and Dyer, E. L. (2023). Half-Hop : A graph upsampling approach for slowing down message passing.

[Azzolin et al., 2022] Azzolin, S., Longa, A., Barbiero, P., Lio, P., and Passerini, A. (2022). Global Explainability of GNNs via Logic Combination of Learned Concepts.

[Azzolin et al., 2023] Azzolin, S., Longa, A., Barbiero, P., Liò, P., and Passerini, A. (2023). Global Explainability of GNNs via Logic Combination of Learned Concepts. Technical Report arXiv :2210.07147, arXiv.

[Ba et al., 2016] Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer Normalization. Technical Report arXiv :1607.06450, arXiv.

[Bacciu et al., 2019] Bacciu, D., Errica, F., and Micheli, A. (2019). Contextual Graph Markov Model : A Deep and Generative Approach to Graph Processing.

[Bacciu and Numeroso, 2022] Bacciu, D. and Numeroso, D. (2022). Explaining Deep Graph Networks via Input Perturbation. *IEEE transactions on neural networks and learning systems*, PP.

[Bach et al., 2015a] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015a). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE*, 10(7) :e0130140.

[Bach et al., 2015b] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015b). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7) :e0130140.

[Bachmann et al., 2020] Bachmann, G., Becigneul, G., and Ganea, O. (2020). Constant Curvature Graph Convolutional Networks. In *Proceedings of the 37th International Conference on Machine Learning*, pages 486–496. PMLR.

[Badgeley et al., 2019] Badgeley, M. A., Zech, J. R., Oakden-Rayner, L., Glicksberg, B. S., Liu, M., Gale, W., McConnell, M. V., Percha, B., Snyder, T. M., and Dudley, J. T. (2019). Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digital Medicine*, 2 :31.

[Baek et al., 2020] Baek, J., Kang, M., and Hwang, S. J. (2020). Accurate Learning of Graph Representations with Graph Multiset Pooling.

[Baek et al., 2021] Baek, J., Kang, M., and Hwang, S. J. (2021). Accurate Learning of Graph Representations with Graph Multiset Pooling. Technical Report arXiv :2102.11533, arXiv.

[Bahri et al., 2021] Bahri, M., Bahl, G., and Zafeiriou, S. (2021). Binary Graph Neural Networks.

[Bai et al., 2020a] Bai, S., Zhang, F., and Torr, P. H. S. (2020a). Hypergraph Convolution and Hypergraph Attention. Technical Report arXiv :1901.08150, arXiv.

[Bai et al., 2020b] Bai, Y., Ding, H., Bian, S., Chen, T., Sun, Y., and Wang, W. (2020b). SimGNN : A Neural Network Approach to Fast Graph Similarity Computation. Technical Report arXiv :1808.05689, arXiv.

[Bajaj et al., 2022] Bajaj, M., Chu, L., Xue, Z. Y., Pei, J., Wang, L., Lam, P. C.-H., and Zhang, Y. (2022). Robust Counterfactual Explanations on Graph Neural Networks. Technical Report arXiv :2107.04086, arXiv.

[Balcilar et al., 2021] Balcilar, M., Heroux, P., Gauzere, B., Vasseur, P., Adam, S., and Honeine, P. (2021). Breaking the Limits of Message Passing Graph Neural Networks. In *Proceedings of the 38th International Conference on Machine Learning*, pages 599–608. PMLR.

[Balcilar et al., 2020] Balcilar, M., Renton, G., Héroux, P., Gaüzère, B., Adam, S., and Honeine, P. (2020). Analyzing the Expressive Power of Graph Neural Networks in a Spectral Perspective.

[Baldassarre and Azizpour, 2019a] Baldassarre, F. and Azizpour, H. (2019a). Explainability Techniques for Graph Convolutional Networks. Technical Report arXiv :1905.13686, arXiv.

[Baldassarre and Azizpour, 2019b] Baldassarre, F. and Azizpour, H. (2019b). Explainability Techniques for Graph Convolutional Networks.

[Baldassarre et al., 2020] Baldassarre, F., Smith, K., Sullivan, J., and Azizpour, H. (2020). Explanation-Based Weakly-Supervised Learning of Visual Relations with Graph Networks. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *Computer Vision – ECCV 2020*, volume 12373, pages 612–630. Springer International Publishing, Cham.

[Barbiero et al., 2022] Barbiero, P., Ciravegna, G., Giannini, F., Lió, P., Gori, M., and Melacci, S. (2022). Entropy-based Logic Explanations of Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6) :6046–6054.

[Barocas et al., 2020] Barocas, S., Selbst, A. D., and Raghavan, M. (2020). The Hidden Assumptions Behind Counterfactual Explanations and Principal Reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 80–89.

[Barredo-Arrieta and Del Ser, 2020] Barredo-Arrieta, A. and Del Ser, J. (2020). Plausible Counterfactuals : Auditing Deep Learning Classifiers with Realistic Adversarial Examples. Technical Report arXiv :2003.11323, arXiv.

[Bartunov et al., 2022] Bartunov, S., Fuchs, F. B., and Lillicrap, T. (2022). Equilibrium Aggregation : Encoding Sets via Optimization. Technical Report arXiv :2202.12795, arXiv.

[Bastos et al., 2022] Bastos, A., Nadgeri, A., Singh, K., Kanezashi, H., Suzumura, T., and Mulang', I. O. (2022). How Expressive are Transformers in Spectral Domain for Graphs ? *Transactions on Machine Learning Research*.

[Battaglia et al., 2018a] Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., Dyer, C., Heess, N., Wierstra, D., Kohli, P., Botvinick, M., Vinyals, O., Li, Y., and Pascanu, R. (2018a). Relational inductive biases, deep learning, and graph networks.

[Battaglia et al., 2018b] Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., Dyer, C., Heess, N., Wierstra, D., Kohli, P., Botvinick, M., Vinyals, O., Li, Y., and Pascanu, R. (2018b). Relational inductive biases, deep learning, and graph networks. Technical Report arXiv :1806.01261, arXiv.

[Battaglia et al., 2016] Battaglia, P. W., Pascanu, R., Lai, M., Rezende, D., and Kavukcuoglu, K. (2016). Interaction Networks for Learning about Objects, Relations and Physics.

[Bau et al., 2019] Bau, A., Durrani, N., Belinkov, Y., Dalvi, F., Sajjad, H., and Glass, J. (2019). IDENTIFYING AND CONTROLLING IMPORTANT NEURONS IN NEURAL MACHINE TRANSLATION.

[Bau et al., 2017] Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network Dissection : Quantifying Interpretability of Deep Visual Representations. Technical Report arXiv :1704.05796, arXiv.

[Bayani and Mitsch, 2022] Bayani, D. and Mitsch, S. (2022). Fanoos : Multi-Resolution, Multi-Strength, Interactive Explanations for Learned Systems. Technical Report arXiv :2006.12453, arXiv.

[Bazhenov et al., 2023] Bazhenov, G., Kuznedelev, D., Malinin, A., Babenko, A., and Prokhorenkova, L. (2023). Evaluating Robustness and Uncertainty of Graph Models Under Structural Distributional Shifts. Technical Report arXiv :2302.13875, arXiv.

[Bechtel and Abrahamsen, 2005] Bechtel, W. and Abrahamsen, A. (2005). Explanation : A mechanist alternative. *Studies in History and Philosophy of Science Part C : Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2) :421–441.

[Beck et al., 2018] Beck, D., Haffari, G., and Cohn, T. (2018). Graph-to-Sequence Learning using Gated Graph Neural Networks. Technical Report arXiv :1806.09835, arXiv.

[Becker and Hinton, 1992] Becker, S. and Hinton, G. E. (1992). Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356) :161–163.

[Beckmann and Smith, 2005] Beckmann, C. F. and Smith, S. M. (2005). Tensorial extensions of independent component analysis for multisubject FMRI analysis. *NeuroImage*, 25(1) :294–311.

[Belbute-Peres et al., 2020] Belbute-Peres, F. D. A., Economon, T., and Kolter, Z. (2020). Combining Differentiable PDE Solvers and Graph Neural Networks for Fluid Flow Prediction. In *Proceedings of the 37th International Conference on Machine Learning*, pages 2402–2411. PMLR.

[Bénard et al., 2021] Bénard, C., Biau, G., da Veiga, S., and Scornet, E. (2021). Interpretable Random Forests via Rule Extraction. Technical Report arXiv :2004.14841, arXiv.

[Bengio et al., 2021] Bengio, E., Jain, M., Korablyov, M., Precup, D., and Bengio, Y. (2021). Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation.

[Bengio, ] Bengio, Y. Learning Deep Architectures for AI.

[Bengio, 2009] Bengio, Y. (2009). Visualizing higher-layer features of a deep network.

[Bengio et al., 2007] Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007). Greedy Layer-Wise Training of Deep Networks. In Schölkopf, B., Platt, J., and Hofmann, T., editors, *Advances in Neural Information Processing Systems 19*, pages 153–160. The MIT Press.

[Bengio and LeCun, ] Bengio, Y. and LeCun, Y. Scaling Learning Algorithms towards AI.

[Bengio and Monperrus, 2004] Bengio, Y. and Monperrus, M. (2004). Non-Local Manifold Tangent Learning. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press.

[Bertossi, 2020] Bertossi, L. (2020). An ASP-Based Approach to Counterfactual Explanations for Classification. Technical Report arXiv :2004.13237, arXiv.

[Besserve et al., 2019] Besserve, M., Mehrjou, A., Sun, R., and Schölkopf, B. (2019). Counterfactuals uncover the modular structure of deep generative models. Technical Report arXiv :1812.03253, arXiv.

[Bevilacqua et al., 2021] Bevilacqua, B., Frasca, F., Lim, D., Srinivasan, B., Cai, C., Balamurugan, G., Bronstein, M. M., and Maron, H. (2021). Equivariant Subgraph Aggregation Networks.

[Bhatt et al., 2020] Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., and Eckersley, P. (2020). Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 648–657, Barcelona Spain. ACM.

[Bianchi et al., 2020] Bianchi, F. M., Grattarola, D., and Alippi, C. (2020). Spectral Clustering with Graph Neural Networks for Graph Pooling. In *Proceedings of the 37th International Conference on Machine Learning*, pages 874–883. PMLR.

[Bianchi et al., 2021] Bianchi, F. M., Grattarola, D., Livi, L., and Alippi, C. (2021). Graph Neural Networks with convolutional ARMA filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

[Bianchi et al., 2022a] Bianchi, F. M., Grattarola, D., Livi, L., and Alippi, C. (2022a). Graph Neural Networks With Convolutional ARMA Filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7) :3496–3507.

[Bianchi et al., 2022b] Bianchi, F. M., Grattarola, D., Livi, L., and Alippi, C. (2022b). Hierarchical Representation Learning in Graph Neural Networks With Node Decimation Pooling. *IEEE Transactions on Neural Networks and Learning Systems*, 33(5) :2195–2207.

[Bica et al., 2020] Bica, I., Jarrett, D., Hüyük, A., and van der Schaar, M. (2020). Learning "What-if" Explanations for Sequential Decision-Making.

[Binder et al., 2016] Binder, A., Montavon, G., Bach, S., Müller, K.-R., and Samek, W. (2016). Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York.

[BK et al., 2021] BK, V., Ganesan, B., Saxena, A., Sharma, D., and Agarwal, A. (2021). Towards Automated Evaluation of Explanations in Graph Neural Networks. Technical Report arXiv :2106.11864, arXiv.

[Black et al., 2022] Black, E., Wang, Z., Datta, A., and Fredrikson, M. (2022). CONSISTENT COUNTERFACTUALS FOR DEEP MODELS.

[Blanc et al., ] Blanc, G., Koch, C., Lange, J., and Tan, L.-Y. A query-optimal algorithm for finding counterfactuals.

[Bo et al., 2022] Bo, D., Shi, C., Wang, L., and Liao, R. (2022). Specformer : Spectral Graph Neural Networks Meet Transformers.

[Bo et al., 2021] Bo, D., Wang, X., Shi, C., and Shen, H. (2021). Beyond Low-frequency Information in Graph Convolutional Networks. Technical Report arXiv :2101.00797, arXiv.

[Bodnar et al., 2021] Bodnar, C., Frasca, F., Wang, Y., Otter, N., Montufar, G. F., Lió, P., and Bronstein, M. (2021). Weisfeiler and Lehman Go Topological : Mes-

sage Passing Simplicial Networks. In *Proceedings of the 38th International Conference on Machine Learning*, pages 1026–1037. PMLR.

[Bodria et al., 2023] Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., and Rinzivillo, S. (2023). Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 37(5) :1719–1778.

[Bogo et al., 2014] Bogo, F., Romero, J., Loper, M., and Black, M. J. (2014). FAUST : Dataset and Evaluation for 3D Mesh Registration. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3794–3801, Columbus, OH, USA. IEEE.

[Bogo et al., 2017] Bogo, F., Romero, J., Pons-Moll, G., and Black, M. J. (2017). Dynamic FAUST : Registering Human Bodies in Motion. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5573–5582, Honolulu, HI. IEEE.

[Bojchevski et al., 2020] Bojchevski, A., Gasteiger, J., Perozzi, B., Kapoor, A., Blais, M., Rózemberczki, B., Lukasik, M., and Günnemann, S. (2020). Scaling Graph Neural Networks with Approximate PageRank. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2464–2473.

[Bojchevski and Günnemann, 2018] Bojchevski, A. and Günnemann, S. (2018). Deep Gaussian Embedding of Graphs : Unsupervised Inductive Learning via Ranking. Technical Report arXiv :1707.03815, arXiv.

[Bojchevski and Günnemann, 2019] Bojchevski, A. and Günnemann, S. (2019). Adversarial Attacks on Node Embeddings via Graph Poisoning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 695–704. PMLR.

[Bojchevski et al., 2018] Bojchevski, A., Shchur, O., Zügner, D., and Günnemann, S. (2018). NetGAN : Generating Graphs via Random Walks. Technical Report arXiv :1803.00816, arXiv.

[Bonabi Mobaraki and Khan, 2023] Bonabi Mobaraki, E. and Khan, A. (2023). A Demonstration of Interpretability Methods for Graph Neural Networks. In *Proceedings of the 6th Joint Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)*, pages 1–5, Seattle WA USA. ACM.

[Bonet et al., 2022] Bonet, E. R., Do, T. H., Qin, X., Hofman, J., Manna, V. P. L., Philips, W., and Deligiannis, N. (2022). Explaining Graph Neural Networks With Topology-Aware Node Selection : Application in Air Quality Inference. *IEEE Transactions on Signal and Information Processing over Networks*, 8 :499–513.

[Bonnet et al., 2023] Bonnet, F., Mazari, A. J., Cinnella, P., and Gallinari, P. (2023). AirfRANS : High Fidelity Computational Fluid Dynamics Dataset for Approximating Reynolds-Averaged Navier-Stokes Solutions. Technical Report arXiv :2212.07564, arXiv.

[Bordes et al., ] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. Translating Embeddings for Modeling Multi-relational Data.

[Bordes et al., 2013] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. (2013). Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

[Borgwardt et al., 2005] Borgwardt, K. M., Ong, C. S., Schonauer, S., Vishwanathan, S. V. N., Smola, A. J., and Kriegel, H.-P. (2005). Protein function prediction via graph kernels. *Bioinformatics*, 21(Suppl 1) :i47–i56.

[Bos, ] Bos, D. O. EEG-based Emotion Recognition. page 18.

[Bourdev and Malik, 2009] Bourdev, L. and Malik, J. (2009). Poselets : Body part detectors trained using 3D human pose annotations. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1365–1372, Kyoto. IEEE.

[Breiman et al., 1984] Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. Taylor & Francis.

[Bresson and Laurent, 2018] Bresson, X. and Laurent, T. (2018). Residual Gated Graph ConvNets. Technical Report arXiv :1711.07553, arXiv.

[Brin and Page, 1998] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1) :107–117.

[Brockschmidt, 2020] Brockschmidt, M. (2020). GNN-FiLM : Graph Neural Networks with Feature-wise Linear Modulation. Technical Report arXiv :1906.12192, arXiv.

[Brody et al., 2021] Brody, S., Alon, U., and Yahav, E. (2021). How Attentive are Graph Attention Networks ?

[Brody et al., 2022] Brody, S., Alon, U., and Yahav, E. (2022). How Attentive are Graph Attention Networks ? Technical Report arXiv :2105.14491, arXiv.

[Bronstein et al., 2021] Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. (2021). Geometric Deep Learning : Grids, Groups, Graphs, Geodesics, and Gauges.

[Bronstein et al., 2017] Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. (2017). Geometric deep learning : Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4) :18–42.

[Bruna et al., 2013] Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. (2013). Spectral Networks and Locally Connected Networks on Graphs.

[Bruna et al., 2014] Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. (2014). Spectral Networks and Locally Connected Networks on Graphs. Technical Report arXiv :1312.6203, arXiv.

[Bui et al., 2023] Bui, T.-C., Le, V.-D., and Li, W.-S. (2023). Generating Real-Time Explanations for GNNs via Multiple Specialty Learners and Online Knowledge Distillation. *IEEE access : practical innovations, open solutions*, 11 :40790–40808.

[Bui et al., 2022] Bui, T.-C., Le, V.-D., Li, W.-S., and Cha, S. K. (2022). INGREX : An Interactive Explanation Framework for Graph Neural Networks. Technical Report arXiv :2211.01548, arXiv.

[Bulat and Tzimiropoulos, 2017] Bulat, A. and Tzimiropoulos, G. (2017). How Far are We from Solving the 2D & 3D Face Alignment Problem ? (and a Dataset of 230,000 3D Facial Landmarks). In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1021–1030, Venice. IEEE.

[Busbridge et al., 2019] Busbridge, D., Sherburn, D., Cavallo, P., and Hammerla, N. Y. (2019). Relational Graph Attention Networks. Technical Report arXiv :1904.05811, arXiv.

[Buterez et al., 2022] Buterez, D., Janet, J. P., Kiddle, S. J., Oglic, D., and Liò, P. (2022). Graph Neural Networks with Adaptive Readouts. Technical Report arXiv :2211.04952, arXiv.

[Byrne, 2019] Byrne, R. M. J. (2019). Counterfactuals in Explainable Artificial Intelligence (XAI) : Evidence from Human Reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 6276–6282, Macao, China. International Joint Conferences on Artificial Intelligence Organization.

[Cai and Wang, 2022] Cai, C. and Wang, Y. (2022). A simple yet effective baseline for non-attributed graph classification. Technical Report arXiv :1811.03508, arXiv.

[Cai et al., 2021a] Cai, T., Luo, S., Xu, K., He, D., Liu, T.-Y., and Wang, L. (2021a). GraphNorm : A Principled Approach to Accelerating Graph Neural Network Training. In *Proceedings of the 38th International Conference on Machine Learning*, pages 1204–1215. PMLR.

[Cai et al., 2021b] Cai, Z., Yan, X., Wu, Y., Ma, K., Cheng, J., and Yu, F. (2021b). DGCL : An efficient communication library for distributed GNN training. In *Proceedings of the Sixteenth European Conference on Computer Systems*, pages 130–144, Online Event United Kingdom. ACM.

[Calhoun et al., 2001] Calhoun, V. D., Adali, T., Pearlson, G. D., and Pekar, J. J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Human Brain Mapping*, 14(3) :140–151.

[Cangea et al., 2018] Cangea, C., Veličković, P., Jovanović, N., Kipf, T., and Liò, P. (2018). Towards Sparse Hierarchical Graph Classifiers. Technical Report arXiv :1811.01287, arXiv.

[Cao et al., 2016a] Cao, S., Lu, W., and Xu, Q. (2016a). Deep Neural Networks for Learning Graph Representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

[Cao et al., 2016b] Cao, S., Lu, W., and Xu, Q. (2016b). Deep Neural Networks for Learning Graph Representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

[Carlson et al., 2010] Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E., and Mitchell, T. (2010). Toward an Architecture for Never-Ending Language Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1) :1306–1313.

[Carton, ] Carton, O. Langages formels, Calculabilité et Complexité. page 179.

[Caruana et al., ] Caruana, R., Kangarloo, H., David, J., Dionisio, N., Sinha, U., and Johnson, D. Case-Based Explanation of Non-Case-Based Learning Methods.

[Caruana et al., 2015] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). Intelligible Models for HealthCare : Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730, Sydney NSW Australia. ACM.

[Carvalho et al., 2019] Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine Learning Interpretability : A Survey on Methods and Metrics. *Electronicsweek*, 8(8) :832.

[Cauchy, 1847] Cauchy, M. A. (1847). Me´thode ge´ne´rale pour la re´solution des syste‘mes d'e´quations simultane´es. page 3.

[Cen et al., 2023] Cen, Y., Hou, Z., Wang, Y., Chen, Q., Luo, Y., Yu, Z., Zhang, H., Yao, X., Zeng, A., Guo, S., Dong, Y., Yang, Y., Zhang, P., Dai, G., Wang, Y., Zhou, C., Yang, H., and Tang, J. (2023). CogDL : A Comprehensive Library for Graph Deep Learning.

[Chakaravarthy et al., 2021] Chakaravarthy, V. T., Pandian, S. S., Raje, S., Sabharwal, Y., Suzumura, T., and Ubaru, S. (2021). Efficient scaling of dynamic graph neural networks. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–15, St. Louis Missouri. ACM.

[Chalasani et al., 2020] Chalasani, P., Chen, J., Chowdhury, A. R., Jha, S., and Wu, X. (2020). Concise Explanations of Neural Networks using Adversarial Training. Technical Report arXiv :1810.06583, arXiv.

[Chalkiadakis, ] Chalkiadakis, I. A brief survey of visualization methods for deep learning models from the perspective of Explainable AI.

[Chamberlain et al., 2021] Chamberlain, B., Rowbottom, J., Gorinova, M. I., Bronstein, M., Webb, S., and Rossi, E. (2021). GRAND : Graph Neural Diffusion. In *Proceedings of the 38th International Conference on Machine Learning*, pages 1407–1418. PMLR.

[Chami et al., 2019] Chami, I., Ying, Z., Ré, C., and Leskovec, J. (2019). Hyperbolic Graph Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[Chan et al., 2022] Chan, A., Xu, J., Long, B., Sanyal, S., Gupta, T., and Ren, X. (2022). SalKG : Learning From Knowledge Graph Explanations for Commonsense Reasoning. Technical Report arXiv :2104.08793, arXiv.

[Chang et al., 2019] Chang, C.-H., Creager, E., Goldenberg, A., and Duvenaud, D. (2019). Explaining Image Classifiers by Counterfactual Generation. Technical Report arXiv :1807.08024, arXiv.

[Chang et al., 2021a] Chang, C.-H., Tan, S., Lengerich, B., Goldenberg, A., and Caruana, R. (2021a). How Interpretable and Trustworthy are GAMs ? Technical Report arXiv :2006.06466, arXiv.

[Chang et al., 2021b] Chang, H., Rong, Y., Xu, T., Bian, Y., Zhou, S., Wang, X., Huang, J., and Zhu, W. (2021b). Not All Low-Pass Filters are Robust in Graph Convolutional Networks.

[Chang et al., ] Chang, J., Gu, J., Wang, L., Meng, G., Xiang, S., and Pan, C. Structure-Aware Convolutional Neural Networks.

[Chapman-Rounds et al., 2021] Chapman-Rounds, M., Bhatt, U., Pazos, E., Schulz, M.-A., and Georgatzis, K. (2021). FIMAP : Feature Importance by Minimal Adversarial Perturbation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13) :11433–11441.

[Chapman-Rounds et al., 2019] Chapman-Rounds, M., Schulz, M.-A., Pazos, E., and Georgatzis, K. (2019). EMAP : Explanation by Minimal Adversarial Perturbation. Technical Report arXiv :1912.00872, arXiv.

[Chater and Oaksford, 2005] Chater, N. and Oaksford, M. (2005). Mental Mechanisms : Speculations on Human Causal Learning and Reasoning. In *Information Sampling and Adaptive Cognition*, pages 210–236. Cambridge University Press, New York, NY, US.

[Chatzianastasis et al., 2023] Chatzianastasis, M., Vazirgiannis, M., and Zhang, Z. (2023). Explainable Multilayer Graph Neural Network for Cancer Gene Prediction. Technical Report arXiv :2301.08831, arXiv.

[Chawla and Wang, 2017] Chawla, N. and Wang, W., editors (2017). *Proceedings of the 2017 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

[Chefer et al., 2021] Chefer, H., Gur, S., and Wolf, L. (2021). Transformer Interpretability Beyond Attention Visualization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791, Nashville, TN, USA. IEEE.

[Chen et al., 2021a] Chen, C., Li, K., Zou, X., and Li, Y. (2021a). DyGNN : Algorithm and Architecture Support of Dynamic Pruning for Graph Neural Networks. In *2021 58th ACM/IEEE Design Automation Conference (DAC)*, pages 1201–1206.

[Chen et al., a] Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., and Su, J. K. This Looks Like That : Deep Learning for Interpretable Image Recognition.

[Chen et al., 2020a] Chen, D., Lin, Y., Li, W., Li, P., Zhou, J., and Sun, X. (2020a). Measuring and Relieving the Over-Smoothing Problem for Graph Neural Networks from the Topological View. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04) :3438–3445.

[Chen et al., 2022a] Chen, J., Gan, C., Cheng, S., Zhou, H., Xiao, Y., and Li, L. (2022a). Unsupervised Editing for Counterfactual Stories. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10) :10473–10481.

[Chen and Jordan, 2019] Chen, J. and Jordan, M. I. (2019). LS-Tree : Model Interpretation When the Data Are Linguistic.

[Chen et al., 2018a] Chen, J., Ma, T., and Xiao, C. (2018a). FastGCN : Fast Learning with Graph Convolutional Networks via Importance Sampling.

[Chen et al., b] Chen, J., Song, L., Wainwright, M. J., and Jordan, M. I. Learning to Explain : An Information-Theoretic Perspective on Model Interpretation.

[Chen et al., 2018b] Chen, J., Song, L., Wainwright, M. J., and Jordan, M. I. (2018b). L-Shapley and C-Shapley : Efficient Model Interpretation for Structured Data. Technical Report arXiv :1808.02610, arXiv.

[Chen et al., 2018c] Chen, J., Song, L., Wainwright, M. J., and Jordan, M. I. (2018c). Learning to Explain : An Information-Theoretic Perspective on Model Interpretation. Technical Report arXiv :1802.07814, arXiv.

[Chen et al., 2018d] Chen, J., Zhu, J., and Song, L. (2018d). Stochastic Training of Graph Convolutional Networks with Variance Reduction. Technical Report arXiv :1710.10568, arXiv.

[Chen et al., 2020b] Chen, L., Chen, Z., and Bruna, J. (2020b). On Graph Neural Networks versus Graph-Augmented MLPs. Technical Report arXiv :2010.15116, arXiv.

[Chen et al., 2020c] Chen, M., Wei, Z., Ding, B., Li, Y., Yuan, Y., Du, X., and Wen, J.-R. (2020c). Scalable Graph Neural Networks via Bidirectional Propagation. In *Advances in Neural Information Processing Systems*, volume 33, pages 14556–14566. Curran Associates, Inc.

[Chen et al., 2020d] Chen, M., Wei, Z., Huang, Z., Ding, B., and Li, Y. (2020d). Simple and Deep Graph Convolutional Networks. Technical Report arXiv :2007.02133, arXiv.

[Chen and Zhao, 2022] Chen, S. and Zhao, Q. (2022). REX : Reasoning-aware and Grounded Explanation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15565–15574, New Orleans, LA, USA. IEEE.

[Chen et al., 2018e] Chen, X., Li, L.-J., Fei-Fei, L., and Gupta, A. (2018e). Iterative Visual Reasoning Beyond Convolutions. Technical Report arXiv :1803.11189, arXiv.

[Chen et al., 2020e] Chen, X., Wang, Y., Xie, X., Hu, X., Basak, A., Liang, L., Yan, M., Deng, L., Ding, Y., Du, Z., Chen, Y., and Xie, Y. (2020e). Rubik : A Hierarchical Architecture for Efficient Graph Learning.

[Chen et al., 2020f] Chen, X., Wang, Y., Xie, X., Hu, X., Basak, A., Liang, L., Yan, M., Deng, L., Ding, Y., Du, Z., Chen, Y., and Xie, Y. (2020f). Rubik : A Hierarchical Architecture for Efficient Graph Learning. Technical Report arXiv :2009.12495, arXiv.

[Chen et al., 2021b] Chen, Y., Bian, Y., Xiao, X., Rong, Y., Xu, T., and Huang, J. (2021b). On Self-Distilling Graph Neural Network. volume 3, pages 2278–2284.

[Chen et al., 2020g] Chen, Z., Chen, L., Villar, S., and Bruna, J. (2020g). Can Graph Neural Networks Count Substructures ? In *Advances in Neural Information Processing Systems*, volume 33, pages 10383–10395. Curran Associates, Inc.

[Chen et al., 2023] Chen, Z., Gao, Q., Bosselut, A., Sabharwal, A., and Richardson, K. (2023). DISCO : Distilling Counterfactuals with Large Language Models. Technical Report arXiv :2212.10534, arXiv.

[Chen et al., 2022b] Chen, Z., Silvestri, F., Wang, J., Zhu, H., Ahn, H., and Tolomei, G. (2022b). R E LAX : Reinforcement Learning Agent Explainer for Arbitrary Predictive Models. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 252–261, Atlanta GA USA. ACM.

[Chen et al., 2020h] Chen, Z., Yan, M., Zhu, M., Deng, L., Li, G., Li, S., and Xie, Y. (2020h). fuseGNN : Accelerating graph convolutional neural network training on GPGPU. In *Proceedings of the 39th International Conference on Computer-Aided Design*, pages 1–9, Virtual Event USA. ACM.

[Cheng et al., 2020a] Cheng, F., Ming, Y., and Qu, H. (2020a). DECE : Decision Explorer with Counterfactual Explanations for Machine Learning Models. Technical Report arXiv :2008.08353, arXiv.

[Cheng et al., 2020b] Cheng, X., Rao, Z., Chen, Y., and Zhang, Q. (2020b). Explaining Knowledge Distillation by Quantifying the Knowledge. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12922–12932, Seattle, WA, USA. IEEE.

[Chereda et al., 2021] Chereda, H., Bleckmann, A., Menck, K., Perera-Bel, J., Stegmaier, P., Auer, F., Kramer, F., Leha, A., and Beißbarth, T. (2021). Explaining decisions of graph convolutional neural networks : Patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer. *Genome Medicine*, 13(1) :42.

[Chiang et al., 2019a] Chiang, W.-L., Liu, X., Si, S., Li, Y., Bengio, S., and Hsieh, C.-J. (2019a). Cluster-GCN : An Efficient Algorithm for Training Deep and Large Graph Convolutional Networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 257–266.

[Chiang et al., 2019b] Chiang, W.-L., Liu, X., Si, S., Li, Y., Bengio, S., and Hsieh, C.-J. (2019b). Cluster-GCN : An Efficient Algorithm for Training Deep and

Large Graph Convolutional Networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 257–266.

[Chien et al., 2020] Chien, E., Peng, J., Li, P., and Milenkovic, O. (2020). Adaptive Universal Generalized PageRank Graph Neural Network.

[Cho et al., 2021] Cho, H., Oh, Y., and Jeon, E. (2021). SEEN : Sharpening Explanations for Graph Neural Networks using Explanations from Neighborhoods. Technical Report arXiv :2106.08532, arXiv.

[Cho et al., 2013] Cho, M., Alahari, K., and Ponce, J. (2013). Learning Graphs to Match. In *2013 IEEE International Conference on Computer Vision*, pages 25–32, Sydney, Australia. IEEE.

[Choi et al., 2017] Choi, E., Bahadori, M. T., Song, L., Stewart, W. F., and Sun, J. (2017). GRAM : Graph-based Attention Model for Healthcare Representation Learning. Technical Report arXiv :1611.07012, arXiv.

[Choudhury et al., 2020] Choudhury, S., Bilbrey, J. A., Ward, L., Xantheas, S. S., Foster, I., Heindel, J. P., Blaiszik, B., and Schwarting, M. E. (2020). HydroNet : Benchmark Tasks for Preserving Intermolecular Interactions and Structural Motifs in Predictive and Generative Models for Molecular Data. Technical Report arXiv :2012.00131, arXiv.

[Church, 1936a] Church, A. (1936a). A note on the Entscheidungsproblem. *The Journal of Symbolic Logic*, 1(1) :40–41.

[Church, 1936b] Church, A. (1936b). An Unsolvable Problem of Elementary Number Theory. *American Journal of Mathematics*, 58(2) :345.

[Clough et al., 2019a] Clough, J. R., Oksuz, I., Puyol-Antón, E., Ruijsink, B., King, A. P., and Schnabel, J. A. (2019a). Global and Local Interpretability for Cardiac MRI Classification. In Shen, D., Liu, T., Peters, T. M., Staib, L. H., Essert, C., Zhou, S., Yap, P.-T., and Khan, A., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, volume 11767, pages 656–664. Springer International Publishing, Cham.

[Clough et al., 2019b] Clough, J. R., Oksuz, I., Puyol-Antón, E., Ruijsink, B., King, A. P., and Schnabel, J. A. (2019b). Global and local interpretability for cardiac MRI classification : 22nd International Conference on Medical Image Computing

and Computer-Assisted Intervention, MICCAI 2019. In Shen, D., Yap, P.-T., Liu, T., Peters, T. M., Khan, A., Staib, L. H., Essert, C., and Zhou, S., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019 - 22nd International Conference, Proceedings*, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 656–664. SPRINGER.

[Colin et al., ] Colin, J., Fel, T., Cadène, R., and Serre, T. What I Cannot Predict, I Do Not Understand : A Human-Centered Evaluation Framework for Explainability Methods.

[Cong et al., 2020] Cong, W., Forsati, R., Kandemir, M., and Mahdavi, M. (2020). Minimal Variance Sampling with Provable Guarantees for Fast Training of Graph Neural Networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1393–1403, Virtual Event CA USA. ACM.

[Cong et al., 2021] Cong, W., Ramezani, M., and Mahdavi, M. (2021). On Provable Benefits of Depth in Training Graph Convolutional Networks.

[Cong et al., 2022] Cong, W., Zhang, S., Kang, J., Yuan, B., Wu, H., Zhou, X., Tong, H., and Mahdavi, M. (2022). Do We Really Need Complicated Model Architectures For Temporal Networks ?

[Cong et al., 2023] Cong, W., Zhang, S., Kang, J., Yuan, B., Wu, H., Zhou, X., Tong, H., and Mahdavi, M. (2023). Do We Really Need Complicated Model Architectures For Temporal Networks ? Technical Report arXiv :2302.11636, arXiv.

[Corso et al., 2020] Corso, G., Cavalleri, L., Beaini, D., Liò, P., and Veličković, P. (2020). Principal Neighbourhood Aggregation for Graph Nets. Technical Report arXiv :2004.05718, arXiv.

[Cosmo et al., 2016] Cosmo, L., Rodolà, E., Bronstein, M. M., Torsello, A., Cremers, D., and Sahilliog, Y. (2016). SHREC'16 : Partial Matching of Deformable Shapes.

[Courville et al., ] Courville, A., Bengio, Y., and Goodfellow, I. *Deep Learning*.

[Covert and Lee, 2021] Covert, I. and Lee, S.-I. (2021). Improving KernelSHAP : Practical Shapley Value Estimation via Linear Regression. Technical Report arXiv :2012.01536, arXiv.

[Crabbé, ] Crabbé, J. Concept Activation Regions : A Generalized Framework For Concept-Based Explanations.

[Crabbé et al., ] Crabbé, J., Curth, A., and Bica, I. Benchmarking Heterogeneous Treatment Effect Models through the Lens of Interpretability.

[Crabbé and van der Schaar, 2021] Crabbé, J. and van der Schaar, M. (2021). Explaining Time Series Predictions with Dynamic Masks. Technical Report arXiv :2106.05303, arXiv.

[Crabbe et al., ] Crabbe, J., Zame, W. R., and Zhang, Y. Learning outside the Black-Box : The pursuit of interpretable models.

[Crawford, ] Crawford, Jamie Morgenstern, B. V. J. W. V. H. W. H. D. I. K. T. G. Datasheets for Datasets. https ://cacm.acm.org/magazines/2021/12/256932-datasheets-for-datasets/abstract.

[Cucala et al., 2022] Cucala, D. T., Kostylev, E. V., Grau, B. C., and Motik, B. (2022). EXPLAINABLE GNN-BASED MODELS OVER KNOWLEDGE GRAPHS.

[Cybenko, 1989] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4) :303–314.

[Dabkowski and Gal, ] Dabkowski, P. and Gal, Y. Real Time Image Saliency for Black Box Classifiers.

[Dabkowski and Gal, 2017] Dabkowski, P. and Gal, Y. (2017). Real Time Image Saliency for Black Box Classifiers. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[Dağlarli, 2020] Dağlarli, E. (2020). Explainable Artificial Intelligence (xAI) Approaches and Deep Meta-Learning Models. In *Advances and Applications in Deep Learning*. IntechOpen.

[Dai et al., ] Dai, H., Kozareva, Z., Dai, B., Smola, A. J., and Song, L. Learning Steady-States of Iterative Algorithms over Graphs.

[Dalvi et al., 2018] Dalvi, F., Durrani, N., Sajjad, H., Belinkov, Y., Bau, A., and Glass, J. (2018). What Is One Grain of Sand in the Desert ? Analyzing Individual Neurons in Deep NLP Models.

[Dandl et al., 2020] Dandl, S., Molnar, C., Binder, M., and Bischl, B. (2020). Multi-Objective Counterfactual Explanations. volume 12269, pages 448–469.

[Dasgupta et al., 2022] Dasgupta, S., Frost, N., and Moshkovitz, M. (2022). Framework for Evaluating Faithfulness of Local Explanations. Technical Report arXiv :2202.00734, arXiv.

[Dash et al., ] Dash, S., Gunluk, O., and Wei, D. Boolean Decision Rules via Column Generation.

[Dash et al., 2015] Dash, S., Malioutov, D. M., and Varshney, K. R. (2015). Learning interpretable classification rules using sequential rowsampling. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3337–3341.

[Dasoulas et al., 2021] Dasoulas, G., Scaman, K., and Virmaux, A. (2021). Lipschitz normalization for self-attention layers with application to graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, pages 2456–2466. PMLR.

[Datta et al., 2016] Datta, A., Sen, S., and Zick, Y. (2016). Algorithmic Transparency via Quantitative Input Influence : Theory and Experiments with Learning Systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 598–617.

[Datta et al., 2022] Datta, D., Chen, F., and Ramakrishnan, N. (2022). Framing Algorithmic Recourse for Anomaly Detection. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 283–293.

[De Cao and Kipf, 2022] De Cao, N. and Kipf, T. (2022). MolGAN : An implicit generative model for small molecular graphs. Technical Report arXiv :1805.11973, arXiv.

[de Oliveira and Martens, 2021] de Oliveira, R. M. B. and Martens, D. (2021). A Framework and Benchmarking Study for Counterfactual Generating Methods on Tabular Data. *Applied Sciences*, 11(16) :7274.

[Defferrard et al., 2016] Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

[Defferrard et al., 2017] Defferrard, M., Bresson, X., and Vandergheynst, P. (2017). Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. Technical Report arXiv :1606.09375, arXiv.

[Delalleau and Bengio, 2011] Delalleau, O. and Bengio, Y. (2011). Shallow vs. Deep Sum-Product Networks. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

[Delaney et al., 2021] Delaney, E., Greene, D., and Keane, M. T. (2021). Instance-based Counterfactual Explanations for Time Series Classification. Technical Report arXiv :2009.13211, arXiv.

[Deng et al., 2018] Deng, H., Birdal, T., and Ilic, S. (2018). PPFNet : Global Context Aware Local Features for Robust 3D Point Matching. Technical Report arXiv :1802.02669, arXiv.

[Deng and Zhang, 2021] Deng, X. and Zhang, Z. (2021). Graph-Free Knowledge Distillation for Graph Neural Networks.

[Derr et al., 2018] Derr, T., Ma, Y., and Tang, J. (2018). Signed Graph Convolutional Network. Technical Report arXiv :1808.06354, arXiv.

[Dettmers et al., 2018] Dettmers, T., Minervini, P., Stenetorp, P., and Riedel, S. (2018). Convolutional 2D Knowledge Graph Embeddings. Technical Report arXiv :1707.01476, arXiv.

[Deutsch et al., 2019] Deutsch, A., Xu, Y., Wu, M., and Lee, V. (2019). TigerGraph : A Native MPP Graph Database. Technical Report arXiv :1901.08248, arXiv.

[Deutsch and Penrose, 1997] Deutsch, D. and Penrose, R. (1997). Quantum theory, the Church–Turing principle and the universal quantum computer. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 400(1818) :97–117.

[Dhillon et al., 2007] Dhillon, I. S., Guan, Y., and Kulis, B. (2007). Weighted Graph Cuts without Eigenvectors A Multilevel Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11) :1944–1957.

[Dhurandhar et al., 2018] Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., and Das, P. (2018). Explanations based on the Missing : To-

wards Contrastive Explanations with Pertinent Negatives. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

[Dhurandhar et al., 2022] Dhurandhar, A., Natesan Ramamurthy, K., and Shanmugam, K. (2022). Is this the Right Neighborhood ? Accurate and Query Efficient Model Agnostic Explanations. *Advances in Neural Information Processing Systems*, 35 :9499–9511.

[Dhurandhar et al., 2019] Dhurandhar, A., Pedapati, T., Balakrishnan, A., Chen, P.-Y., Shanmugam, K., and Puri, R. (2019). Model Agnostic Contrastive Explanations for Structured Data. Technical Report arXiv :1906.00117, arXiv.

[Dhurandhar et al., ] Dhurandhar, A., Shanmugam, K., Luss, R., and Olsen, P. A. Improving Simple Models with Confidence Profiles.

[Díaz-Rodríguez et al., 2023] Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., de Prado, M. L., Herrera-Viedma, E., and Herrera, F. (2023). Connecting the Dots in Trustworthy Artificial Intelligence : From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation. Technical Report arXiv :2305.02231, arXiv.

[Diehl, 2019] Diehl, F. (2019). Edge Contraction Pooling for Graph Neural Networks. Technical Report arXiv :1905.10990, arXiv.

[Ding et al., 2021] Ding, M., Kong, K., Li, J., Zhu, C., Dickerson, J. P., Huang, F., and Goldstein, T. (2021). VQ-GNN : A Universal Framework to Scale up Graph Neural Networks using Vector Quantization.

[Dombrowski et al., 2019] Dombrowski, A.-K., Alber, M., Anders, C. J., Ackermann, M., Müller, K.-R., and Kessel, P. (2019). Explanations can be manipulated and geometry is to blame.

[Dominguez-Olmedo et al., ] Dominguez-Olmedo, R., Karimi, A.-H., and Schölkopf, B. On the Adversarial Robustness of Causal Algorithmic Recourse.

[Dong et al., 2021a] Dong, J., Zheng, D., Yang, L. F., and Karypis, G. (2021a). Global Neighbor Sampling for Mixed CPU-GPU Training on Giant Graphs. Technical Report arXiv :2106.06150, arXiv.

[Dong et al., 2017] Dong, Y., Chawla, N. V., and Swami, A. (2017). Metapath2vec : Scalable Representation Learning for Heterogeneous Networks. In *Proceedings of*

*the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 135–144, Halifax NS Canada. ACM.

[Dong et al., 2021b] Dong, Y., Ding, K., Jalaian, B., Ji, S., and Li, J. (2021b). AdaGNN : Graph Neural Networks with Adaptive Frequency Response Filter. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 392–401, Virtual Event Queensland Australia. ACM.

[Donnat et al., 2018] Donnat, C., Zitnik, M., Hallac, D., and Leskovec, J. (2018). Learning Structural Node Embeddings via Diffusion Wavelets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1320–1329, London United Kingdom. ACM.

[Doshi-Velez and Kim, 2017a] Doshi-Velez, F. and Kim, B. (2017a). Towards A Rigorous Science of Interpretable Machine Learning.

[Doshi-Velez and Kim, 2017b] Doshi-Velez, F. and Kim, B. (2017b). Towards A Rigorous Science of Interpretable Machine Learning. Technical Report arXiv :1702.08608, arXiv.

[Došilović et al., 2018] Došilović, F. K., Brčić, M., and Hlupić, N. (2018). Explainable artificial intelligence : A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215.

[Dou et al., 2021] Dou, Y., Shu, K., Xia, C., Yu, P. S., and Sun, L. (2021). User Preference-aware Fake News Detection. Technical Report arXiv :2104.12259, arXiv.

[Downs et al., ] Downs, M., Chu, J. L., Yacoby, Y., Doshi-Velez, F., and Pan, W. CRUDS : Counterfactual Recourse Using Disentangled Subspaces.

[Du et al., 2018] Du, J., Zhang, S., Wu, G., Moura, J. M. F., and Kar, S. (2018). Topology Adaptive Graph Convolutional Networks. Technical Report arXiv :1710.10370, arXiv.

[Durrani et al., 2020] Durrani, N., Sajjad, H., Dalvi, F., and Belinkov, Y. (2020). Analyzing Individual Neurons in Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4865–4880, Online. Association for Computational Linguistics.

[Duval and Malliaros, 2021] Duval, A. and Malliaros, F. D. (2021). GraphSVX : Shapley Value Explanations for Graph Neural Networks.

[Duvenaud et al., ] Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gomez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints.

[Duvenaud et al., 2015] Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. P. (2015). Convolutional Networks on Graphs for Learning Molecular Fingerprints. Technical Report arXiv :1509.09292, arXiv.

[Dwivedi et al., 2023a] Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., and Ranjan, R. (2023a). Explainable AI (XAI) : Core Ideas, Techniques, and Solutions. *ACM Computing Surveys*, 55(9) :1–33.

[Dwivedi et al., 2022a] Dwivedi, V. P., Joshi, C. K., Luu, A. T., Laurent, T., Bengio, Y., and Bresson, X. (2022a). Benchmarking Graph Neural Networks. Technical Report arXiv :2003.00982, arXiv.

[Dwivedi et al., 2022b] Dwivedi, V. P., Luu, A. T., Laurent, T., Bengio, Y., and Bresson, X. (2022b). Graph Neural Networks with Learnable Structural and Positional Representations. Technical Report arXiv :2110.07875, arXiv.

[Dwivedi et al., 2023b] Dwivedi, V. P., Rampášek, L., Galkin, M., Parviz, A., Wolf, G., Luu, A. T., and Beaini, D. (2023b). Long Range Graph Benchmark. Technical Report arXiv :2206.08164, arXiv.

[E et al., 2020] E, D., Kw, J., Mr, S., and A, K. (2020). Sex classification using long-range temporal dependence of resting-state functional MRI time series. *Human brain mapping*, 41(13).

[Errica et al., 2019] Errica, F., Podda, M., Bacciu, D., and Micheli, A. (2019). A Fair Comparison of Graph Neural Networks for Graph Classification.

[Esposito et al., 2005] Esposito, F., Scarabino, T., Hyvarinen, A., Himberg, J., Formisano, E., Comani, S., Tedeschi, G., Goebel, R., Seifritz, E., and Di Salle, F. (2005). Independent component analysis of fMRI group studies by self-organizing clustering. *NeuroImage*, 25(1) :193–205.

[Etmann et al., 2019] Etmann, C., Lunz, S., Maass, P., and Schönlieb, C.-B. (2019). On the Connection Between Adversarial Robustness and Saliency Map Interpretability. Technical Report arXiv :1905.04172, arXiv.

[Euclide, ] Euclide. *Elements*.

[Faber et al., 2021] Faber, L., K. Moghaddam, A., and Wattenhofer, R. (2021). When Comparing to Ground Truth is Wrong : On Evaluating GNN Explanation Methods. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 332–341, Virtual Event Singapore. ACM.

[Faber et al., 2020] Faber, L., Moghaddam, A. K., and Wattenhofer, R. (2020). Contrastive Graph Neural Network Explanation. Technical Report arXiv :2010.13663, arXiv.

[Fan et al., 2022] Fan, W., Jin, W., Liu, X., Xu, H., Tang, X., Wang, S., Li, Q., Tang, J., Wang, J., and Aggarwal, C. (2022). Jointly Attacking Graph Neural Network and its Explanations. Technical Report arXiv :2108.03388, arXiv.

[Fang et al., 2022] Fang, J., Liu, W., Zhang, A., Wang, X., He, X., Wang, K., and Chua, T.-S. (2022). On Regularization for Explaining Graph Neural Networks : An Information Theory Perspective.

[Fang et al., 2023] Fang, J., Wang, X., Zhang, A., Liu, Z., He, X., and Chua, T.-S. (2023). Cooperative Explanations of Graph Neural Networks. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 616–624, Singapore Singapore. ACM.

[Fang and Choromanska, 2022] Fang, S. and Choromanska, A. (2022). Backdoor Attacks on the DNN Interpretation System. Technical Report arXiv :2011.10698, arXiv.

[Feng et al., 2022a] Feng, A., You, C., Wang, S., and Tassiulas, L. (2022a). KerGNNs : Interpretable Graph Neural Networks with Graph Kernels. Technical Report arXiv :2201.00491, arXiv.

[Feng et al., 2020a] Feng, B., Wang, Y., Li, X., Yang, S., Peng, X., and Ding, Y. (2020a). SGQuant : Squeezing the Last Bit on Graph Neural Networks with Specialized Quantization. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1044–1052.

[Feng et al., 2022b] Feng, Q., Liu, N., Yang, F., Tang, R., Du, M., and Hu, X. (2022b). DEGREE : DECOMPOSITION BASED EXPLANATION FOR GRAPH NEURAL NETWORKS.

[Feng et al., 2020b] Feng, W., Zhang, J., Dong, Y., Han, Y., Luan, H., Xu, Q., Yang, Q., Kharlamov, E., and Tang, J. (2020b). Graph Random Neural Networks for Semi-Supervised Learning on Graphs. In *Advances in Neural Information Processing Systems*, volume 33, pages 22092–22103. Curran Associates, Inc.

[Feng et al., 2019] Feng, Y., You, H., Zhang, Z., Ji, R., and Gao, Y. (2019). Hypergraph Neural Networks. Technical Report arXiv :1809.09401, arXiv.

[Fernández et al., 2020] Fernández, R. R., Martín de Diego, I., Aceña, V., Fernández-Isabel, A., and Moguerza, J. M. (2020). Random forest explainability using counterfactual sets. *Information Fusion*, 63 :196–207.

[Fernández-Loría et al., 2021] Fernández-Loría, C., Provost, F., and Han, X. (2021). Explaining Data-Driven Decisions made by AI Systems : The Counterfactual Approach. Technical Report arXiv :2001.07417, arXiv.

[Fey, 2019] Fey, M. (2019). Just Jump : Dynamic Neighborhood Aggregation in Graph Neural Networks. Technical Report arXiv :1904.04849, arXiv.

[Fey and Lenssen, 2019] Fey, M. and Lenssen, J. E. (2019). Fast Graph Representation Learning with PyTorch Geometric.

[Fey et al., 2021a] Fey, M., Lenssen, J. E., Weichert, F., and Leskovec, J. (2021a). GNNAutoScale : Scalable and Expressive Graph Neural Networks via Historical Embeddings.

[Fey et al., 2021b] Fey, M., Lenssen, J. E., Weichert, F., and Leskovec, J. (2021b). GNNAutoScale : Scalable and Expressive Graph Neural Networks via Historical Embeddings.

[Fey et al., 2021c] Fey, M., Lenssen, J. E., Weichert, F., and Leskovec, J. (2021c). GNNAutoScale : Scalable and Expressive Graph Neural Networks via Historical Embeddings. In *Proceedings of the 38th International Conference on Machine Learning*, pages 3294–3304. PMLR.

[Fey et al., 2018] Fey, M., Lenssen, J. E., Weichert, F., and Müller, H. (2018). SplineCNN : Fast Geometric Deep Learning with Continuous B-Spline Kernels. Technical Report arXiv :1711.08920, arXiv.

[Fiedler and Juslin, 2005] Fiedler, K. and Juslin, P. (2005). Information Sampling and Adaptive Cognition.

[Finkelstein et al., ] Finkelstein, M., Liu, L., Schlot, N. L., Kolumbus, Y., Parkes, D. C., Rosenschein, J. S., and Keren, S. Explainable Reinforcement Learning via Model Transforms.

[Fisher et al., 2019] Fisher, A., Rudin, C., and Dominici, F. (2019). All Models are Wrong, but Many are Useful : Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. Technical Report arXiv :1801.01489, arXiv.

[Fong et al., 2019] Fong, R., Patrick, M., and Vedaldi, A. (2019). Understanding Deep Networks via Extremal Perturbations and Smooth Masks. Technical Report arXiv :1910.08485, arXiv.

[Fong and Vedaldi, 2019] Fong, R. and Vedaldi, A. (2019). Explanations for Attributing Deep Neural Network Predictions. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R., editors, *Explainable AI : Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pages 149–167. Springer International Publishing, Cham.

[Fong and Vedaldi, 2017] Fong, R. C. and Vedaldi, A. (2017). Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3449–3457, Venice. IEEE.

[Fout et al., ] Fout, A., Byrd, J., Shariat, B., and Ben-Hur, A. Protein Interface Prediction using Graph Convolutional Networks.

[Frasca et al., 2020a] Frasca, F., Rossi, E., Eynard, D., Chamberlain, B., Bronstein, M., and Monti, F. (2020a). SIGN : Scalable Inception Graph Neural Networks. Technical Report arXiv :2004.11198, arXiv.

[Frasca et al., 2020b] Frasca, F., Rossi, E., Eynard, D., Chamberlain, B., Bronstein, M., and Monti, F. (2020b). SIGN : Scalable Inception Graph Neural Networks.

[Freiesleben, 2022] Freiesleben, T. (2022). The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples. *Minds and Machines*, 32(1) :77–109.

[Freitas and Freitas, ] Freitas, A. A. and Freitas, A. A. Comprehensible Classification Models – a position paper. 15(1).

[Freitas et al., ] Freitas, S., Neil, J., Dong, Y., and Chau, D. H. A Large-Scale Database for Graph Representation Learning.

[Frey et al., 2021] Frey, C. M. M., Ma, Y., and Schubert, M. (2021). SEA : Graph Shell Attention in Graph Neural Networks. Technical Report arXiv :2110.10674, arXiv.

[Friedman, 2001] Friedman, J. H. (2001). Greedy function approximation : A gradient boosting machine. *The Annals of Statistics*, 29(5) :1189–1232.

[Frosst and Hinton, 2017] Frosst, N. and Hinton, G. (2017). Distilling a Neural Network Into a Soft Decision Tree. Technical Report arXiv :1711.09784, arXiv.

[Frye et al., 2020] Frye, C., de Mijolla, D., Begley, T., Cowton, L., Stanley, M., and Feige, I. (2020). Shapley explainability on the data manifold.

[Frye et al., ] Frye, C., Rowat, C., and Feige, I. Asymmetric Shapley values : Incorporating causal knowledge into model-agnostic explainability.

[Fu et al., 2022a] Fu, G., Zhao, P., and Bian, Y. (2022a). $p$-Laplacian Based Graph Neural Networks. In *Proceedings of the 39th International Conference on Machine Learning*, pages 6878–6917. PMLR.

[Fu et al., 2022b] Fu, Q., Ji, Y., and Huang, H. H. (2022b). TLPGNN : A Lightweight Two-Level Parallelism Paradigm for Graph Neural Network Computation on GPU. In *Proceedings of the 31st International Symposium on High-Performance Parallel and Distributed Computing*, pages 122–134, Minneapolis MN USA. ACM.

[Fu et al., 2022c] Fu, Q., Ji, Y., and Huang, H. H. (2022c). TLPGNN : A Lightweight Two-Level Parallelism Paradigm for Graph Neural Network Computation on GPU. In *Proceedings of the 31st International Symposium on High-Performance Parallel and Distributed Computing*, pages 122–134, Minneapolis MN USA. ACM.

[Fu et al., 2020a] Fu, T.-J., Wang, X. E., Peterson, M. F., Grafton, S. T., Eckstein, M. P., and Wang, W. Y. (2020a). Counterfactual Vision-and-Language Navigation via Adversarial Path Sampler. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *Computer Vision – ECCV 2020*, volume 12351, pages 71–86. Springer International Publishing, Cham.

[Fu et al., 2020b] Fu, X., Zhang, J., Meng, Z., and King, I. (2020b). MAGNN : Metapath Aggregated Graph Neural Network for Heterogeneous Graph Embedding. In *Proceedings of The Web Conference 2020*, pages 2331–2341, Taipei Taiwan. ACM.

[Funke et al., 2020] Funke, T., Khosla, M., and Anand, A. (2020). Hard Masking for Explaining Graph Neural Networks.

[Futoma et al., ] Futoma, J., Hughes, M. C., and Doshi-Velez, F. POPCORN : Partially Observed Prediction Constrained Reinforcement Learning.

[Gallicchio and Micheli, 2010] Gallicchio, C. and Micheli, A. (2010). Graph Echo State Networks. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

[Gama et al., 2018] Gama, F., Ribeiro, A., and Bruna, J. (2018). Diffusion Scattering Transforms on Graphs.

[Gama et al., 2019] Gama, F., Ribeiro, A., and Bruna, J. (2019). Stability of Graph Scattering Transforms. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[Gandhi and Iyer, ] Gandhi, S. and Iyer, A. P. P3 : Distributed Deep Graph Learning at Scale.

[Gao et al., 2019a] Gao, F., Wolf, G., and Hirn, M. (2019a). Geometric Scattering for Graph Data Analysis. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2122–2131. PMLR.

[Gao and Ji, 2019a] Gao, H. and Ji, S. (2019a). Graph Representation Learning via Hard and Channel-Wise Attention Networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 741–749, Anchorage AK USA. ACM.

[Gao and Ji, 2019b] Gao, H. and Ji, S. (2019b). Graph Representation Learning via Hard and Channel-Wise Attention Networks.

[Gao et al., 2018] Gao, H., Wang, Z., and Ji, S. (2018). Large-Scale Learnable Graph Convolutional Networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1416–1424.

[Gao et al., 2021a] Gao, J., Wang, X., Wang, Y., Yan, Y., and Xie, X. (2021a). Learning Groupwise Explanations for Black-Box Models. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 2396–2402, Montreal, Canada. International Joint Conferences on Artificial Intelligence Organization.

[Gao et al., 2022] Gao, X., Dai, W., Li, C., Xiong, H., and Frossard, P. (2022). iPool—Information-Based Pooling in Hierarchical Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9) :5032–5044.

[Gao et al., 2023] Gao, Y., Chen, R., Qin, W., Wei, L., and Zhou, C. (2023). Learning from explainable data-driven tunneling graphs : A spatio-temporal graph convolutional network for clogging detection. *Automation in Construction*, 147 :104741.

[Gao et al., 2021b] Gao, Y., Sun, T., Bhatt, R., Yu, D., Hong, S., and Zhao, L. (2021b). GNES : Learning to Explain Graph Neural Networks. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 131–140, Auckland, New Zealand. IEEE.

[Gao et al., 2019b] Gao, Y., Yang, H., Zhang, P., Zhou, C., and Hu, Y. (2019b). GraphNAS : Graph Neural Architecture Search with Reinforcement Learning.

[Gasteiger et al., 2018] Gasteiger, J., Bojchevski, A., and Günnemann, S. (2018). Predict then Propagate : Graph Neural Networks meet Personalized PageRank.

[Gasteiger et al., 2022a] Gasteiger, J., Bojchevski, A., and Günnemann, S. (2022a). Predict then Propagate : Graph Neural Networks meet Personalized PageRank. Technical Report arXiv :1810.05997, arXiv.

[Gasteiger et al., 2022b] Gasteiger, J., Giri, S., Margraf, J. T., and Günnemann, S. (2022b). Fast and Uncertainty-Aware Directional Message Passing for Non-Equilibrium Molecules. Technical Report arXiv :2011.14115, arXiv.

[Gasteiger et al., 2022c] Gasteiger, J., Groß, J., and Günnemann, S. (2022c). Directional Message Passing for Molecular Graphs. Technical Report arXiv :2003.03123, arXiv.

[Gasteiger et al., 2019] Gasteiger, J., Weiß enberger, S., and Günnemann, S. (2019). Diffusion Improves Graph Learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[Gat et al., 2022] Gat, I., Calderon, N., Reichart, R., and Hazan, T. (2022). A Functional Information Perspective on Model Interpretation. Technical Report arXiv :2206.05700, arXiv.

[Ge et al., 2021] Ge, Y., Xiao, Y., Xu, Z., Zheng, M., Karanam, S., Chen, T., Itti, L., and Wu, Z. (2021). A Peek Into the Reasoning of Neural Networks : Interpreting with Structural Visual Concepts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2195–2204, Nashville, TN, USA. IEEE.

[Geerts and Reutter, 2021] Geerts, F. and Reutter, J. L. (2021). Expressiveness and Approximation Properties of Graph Neural Networks.

[Geng et al., 2020] Geng, T., Li, A., Shi, R., Wu, C., Wang, T., Li, Y., Haghi, P., Tumeo, A., Che, S., Reinhardt, S., and Herbordt, M. C. (2020). AWB-GCN : A Graph Convolutional Network Accelerator with Runtime Workload Rebalancing. In *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pages 922–936.

[Geng et al., 2021] Geng, T., Wu, C., Zhang, Y., Tan, C., Xie, C., You, H., Herbordt, M., Lin, Y., and Li, A. (2021). I-GCN : A Graph Convolutional Network Accelerator with Runtime Locality Enhancement through Islandization. In *MICRO-54 : 54th Annual IEEE/ACM International Symposium on Microarchitecture*, pages 1051–1063, Virtual Event Greece. ACM.

[Geng et al., 2019] Geng, X., Li, Y., Wang, L., Zhang, L., Yang, Q., Ye, J., and Liu, Y. (2019). Spatiotemporal Multi-Graph Convolution Network for Ride-Hailing Demand Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01) :3656–3663.

[Geometric, 2022] Geometric, P. (2022). A Principled Approach to Aggregations.

[Geometric, 2023] Geometric, P. (2023). Graph Machine Learning Explainability with PyG.

[Gevrey et al., 2003] Gevrey, M., Dimopoulos, I., and Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological Modelling*, 160(3) :249–264.

[Ghandeharioun et al., 2022] Ghandeharioun, A., Kim, B., Li, C.-L., Jou, B., Eoff, B., and Picard, R. W. (2022). DISSECT : Disentangled Simultaneous Explanations via Concept Traversals.

[Ghazimatin et al., 2020] Ghazimatin, A., Balalau, O., Roy, R. S., and Weikum, G. (2020). PRINCE : Provider-side Interpretability with Counterfactual Explanations in Recommender Systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 196–204.

[Ghorbani et al., 2019a] Ghorbani, A., Abid, A., and Zou, J. (2019a). Interpretation of Neural Networks Is Fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01) :3681–3688.

[Ghorbani et al., 2019b] Ghorbani, A., Abid, A., and Zou, J. (2019b). Interpretation of Neural Networks Is Fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01) :3681–3688.

[Ghorbani et al., 2019c] Ghorbani, A., Wexler, J., Zou, J., and Kim, B. (2019c). Towards Automatic Concept-based Explanations.

[Ghorbani et al., 2019d] Ghorbani, A., Wexler, J., Zou, J., and Kim, B. (2019d). Towards Automatic Concept-based Explanations. Technical Report arXiv :1902.03129, arXiv.

[Ghorbani et al., ] Ghorbani, A., Wexler, J., Zou, J. Y., and Kim, B. Towards Automatic Concept-based Explanations.

[Gilmer et al., 2017a] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017a). Neural Message Passing for Quantum Chemistry.

[Gilmer et al., 2017b] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017b). Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1263–1272. PMLR.

[Glennan, 2002] Glennan, S. (2002). Rethinking Mechanistic Explanation. *Philosophy of Science*, 69(S3) :S342–S353.

[Glorot and Bengio, ] Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks.

[Gödel, ] Gödel, K. On Formally Undecidable Propositions of Principia Mathematica And Related Systems.

[Goh et al., 2021] Goh, G. S. W., Lapuschkin, S., Weber, L., Samek, W., and Binder, A. (2021). Understanding Integrated Gradients with SmoothTaylor for Deep Neural Network Attribution. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4949–4956.

[Goldstone et al., 2016] Goldstone, A., Mayhew, S. D., Przezdzik, I., Wilson, R. S., Hale, J. R., and Bagshaw, A. P. (2016). Gender Specific Re-organization of Resting-State Networks in Older Age. *Frontiers in Aging Neuroscience*, 8.

[Gomez et al., 2020] Gomez, O., Holter, S., Yuan, J., and Bertini, E. (2020). ViCE : Visual Counterfactual Explanations for Machine Learning Models. Technical Report arXiv :2003.02428, arXiv.

[Gómez-Bombarelli et al., 2018] Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2) :268–276.

[Gong et al., 2009] Gong, G., Rosa-Neto, P., Carbonell, F., Chen, Z. J., He, Y., and Evans, A. C. (2009). Age- and Gender-Related Differences in the Cortical Anatomical Network. *Journal of Neuroscience*, 29(50) :15684–15693.

[Gong et al., 2022] Gong, Z., Ji, H., Yao, Y., Fletcher, C. W., Hughes, C. J., and Torrellas, J. (2022). Graphite : Optimizing graph neural networks on CPUs through cooperative software-hardware techniques. In *Proceedings of the 49th Annual International Symposium on Computer Architecture*, pages 916–931, New York New York. ACM.

[Goodfellow et al., 2015] Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples.

[Goodman and Flaxman, 2017] Goodman, B. and Flaxman, S. (2017). European Union Regulations on Algorithmic Decision Making and a "Right to Explanation". *AI Magazine*, 38(3) :50–57.

[Gori et al., 2005] Gori, M., Monfardini, G., and Scarselli, F. (2005). A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734 vol. 2.

[Goyal et al., 2019] Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., and Lee, S. (2019). Counterfactual Visual Explanations. Technical Report arXiv :1904.07451, arXiv.

[Grattarola and Alippi, 2020] Grattarola, D. and Alippi, C. (2020). Graph Neural Networks in TensorFlow and Keras with Spektral.

[Gravina et al., 2022] Gravina, A., Bacciu, D., and Gallicchio, C. (2022). Anti-Symmetric DGN : A stable architecture for Deep Graph Networks.

[Grover and Leskovec, 2016] Grover, A. and Leskovec, J. (2016). Node2vec : Scalable Feature Learning for Networks. Technical Report arXiv :1607.00653, arXiv.

[Gu and Dong, 2021] Gu, J. and Dong, C. (2021). Interpreting Super-Resolution Networks with Local Attribution Maps. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9195–9204, Nashville, TN, USA. IEEE.

[Guerdan et al., 2021] Guerdan, L., Raymond, A., and Gunes, H. (2021). Toward Affective XAI : Facial Affect Analysis for Understanding Explainable Human-AI Interactions. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3789–3798, Montreal, BC, Canada. IEEE.

[Guerrero et al., 2018] Guerrero, P., Kleiman, Y., Ovsjanikov, M., and Mitra, N. J. (2018). PCPNET : Learning Local Shape Properties from Raw Point Clouds. *Computer Graphics Forum*, 37(2) :75–85.

[Gui et al., 2023] Gui, S., Yuan, H., Wang, J., Lao, Q., Li, K., and Ji, S. (2023). FlowX : Towards Explainable Graph Neural Networks via Message Flows.

[Guidotti, 2022] Guidotti, R. (2022). Counterfactual explanations and how to find them : Literature review and benchmarking. *Data Mining and Knowledge Discovery.*

[Guidotti et al., 2019] Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., and Turini, F. (2019). Factual and Counterfactual Explanations for Black Box Decision Making. *IEEE Intelligent Systems*, 34(6) :14–23.

[Gulcehre et al., 2018] Gulcehre, C., Denil, M., Malinowski, M., Razavi, A., Pascanu, R., Hermann, K. M., Battaglia, P., Bapst, V., Raposo, D., Santoro, A., and de Freitas, N. (2018). Hyperbolic Attention Networks.

[Guo et al., 2022] Guo, K., Zhou, K., Hu, X., Li, Y., Chang, Y., and Wang, X. (2022). Orthogonal Graph Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4) :3996–4004.

[Guo et al., 2018] Guo, M., Chou, E., Huang, D.-A., Song, S., Yeung, S., and Fei-Fei, L. (2018). Neural Graph Matching Networks for Fewshot 3D Action Recognition. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, volume 11205, pages 673–689. Springer International Publishing, Cham.

[Guo et al., 2019] Guo, S., Lin, Y., Feng, N., Song, C., and Wan, H. (2019). Attention Based Spatial-Temporal Graph Convolutional Networks for Traffic Flow Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01) :922–929.

[Guo and Pagnoni, 2008] Guo, Y. and Pagnoni, G. (2008). A unified framework for group independent component analysis for multi-subject fMRI data. *NeuroImage*, 42(3) :1078–1093.

[Gurumoorthy et al., 2019] Gurumoorthy, K. S., Dhurandhar, A., Cecchi, G., and Aggarwal, C. (2019). Efficient Data Representation by Selecting Prototypes with Importance Weights. Technical Report arXiv :1707.01212, arXiv.

[Gururangan et al., 2018] Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (2018). Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

[Hagmann, 2005] Hagmann, P. (2005). *From Diffusion MRI to Brain Connectomics*. PhD thesis, EPFL, Lausanne.

[Ham et al., 2016] Ham, B., Cho, M., Schmid, C., and Ponce, J. (2016). Proposal Flow. Technical Report arXiv :1511.05065, arXiv.

[Hamilton et al., ] Hamilton, W., Ying, Z., and Leskovec, J. Inductive Representation Learning on Large Graphs.

[Hamilton et al., 2017] Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[Hamilton et al., 2018] Hamilton, W. L., Ying, R., and Leskovec, J. (2018). Inductive Representation Learning on Large Graphs. Technical Report arXiv :1706.02216, arXiv.

[Han et al., ] Han, T., Srinivas, S., and Lakkaraju, H. Which Explanation Should I Choose ? A Function Approximation Perspective to Characterizing Post Hoc Explanations.

[Han et al., 2021] Han, T., Tu, W.-W., and Li, Y.-F. (2021). Explanation Consistency Training : Facilitating Consistency-Based Semi-Supervised Learning with Interpretability. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9) :7639–7646.

[Han and Zhao, 2022] Han, X. and Zhao, Y. (2022). Interpretable Graph Reservoir Computing With the Temporal Pattern Attention. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15.

[Hancox-Li, 2020] Hancox-Li, L. (2020). Robustness in machine learning explanations : Does it matter ? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 640–647, Barcelona Spain. ACM.

[Hansen and Rieger, 2019] Hansen, L. K. and Rieger, L. (2019). Interpretability in Intelligent Systems – A New Concept ? In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R., editors, *Explainable AI : Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pages 41–49. Springer International Publishing, Cham.

[Hasanzadeh et al., 2020] Hasanzadeh, A., Hajiramezanali, E., Boluki, S., Zhou, M., Duffield, N., Narayanan, K., and Qian, X. (2020). Bayesian Graph Neural Net-

works with Adaptive Connection Sampling. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4094–4104. PMLR.

[Hase and Bansal, 2022] Hase, P. and Bansal, M. (2022). When Can Models Learn From Explanations ? A Formal Framework for Understanding the Roles of Explanation Data. In *Proceedings of the First Workshop on Learning with Natural Language Supervision*, pages 29–39, Dublin, Ireland. Association for Computational Linguistics.

[Hashemi and Fathi, 2020] Hashemi, M. and Fathi, A. (2020). PermuteAttack : Counterfactual Explanation of Machine Learning Credit Scorecards. Technical Report arXiv :2008.10138, arXiv.

[Hassani and Khasahmadi, 2020] Hassani, K. and Khasahmadi, A. H. (2020). Contrastive Multi-View Representation Learning on Graphs. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4116–4126. PMLR.

[Hastad, 1986] Hastad, J. (1986). Almost optimal lower bounds for small depth circuits. In *Proceedings of the Eighteenth Annual ACM Symposium on Theory of Computing - STOC '86*, pages 6–20, Berkeley, California, United States. ACM Press.

[Håstad and Goldmann, 1991] Håstad, J. and Goldmann, M. (1991). On the power of small-depth threshold circuits. *computational complexity*, 1(2) :113–129.

[He et al., 2022a] He, D., Liang, C., Liu, H., Wen, M., Jiao, P., and Feng, Z. (2022a). Block Modeling-Guided Graph Convolutional Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4) :4022–4029.

[He et al., 2022b] He, H., Ji, Y., and Huang, H. H. (2022b). Illuminati : Towards Explaining Graph Neural Networks for Cybersecurity Analysis. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 74–89.

[He et al., 2021a] He, M., Wei, Z., Huang, Z., and Xu, H. (2021a). BernNet : Learning Arbitrary Graph Spectral Filters via Bernstein Approximation.

[He et al., 2021b] He, T., Ong, Y.-S., and Bai, L. (2021b). Learning Conjoint Attentions for Graph Neural Nets.

[He et al., 2022c] He, W., Vu, M. N., Jiang, Z., and Thai, M. T. (2022c). An Explainer for Temporal Graph Neural Networks. Technical Report arXiv :2209.00807, arXiv.

[He et al., 2020] He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., and Wang, M. (2020). LightGCN : Simplifying and Powering Graph Convolution Network for Recommendation. Technical Report arXiv :2002.02126, arXiv.

[He et al., 2021c] He, Y., Wang, Y., Liu, C., Li, H., and Li, X. (2021c). TARe : Task-Adaptive in-situ ReRAM Computing for Graph Learning. In *2021 58th ACM/IEEE Design Automation Conference (DAC)*, pages 577–582.

[Hebb, 1949] Hebb, D. O. (1949). *The Organization of Behavior ; a Neuropsychological Theory*. The Organization of Behavior ; a Neuropsychological Theory. Wiley, Oxford, England.

[Hempel and Oppenheim, 1948] Hempel, C. G. and Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, 15(2) :135–175.

[Henderson et al., 2021] Henderson, R., Clevert, D.-A., and Montanari, F. (2021). Improving Molecular Graph Neural Network Explainability with Orthonormalization and Induced Sparsity. Technical Report arXiv :2105.04854, arXiv.

[Hendricks et al., 2018a] Hendricks, L. A., Hu, R., Darrell, T., and Akata, Z. (2018a). Generating Counterfactual Explanations with Natural Language. Technical Report arXiv :1806.09809, arXiv.

[Hendricks et al., 2018b] Hendricks, L. A., Hu, R., Darrell, T., and Akata, Z. (2018b). Grounding Visual Explanations. Technical Report arXiv :1807.09685, arXiv.

[Herath et al., 2022] Herath, J. D., Wakodikar, P. P., Yang, P., and Yan, G. (2022). CFGExplainer : Explaining Graph Neural Network-Based Malware Classification from Control Flow Graphs. In *2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 172–184, Baltimore, MD, USA. IEEE.

[Himmelhuber et al., ] Himmelhuber, A., Grimm, S., Zillner, S., Ringsquandl, M., Joblin, M., and Runkler, T. A New Concept for Explaining Graph Neural Networks.

[Hinton, 2012] Hinton, G. (2012). Rmsprop_hinton_2012.

[Hinton et al., 2015] Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the Knowledge in a Neural Network.

[Hinton et al., 2006] Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7) :1527–1554.

[Hoernle et al., 2020] Hoernle, N., Gal, K., Grosz, B., Lyons, L., Ren, A., and Rubin, A. (2020). Interpretable Models for Understanding Immersive Simulations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 2319–2325, Yokohama, Japan. International Joint Conferences on Artificial Intelligence Organization.

[Hoffman et al., 2018] Hoffman, R., Miller, T., Mueller, S. T., Klein, G., and Clancey, W. J. (2018). Explaining Explanation, Part 4 : A Deep Dive on Deep Nets. *IEEE Intelligent Systems*, 33(3) :87–95.

[Hoffman and Klein, 2017] Hoffman, R. R. and Klein, G. (2017). Explaining Explanation, Part 1 : Theoretical Foundations. *IEEE Intelligent Systems*, 32(3) :68–73.

[Hohman et al., 2020] Hohman, F., Conlen, M., Heer, J., and Chau, D. H. P. (2020). Communicating with Interactive Articles. *Distill*, 5(9) :e28.

[Hohman et al., 2019] Hohman, F., Kahng, M., Pienta, R., and Chau, D. H. (2019). Visual Analytics in Deep Learning : An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 25(8) :2674–2693.

[Holdijk et al., 2021] Holdijk, L., Boon, M., Henckens, S., and de Jong, L. (2021). [Re] Parameterized Explainer for Graph Neural Network.

[Holzinger et al., 2021] Holzinger, A., Malle, B., Saranti, A., and Pfeifer, B. (2021). Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI. *Information Fusion*, 71 :28–37.

[Homberg et al., 2023] Homberg, S., Janosch, M., Morris, G. M., and Koch, O. (2023). Interpreting Graph Neural Networks with Myerson Values for Cheminformatics Approaches.

[Hong et al., 2019] Hong, S., Yang, D., Choi, J., and Lee, H. (2019). Interpretable Text-to-Image Synthesis with Hierarchical Semantic Layout Generation. In Sa-

mek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R., editors, *Explainable AI : Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pages 77–95. Springer International Publishing, Cham.

[Hooker et al., ] Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B. A Benchmark for Interpretability Methods in Deep Neural Networks.

[Horn, 1987] Horn, B. K. P. (1987). Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America A*, 4(4) :629.

[Hornik, 1991] Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2) :251–257.

[Hoshen, 2018] Hoshen, Y. (2018). VAIN : Attentional Multi-agent Predictive Modeling. Technical Report arXiv :1706.06122, arXiv.

[Hou et al., 2019] Hou, Y., Zhang, J., Cheng, J., Ma, K., Ma, R. T. B., Chen, H., and Yang, M.-C. (2019). Measuring and Improving the Use of Graph Information in Graph Neural Networks.

[Hoyt, 2023] Hoyt, C. T. (2023). Class Resolver.

[Hsieh et al., 2021] Hsieh, C.-Y., Yeh, C.-K., Liu, X., Ravikumar, P., Kim, S., Kumar, S., and Hsieh, C.-J. (2021). Evaluations and Methods for Explanation through Robustness Analysis. Technical Report arXiv :2006.00442, arXiv.

[Hu et al., 2020a] Hu, J., Li, T., and Dong, S. (2020a). GCN-LRP explanation : Exploring latent attention of graph convolutional networks. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

[Hu et al., ] Hu, R., Chau, S. L., Huertas, J. F., and Sejdinovic, D. Explaining Preferences with Shapley Values.

[Hu et al., 2020b] Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. (2020b). Open Graph Benchmark : Datasets for Machine Learning on Graphs. In *Advances in Neural Information Processing Systems*, volume 33, pages 22118–22133. Curran Associates, Inc.

[Hu et al., 2021] Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. (2021). Open Graph Benchmark : Datasets for Machine Learning on Graphs. Technical Report arXiv :2005.00687, arXiv.

[Hu* et al., 2019] Hu*, W., Liu*, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. (2019). Strategies for Pre-training Graph Neural Networks.

[Hu et al., 2020c] Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. (2020c). Strategies for Pre-training Graph Neural Networks. Technical Report arXiv :1905.12265, arXiv.

[Hu et al., 2019] Hu, W., Miyato, T., Tokui, S., Matsumoto, E., and Sugiyama, M. (2019). Unsupervised Discrete Representation Learning. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R., editors, *Explainable AI : Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pages 97–119. Springer International Publishing, Cham.

[Hu et al., 2020d] Hu, Y., Ye, Z., Wang, M., Yu, J., Zheng, D., Li, M., Zhang, Z., Zhang, Z., and Wang, Y. (2020d). FeatGraph : A Flexible and Efficient Backend for Graph Neural Network Systems. Technical Report arXiv :2008.11359, arXiv.

[Hu et al., 2020e] Hu, Z., Dong, Y., Wang, K., and Sun, Y. (2020e). Heterogeneous Graph Transformer. In *Proceedings of The Web Conference 2020*, pages 2704–2710, Taipei Taiwan. ACM.

[Huang et al., 2020a] Huang, G., Dai, G., Wang, Y., and Yang, H. (2020a). GE-SpMM : General-purpose Sparse Matrix-Matrix Multiplication on GPUs for Graph Neural Networks.

[Huang et al., 2021] Huang, K., Zhai, J., Zheng, Z., Yi, Y., and Shen, X. (2021). Understanding and bridging the gaps in current GNN performance optimizations. In *Proceedings of the 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, pages 119–132, Virtual Event Republic of Korea. ACM.

[Huang et al., 2022a] Huang, L., Zhang, Z., Du, Z., Li, S., Zheng, H., Xie, Y., and Tan, N. (2022a). EPQuant : A Graph Neural Network compression approach based on product quantization. *Neurocomputing*, 503 :49–61.

[Huang et al., 2020b] Huang, Q., He, H., Singh, A., Lim, S.-N., and Benson, A. R. (2020b). Combining Label Propagation and Simple Models Out-performs Graph Neural Networks. Technical Report arXiv :2010.13993, arXiv.

[Huang et al., 2022b] Huang, Q., Yamada, M., Tian, Y., Singh, D., and Chang, Y. (2022b). GraphLIME : Local Interpretable Model Explanations for Graph Neural Networks. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–6.

[Huang et al., 2020c] Huang, Q., Yamada, M., Tian, Y., Singh, D., Yin, D., and Chang, Y. (2020c). GraphLIME : Local Interpretable Model Explanations for Graph Neural Networks. Technical Report arXiv :2001.06216, arXiv.

[Huang et al., 2018] Huang, W., Zhang, T., Rong, Y., and Huang, J. (2018). Adaptive Sampling Towards Fast Graph Representation Learning. Technical Report arXiv :1809.05343, arXiv.

[Huang et al., 2023] Huang, X., Yang, Y., Wang, Y., Wang, C., Zhang, Z., Xu, J., Chen, L., and Vazirgiannis, M. (2023). DGraph : A Large-Scale Financial Dataset for Graph Anomaly Detection. Technical Report arXiv :2207.03579, arXiv.

[Huang and Kong, 2022] Huang, Y. and Kong, A. W.-K. (2022). TRANSFERABLE ADVERSARIAL ATTACK BASED ON INTEGRATED GRADIENTS.

[Huang et al., 2022c] Huang, Y., Zheng, L., Yao, P., Wang, Q., Liao, X., Jin, H., and Xue, J. (2022c). Accelerating Graph Convolutional Networks Using Crossbar-based Processing-In-Memory Architectures. In *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 1029–1042.

[Hurley and Rickard, 2009] Hurley, N. and Rickard, S. (2009). Comparing Measures of Sparsity. *IEEE Transactions on Information Theory*, 55(10) :4723–4741.

[Huysmans et al., 2011] Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., and Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1) :141–154.

[Hüyük et al., 2020] Hüyük, A., Jarrett, D., Tekin, C., and van der Schaar, M. (2020). Explaining by Imitating : Understanding Decisions by Interpretable Policy Learning.

[Hyvarinen, 1999] Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3) :626–634.

[Iakovlev et al., 2020] Iakovlev, V., Heinonen, M., and Lähdesmäki, H. (2020). Learning continuous-time PDEs from sparse data with graph neural networks.

[Inouye et al., ] Inouye, D. I., Leqi, L., Kim, J. S., Aragam, B., and Ravikumar, P. Automated Dependence Plots.

[Ioffe and Szegedy, 2015] Ioffe, S. and Szegedy, C. (2015). Batch Normalization : Accelerating Deep Network Training by Reducing Internal Covariate Shift.

[Ismail et al., ] Ismail, A. A., Feizi, S., and Bravo, H. C. Improving Deep Learning Interpretability by Saliency Guided Training.

[Isufi et al., 2022] Isufi, E., Gama, F., and Ribeiro, A. (2022). EdgeNets : Edge Varying Graph Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11) :7457–7473.

[Izza and Marques-Silva, 2021] Izza, Y. and Marques-Silva, J. (2021). On Explaining Random Forests with SAT. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 2584–2591, Montreal, Canada. International Joint Conferences on Artificial Intelligence Organization.

[Jacovi and Goldberg, 2020a] Jacovi, A. and Goldberg, Y. (2020a). Towards Faithfully Interpretable NLP Systems : How should we define and evaluate faithfulness ?

[Jacovi and Goldberg, 2020b] Jacovi, A. and Goldberg, Y. (2020b). Towards Faithfully Interpretable NLP Systems : How Should We Define and Evaluate Faithfulness ? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

[Jacovi and Goldberg, 2021a] Jacovi, A. and Goldberg, Y. (2021a). Aligning Faithful Interpretations with their Social Attribution. *Transactions of the Association for Computational Linguistics*, 9 :294–310.

[Jacovi and Goldberg, 2021b] Jacovi, A. and Goldberg, Y. (2021b). Aligning Faithful Interpretations with their Social Attribution. *Transactions of the Association for Computational Linguistics*, 9 :294–310.

[Janizek et al., 2020] Janizek, J. D., Sturmfels, P., and Lee, S.-I. (2020). Explaining Explanations : Axiomatic Feature Interactions for Deep Networks. Technical Report arXiv :2002.04138, arXiv.

[Janzing et al., ] Janzing, D., Minorics, L., and Blobaum, P. Feature relevance quantification in explainable AI : A causal problem.

[Jaume et al., 2021] Jaume, G., Pati, P., Bozorgtabar, B., Foncubierta-Rodríguez, A., Feroce, F., Anniciello, A. M., Rau, T., Thiran, J.-P., Gabrani, M., and Goksel, O. (2021). Quantifying Explainers of Graph Neural Networks in Computational Pathology. Technical Report arXiv :2011.12646, arXiv.

[Jaume et al., 2020] Jaume, G., Pati, P., Foncubierta-Rodriguez, A., Feroce, F., Scognamiglio, G., Anniciello, A. M., Thiran, J.-P., Goksel, O., and Gabrani, M. (2020). Towards Explainable Graph Representations in Digital Pathology. Technical Report arXiv :2007.00311, arXiv.

[Jaynes et al., 2003] Jaynes, E. T., Bretthorst, G. L., and EBSCO Publishing (2003). *Probability Theory : The Logic of Science*. Cambridge University Press, Cambridge.

[Jeanneret et al., 2022] Jeanneret, G., Simon, L., and Jurie, F. (2022). Diffusion Models for Counterfactual Explanations. Technical Report arXiv :2203.15636, arXiv.

[Jeong et al., ] Jeong, H., Lee, S., Hwang, S. J., and Son, S. Learning to Generate Inversion-Resistant Model Explanations.

[Jethani et al., 2022] Jethani, N., Sudarshan, M., Covert, I., Lee, S.-I., and Ranganath, R. (2022). FastSHAP : Real-Time Shapley Value Estimation.

[Jeyakumar et al., ] Jeyakumar, J. V., Noor, J., Cheng, Y.-H., Garcia, L., and Srivastava, M. How Can I Explain This to You ? An Empirical Study of Deep Neural Network Explanation Methods.

[Ji et al., 2022] Ji, C., Wang, R., and Wu, H. (2022). Perturb more, trap more : Understanding behaviors of graph neural networks. *Neurocomputing*, 493 :59–75.

[Jia et al., ] Jia, Z., Lin, S., Gao, M., Zaharia, M., and Aiken, A. Improving the Accuracy, Scalability, and Performance of Graph Neural Networks with Roc.

[Jiang and Li, 2022] Jiang, B. and Li, Z. (2022). Defending Against Backdoor Attack on Graph Nerual Network by Explainability. Technical Report arXiv :2209.02902, arXiv.

[Jiang et al., 2022] Jiang, J., Leofante, F., Rago, A., and Toni, F. (2022). Formalising the Robustness of Counterfactual Explanations for Neural Networks. Technical Report arXiv :2208.14878, arXiv.

[Jiménez-Luna et al., 2021] Jiménez-Luna, J., Skalic, M., Weskamp, N., and Schneider, G. (2021). Coloring Molecules with Explainable Artificial Intelligence for Preclinical Relevance Assessment. *Journal of Chemical Information and Modeling*, 61(3) :1083–1094.

[Jin et al., 2021] Jin, D., Yu, Z., Huo, C., Wang, R., Wang, X., He, D., and Han, J. (2021). Universal Graph Convolutional Networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 10654–10664. Curran Associates, Inc.

[Jin et al., 2022] Jin, W., Li, X., and Hamarneh, G. (2022). Evaluating Explainable AI on a Multi-Modal Medical Imaging Task : Can Existing Algorithms Fulfill Clinical Requirements ? *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11) :11945–11953.

[Jin et al., 2020a] Jin, W., Qu, M., Jin, X., and Ren, X. (2020a). Recurrent Event Network : Autoregressive Structure Inference over Temporal Knowledge Graphs. Technical Report arXiv :1904.05530, arXiv.

[Jin et al., 2020b] Jin, X., Wei, Z., Du, J., Xue, X., and Ren, X. (2020b). Towards Hierarchical Importance Attribution : Explaining Compositional Semantics for Neural Sequence Models.

[Johnson, 2017] Johnson, D. D. (2017). LEARNING GRAPHICAL STATE TRANSITIONS.

[Johnson et al., 2018] Johnson, J., Gupta, A., and Fei-Fei, L. (2018). Image Generation from Scene Graphs. Technical Report arXiv :1804.01622, arXiv.

[Joshi et al., 2022] Joshi, C. K., Liu, F., Xun, X., Lin, J., and Foo, C.-S. (2022). On Representation Knowledge Distillation for Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12.

[Joshi et al., 2019] Joshi, S., Koyejo, O., Vijitbenjaronk, W., Kim, B., and Ghosh, J. (2019). Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems. Technical Report arXiv :1907.09615, arXiv.

[Kaler et al., 2022] Kaler, T., Stathas, N., Ouyang, A., Iliopoulos, A.-S., Schardl, T. B., Leiserson, C. E., and Chen, J. (2022). Accelerating Training and Inference of Graph Neural Networks with Fast Sampling and Pipelining.

[Kamal et al., 2023] Kamal, A., Vincent, E., Plantevit, M., and Robardet, C. (2023). Improving the Quality of Rule-Based GNN Explanations. In Koprinska, I., Mignone, P., Guidotti, R., Jaroszewicz, S., Fröning, H., Gullo, F., Ferreira, P. M., Roqueiro, D., Ceddia, G., Nowaczyk, S., Gama, J., Ribeiro, R., Gavaldà, R., Masciari, E., Ras, Z., Ritacco, E., Naretto, F., Theissler, A., Biecek, P., Verbeke, W., Schiele, G., Pernkopf, F., Blott, M., Bordino, I., Danesi, I. L., Ponti, G., Severini, L., Appice, A., Andresini, G., Medeiros, I., Graça, G., Cooper, L., Ghazaleh, N., Richiardi, J., Saldana, D., Sechidis, K., Canakoglu, A., Pido, S., Pinoli, P., Bifet, A., and Pashami, S., editors, *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, volume 1752, pages 467–482. Springer Nature Switzerland, Cham.

[Kanamori et al., 2020] Kanamori, K., Takagi, T., Kobayashi, K., and Arimura, H. (2020). DACE : Distribution-Aware Counterfactual Explanation by Mixed-Integer Linear Optimization. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 2855–2862, Yokohama, Japan. International Joint Conferences on Artificial Intelligence Organization.

[Kanamori et al., ] Kanamori, K., Takagi, T., Kobayashi, K., and Ike, Y. Counterfactual Explanation Trees : Transparent and Consistent Actionable Recourse with Decision Trees.

[Kanehira et al., 2019] Kanehira, A., Takemoto, K., Inayoshi, S., and Harada, T. (2019). Multimodal Explanations by Predicting Counterfactuality in Videos. Technical Report arXiv :1812.01263, arXiv.

[Kapishnikov et al., 2019] Kapishnikov, A., Bolukbasi, T., Viegas, F., and Terry, M. (2019). XRAI : Better Attributions Through Regions. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4947–4956, Seoul, Korea (South). IEEE.

[Karimi et al., 2020] Karimi, A.-H., Barthe, G., Balle, B., and Valera, I. (2020). Model-Agnostic Counterfactual Explanations for Consequential Decisions.

[Karimi et al., 2021] Karimi, A.-H., Barthe, G., Schölkopf, B., and Valera, I. (2021). A survey of algorithmic recourse : Definitions, formulations, solutions, and prospects.

[Karimi et al., 2023] Karimi, A.-H., Barthe, G., Schölkopf, B., and Valera, I. (2023). A Survey of Algorithmic Recourse : Contrastive Explanations and Consequential Recommendations. *ACM Computing Surveys*, 55(5) :1–29.

[Kasanishi et al., 2021] Kasanishi, T., Wang, X., and Yamasaki, T. (2021). Edge-Level Explanations for Graph Neural Networks by Extending Explainability Methods for Convolutional Neural Networks. Technical Report arXiv :2111.00722, arXiv.

[Kaur et al., 2020] Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Wortman Vaughan, J. (2020). Interpreting Interpretability : Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Honolulu HI USA. ACM.

[Kaushik et al., 2020] Kaushik, D., Hovy, E., and Lipton, Z. C. (2020). LEARNING THE DIFFERENCE THAT MAKES A DIFFER- ENCE WITH COUNTERFACTUALLY-AUGMENTED DATA.

[Kazi et al., 2021] Kazi, A., Farghadani, S., and Navab, N. (2021). IA-GCN : Interpretable Attention based Graph Convolutional Network for Disease prediction. Technical Report arXiv :2103.15587, arXiv.

[Keane and Smyth, 2020] Keane, M. T. and Smyth, B. (2020). Good Counterfactuals and Where to Find Them : A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI). Technical Report arXiv :2005.13997, arXiv.

[Kearnes et al., 2016] Kearnes, S., McCloskey, K., Berndl, M., Pande, V., and Riley, P. (2016). Molecular Graph Convolutions : Moving Beyond Fingerprints. *Journal of Computer-Aided Molecular Design*, 30(8) :595–608.

[Keil, 2006] Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, 57 :227–254.

[Khakzar et al., 2022] Khakzar, A., Khorsandi, P., Nobahari, R., and Navab, N. (2022). Do Explanations Explain ? Model Knows Best. In *2022 IEEE/CVF*

*Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10234–10243, New Orleans, LA, USA. IEEE.

[Khasahmadi et al., 2020] Khasahmadi, A. H., Hassani, K., Moradi, P., Lee, L., and Morris, Q. (2020). Memory-Based Graph Networks. Technical Report arXiv :2002.09518, arXiv.

[Khorram and Fuxin, 2022] Khorram, S. and Fuxin, L. (2022). Cycle-Consistent Counterfactuals by Latent Transformations. Technical Report arXiv :2203.15064, arXiv.

[Kim et al., 2016a] Kim, B., Khanna, R., and Koyejo, O. O. (2016a). Examples are not enough, learn to criticize ! Criticism for Interpretability. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

[Kim et al., 2016b] Kim, B., Khanna, R., and Koyejo, O. O. (2016b). Examples are not enough, learn to criticize ! Criticism for Interpretability. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

[Kim et al., ] Kim, B., Rudin, C., and Shah, J. A. The Bayesian Case Model : A Generative Approach for Case-Based Reasoning and Prototype Classification.

[Kim et al., 2018a] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and Sayres, R. (2018a). Interpretability Beyond Feature Attribution : Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, pages 2668–2677. PMLR.

[Kim et al., 2018b] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and Sayres, R. (2018b). Interpretability Beyond Feature Attribution : Quantitative Testing with Concept Activation Vectors (TCAV). Technical Report arXiv :1711.11279, arXiv.

[Kim and Oh, 2020] Kim, D. and Oh, A. (2020). How to Find Your Friendly Neighborhood : Graph Attention Design with Self-Supervision.

[Kim et al., 2020] Kim, H., Shin, S., Jang, J., Song, K., Joo, W., Kang, W., and Moon, I.-C. (2020). Counterfactual Fairness with Disentangled Causal Effect Variational Autoencoder. Technical Report arXiv :2011.11878, arXiv.

[Kim et al., 2022] Kim, J. S., Plumb, G., and Talwalkar, A. (2022). Sanity Simulations for Saliency Methods.

[Kindermans et al., 2017] Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. (2017). The (Un)reliability of saliency methods.

[Kindermans et al., 2019] Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. (2019). The (Un)reliability of Saliency Methods. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R., editors, *Explainable AI : Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pages 267–280. Springer International Publishing, Cham.

[Kindermans et al., 2016] Kindermans, P.-J., Schütt, K., Müller, K.-R., and Dähne, S. (2016). Investigating the influence of noise and distractors on the interpretation of neural networks. Technical Report arXiv :1611.07270, arXiv.

[Kingma and Ba, 2017] Kingma, D. P. and Ba, J. (2017). Adam : A Method for Stochastic Optimization.

[Kiningham et al., 2020] Kiningham, K., Re, C., and Levis, P. (2020). GRIP : A Graph Neural Network Accelerator Architecture.

[Kipf and Welling, 2016a] Kipf, T. N. and Welling, M. (2016a). Semi-Supervised Classification with Graph Convolutional Networks.

[Kipf and Welling, 2016b] Kipf, T. N. and Welling, M. (2016b). Variational Graph Auto-Encoders. Technical Report arXiv :1611.07308, arXiv.

[Kipf and Welling, 2017] Kipf, T. N. and Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. *International Conference on Learning Representations*.

[Kiviniemi et al., 2009] Kiviniemi, V., Starck, T., Remes, J., Long, X., Nikkinen, J., Haapea, M., Veijola, J., Moilanen, I., Isohanni, M., Zang, Y.-F., and Tervonen, O. (2009). Functional segmentation of the brain cortex using high model order group PICA. *Human Brain Mapping*, 30(12) :3865–3886.

[Klein, 1893] Klein, F. (1893). Vergleichende Betrachtungen über neuere geometrische Forschungen. *Mathematische Annalen*, 43(1) :63–100.

[Klein, 2018] Klein, G. (2018). Explaining Explanation, Part 3 : The Causal Landscape. *IEEE Intelligent Systems*, 33(2) :83–88.

[Kleinberg et al., 2016] Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores.

[Knyazev et al., 2019] Knyazev, B., Taylor, G. W., and Amer, M. R. (2019). Understanding Attention and Generalization in Graph Neural Networks. Technical Report arXiv :1905.02850, arXiv.

[Ko et al., ] Ko, C.-Y., Mohapatra, J., Liu, S., Chen, P.-Y., Daniel, L., and Weng, T.-W. Revisiting Contrastive Learning through the Lens of Neighborhood Component Analysis : An Integrated Framework.

[Koh and Liang, 2020a] Koh, P. W. and Liang, P. (2020a). Understanding Black-box Predictions via Influence Functions.

[Koh and Liang, 2020b] Koh, P. W. and Liang, P. (2020b). Understanding Black-box Predictions via Influence Functions. Technical Report arXiv :1703.04730, arXiv.

[Korikov et al., 2021] Korikov, A., Shleyfman, A., and Beck, J. C. (2021). Counterfactual Explanations for Optimization-Based Decisions in the Context of the GDPR. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 4097–4103, Montreal, Canada. International Joint Conferences on Artificial Intelligence Organization.

[Kowal, ] Kowal, D. R. Bayesian subset selection and variable importance for interpretable prediction and classification.

[Krause et al., 2016] Krause, J., Perer, A., and Ng, K. (2016). Interacting with Predictions : Visual Inspection of Black-box Machine Learning Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5686–5697, San Jose California USA. ACM.

[Kreuzer et al., 2021] Kreuzer, D., Beaini, D., Hamilton, W. L., Létourneau, V., and Tossou, P. (2021). Rethinking Graph Transformers with Spectral Attention.

[Krizhevsky et al., 2017] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6) :84–90.

[Kuang et al., 2021] Kuang, W., Wang, Z., Li, Y., Wei, Z., and Ding, B. (2021). Coarformer : Transformer for large graph via graph coarsening.

[Kulesza et al., 2013] Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., and Wong, W.-K. (2013). Too much, too little, or just right ? Ways explanations impact end users' mental models. *Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC*, pages 3–10.

[Kumar et al., ] Kumar, I. E., Scheidegger, C., Venkatasubramanian, S., and Friedd-ler, S. A. Shapley Residuals : Quantifying the limits of the Shapley value for explanations.

[Kumar et al., 2019] Kumar, S., Zhang, X., and Leskovec, J. (2019). Predicting Dynamic Embedding Trajectory in Temporal Interaction Networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1269–1278, Anchorage AK USA. ACM.

[Kusner et al., 2018] Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. (2018). Counterfactual Fairness.

[L et al., 2014] L, G., G, S.-K., Cf, B., Ej, A., G, D., Ce, S., E, Z., Kp, E., N, F., Ce, M., S, M., J, X., E, Y., G, B., K, U., Kl, M., and Sm, S. (2014). ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *NeuroImage*, 95.

[La Malfa et al., 2021] La Malfa, E., Zbrzezny, A., Michelmore, R., Paoletti, N., and Kwiatkowska, M. (2021). On Guaranteed Optimal Robust Explanations for NLP Models. Technical report.

[Laird et al., 2023] Laird, E., Madushanka, A., Kraka, E., and Clark, C. (2023). XInsight : Revealing Model Insights for GNNs with Flow-based Explanations. Technical Report arXiv :2306.04791, arXiv.

[Lakkaraju et al., 2019] Lakkaraju, H., Kamar, E., Caruana, R., and Leskovec, J. (2019). Faithful and Customizable Explanations of Black Box Models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 131–138, Honolulu HI USA. ACM.

[Landrieu and Simonovsky, 2018] Landrieu, L. and Simonovsky, M. (2018). Large-scale Point Cloud Semantic Segmentation with Superpoint Graphs. Technical Report arXiv :1711.09869, arXiv.

[Laugel et al., 2017] Laugel, T., Lesot, M.-J., Marsala, C., Renard, X., and Detyniecki, M. (2017). Inverse Classification for Comparison-based Interpretability in Machine Learning. Technical Report arXiv :1712.08443, arXiv.

[Laugel et al., 2018] Laugel, T., Lesot, M.-J., Marsala, C., Renard, X., and Detyniecki, M. (2018). Comparison-Based Inverse Classification for Interpretability in Machine Learning. In Medina, J., Ojeda-Aciego, M., Verdegay, J. L., Pelta, D. A., Cabrera, I. P., Bouchon-Meunier, B., and Yager, R. R., editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations*, volume 853, pages 100–111. Springer International Publishing, Cham.

[Laugel et al., 2019] Laugel, T., Lesot, M.-J., Marsala, C., Renard, X., and Detyniecki, M. (2019). The Dangers of Post-hoc Interpretability : Unjustified Counterfactual Explanations. *International Joint Conference on Artificial Intelligence*, pages 2801–2807.

[Laugel et al., 2020] Laugel, T., Lesot, M.-J., Marsala, C., Renard, X., and Detyniecki, M. (2020). Unjustified Classification Regions and Counterfactual Explanations in Machine Learning. In Brefeld, U., Fromont, E., Hotho, A., Knobbe, A., Maathuis, M., and Robardet, C., editors, *Machine Learning and Knowledge Discovery in Databases*, volume 11907, pages 37–54. Springer International Publishing, Cham.

[Le et al., 2011] Le, Q. V., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G. S., Dean, J., and Ng, A. Y. (2011). Building high-level features using large scale unsupervised learning.

[Leavitt and Morcos, 2020] Leavitt, M. L. and Morcos, A. (2020). Towards falsifiable interpretability research.

[Lecun et al., 1998] Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (Nov./1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11) :2278–2324.

[Lee et al., 2019a] Lee, J., Lee, I., and Kang, J. (2019a). Self-Attention Graph Pooling. Technical Report arXiv :1904.08082, arXiv.

[Lee et al., 2019b] Lee, J., Lee, I., and Kang, J. (2019b). Self-Attention Graph Pooling.

[Lee et al., 2018a] Lee, J. B., Rossi, R., and Kong, X. (2018a). Graph Classification using Structural Attention. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1666–1674, London United Kingdom. ACM.

[Lee et al., 2018b] Lee, J. B., Rossi, R. A., Kim, S., Ahmed, N. K., and Koh, E. (2018b). Attention Models in Graphs : A Survey. Technical Report arXiv :1807.07984, arXiv.

[Lee et al., ] Lee, J. R., Kim, S., Park, I., Eo, T., and Hwang, D. Relevance-CAM : Your Model Already Knows Where To Look.

[Lei et al., 2017] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2017). Distribution-Free Predictive Inference For Regression. Technical Report arXiv :1604.04173, arXiv.

[Letham et al., 2015a] Letham, B., Rudin, C., McCormick, T. H., and Madigan, D. (2015a). Interpretable classifiers using rules and Bayesian analysis : Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3).

[Letham et al., 2015b] Letham, B., Rudin, C., McCormick, T. H., and Madigan, D. (2015b). Interpretable classifiers using rules and Bayesian analysis : Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3).

[Levie et al., 2018] Levie, R., Monti, F., Bresson, X., and Bronstein, M. M. (2018). CayleyNets : Graph Convolutional Neural Networks with Complex Rational Spectral Filters. Technical Report arXiv :1705.07664, arXiv.

[Levie et al., 2019] Levie, R., Monti, F., Bresson, X., and Bronstein, M. M. (2019). CayleyNets : Graph Convolutional Neural Networks With Complex Rational Spectral Filters. *IEEE Transactions on Signal Processing*, 67(1) :97–109.

[Levin et al., ] Levin, R., Borgnia, E., Goldblum, M., Shu, M., Huang, F., and Goldstein, T. Where do Models go Wrong ? Parameter-Space Saliency Maps for Explainability.

[Lewis, 1973] Lewis, D. K. (1973). Counterfactuals.

[Li et al., 2019a] Li, B., Li, X., Zhang, Z., and Wu, F. (2019a). Spatio-Temporal Graph Routing for Skeleton-Based Action Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01) :8561–8568.

[Li et al., 2022a] Li, C., Lou, J., Liu, S., Chen, Z., and Yuan, X. (2022a). Shapley Explainer - An Interpretation Method for GNNs Used in SDN. In *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, pages 5534–5540.

[Li et al., 2021a] Li, G., Müller, M., Ghanem, B., and Koltun, V. (2021a). Training Graph Neural Networks with 1000 Layers. In *Proceedings of the 38th International Conference on Machine Learning*, pages 6437–6449. PMLR.

[Li et al., 2022b] Li, G., Müller, M., Ghanem, B., and Koltun, V. (2022b). Training Graph Neural Networks with 1000 Layers.

[Li et al., 2022c] Li, G., Müller, M., Ghanem, B., and Koltun, V. (2022c). Training Graph Neural Networks with 1000 Layers. Technical Report arXiv :2106.07476, arXiv.

[Li et al., 2019b] Li, G., Müller, M., Thabet, A., and Ghanem, B. (2019b). DeepGCNs : Can GCNs Go as Deep as CNNs ? Technical Report arXiv :1904.03751, arXiv.

[Li et al., 2020a] Li, G., Xiong, C., Thabet, A., and Ghanem, B. (2020a). DeeperGCN : All You Need to Train Deeper GCNs. Technical Report arXiv :2006.07739, arXiv.

[Li et al., 2021b] Li, J., Louri, A., Karanth, A., and Bunescu, R. (2021b). GCNAX : A Flexible and Energy-efficient Accelerator for Graph Convolutional Neural Networks. In *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 775–788.

[Li et al., 2020b] Li, M., Chen, S., Zhang, Y., and Tsang, I. (2020b). Graph Cross Networks with Vertex Infomax Pooling. In *Advances in Neural Information Processing Systems*, volume 33, pages 14093–14105. Curran Associates, Inc.

[Li et al., 2022d] Li, M., Guo, X., Wang, Y., Wang, Y., and Lin, Z. (2022d). G$ˆ2$CN : Graph Gaussian Convolution Networks with Concentrated Graph Filters. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12782–12796. PMLR.

[Li et al., 2020c] Li, M., Ma, Z., Wang, Y. G., and Zhuang, X. (2020c). Fast Haar Transforms for Graph Neural Networks. *Neural Networks*, 128 :188–198.

[Li et al., 2020d] Li, Q., Cummings, R., and Mintz, Y. (2020d). Optimal Local Explainer Aggregation for Interpretable Prediction. Technical Report arXiv :2003.09466, arXiv.

[Li et al., 2018a] Li, Q., Han, Z., and Wu, X.-M. (2018a). Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning. Technical Report arXiv :1801.07606, arXiv.

[Li et al., 2018b] Li, R., Wang, S., Zhu, F., and Huang, J. (2018b). Adaptive Graph Convolutional Neural Networks. Technical Report arXiv :1801.03226, arXiv.

[Li et al., 2022e] Li, S., Niu, D., Wang, Y., Han, W., Zhang, Z., Guan, T., Guan, Y., Liu, H., Huang, L., Du, Z., Xue, F., Fang, Y., Zheng, H., and Xie, Y. (2022e). Hyperscale FPGA-as-a-service architecture for large-scale distributed graph neural network. In *Proceedings of the 49th Annual International Symposium on Computer Architecture*, pages 946–961, New York New York. ACM.

[Li et al., 2023a] Li, W., Chen, K., Liu, S., Huang, W., Zhang, H., Yingjie, T., Yun, S., and Song, M. (2023a). Message-passing Selection : Towards Interpretable GNNs for Graph Classification.

[Li et al., 2022f] Li, W., Li, Y., Li, Z., Hao, J., and Pang, Y. (2022f). DAG Matters ! GFlowNets Enhanced Explainer for Graph Neural Networks.

[Li et al., ] Li, X., Xiong, H., Li, X., Wu, X., Chen, Z., and Dou, D. InterpretDL : Explaining Deep Models in PaddlePaddle.

[Li et al., 2020e] Li, X.-H., Shi, Y., Li, H., Bai, W., Song, Y., Cao, C. C., and Chen, L. (2020e). Quantitative Evaluations on Saliency Methods : An Experimental Study. Technical Report arXiv :2012.15616, arXiv.

[Li et al., 2018c] Li, Y., Bu, R., Sun, M., Wu, W., Di, X., and Chen, B. (2018c). PointCNN : Convolution On $\mathcal{X}$-Transformed Points. Technical Report arXiv :1801.07791, arXiv.

[Li et al., 2019c] Li, Y., Gu, C., Dullien, T., Vinyals, O., and Kohli, P. (2019c). Graph Matching Networks for Learning the Similarity of Graph Structured Objects. Technical Report arXiv :1904.12787, arXiv.

[Li et al., 2022g] Li, Y., Liu, L., Wang, G., Du, Y., and Chen, P. (2022g). EGNN : Constructing explainable graph neural networks via knowledge distillation. *Knowledge-Based Systems*, 241 :108345.

[Li et al., 2017] Li, Y., Tarlow, D., Brockschmidt, M., and Zemel, R. (2017). Gated Graph Sequence Neural Networks. Technical Report arXiv :1511.05493, arXiv.

[Li et al., 2022h] Li, Y., Verma, S., Yang, S., Zhou, J., and Chen, F. (2022h). Are Graph Neural Network Explainers Robust to Graph Noises ? In Aziz, H., Corrêa, D., and French, T., editors, *AI 2022 : Advances in Artificial Intelligence*, Lecture Notes in Computer Science, pages 161–174, Cham. Springer International Publishing.

[Li et al., 2018d] Li, Y., Yu, R., Shahabi, C., and Liu, Y. (2018d). Diffusion Convolutional Recurrent Neural Network : Data-Driven Traffic Forecasting. Technical Report arXiv :1707.01926, arXiv.

[Li et al., 2023b] Li, Y., Zhou, J., Verma, S., and Chen, F. (2023b). A Survey of Explainable Graph Neural Networks : Taxonomy and Evaluation Metrics. Technical Report arXiv :2207.12599, arXiv.

[Li et al., 2020f] Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Stuart, A., Bhattacharya, K., and Anandkumar, A. (2020f). Multipole Graph Neural Operator for Parametric Partial Differential Equations. In *Advances in Neural Information Processing Systems*, volume 33, pages 6755–6766. Curran Associates, Inc.

[Lian et al., 2020] Lian, Z., Tao, J., Liu, B., Huang, J., Yang, Z., and Li, R. (2020). Context-Dependent Domain Adversarial Neural Network for Multimodal Emotion Recognition. In *Interspeech 2020*, pages 394–398. ISCA.

[Liang et al., 2020a] Liang, H., Ouyang, Z., Zeng, Y., Su, H., He, Z., Xia, S.-T., Zhu, J., and Zhang, B. (2020a). Training Interpretable Convolutional Neural Networks by Differentiating Class-Specific Filters. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *Computer Vision – ECCV 2020*, volume 12347, pages 622–638. Springer International Publishing, Cham.

[Liang et al., 2020b] Liang, J., Gurukar, S., and Parthasarathy, S. (2020b). MILE : A Multi-Level Framework for Scalable Graph Embedding. Technical Report arXiv :1802.09612, arXiv.

[Liang et al., 2020c] Liang, S., Wang, Y., Liu, C., He, L., Li, H., and Li, X. (2020c). EnGN : A High-Throughput and Energy-Efficient Accelerator for Large Graph Neural Networks.

[Liang et al., ] Liang, X., Hu, Z., Zhang, H., Lin, L., and Xing, E. P. Symbolic Graph Reasoning Meets Convolutions.

[Liao et al., ] Liao, Q. V., Bellamy, R. K. E., Muller, M., and Candello, H. Human-AI Collaboration : Towards Socially-Guided Machine Learning.

[Liao et al., 2018] Liao, R., Zhao, Z., Urtasun, R., and Zemel, R. (2018). Lanczos-Net : Multi-Scale Deep Graph Convolutional Networks.

[Liao et al., 2019] Liao, R., Zhao, Z., Urtasun, R., and Zemel, R. S. (2019). LANCZOSNET : MULTI-SCALE DEEP GRAPH CONVO- LUTIONAL NETWORKS.

[Lim and Dey, 2009] Lim, B. Y. and Dey, A. K. (2009). Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing*, pages 195–204, Orlando Florida USA. ACM.

[Lim et al., 2021a] Lim, D., Hohne, F., Li, X., Huang, S. L., Gupta, V., Bhalerao, O., and Lim, S.-N. (2021a). Large Scale Learning on Non-Homophilous Graphs : New Benchmarks and Strong Simple Methods. Technical Report arXiv :2110.14446, arXiv.

[Lim et al., 2021b] Lim, D., Lee, H., and Kim, S. (2021b). Building Reliable Explanations of Unreliable Neural Networks : Locally Smoothing Perspective of Model Interpretation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6464–6473, Nashville, TN, USA. IEEE.

[Lin et al., 2020a] Lin, C., Sun, G. J., Bulusu, K. C., Dry, J. R., and Hernandez, M. (2020a). Graph Neural Networks Including Sparse Interpretability. Technical Report arXiv :2007.00119, arXiv.

[Lin et al., ] Lin, W., Lan, H., and Li, B. Generative Causal Explanations for Graph Neural Networks.

[Lin et al., 2022a] Lin, W., Lan, H., Wang, H., and Li, B. (2022a). OrphicX : A Causality-Inspired Latent Variable Model for Interpreting Graph Neural Net-

works. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13719–13728, New Orleans, LA, USA. IEEE.

[Lin et al., 2022b] Lin, Y., Wang, A. S., Undersander, E., and Rai, A. (2022b). Efficient and Interpretable Robot Manipulation with Graph Neural Networks. Technical Report arXiv :2102.13177, arXiv.

[Lin et al., 2020b] Lin, Y.-S., Lee, W.-C., and Celik, Z. B. (2020b). What Do You See ? Evaluation of Explainable Artificial Intelligence (XAI) Interpretability through Neural Backdoors. Technical Report arXiv :2009.10639, arXiv.

[Lin et al., 2020c] Lin, Z., Li, C., Miao, Y., Liu, Y., and Xu, Y. (2020c). PaGraph : Scaling GNN training on large graphs via computation-aware caching. In *Proceedings of the 11th ACM Symposium on Cloud Computing*, pages 401–415, Virtual Event USA. ACM.

[Lipton, 2018] Lipton, Z. C. (2018). The Mythos of Model Interpretability : In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3) :31–57.

[Liu et al., 2020a] Liu, H., Lu, S., Chen, X., and He, B. (2020a). G $^{3}$ : When graph neural networks meet parallel graph processing systems on GPUs. *Proceedings of the VLDB Endowment*, 13(12) :2813–2816.

[Liu et al., 2023] Liu, H., Zheng, K., Yu, S., and Chen, B. (2023). Towards a More Stable and General Subgraph Information Bottleneck. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

[Liu et al., 2020b] Liu, M., Gao, H., and Ji, S. (2020b). Towards Deeper Graph Neural Networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 338–348, Virtual Event CA USA. ACM.

[Liu et al., 2021a] Liu, M., Luo, Y., Wang, L., Xie, Y., Yuan, H., Gui, S., Yu, H., Xu, Z., Zhang, J., Liu, Y., Yan, K., Liu, H., Fu, C., Oztekin, B., Zhang, X., and Ji, S. (2021a). DIG : A Turnkey Library for Diving into Graph Deep Learning Research.

[Liu et al., 2021b] Liu, M., Luo, Y., Wang, L., Xie, Y., Yuan, H., Gui, S., Yu, H., Xu, Z., Zhang, J., Liu, Y., Yan, K., Liu, H., Fu, C., Oztekin, B., Zhang, X., and Ji, S. (2021b). DIG : A Turnkey Library for Diving into Graph Deep Learning Research.

[Liu et al., 2019] Liu, Q., Nickel, M., and Kiela, D. (2019). Hyperbolic Graph Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[Liu et al., ] Liu, W., Li, R., Zheng, M., Karanam, S., Wu, Z., Bhanu, B., Radke, R. J., and Camps, O. Towards Visually Explaining Variational Autoencoders.

[Liu et al., 2021c] Liu, X., Ding, J., Jin, W., Xu, H., Ma, Y., Liu, Z., and Tang, J. (2021c). Graph Neural Networks with Adaptive Residual. In *Advances in Neural Information Processing Systems*, volume 34, pages 9720–9733. Curran Associates, Inc.

[Liu et al., 2021d] Liu, X., Jin, W., Ma, Y., Li, Y., Liu, H., Wang, Y., Yan, M., and Tang, J. (2021d). Elastic Graph Neural Networks. In *Proceedings of the 38th International Conference on Machine Learning*, pages 6837–6849. PMLR.

[Liu et al., 2018a] Liu, X., Luo, Z., and Huang, H. (2018a). Jointly Multiple Events Extraction via Attention-based Graph Information Aggregation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256.

[Liu et al., 2020c] Liu, Y., Hildebrandt, M., Joblin, M., Ringsquandl, M., Raissouni, R., and Tresp, V. (2020c). Neural Multi-Hop Reasoning With Logical Rules on Biomedical Knowledge Graphs.

[Liu et al., 2022a] Liu, Y., Zhang, X., and Xie, S. (2022a). A Differential Geometric View and Explainability of GNN on Evolving Graphs.

[Liu et al., 2022b] Liu, Y., Zhang, X., and Xie, S. (2022b). Trade less Accuracy for Fairness and Trade-off Explanation for GNN. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 4681–4690.

[Liu et al., 2018b] Liu, Z., Chen, C., Li, L., Zhou, J., Li, X., Song, L., and Qi, Y. (2018b). GeniePath : Graph Neural Networks with Adaptive Receptive Paths. Technical Report arXiv :1802.00910, arXiv.

[Liu et al., 2022c] Liu, Z., Zhou, K., Yang, F., Li, L., Chen, R., and Hu, X. (2022c). EXACT : SCALABLE GRAPH NEURAL NETWORKS TRAINING VIA EXTREME ACTIVATION COMPRESSION.

[Longa et al., 2023] Longa, A., Azzolin, S., Santin, G., Cencetti, G., Liò, P., Lepri, B., and Passerini, A. (2023). Explaining the Explainers in Graph Neural Networks : A Comparative Study. Technical Report arXiv :2210.15304, arXiv.

[Lou et al., 2020] Lou, A., Katsman, I., Jiang, Q., Belongie, S., Lim, S.-N., and Sa, C. D. (2020). Differentiating through the Fréchet Mean. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6393–6403. PMLR.

[Loukas, 2019] Loukas, A. (2019). What graph neural networks cannot learn : Depth vs width.

[Loveland et al., 2021] Loveland, D., Liu, S., Kailkhura, B., Hiszpanski, A., and Han, Y. (2021). Reliable Graph Neural Network Explanations Through Adversarial Training. Technical Report arXiv :2106.13427, arXiv.

[Lu and Li, 2020] Lu, Y.-J. and Li, C.-T. (2020). GCAN : Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media. Technical Report arXiv :2004.11648, arXiv.

[Luan et al., 2019] Luan, S., Zhao, M., Chang, X.-W., and Precup, D. (2019). Break the Ceiling : Stronger Multi-scale Deep Graph Convolutional Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[Lucic et al., 2020] Lucic, A., Haned, H., and de Rijke, M. (2020). Why Does My Model Fail ? Contrastive Local Explanations for Retail Forecasting. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 90–98.

[Lucic et al., 2021] Lucic, A., Oosterhuis, H., Haned, H., and de Rijke, M. (2021). FOCUS : Flexible Optimizable Counterfactual Explanations for Tree Ensembles.

[Lucic et al., 2022] Lucic, A., ter Hoeve, M., Tolomei, G., de Rijke, M., and Silvestri, F. (2022). CF-GNNExplainer : Counterfactual Explanations for Graph Neural Networks.

[Lukovnikov and Fischer, 2021] Lukovnikov, D. and Fischer, A. (2021). Improving Breadth-Wise Backpropagation in Graph Neural Networks Helps Learning Long-Range Dependencies. In *Proceedings of the 38th International Conference on Machine Learning*, pages 7180–7191. PMLR.

[Lundberg and Lee, 2017] Lundberg, S. M. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[Lundstrom et al., 2022] Lundstrom, D., Huang, T., and Razaviyayn, M. (2022). A Rigorous Study of Integrated Gradients Method and Extensions to Internal Neuron Attributions. Technical Report arXiv :2202.11912, arXiv.

[Luo et al., ] Luo, D., Cheng, W., Xu, D., Yu, W., Chen, H., Zhang, X., and Zong, B. Parameterized Explainer for Graph Neural Network.

[Luo et al., 2020] Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H., and Zhang, X. (2020). Parameterized Explainer for Graph Neural Network. In *Advances in Neural Information Processing Systems*, volume 33, pages 19620–19631. Curran Associates, Inc.

[Luo et al., 2021] Luo, G., Li, J., Peng, H., Yang, C., Sun, L., Yu, P. S., and He, L. (2021). Graph Entropy Guided Node Embedding Dimension Selection for Graph Neural Networks. volume 3, pages 2767–2774.

[Luss et al., 2021] Luss, R., Chen, P.-Y., Dhurandhar, A., Sattigeri, P., Zhang, Y., Shanmugam, K., and Tu, C.-C. (2021). Leveraging Latent Features for Local Explanations. Technical Report arXiv :1905.12698, arXiv.

[Lv et al., 2022] Lv, G., Chen, L., and Cao, C. C. (2022). On Glocal Explainability of Graph Neural Networks. In Bhattacharya, A., Lee Mong Li, J., Agrawal, D., Reddy, P. K., Mohania, M., Mondal, A., Goyal, V., and Uday Kiran, R., editors, *Database Systems for Advanced Applications*, Lecture Notes in Computer Science, pages 648–664, Cham. Springer International Publishing.

[Lv et al., 2021] Lv, Q., Ding, M., Liu, Q., Chen, Y., Feng, W., He, S., Zhou, C., Jiang, J., Dong, Y., and Tang, J. (2021). Are we really making much progress ? : Revisiting, benchmarking and refining heterogeneous graph neural networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1150–1160, Virtual Event Singapore. ACM.

[M et al., 2002] M, S., F, K., and H, B. (2002). ICA of fMRI group study data. *NeuroImage*, 16(3 Pt 1).

[M et al., 2010] M, Y., T, E., Aj, L., and A, L. (2010). Subcortical functional connectivity and verbal episodic memory in healthy elderly–a resting state fMRI study. *NeuroImage*, 52(1).

[Ma et al., a] Ma, J., Guo, R., Mishra, S., Zhang, A., and Li, J. CLEAR : Generative Counterfactual Explanations on Graphs.

[Ma et al., 2019a] Ma, J., Tang, W., Zhu, J., and Mei, Q. (2019a). A Flexible Generative Framework for Graph-based Semi-supervised Learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[Ma et al., b] Ma, L., Yang, Z., Wu, M., Miao, Y., Zhou, L., Xue, J., and Dai, Y. NeuGraph : Parallel Deep Neural Network Computation on Large Graphs.

[Ma et al., c] Ma, T., Chen, J., and Xiao, C. Constrained Generation of Semantically Valid Graphs via Regularizing Variational Autoencoders.

[Ma et al., 2018a] Ma, T., Chen, J., and Xiao, C. (2018a). Constrained Generation of Semantically Valid Graphs via Regularizing Variational Autoencoders. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

[Ma et al., 2021a] Ma, Y., Liu, X., Zhao, T., Liu, Y., Tang, J., and Shah, N. (2021a). A Unified View on Graph Neural Networks as Graph Signal Denoising. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1202–1211, Virtual Event Queensland Australia. ACM.

[Ma et al., 2019b] Ma, Y., Wang, S., Aggarwal, C. C., and Tang, J. (2019b). Graph Convolutional Networks with EigenPooling. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 723–731, Anchorage AK USA. ACM.

[Ma et al., 2018b] Ma, Y., Wang, S., Aggarwal, C. C., Yin, D., and Tang, J. (2018b). Multi-dimensional Graph Convolutional Networks. Technical Report arXiv :1808.06099, arXiv.

[Ma et al., 2021b] Ma, Z., Gu, H., and Liu, Z. (2021b). Understanding Drug Abuse Social Network Using Weighted Graph Neural Networks Explainer. In Gervasi,

O., Murgante, B., Misra, S., Garau, C., Blečić, I., Taniar, D., Apduhan, B. O., Rocha, A. M. A. C., Tarantino, E., and Torre, C. M., editors, *Computational Science and Its Applications – ICCSA 2021*, Lecture Notes in Computer Science, pages 52–61, Cham. Springer International Publishing.

[Ma et al., 2021c] Ma, Z., Xuan, J., Wang, Y. G., Li, M., and Lio, P. (2021c). Path Integral Based Convolution and Pooling for Graph Neural Networks. *Journal of Statistical Mechanics : Theory and Experiment*, 2021(12) :124011.

[Macdonald et al., ] Macdonald, J., Besancon, M., and Pokutta, S. Interpretable Neural Networks with Frank-Wolfe : Sparse Relevance Maps and Relevance Orderings.

[Madaan et al., 2021] Madaan, N., Padhi, I., Panwar, N., and Saha, D. (2021). Generate Your Counterfactuals : Towards Controlled Counterfactual Generation for Text. Technical Report arXiv :2012.04698, arXiv.

[Madsen et al., 2023] Madsen, A., Reddy, S., and Chandar, S. (2023). Post-hoc Interpretability for Neural NLP : A Survey. *ACM Computing Surveys*, 55(8) :1–42.

[Magister et al., 2021] Magister, L. C., Kazhdan, D., Singh, V., and Liò, P. (2021). GCExplainer : Human-in-the-Loop Concept-based Explanations for Graph Neural Networks. Technical Report arXiv :2107.11889, arXiv.

[Mahendran and Vedaldi, 2016] Mahendran, A. and Vedaldi, A. (2016). Salient Deconvolutional Networks. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, volume 9910, pages 120–135. Springer International Publishing, Cham.

[Maleki et al., 2014] Maleki, S., Tran-Thanh, L., Hines, G., Rahwan, T., and Rogers, A. (2014). Bounding the Estimation Error of Sampling-based Shapley Value Approximation. Technical Report arXiv :1306.4265, arXiv.

[Marcheggiani et al., 2018] Marcheggiani, D., Bastings, J., and Titov, I. (2018). Exploiting Semantics in Neural Machine Translation with Graph Convolutional Networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, pages 486–492, New Orleans, Louisiana. Association for Computational Linguistics.

[Mardaoui and Garreau, 2021] Mardaoui, D. and Garreau, D. (2021). An Analysis of LIME for Text Data. Technical Report arXiv :2010.12487, arXiv.

[Markowitz et al., 2021] Markowitz, E., Balasubramanian, K., Mirtaheri, M., Abu-El-Haija, S., Perozzi, B., Steeg, G. V., and Galstyan, A. (2021). Graph Traversal with Tensor Functionals : A Meta-Algorithm for Scalable Learning.

[Marques-Silva et al., 2021] Marques-Silva, J., Gerspacher, T., Cooper, M., Ignatiev, A., and Narodytska, N. (2021). Explanations for Monotonic Classifiers. Technical Report arXiv :2106.00154, arXiv.

[Martens and Provost, 2014] Martens, D. and Provost, F. (2014). Explaining Data-Driven Document Classifications. *MIS Quarterly*, 38(1) :73–100.

[McCLELLAND et al., ] McCLELLAND, J. L., Rumelhart, D. E., and Hinton, G. E. Parallel Distributed Processing.

[McCorduck, 2004] McCorduck, P. (2004). *Machines Who Think : A Personal Inquiry into the History and Prospects of Artificial Intelligence*. A.K. Peters, Natick, Mass, 25th anniversary update edition.

[McCulloch and Pitts, 1943] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4) :115–133.

[Md et al., 2021] Md, V., Misra, S., Ma, G., Mohanty, R., Georganas, E., Heinecke, A., Kalamkar, D., Ahmed, N. K., and Avancha, S. (2021). DistGNN : Scalable distributed training for large-scale graph neural networks. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–14, St. Louis Missouri. ACM.

[Melis and Jaakkola, a] Melis, D. A. and Jaakkola, T. Towards Robust Interpretability with Self-Explaining Neural Networks.

[Melis and Jaakkola, b] Melis, D. A. and Jaakkola, T. Towards Robust Interpretability with Self-Explaining Neural Networks.

[Mernyei and Cangea, 2022] Mernyei, P. and Cangea, C. (2022). Wiki-CS : A Wikipedia-Based Benchmark for Graph Neural Networks. Technical Report arXiv :2007.02901, arXiv.

[Mf et al., 2013] Mf, G., Sn, S., Ja, W., Ts, C., B, F., Jl, A., J, X., S, J., M, W., Jr, P., Dc, V. E., and M, J. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, 80.

[Micheli, 2009] Micheli, A. (2009). Neural Network for Graphs : A Contextual Constructive Approach. *IEEE Transactions on Neural Networks*, 20(3) :498–511.

[Miller, 2018] Miller, T. (2018). Explanation in Artificial Intelligence : Insights from the Social Sciences. Technical Report arXiv :1706.07269, arXiv.

[Miller, 2019a] Miller, T. (2019a). Explanation in artificial intelligence : Insights from the social sciences. *Artificial Intelligence*, 267 :1–38.

[Miller, 2019b] Miller, T. (2019b). Explanation in artificial intelligence : Insights from the social sciences. *Artificial Intelligence*, 267 :1–38.

[Miller, 2021] Miller, T. (2021). Contrastive Explanation : A Structural-Model Approach. *The Knowledge Engineering Review*, 36 :e14.

[Min et al., 2020] Min, Y., Wenkel, F., and Wolf, G. (2020). Scattering GCN : Overcoming Oversmoothness in Graph Convolutional Networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 14498–14508. Curran Associates, Inc.

[Min et al., 2021] Min, Y., Wenkel, F., and Wolf, G. (2021). Geometric Scattering Attention Networks. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8518–8522.

[Ming et al., 2018] Ming, Y., Qu, H., and Bertini, E. (2018). RuleMatrix : Visualizing and Understanding Classifiers with Rules. Technical Report arXiv :1807.06228, arXiv.

[Mishra et al., 2023] Mishra, S., Dutta, S., Long, J., and Magazzeni, D. (2023). A Survey on the Robustness of Feature Importance and Counterfactual Explanations. Technical Report arXiv :2111.00358, arXiv.

[Mitchell et al., 2019] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229.

[Mitchell, 1997] Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Series in Computer Science. McGraw-Hill, New York.

[Mittelstadt et al., 2019] Mittelstadt, B., Russell, C., and Wachter, S. (2019). Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 279–288.

[Mo et al., 2021] Mo, X., Xing, Y., and Lv, C. (2021). Heterogeneous Edge-Enhanced Graph Attention Network For Multi-Agent Trajectory Prediction. Technical Report arXiv :2106.07161, arXiv.

[Mohri et al., 2018] Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of Machine Learning*. Adaptive Computation and Machine Learning. The MIT Press, Cambridge, Massachusetts London, England, second edition edition.

[Molnar, 2022] Molnar, C. (2022). *Interpretable Machine Learning : A Guide for Making Black Box Models Explainable*. Christoph Molnar, Munich, Germany, second edition edition.

[Molnar et al., 2020] Molnar, C., Casalicchio, G., and Bischl, B. (2020). Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. In Koprinska, I., Kamp, M., Appice, A., Loglisci, C., Antonie, L., Zimmermann, A., Guidotti, R., Özgöbek, Ö., Ribeiro, R. P., Gavaldà, R., Gama, J., Adilova, L., Krishnamurthy, Y., Ferreira, P. M., Malerba, D., Medeiros, I., Ceci, M., Manco, G., Masciari, E., Ras, Z. W., Christen, P., Ntoutsi, E., Schubert, E., Zimek, A., Monreale, A., Biecek, P., Rinzivillo, S., Kille, B., Lommatzsch, A., and Gulla, J. A., editors, *ECML PKDD 2020 Workshops*, Communications in Computer and Information Science, pages 417–431, Cham. Springer International Publishing.

[Montavon, 2019] Montavon, G. (2019). Gradient-Based Vs. Propagation-Based Explanations : An Axiomatic Comparison. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R., editors, *Explainable AI : Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pages 253–265. Springer International Publishing, Cham.

[Montavon et al., 2019] Montavon, G., Binder, A., Lapuschkin, S., Samek, W., and Müller, K.-R. (2019). Layer-Wise Relevance Propagation : An Overview. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R., editors,

*Explainable AI : Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pages 193–209. Springer International Publishing, Cham.

[Montavon et al., 2017] Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65 :211–222.

[Montavon et al., 2018] Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73 :1–15.

[Monti et al., 2016] Monti, F., Boscaini, D., Masci, J., Rodolà, E., Svoboda, J., and Bronstein, M. M. (2016). Geometric deep learning on graphs and manifolds using mixture model CNNs. Technical Report arXiv :1611.08402, arXiv.

[Monti et al., 2017] Monti, F., Bronstein, M. M., and Bresson, X. (2017). Geometric Matrix Completion with Recurrent Multi-Graph Neural Networks. Technical Report arXiv :1704.06803, arXiv.

[Monti et al., 2018] Monti, F., Otness, K., and Bronstein, M. M. (2018). MOTIFNET : A MOTIF-BASED GRAPH CONVOLUTIONAL NETWORK FOR DIRECTED GRAPHS. In *2018 IEEE Data Science Workshop (DSW)*, pages 225–228.

[Moreira et al., 2022] Moreira, C., Chou, Y.-L., Hsieh, C., Ouyang, C., Jorge, J., and Pereira, J. M. (2022). Benchmarking Counterfactual Algorithms for XAI : From White Box to Black Box. Technical Report arXiv :2203.02399, arXiv.

[Morris et al., 2019] Morris, C., Ritzert, M., Fey, M., Hamilton, W. L., Lenssen, J. E., Rattan, G., and Grohe, M. (2019). Weisfeiler and Leman Go Neural : Higher-Order Graph Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01) :4602–4609.

[Morris et al., 2021] Morris, C., Ritzert, M., Fey, M., Hamilton, W. L., Lenssen, J. E., Rattan, G., and Grohe, M. (2021). Weisfeiler and Leman Go Neural : Higher-order Graph Neural Networks. Technical Report arXiv :1810.02244, arXiv.

[Moshkovitz et al., 2021] Moshkovitz, M., Yang, Y.-Y., and Chaudhuri, K. (2021). Connecting Interpretability and Robustness in Decision Trees through Separation. Technical Report arXiv :2102.07048, arXiv.

[Mothilal et al., 2021] Mothilal, R. K., Mahajan, D., Tan, C., and Sharma, A. (2021). Towards Unifying Feature Attribution and Counterfactual Explanations : Different Means to the Same End. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 652–663.

[Mothilal et al., 2020] Mothilal, R. K., Sharma, A., and Tan, C. (2020). Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617.

[Mueller et al., 2019] Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., and Klein, G. (2019). Explanation in Human-AI Systems : A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI. Technical Report arXiv :1902.01876, arXiv.

[Mueller et al., 2021] Mueller, S. T., Veinott, E. S., Hoffman, R. R., Klein, G., Alam, L., Mamun, T., and Clancey, W. J. (2021). Principles of Explanation in Human-AI Systems. Technical Report arXiv :2102.04972, arXiv.

[Murdoch et al., 2019] Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 116(44) :22071–22080.

[Murphy, ] Murphy, K. P. *Probabilistic Machine Learning : Advanced Topics*.

[Murphy, 2012] Murphy, K. P. (2012). *Machine Learning : A Probabilistic Perspective*. Adaptive Computation and Machine Learning Series. MIT Press, Cambridge, MA.

[Murphy et al., 2019] Murphy, R., Srinivasan, B., Rao, V., and Ribeiro, B. (2019). Relational Pooling for Graph Representations. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4663–4673. PMLR.

[Nair et al., 2021] Nair, R., Mattetti, M., Daly, E., Wei, D., Alkan, O., and Zhang, Y. (2021). What Changed ? Interpretable Model Comparison. In *Proceedings of*

*the Thirtieth International Joint Conference on Artificial Intelligence*, pages 2855–2861, Montreal, Canada. International Joint Conferences on Artificial Intelligence Organization.

[Nam et al., 2020] Nam, W.-J., Choi, J., and Lee, S.-W. (2020). Interpreting Deep Neural Networks with Relative Sectional Propagation by Analyzing Comparative Gradients and Hostile Activations. Technical Report arXiv :2012.03434, arXiv.

[Nam et al., 2019] Nam, W.-J., Gur, S., Choi, J., Wolf, L., and Lee, S.-W. (2019). Relative Attributing Propagation : Interpreting the Comparative Contributions of Individual Units in Deep Neural Networks. Technical Report arXiv :1904.00605, arXiv.

[Narasimhan et al., 2018] Narasimhan, M., Lazebnik, S., and Schwing, A. G. (2018). Out of the Box : Reasoning with Graph Convolution Nets for Factual Visual Question Answering. Technical Report arXiv :1811.00538, arXiv.

[Nauta et al., 2023] Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., and Seifert, C. (2023). From Anecdotal Evidence to Quantitative Evaluation Methods : A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*, 55(13s) :1–42.

[Neerincx et al., 2018] Neerincx, M. A., van der Waa, J., Kaptein, F., and van Diggelen, J. (2018). Using Perceptual and Cognitive Explanations for Enhanced Human-Agent Team Performance. In *Engineering Psychology and Cognitive Ergonomics : 15th International Conference, EPCE 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings*, pages 204–214, Berlin, Heidelberg. Springer-Verlag.

[Nemirovsky et al., 2021] Nemirovsky, D., Thiebaut, N., Xu, Y., and Gupta, A. (2021). CounteRGAN : Generating Realistic Counterfactuals with Residual Generative Adversarial Nets. Technical Report arXiv :2009.05199, arXiv.

[Nesterov, 1983] Nesterov, Y. (1983). A method for solving the convex programming problem with convergence rate O(1/k$\hat{2}$). *Proceedings of the USSR Academy of Sciences*.

[Ng et al., ] Ng, Y. C., Colombo, N., and Silva, R. Bayesian Semi-supervised Learning with Graph Gaussian Processes.

[Nguyen et al., 2016] Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., and Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks.

[Nguyen et al., 2019] Nguyen, A., Yosinski, J., and Clune, J. (2019). Understanding Neural Networks via Feature Visualization : A Survey. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R., editors, *Explainable AI : Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pages 55–76. Springer International Publishing, Cham.

[Nguyen et al., ] Nguyen, G., Taesiri, M. R., and Nguyen, A. Visual correspondence-based explanations improve AI robustness and human-AI team accuracy.

[Nguyen and Grishman, 2018] Nguyen, T. and Grishman, R. (2018). Graph Convolutional Networks With Argument-Aware Pooling for Event Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

[Nie et al., ] Nie, W., Zhang, Y., and Patel, A. B. A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations.

[Niepert et al., 2016] Niepert, M., Ahmed, M., and Kutzkov, K. (2016). Learning Convolutional Neural Networks for Graphs. Technical Report arXiv :1605.05273, arXiv.

[Nigam et al., 2021] Nigam, A., Pollice, R., Krenn, M., Gomes, G. d. P., and Aspuru-Guzik, A. (2021). Beyond generative models : Superfast traversal, optimization, novelty, exploration and discovery (STONED) algorithm for molecules using SELFIES. *Chemical Science*, 12(20) :7079–7090.

[Noutahi et al., 2020] Noutahi, E., Beaini, D., Horwood, J., Giguère, S., and Tossou, P. (2020). Towards Interpretable Sparse Graph Representation Learning with Laplacian Pooling.

[Noutahi et al., 2019] Noutahi, E., Beani, D., Horwood, J., and Tossou, P. (2019). Towards Interpretable Molecular Graph Representation Learning.

[NT et al., 2021] NT, H., Maehara, T., and Murata, T. (2021). Revisiting Graph Neural Networks : Graph Filtering Perspective. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8376–8383.

[Oh et al., 2019] Oh, S. J., Schiele, B., and Fritz, M. (2019). Towards Reverse-Engineering Black-Box Neural Networks. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R., editors, *Explainable AI : Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pages 121–144. Springer International Publishing, Cham.

[Olah et al., 2017] Olah, C., Mordvintsev, A., and Schubert, L. (2017). Feature Visualization. *Distill*, 2(11) :e7.

[Olshausen and Field, 1996] Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583) :607–609.

[Oono and Suzuki, 2019] Oono, K. and Suzuki, T. (2019). Graph Neural Networks Exponentially Lose Expressive Power for Node Classification.

[Orlichenko et al., 2022] Orlichenko, A., Qu, G., and Wang, Y.-P. (2022). Phenotype guided interpretable graph convolutional network analysis of fMRI data reveals changing brain connectivity during adolescence. In Gimi, B. S. and Krol, A., editors, *Medical Imaging 2022 : Biomedical Applications in Molecular, Structural, and Functional Imaging*, page 43, San Diego, United States. SPIE.

[O'Shaughnessy et al., ] O'Shaughnessy, M., Canal, G., Connor, M., Davenport, M., and Rozell, C. Generative causal explanations of black-box classifiers.

[Ott et al., 2022] Ott, S., Barbosa-Silva, A., and Samwald, M. (2022). LinkExplorer : Predicting, explaining and exploring links in large biomedical knowledge graphs. Technical report, bioRxiv.

[Paleja et al., ] Paleja, R., Ghuy, M., Arachchige, N. R., Jensen, R., and Gombolay, M. The Utility of Explainable AI in Ad Hoc Human-Machine Teaming.

[Pan et al., 2018a] Pan, S., Hu, R., Long, G., Jiang, J., Yao, L., and Zhang, C. (2018a). Adversarially Regularized Graph Autoencoder for Graph Embedding. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 2609–2615, Stockholm, Sweden. International Joint Conferences on Artificial Intelligence Organization.

[Pan et al., 2018b] Pan, S., Hu, R., Long, G., Jiang, J., Yao, L., and Zhang, C. (2018b). Adversarially Regularized Graph Autoencoder for Graph Embedding.

In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 2609–2615, Stockholm, Sweden. International Joint Conferences on Artificial Intelligence Organization.

[Pan et al., 2019] Pan, S., Hu, R., Long, G., Jiang, J., Yao, L., and Zhang, C. (2019). Adversarially Regularized Graph Autoencoder for Graph Embedding. Technical Report arXiv :1802.04407, arXiv.

[Pareja et al., 2019] Pareja, A., Domeniconi, G., Chen, J., Ma, T., Suzumura, T., Kanezashi, H., Kaler, T., Schardl, T. B., and Leiserson, C. E. (2019). EvolveGCN : Evolving Graph Convolutional Networks for Dynamic Graphs. Technical Report arXiv :1902.10191, arXiv.

[Parekh et al., 2022] Parekh, J., Parekh, S., Mozharovskyi, P., d'Alch é-Buc, F., and Richard, G. (2022). Listen to Interpret : Post-hoc Interpretability for Audio Networks with NMF. Preprint, Open Science Framework.

[Park et al., 2021] Park, C. W., Kornbluth, M., Vandermause, J., Wolverton, C., Kozinsky, B., and Mailoa, J. P. (2021). Accurate and scalable graph neural network force field and molecular dynamics with direct force architecture. *npj Computational Materials*, 7(1) :1–9.

[Park, 2023] Park, H. (2023). Providing Post-Hoc Explanation for Node Representation Learning Models Through Inductive Conformal Predictions. *IEEE access : practical innovations, open solutions*, 11 :1202–1212.

[Pawelczyk and Bielawski, ] Pawelczyk, M. and Bielawski, S. CARLA : A Python Library to Benchmark Algorithmic Recourse and Counterfactual Explanation Algorithms.

[Pawelczyk et al., 2020a] Pawelczyk, M., Broelemann, K., and Kasneci, G. (2020a). Learning Model-Agnostic Counterfactual Explanations for Tabular Data. In *Proceedings of The Web Conference 2020*, pages 3126–3132, Taipei Taiwan. ACM.

[Pawelczyk et al., 2020b] Pawelczyk, M., Broelemann, K., and Kasneci, G. (2020b). On Counterfactual Explanations under Predictive Multiplicity. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 809–818. PMLR.

[Pawelczyk et al., 2022] Pawelczyk, M., Datta, T., van-den-Heuvel, J., Kasneci, G., and Lakkaraju, H. (2022). Probabilistically Robust Recourse : Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse. Technical Report arXiv :2203.06768, arXiv.

[Peach et al., 2020] Peach, R. L., Arnaudon, A., Schmidt, J. A., Palasciano, H. A., Bernier, N. R., Jelfs, K., Yaliraki, S., and Barahona, M. (2020). Hcga : Highly Comparative Graph Analysis for network phenotyping. Technical report, bioRxiv.

[Pei et al., 2019] Pei, H., Wei, B., Chang, K. C.-C., Lei, Y., and Yang, B. (2019). Geom-GCN : Geometric Graph Convolutional Networks.

[Peng et al., 2022] Peng, J., Chen, Z., Shao, Y., Shen, Y., Chen, L., and Cao, J. (2022). Sancus : *s* t *a* le *n* ess-aware *c* omm *u* nication-avoiding full-graph decentralized training in large-scale graph neural networks. *Proceedings of the VLDB Endowment*, 15(9) :1937–1950.

[Pereira et al., 2022] Pereira, T. A., Nascimento, E. J. F., Mesquita, D., and Souza, A. H. (2022). ConveXplainer for Graph Neural Networks. In Xavier-Junior, J. C. and Rios, R. A., editors, *Intelligent Systems*, Lecture Notes in Computer Science, pages 588–600, Cham. Springer International Publishing.

[Perlmutter et al., 2020] Perlmutter, M., Gao, F., Wolf, G., and Hirn, M. (2020). Geometric Wavelet Scattering Networks on Compact Riemannian Manifolds. In *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, pages 570–604. PMLR.

[Petsiuk, ] Petsiuk, V. RISE : Randomized Input Sampling for Explanation of Black-box Models.

[Petsiuk et al., 2021] Petsiuk, V., Jain, R., Manjunatha, V., Morariu, V. I., Mehra, A., Ordonez, V., and Saenko, K. (2021). Black-box Explanation of Object Detectors via Saliency Maps. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11438–11447, Nashville, TN, USA. IEEE.

[Pfeifer et al., 2022] Pfeifer, B., Saranti, A., and Holzinger, A. (2022). GNN-SubNet : Disease subnetwork detection with explainable graph neural networks. *Bioinformatics (Oxford, England)*, 38(Supplement_2) :ii120–ii126.

[Pham et al., 2021] Pham, L., Vu, T. H., and Tran, T. A. (2021). Facial Expression Recognition Using Residual Masking Network. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4513–4519.

[Pillai et al., 2022] Pillai, V., Koohpayegani, S. A., Ouligian, A., Fong, D., and Pirsiavash, H. (2022). Consistent Explanations by Contrastive Learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10203–10212, New Orleans, LA, USA. IEEE.

[Pillai and Pirsiavash, 2021] Pillai, V. and Pirsiavash, H. (2021). Explainable Models with Consistent Interpretations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3) :2431–2439.

[Platonov et al., 2023] Platonov, O., Kuznedelev, D., Diskin, M., Babenko, A., and Prokhorenkova, L. (2023). A critical look at the evaluation of GNNs under heterophily : Are we really making progress ? Technical Report arXiv :2302.11640, arXiv.

[Plumb et al., 2019] Plumb, G., Molitor, D., and Talwalkar, A. (2019). Model Agnostic Supervised Local Explanations.

[Plumb et al., 2020] Plumb, G., Terhorst, J., Sankararaman, S., and Talwalkar, A. (2020). Explaining Groups of Points in Low-Dimensional Representations. Technical Report arXiv :2003.01640, arXiv.

[Polyak, 1964] Polyak, B. T. (1964). Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5) :1–17.

[Pope et al., 2019a] Pope, P. E., Kolouri, S., Rostami, M., Martin, C. E., and Hoffmann, H. (2019a). Explainability Methods for Graph Convolutional Neural Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10764–10773, Long Beach, CA, USA. IEEE.

[Pope et al., 2019b] Pope, P. E., Kolouri, S., Rostami, M., Martin, C. E., and Hoffmann, H. (2019b). Explainability Methods for Graph Convolutional Neural Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10764–10773, Long Beach, CA, USA. IEEE.

[Poyiadzi et al., 2020a] Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., and Flach, P. (2020a). FACE : Feasible and Actionable Counterfactual Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350.

[Poyiadzi et al., 2020b] Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., and Flach, P. (2020b). FACE : Feasible and Actionable Counterfactual Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, New York NY USA. ACM.

[Prado-Romero et al., 2023] Prado-Romero, M. A., Prenkaj, B., and Stilo, G. (2023). Revisiting CounteRGAN for Counterfactual Explainability of Graphs.

[Prado-Romero and Stilo, 2022] Prado-Romero, M. A. and Stilo, G. (2022). GRETEL : A unified framework for Graph Counterfactual Explanation Evaluation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4389–4393.

[Qasim et al., 2019] Qasim, S. R., Kieseler, J., Iiyama, Y., and Pierini, M. (2019). Learning representations of irregular particle-detector geometry with distance-weighted graph networks. *The European Physical Journal C*, 79(7) :608.

[Qi et al., 2017a] Qi, C. R., Su, H., Mo, K., and Guibas, L. J. (2017a). PointNet : Deep Learning on Point Sets for 3D Classification and Segmentation. Technical Report arXiv :1612.00593, arXiv.

[Qi et al., 2017b] Qi, C. R., Yi, L., Su, H., and Guibas, L. J. (2017b). PointNet++ : Deep Hierarchical Feature Learning on Point Sets in a Metric Space. Technical Report arXiv :1706.02413, arXiv.

[Qi et al., 2018] Qi, S., Wang, W., Jia, B., Shen, J., and Zhu, S.-C. (2018). Learning Human-Object Interactions by Graph Parsing Neural Networks. Technical Report arXiv :1808.07962, arXiv.

[Qi et al., 2017c] Qi, X., Liao, R., Jia, J., Fidler, S., and Urtasun, R. (2017c). 3D Graph Neural Networks for RGBD Semantic Segmentation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5209–5218, Venice. IEEE.

[Qin et al., 2019] Qin, L., Bosselut, A., Holtzman, A., Bhagavatula, C., Clark, E., and Choi, Y. (2019). Counterfactual Story Reasoning and Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5042–5052, Hong Kong, China. Association for Computational Linguistics.

[Qin et al., 2018] Qin, Z., Yu, F., Liu, C., and Chen, X. (2018). How convolutional neural network see the world - A survey of convolutional neural network visualization methods. Technical Report arXiv :1804.11191, arXiv.

[Qiu et al., 2020] Qiu, J., Chen, Q., Dong, Y., Zhang, J., Yang, H., Ding, M., Wang, K., and Tang, J. (2020). GCC : Graph Contrastive Coding for Graph Neural Network Pre-Training. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1150–1160, Virtual Event CA USA. ACM.

[Qiu et al., 2018] Qiu, J., Tang, J., Ma, H., Dong, Y., Wang, K., and Tang, J. (2018). DeepInf : Social Influence Prediction with Deep Learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2110–2119.

[Ragno et al., 2022] Ragno, A., La Rosa, B., and Capobianco, R. (2022). Prototype-based Interpretable Graph Neural Networks. *IEEE Transactions on Artificial Intelligence*, pages 1–11.

[Rahimi et al., 2018] Rahimi, A., Cohn, T., and Baldwin, T. (2018). Semi-supervised User Geolocation via Graph Convolutional Networks. Technical Report arXiv :1804.08049, arXiv.

[Rahman et al., 2021] Rahman, M. K., Sujon, M. H., and Azad, A. (2021). FusedMM : A Unified SDDMM-SpMM Kernel for Graph Embedding and Graph Neural Networks.

[Rahman et al., ] Rahman, T., Jin, S., and Gogate, V. Look Ma, No Latent Variables : Accurate Cutset Networks via Compilation.

[Raison, 2022] Raison, A. (2022). ScoreCAM GNN : A generalization of an optimal local post-hoc explaining method to any geometric deep learning models.

[Ramakrishnan et al., 2020] Ramakrishnan, G., Lee, Y. C., and Albarghouthi, A. (2020). Synthesizing Action Sequences for Modifying Model Decisions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04) :5462–5469.

[Ramezani et al., 2022] Ramezani, M., Cong, W., Mahdavi, M., Kandemir, M. T., and Sivasubramaniam, A. (2022). LEARN LOCALLY, CORRECT GLOBALLY : A DISTRIBUTED ALGORITHM FOR TRAINING GRAPH NEURAL NETWORKS.

[Rampášek et al., 2023] Rampášek, L., Galkin, M., Dwivedi, V. P., Luu, A. T., Wolf, G., and Beaini, D. (2023). Recipe for a General, Powerful, Scalable Graph Transformer. Technical Report arXiv :2205.12454, arXiv.

[Ranjan et al., 2018] Ranjan, A., Bolkart, T., Sanyal, S., and Black, M. J. (2018). Generating 3D faces using Convolutional Mesh Autoencoders. Technical Report arXiv :1807.10267, arXiv.

[Ranjan et al., 2020] Ranjan, E., Sanyal, S., and Talukdar, P. P. (2020). ASAP : Adaptive Structure Aware Pooling for Learning Hierarchical Graph Representations. Technical Report arXiv :1911.07979, arXiv.

[Ranzato et al., 2007] Ranzato, M., Boureau, Y.-L., Chopra, S., and LeCun, Y. (2007). A Unified Energy-Based Framework for Unsupervised Learning. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, pages 371–379. PMLR.

[Rao et al., 2021a] Rao, J., Zheng, S., and Yang, Y. (2021a). Quantitative Evaluation of Explainable Graph Neural Networks for Molecular Property Prediction. Technical Report arXiv :2107.04119, arXiv.

[Rao et al., 2021b] Rao, S. X., Zhang, S., Han, Z., Zhang, Z., Min, W., Chen, Z., Shan, Y., Zhao, Y., and Zhang, C. (2021b). xFraud : Explainable Fraud Transaction Detection. *Proceedings of the VLDB Endowment*, 15(3) :427–436.

[Rathee et al., 2021] Rathee, M., Zhang, Z., Funke, T., Khosla, M., and Anand, A. (2021). Learnt Sparsification for Interpretable Graph Neural Networks. Technical Report arXiv :2106.12920, arXiv.

[Rathi, 2019] Rathi, S. (2019). Generating Counterfactual and Contrastive Explanations using SHAP. Technical Report arXiv :1906.09293, arXiv.

[Rebuffi et al., 2020] Rebuffi, S.-A., Fong, R., Ji, X., and Vedaldi, A. (2020). There and Back Again : Revisiting Backpropagation Saliency Methods. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8836–8845, Seattle, WA, USA. IEEE.

[Ribeiro et al., 2017] Ribeiro, L. F. R., Savarese, P. H. P., and Figueiredo, D. R. (2017). Struc2vec : Learning Node Representations from Structural Identity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 385–394.

[Ribeiro et al., 2016a] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016a). Model-Agnostic Interpretability of Machine Learning.

[Ribeiro et al., 2016b] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016b). "Why Should I Trust You ?" : Explaining the Predictions of Any Classifier. Technical Report arXiv :1602.04938, arXiv.

[Ribeiro et al., 2018] Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors : High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

[Rivest, 1987] Rivest, R. L. (1987). Learning decision lists. *Machine Learning*, 2(3) :229–246.

[Robeer et al., 2021] Robeer, M., Bex, F., and Feelders, A. (2021). Generating Realistic Natural Language Counterfactuals. In *Findings of the Association for Computational Linguistics : EMNLP 2021*, pages 3611–3625, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Roberts and Yaida, 2022] Roberts, D. A. and Yaida, S. (2022). *The Principles of Deep Learning Theory : An Effective Theory Approach to Understanding Neural Networks*. Cambridge University Press, Cambridge New York, NY Port Melbourne, VIC New Delhi Singapore.

[Robnik-Šikonja and Bohanec, 2018] Robnik-Šikonja, M. and Bohanec, M. (2018). Perturbation-Based Explanations of Prediction Models. In Zhou, J. and Chen, F., editors, *Human and Machine Learning*, pages 159–175, Cham. Springer International Publishing.

[Robnik-Šikonja and Kononenko, 2008] Robnik-Šikonja, M. and Kononenko, I. (2008). Explaining Classifications For Individual Instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5) :589–600.

[Rodriguez et al., 2021] Rodriguez, P., Caccia, M., Lacoste, A., Zamparo, L., Laradji, I., Charlin, L., and Vazquez, D. (2021). Beyond Trivial Counterfactual Explanations with Diverse Valuable Explanations. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1036–1045, Montreal, QC, Canada. IEEE.

[Rong et al., 2019] Rong, Y., Huang, W., Xu, T., and Huang, J. (2019). DropEdge : Towards Deep Graph Convolutional Networks on Node Classification.

[Rong et al., 2022] Rong, Y., Leemann, T., Borisov, V., Kasneci, G., and Kasneci, E. (2022). A Consistent and Efficient Evaluation Strategy for Attribution Methods. Technical Report arXiv :2202.00449, arXiv.

[Rosenblatt, 1958] Rosenblatt, F. (1958). The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6) :386–408.

[Rosenblatt, 1962] Rosenblatt, F. (1962). *Principles of Neurodynamics : Perceptrons and the Theory of Brain Mechanisms*. Spartan Books.

[Ross et al., 2017] Ross, A. S., Hughes, M. C., and Doshi-Velez, F. (2017). Right for the Right Reasons : Training Differentiable Models by Constraining their Explanations. Technical Report arXiv :1703.03717, arXiv.

[Rossi et al., 2020] Rossi, E., Chamberlain, B., Frasca, F., Eynard, D., Monti, F., and Bronstein, M. (2020). Temporal Graph Networks for Deep Learning on Dynamic Graphs. Technical Report arXiv :2006.10637, arXiv.

[Rossi et al., 2023] Rossi, E., Charpentier, B., Di Giovanni, F., Frasca, F., Günnemann, S., and Bronstein, M. (2023). Edge Directionality Improves Learning on Heterophilic Graphs. Technical Report arXiv :2305.10498, arXiv.

[Rossi et al., 2022] Rossi, E., Kenlay, H., Gorinova, M. I., Chamberlain, B. P., Dong, X., and Bronstein, M. (2022). On the Unreasonable Effectiveness of Feature propagation in Learning on Graphs with Missing Node Features. Technical Report arXiv :2111.12128, arXiv.

[Roy et al., ] Roy, N. A., Kim, J., and Rabinowitz, N. Explainability Via Causal Self-Talk.

[Rozemberczki et al., 2021a] Rozemberczki, B., Allen, C., and Sarkar, R. (2021a). Multi-scale Attributed Node Embedding. Technical Report arXiv :1909.13021, arXiv.

[Rozemberczki et al., 2019] Rozemberczki, B., Davies, R., Sarkar, R., and Sutton, C. (2019). GEMSEC : Graph Embedding with Self Clustering. Technical Report arXiv :1802.03997, arXiv.

[Rozemberczki et al., 2021b] Rozemberczki, B., Englert, P., Kapoor, A., Blais, M., and Perozzi, B. (2021b). Pathfinder Discovery Networks for Neural Message Passing.

[Rozemberczki and Sarkar, 2020] Rozemberczki, B. and Sarkar, R. (2020). Characteristic Functions on Graphs : Birds of a Feather, from Statistical Descriptors to Parametric Models. Technical Report arXiv :2005.07959, arXiv.

[Rudin, 2019a] Rudin, C. (2019a). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.

[Rudin, 2019b] Rudin, C. (2019b). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Technical Report arXiv :1811.10154, arXiv.

[Rudin et al., 2022] Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., and Zhong, C. (2022). Interpretable machine learning : Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16(none) :1–85.

[Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088) :533–536.

[Russell, 2019] Russell, C. (2019). Efficient Search for Diverse Coherent Explanations. Technical Report arXiv :1901.04909, arXiv.

[Russell et al., ] Russell, C., Kusner, M. J., Loftus, J., and Silva, R. When Worlds Collide : Integrating Different Counterfactual Assumptions in Fairness.

[Sagonas et al., 2016] Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2016). 300 Faces In-The-Wild Challenge : Database and results. *Image and Vision Computing*, 47 :3–18.

[Sagonas et al., 2013] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2013). A Semi-automatic Methodology for Facial Landmark Annotation. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 896–903, OR, USA. IEEE.

[Saha et al., 2022] Saha, S., Das, M., and Bandyopadhyay, S. (2022). A Model-Centric Explainer for Graph Neural Network based Node Classification. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4434–4438, Atlanta GA USA. ACM.

[Sahoo et al., 2020] Sahoo, S. S., Venugopalan, S., Li, L., Singh, R., and Riley, P. (2020). Scaling Symbolic Methods using Gradients for Neural Model Explanation.

[Salha et al., 2019] Salha, G., Limnios, S., Hennequin, R., Tran, V.-A., and Vazirgiannis, M. (2019). Gravity-Inspired Graph Autoencoders for Directed Link Prediction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 589–598, Beijing China. ACM.

[Salimi-Khorshidi et al., 2014] Salimi-Khorshidi, G., Douaud, G., Beckmann, C. F., Glasser, M. F., Griffanti, L., and Smith, S. M. (2014). Automatic denoising of functional MRI data : Combining independent component analysis and hierarchical fusion of classifiers. *NeuroImage*, 90 :449–468.

[Samangouei et al., 2018] Samangouei, P., Saeedi, A., Nakagawa, L., and Silberman, N. (2018). ExplainGAN : Model Explanation via Decision Boundary Crossing Transformations. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, volume 11214, pages 681–696. Springer International Publishing, Cham.

[Samek et al., 2019] Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R., editors (2019). *Explainable AI : Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*. Springer International Publishing, Cham.

[Samek and Müller, 2019] Samek, W. and Müller, K.-R. (2019). Towards Explainable Artificial Intelligence. In Samek, W., Montavon, G., Vedaldi, A., Hansen,

L. K., and Müller, K.-R., editors, *Explainable AI : Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pages 5–22. Springer International Publishing, Cham.

[Sammani et al., 2022] Sammani, F., Mukherjee, T., and Deligiannis, N. (2022). NLX-GPT : A Model for Natural Language Explanations in Vision and Vision-Language Tasks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8312–8322, New Orleans, LA, USA. IEEE.

[Samoilescu et al., 2021] Samoilescu, R.-F., Van Looveren, A., and Klaise, J. (2021). Model-agnostic and Scalable Counterfactual Explanations via Reinforcement Learning. Technical Report arXiv :2106.02597, arXiv.

[Sanchez-Lengeling et al., 2020] Sanchez-Lengeling, B., Wei, J., Lee, B., Reif, E., Wang, P., Qian, W., McCloskey, K., Colwell, L., and Wiltschko, A. (2020). Evaluating Attribution for Graph Neural Networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 5898–5910. Curran Associates, Inc.

[Sanchez-Lengeling et al., ] Sanchez-Lengeling, B., Wei, J., Lee, B., Reif, E., Wang, P. Y., Qian, W., McCloskey, K., Colwell, L., and Wiltschko, A. Evaluating Attribution for Graph Neural Networks.

[Santoro et al., 2017] Santoro, A., Raposo, D., Barrett, D. G. T., Malinowski, M., Pascanu, R., Battaglia, P., and Lillicrap, T. (2017). A simple neural network module for relational reasoning. Technical Report arXiv :1706.01427, arXiv.

[Sarkar et al., 2022] Sarkar, A., Vijaykeerthy, D., Sarkar, A., and Balasubramanian, V. N. (2022). A Framework for Learning Ante-hoc Explainable Models via Concepts. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10276–10285, New Orleans, LA, USA. IEEE.

[Sato et al., 2019] Sato, R., Yamada, M., and Kashima, H. (2019). Approximation Ratios of Graph Neural Networks for Combinatorial Problems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[Sattarzadeh et al., 2020] Sattarzadeh, S., Sudhakar, M., Lem, A., Mehryar, S., Plataniotis, K. N., Jang, J., Kim, H., Jeong, Y., Lee, S., and Bae, K. (2020). Explaining Convolutional Neural Networks through Attribution-Based Input Sampling and Block-Wise Feature Aggregation. Technical Report arXiv :2010.00672, arXiv.

[Sauer and Geiger, 2021] Sauer, A. and Geiger, A. (2021). COUNTERFACTUAL GENERATIVE NETWORKS.

[Saueressig et al., 2021] Saueressig, C., Berkley, A., Kang, E., Munbodh, R., and Singh, R. (2021). Exploring Graph-Based Neural Networks for Automatic Brain Tumor Segmentation. In Bowles, J., Broccia, G., and Nanni, M., editors, *From Data to Models and Back*, Lecture Notes in Computer Science, pages 18–37, Cham. Springer International Publishing.

[Sayres et al., 2019] Sayres, R., Xu, S., Saensuksopa, T., Le, M., and Webster, D. R. (2019). Assistance from a deep learning system improves diabetic retinopathy assessment in optometrists. *Investigative Ophthalmology & Visual Science*, 60(9) :1433.

[Scarselli et al., 2009] Scarselli, F., Gori, M., Ah Chung Tsoi, Hagenbuchner, M., and Monfardini, G. (2009). The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, 20(1) :61–80.

[Schleich et al., 2021] Schleich, M., Geng, Z., Zhang, Y., and Suciu, D. (2021). GeCo : Quality Counterfactual Explanations in Real Time. Technical Report arXiv :2101.01292, arXiv.

[Schlichtkrull et al., 2017] Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., and Welling, M. (2017). Modeling Relational Data with Graph Convolutional Networks. Technical Report arXiv :1703.06103, arXiv.

[Schlichtkrull et al., 2020] Schlichtkrull, M. S., Cao, N. D., and Titov, I. (2020). Interpreting Graph Neural Networks for NLP With Differentiable Edge Masking.

[Schnake et al., 2022] Schnake, T., Eberle, O., Lederer, J., Nakajima, S., Schütt, K. T., Müller, K.-R., and Montavon, G. (2022). Higher-Order Explanations of Graph Neural Networks via Relevant Walks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11) :7581–7596.

[Schoch et al., ] Schoch, S., Xu, H., and Ji, Y. CS-SHAPLEY : Class-wise Shapley Values for Data Valuation in Classification.

[Schölkopf et al., 2001] Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A Generalized Representer Theorem. In Helmbold, D. and Williamson, B., editors,

*Computational Learning Theory*, Lecture Notes in Computer Science, pages 416–426, Berlin, Heidelberg. Springer.

[Schütt et al., 2017] Schütt, K. T., Kindermans, P.-J., Sauceda, H. E., Chmiela, S., Tkatchenko, A., and Müller, K.-R. (2017). SchNet : A continuous-filter convolutional neural network for modeling quantum interactions. Technical Report arXiv :1706.08566, arXiv.

[Seifert et al., 2017] Seifert, C., Aamir, A., Balagopalan, A., Jain, D., Sharma, A., Grottel, S., and Gumhold, S. (2017). Visualizations of Deep Neural Networks in Computer Vision : A Survey. In Cerquitelli, T., Quercia, D., and Pasquale, F., editors, *Transparent Data Mining for Big and Small Data*, Studies in Big Data, pages 123–144. Springer International Publishing, Cham.

[Selbst et al., 2019] Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pages 59–68. Association for Computing Machinery, Inc.

[Selvaraju et al., 2017] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM : Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

[Selvaraju et al., 2020] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). Grad-CAM : Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2) :336–359.

[Seo et al., 2018] Seo, J., Choe, J., Koo, J., Jeon, S., Kim, B., and Jeon, T. (2018). Noise-adding Methods of Saliency Map as Series of Higher Order Partial Derivative. Technical Report arXiv :1806.03000, arXiv.

[Seo et al., 2016] Seo, Y., Defferrard, M., Vandergheynst, P., and Bresson, X. (2016). Structured Sequence Modeling with Graph Convolutional Recurrent Networks. Technical Report arXiv :1612.07659, arXiv.

[Serra and Niepert, 2023] Serra, G. and Niepert, M. (2023). L2XGNN : Learning to Explain Graph Neural Networks. Technical Report arXiv :2209.14402, arXiv.

[Serrurier et al., 2023] Serrurier, M., Mamalet, F., Fel, T., Béthune, L., and Boissin, T. (2023). On the explainable properties of 1-Lipschitz Neural Networks : An Optimal Transport Perspective. Technical Report arXiv :2206.06854, arXiv.

[Setzu et al., 2021] Setzu, M., Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., and Giannotti, F. (2021). GLocalX – From Local to Global Explanations of Black Box AI Models. https ://arxiv.org/abs/2101.07685v2.

[Shaham et al., 2018] Shaham, U., Stanton, K., Li, H., Nadler, B., Basri, R., and Kluger, Y. (2018). SpectralNet : Spectral Clustering using Deep Neural Networks. Technical Report arXiv :1801.01587, arXiv.

[Shalev-Shwartz and Ben-David, 2014] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning : From Theory to Algorithms*. Cambridge University Press, 1 edition.

[Shan et al., ] Shan, C., Shen, Y., Zhang, Y., Li, X., and Li, D. Reinforcement Learning Enhanced Explainer for Graph Neural Networks.

[Shang et al., 2019] Shang, J., Xiao, C., Ma, T., Li, H., and Sun, J. (2019). GAMENet : Graph Augmented MEmory Networks for Recommending Medication Combination. Technical Report arXiv :1809.01852, arXiv.

[Shannon, 1938] Shannon, C. E. (1938). A symbolic analysis of relay and switching circuits. *Transactions of the American Institute of Electrical Engineers*, 57(12) :713–723.

[Shannon, 1948] Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3) :379–423.

[Shapley, 1951] Shapley, L. S. (1951). Notes on the N-Person Game — II : The Value of an N-Person Game. Technical report, RAND Corporation.

[Sharma et al., 2020] Sharma, S., Henderson, J., and Ghosh, J. (2020). CERTIFAI : Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 166–172.

[Shchur et al., 2019] Shchur, O., Mumme, M., Bojchevski, A., and Günnemann, S. (2019). Pitfalls of Graph Neural Network Evaluation. Technical Report arXiv :1811.05868, arXiv.

[Shen et al., 2021a] Shen, W., Wei, Z., Huang, S., Zhang, B., Chen, P., Zhao, P., and Zhang, Q. (2021a). Verifiability and Predictability : Interpreting Utilities of Network Architectures for Point Cloud Processing. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10698–10707, Nashville, TN, USA. IEEE.

[Shen et al., 2021b] Shen, W., Wei, Z., Huang, S., Zhang, B., Fan, J., Zhao, P., and Zhang, Q. (2021b). Interpretable Compositional Convolutional Neural Networks. Technical Report arXiv :2107.04474, arXiv.

[Shi et al., 2020] Shi, W., Ragunathan, and Rajkumar (2020). Point-GNN : Graph Neural Network for 3D Object Detection in a Point Cloud. Technical Report arXiv :2003.01251, arXiv.

[Shi et al., 2021] Shi, Y., Huang, Z., Feng, S., Zhong, H., Wang, W., and Sun, Y. (2021). Masked Label Prediction : Unified Message Passing Model for Semi-Supervised Classification. Technical Report arXiv :2009.03509, arXiv.

[Shin et al., 2022] Shin, Y.-M., Kim, S.-W., Yoon, E.-B., and Shin, W.-Y. (2022). Prototype-Based Explanations for Graph Neural Networks (Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11) :13047–13048.

[Shitole et al., ] Shitole, V., Fuxin, L., Kahng, M., Tadepalli, P., and Fern, A. One Explanation is Not Enough : Structured Attention Graphs for Image Classification.

[Shotton et al., 2009] Shotton, J., Winn, J., Rother, C., and Criminisi, A. (2009). TextonBoost for Image Understanding : Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context. *International Journal of Computer Vision*, 81(1) :2–23.

[Shrikumar et al., 2017] Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. (2017). Not Just a Black Box : Learning Important Features Through Propagating Activation Differences.

[Shrotri et al., 2022] Shrotri, A. A., Narodytska, N., Ignatiev, A., Meel, K. S., Marques-Silva, J., and Vardi, M. Y. (2022). Constraint-Driven Explanations for Black-Box ML Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8) :8304–8314.

[Shumovskaia et al., 2022] Shumovskaia, V., Ntemos, K., Vlaski, S., and Sayed, A. H. (2022). Explainability and Graph Learning from Social Interactions. *IEEE Transactions on Signal and Information Processing over Networks*, 8 :946–959.

[Sieger et al., 2023] Sieger, L. N., Heindorf, S., Blübaum, L., and Ngomo, A.-C. N. (2023). Counterfactual Explanations for Concepts in $\mathcal{ELH}$. Technical Report arXiv :2301.05109, arXiv.

[Simonovsky and Komodakis, 2017] Simonovsky, M. and Komodakis, N. (2017). Dynamic Edge-Conditioned Filters in Convolutional Neural Networks on Graphs. Technical Report arXiv :1704.02901, arXiv.

[Simonovsky and Komodakis, 2018] Simonovsky, M. and Komodakis, N. (2018). GraphVAE : Towards Generation of Small Graphs Using Variational Autoencoders. In Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., and Maglogiannis, I., editors, *Artificial Neural Networks and Machine Learning – ICANN 2018*, Lecture Notes in Computer Science, pages 412–422, Cham. Springer International Publishing.

[Simonyan et al., 2014a] Simonyan, K., Vedaldi, A., and Zisserman, A. (2014a). Deep Inside Convolutional Networks : Visualising Image Classification Models and Saliency Maps.

[Simonyan et al., 2014b] Simonyan, K., Vedaldi, A., and Zisserman, A. (2014b). Deep Inside Convolutional Networks : Visualising Image Classification Models and Saliency Maps. Technical Report arXiv :1312.6034, arXiv.

[Singh and Lee, 2017] Singh, K. K. and Lee, Y. J. (2017). Hide-and-Seek : Forcing a Network to be Meticulous for Weakly-Supervised Object and Action Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3544–3553.

[Singh and Chen, 2022] Singh, R. and Chen, Y. (2022). Signed Graph Neural Networks : A Frequency Perspective. *Transactions on Machine Learning Research*.

[Singla et al., 2020] Singla, S., Pollack, B., Chen, J., and Batmanghelich, K. (2020). Explanation by Progressive Exaggeration.

[Singla et al., 2019] Singla, S., Wallace, E., Feng, S., and Feizi, S. (2019). Understanding Impacts of High-Order Loss Approximations and Features in Deep Learning Interpretation. Technical Report arXiv :1902.00407, arXiv.

[Sixt et al., 2020] Sixt, L., Schuessler, M., Weiß, P., and Landgraf, T. (2020). Interpretability Through Invertibility : A Deep Convolutional Network With Ideal Counterfactuals And Isosurfaces.

[Slack et al., 2020] Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling LIME and SHAP : Adversarial Attacks on Post hoc Explanation Methods.

[Sm et al., 2013] Sm, S., Cf, B., J, A., Ej, A., J, B., G, D., E, D., Da, F., L, G., Mp, H., M, K., T, L., Kl, M., S, M., S, P., J, P., G, S.-K., Az, S., At, V., Mw, W., J, X., E, Y., K, U., Dc, V. E., Mf, G., and undefined (2013). Resting-state fMRI in the Human Connectome Project. *Neuroimage*, 80 :144–168.

[Smilkov et al., 2017a] Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017a). SmoothGrad : Removing noise by adding noise.

[Smilkov et al., 2017b] Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017b). SmoothGrad : Removing noise by adding noise. Technical Report arXiv :1706.03825, arXiv.

[Smyth and Keane, 2021] Smyth, B. and Keane, M. T. (2021). A Few Good Counterfactuals : Generating Interpretable, Plausible and Diverse Counterfactual Explanations. Technical Report arXiv :2101.09056, arXiv.

[Sohrabizadeh et al., 2022] Sohrabizadeh, A., Chi, Y., and Cong, J. (2022). StreamGCN : Accelerating Graph Convolutional Networks with Streaming Processing. In *2022 IEEE Custom Integrated Circuits Conference (CICC)*, pages 1–8, Newport Beach, CA, USA. IEEE.

[Sokol and Flach, ] Sokol, K. and Flach, P. Counterfactual Explanations of Machine Learning Predictions : Opportunities and Challenges for AI Safety.

[Sokol and Flach, 2020] Sokol, K. and Flach, P. (2020). Explainability Fact Sheets : A Framework for Systematic Assessment of Explainable Approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 56–67.

[Sood and Craven, 2021] Sood, A. and Craven, M. (2021). Feature Importance Explanations for Temporal Black-Box Models. Technical Report arXiv :2102.11934, arXiv.

[Sorkun et al., 2019] Sorkun, M. C., Khetan, A., and Er, S. (2019). AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Scientific Data*, 6(1) :143.

[Souza et al., ] Souza, V. F. C., Cicalese, F., Laber, E. S., and Molinaro, M. Decision Trees with Short Explainable Rules.

[Sperduti and Starita, 1997] Sperduti, A. and Starita, A. (1997). Supervised neural networks for the classification of structures. *IEEE transactions on neural networks*, 8(3) :714–735.

[Spinelli et al., 2022] Spinelli, I., Scardapane, S., and Uncini, A. (2022). A Meta-Learning Approach for Training Explainable Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–9.

[Spooner et al., 2021] Spooner, T., Dervovic, D., Long, J., Shepard, J., Chen, J., and Magazzeni, D. (2021). Counterfactual Explanations for Arbitrary Regression Models. Technical Report arXiv :2106.15212, arXiv.

[Spreitzer and Haned, ] Spreitzer, N. and Haned, H. Evaluating the Practicality of Counterfactual Explanations.

[Springenberg et al., 2015] Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2015). Striving for Simplicity : The All Convolutional Net.

[Srinivas and Fleuret, ] Srinivas, S. and Fleuret, F. Full-Gradient Representation for Neural Network Visualization.

[Srinivas and Fleuret, 2020] Srinivas, S. and Fleuret, F. (2020). Rethinking the Role of Gradient-based Attribution Methods for Model Interpretability.

[Srivastava et al., 2014] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56) :1929–1958.

[Stadler et al., 2021] Stadler, M., Charpentier, B., Geisler, S., Zügner, D., and Günnemann, S. (2021). Graph Posterior Network : Bayesian Predictive Uncertainty

for Node Classification. In *Advances in Neural Information Processing Systems*, volume 34, pages 18033–18048. Curran Associates, Inc.

[Staff, 2020] Staff, P. A. I. (2020). Explainable AI in Practice Falls Short of Transparency Goals. https ://partnershiponai.org/xai-in-practice/.

[Stalder et al., ] Stalder, S., Perraudin, N., Achanta, R., Perez-Cruz, F., and Volpi, M. What You See is What You Classify : Black Box Attributions.

[Sterling and Irwin, 2015] Sterling, T. and Irwin, J. J. (2015). ZINC 15 – Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling*, 55(11) :2324–2337.

[Štrumbelj and Kononenko, 2014] Štrumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3) :647–665.

[Sturmfels et al., 2020] Sturmfels, P., Lundberg, S., and Lee, S.-I. (2020). Visualizing the Impact of Feature Attribution Baselines. *Distill*, 5(1) :e22.

[Su et al., 2019] Su, J., Vargas, D. V., and Kouichi, S. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5) :828–841.

[Sun et al., 2021a] Sun, J., Cheng, Z., Zuberi, S., Perez, F., and Volkovs, M. (2021a). HGCF : Hyperbolic Graph Convolution Networks for Collaborative Filtering. In *Proceedings of the Web Conference 2021*, pages 593–601, Ljubljana Slovenia. ACM.

[Sun et al., 2020a] Sun, K., Koniusz, P., and Wang, Z. (2020a). Fisher-Bures Adversary Graph Convolutional Networks. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, pages 465–475. PMLR.

[Sun et al., 2020b] Sun, K., Zhu, Z., and Lin, Z. (2020b). AdaGCN : Adaboosting Graph Convolutional Networks into Deep Models.

[Sun et al., 2022] Sun, L., Dou, Y., Yang, C., Wang, J., Liu, Y., Yu, P. S., He, L., and Li, B. (2022). Adversarial Attack and Defense on Graph Data : A Survey. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–20.

[Sun et al., 2021b] Sun, L., Zhang, Z., Zhang, J., Wang, F., Peng, H., Su, S., and Yu, P. S. (2021b). Hyperbolic Variational Graph Neural Network for Modeling

Dynamic Graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5) :4375–4383.

[Sun et al., 2021c] Sun, Q., Li, J., Peng, H., Wu, J., Ning, Y., Yu, P. S., and He, L. (2021c). SUGAR : Subgraph Neural Network with Reinforcement Pooling and Self-Supervised Mutual Information Mechanism. In *Proceedings of the Web Conference 2021*, pages 2081–2091, Ljubljana Slovenia. ACM.

[Sun et al., 2021d] Sun, Y., Valente, A., Liu, S., and Wang, D. (2021d). Preserve, Promote, or Attack ? GNN Explanation via Topology Perturbation. Technical Report arXiv :2103.13944, arXiv.

[Sun et al., 2019] Sun, Z., Deng, Z.-H., Nie, J.-Y., and Tang, J. (2019). RotatE : Knowledge Graph Embedding by Relational Rotation in Complex Space. Technical Report arXiv :1902.10197, arXiv.

[Sun et al., 2017] Sun, Z., Hu, W., and Li, C. (2017). Cross-lingual Entity Alignment via Joint Attribute-Preserving Embedding. Technical Report arXiv :1708.05045, arXiv.

[Sundararajan and Najmi, 2020] Sundararajan, M. and Najmi, A. (2020). The many Shapley values for model explanation. Technical Report arXiv :1908.08474, arXiv.

[Sundararajan et al., 2017a] Sundararajan, M., Taly, A., and Yan, Q. (2017a). Axiomatic Attribution for Deep Networks.

[Sundararajan et al., 2017b] Sundararajan, M., Taly, A., and Yan, Q. (2017b). Axiomatic Attribution for Deep Networks. Technical Report arXiv :1703.01365, arXiv.

[Sundararajan et al., 2019] Sundararajan, M., Xu, S., Taly, A., Sayres, R., and Najmi, A. (2019). Exploring Principled Visualizations for Deep Network Attributions. *Los Angeles*.

[Susnjara et al., 2015] Susnjara, A., Perraudin, N., Kressner, D., and Vandergheynst, P. (2015). Accelerated filtering on graphs using Lanczos method. Technical Report arXiv :1509.04537, arXiv.

[Sutton and Barto, 1998] Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning : An Introduction (2nd Edition)*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, Mass.

[Szegedy et al., 2014] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks.

[Tailor et al., 2021a] Tailor, S. A., Fernandez-Marques, J., and Lane, N. D. (2021a). Degree-Quant : Quantization-Aware Training for Graph Neural Networks.

[Tailor et al., 2021b] Tailor, S. A., Opolka, F., Lio, P., and Lane, N. D. (2021b). Do We Need Anisotropic Graph Neural Networks ?

[Tailor et al., 2022a] Tailor, S. A., Opolka, F. L., Liò, P., and Lane, N. D. (2022a). Do We Need Anisotropic Graph Neural Networks ? Technical Report arXiv :2104.01481, arXiv.

[Tailor et al., 2022b] Tailor, S. A., Opolka, F. L., Liò, P., and Lane, N. D. (2022b). Do We Need Anisotropic Graph Neural Networks ?

[Tan et al., 2022] Tan, J., Geng, S., Fu, Z., Ge, Y., Xu, S., Li, Y., and Zhang, Y. (2022). Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning. In *Proceedings of the ACM Web Conference 2022*, pages 1018–1027.

[Tavares et al., ] Tavares, Z., Koppel, J., Zhang, X., Das, R., and Solar-Lezama, A. A Language for Counterfactual Generative Models.

[Te et al., 2018] Te, G., Hu, W., Guo, Z., and Zheng, A. (2018). RGCNN : Regularized Graph CNN for Point Cloud Segmentation. Technical Report arXiv :1806.02952, arXiv.

[Tekkesinoglu et al., 2020] Tekkesinoglu, S., Kampik, T., and Främling, K. (2020). Py-CIU : A Python Library for Explaining Machine Learning Predictions Using Contextual Importance and Utility.

[Teney et al., 2020] Teney, D., Abbasnedjad, E., and Van Den Hengel, A. (2020). Learning What Makes a Difference from Counterfactual Examples and Gradient Supervision. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *Computer Vision – ECCV 2020*, volume 12355, pages 580–599. Springer International Publishing, Cham.

[Tenney et al., 2020] Tenney, I., Wexler, J., Bastings, J., Bolukbasi, T., Coenen, A., Gehrmann, S., Jiang, E., Pushkarna, M., Radebaugh, C., Reif, E., and Yuan, A. (2020). The Language Interpretability Tool : Extensible, Interactive Visualizations

and Analysis for NLP Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 107–118, Online. Association for Computational Linguistics.

[Tesic and Hahn, 2022] Tesic, M. and Hahn, U. (2022). Can counterfactual explanations of AI systems' predictions skew lay users' causal intuitions about the world ? If so, can we correct for that ? *Patterns*, 3(12) :100635.

[Teufel et al., 2022] Teufel, J., Torresi, L., Reiser, P. N., and Friederich, P. (2022). MEGAN : Multi Explanation Graph Attention Network.

[Thekumparampil et al., 2018] Thekumparampil, K. K., Wang, C., Oh, S., and Li, L.-J. (2018). Attention-based Graph Neural Network for Semi-supervised Learning. Technical Report arXiv :1803.03735, arXiv.

[Thorpe et al., ] Thorpe, J., Qiao, Y., Eyolfson, J., Teng, S., Hu, G., Jia, Z., Wei, J., Vora, K., Netravali, R., Kim, M., and Xu, G. H. Dorylus : Affordable, Scalable, and Accurate GNN Training with Distributed CPU Servers and Serverless Threads.

[Tian et al., 2020a] Tian, C., Ma, L., Yang, Z., and Dai, Y. (2020a). PCGCN : Partition-Centric Processing for Accelerating Graph Convolutional Network. In *2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 936–945.

[Tian et al., 2020b] Tian, C., Ma, L., Yang, Z., and Dai, Y. (2020b). PCGCN : Partition-Centric Processing for Accelerating Graph Convolutional Network. In *2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 936–945.

[Tian et al., 2023] Tian, Y., Wang, X., Yao, X., Liu, H., and Yang, Y. (2023). Predicting molecular properties based on the interpretable graph neural network with multistep focus mechanism. *Briefings in Bioinformatics*, 24(1) :bbac534.

[Tibshirani and Friedman, ] Tibshirani, S. and Friedman, H. *The Elements of Statistical Learning*.

[Togninalli et al., 2019] Togninalli, M., Ghisu, E., Llinares-López, F., Rieck, B., and Borgwardt, K. (2019). Wasserstein Weisfeiler-Lehman Graph Kernels. Technical Report arXiv :1906.01277, arXiv.

[Tolkachev et al., 2022] Tolkachev, G., Mell, S., Zdancewic, S., and Bastani, O. (2022). Counterfactual Explanations for Natural Language Interfaces. Technical Report arXiv :2204.13192, arXiv.

[Tolomei et al., 2017] Tolomei, G., Silvestri, F., Haines, A., and Lalmas, M. (2017). Interpretable Predictions of Tree-based Ensembles via Actionable Feature Tweaking. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 465–474.

[Tomsett et al., ] Tomsett, R., Braines, D., Harborne, D., Preece, A., and Chakraborty, S. Interpretable to Whom ? A Role-based Model for Analyzing Interpretable Machine Learning Systems.

[Tomsett et al., 2020] Tomsett, R., Harborne, D., Chakraborty, S., Gurram, P., and Preece, A. (2020). Sanity Checks for Saliency Metrics. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04) :6021–6029.

[Tong et al., 2020] Tong, Z., Liang, Y., Sun, C., Li, X., Rosenblum, D., and Lim, A. (2020). Digraph Inception Convolutional Networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 17907–17918. Curran Associates, Inc.

[Topping et al., 2021] Topping, J., Giovanni, F. D., Chamberlain, B. P., Dong, X., and Bronstein, M. M. (2021). Understanding over-squashing and bottlenecks on graphs via curvature.

[Tripathy et al., 2020] Tripathy, A., Yelick, K., and Buluc, A. (2020). Reducing Communication in Graph Neural Network Training. Technical Report arXiv :2005.03300, arXiv.

[Trivedi et al., 2017] Trivedi, R., Dai, H., Wang, Y., and Song, L. (2017). Know-Evolve : Deep Temporal Reasoning for Dynamic Knowledge Graphs. Technical Report arXiv :1705.05742, arXiv.

[Trouillon et al., 2016] Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., and Bouchard, G. (2016). Complex Embeddings for Simple Link Prediction. Technical Report arXiv :1606.06357, arXiv.

[Tsang et al., 2020] Tsang, M., Cheng, D., Liu, H., Feng, X., Zhou, E., and Liu, Y. (2020). FEATURE INTERACTION INTERPRETABILITY : A CASE FOR

EXPLAINING AD-RECOMMENDATION SYSTEMS VIA NEURAL INTERACTION DETECTION.

[Tsitsulin et al., 2023] Tsitsulin, A., Palowitch, J., Perozzi, B., and Müller, E. (2023). Graph Clustering with Graph Neural Networks. Technical Report arXiv :2006.16904, arXiv.

[Tu et al., 2018] Tu, K., Cui, P., Wang, X., Yu, P. S., and Zhu, W. (2018). Deep Recursive Network Embedding with Regular Equivalence. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2357–2366, London United Kingdom. ACM.

[Turing, 1937a] Turing, A. M. (1937a). Computability and λ-definability. *The Journal of Symbolic Logic*, 2(4) :153–163.

[Turing, 1937b] Turing, A. M. (1937b). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42(1) :230–265.

[TURING, 1950] TURING, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236) :433–460.

[Ulyanov et al., 2017] Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2017). Instance Normalization : The Missing Ingredient for Fast Stylization. Technical Report arXiv :1607.08022, arXiv.

[Upadhyay et al., 2021] Upadhyay, S., Joshi, S., and Lakkaraju, H. (2021). Towards Robust and Reliable Algorithmic Recourse. Technical Report arXiv :2102.13620, arXiv.

[Ustun and Rudin, 2016] Ustun, B. and Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3) :349–391.

[Ustun et al., 2019] Ustun, B., Spangher, A., and Liu, Y. (2019). Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, Atlanta GA USA. ACM.

[Ustun et al., 2014] Ustun, B., Tracà, S., and Rudin, C. (2014). Supersparse Linear Integer Models for Interpretable Classification.

[Valsesia et al., 2021] Valsesia, D., Fracastoro, G., and Magli, E. (2021). Don't stack layers in graph neural networks, wire them randomly.

[Van Der Vaart and Wellner, 1996] Van Der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer, New York, NY.

[van der Waa et al., 2018] van der Waa, J., Robeer, M., van Diggelen, J., Brinkhuis, M., and Neerincx, M. (2018). Contrastive Explanations with Local Foil Trees. Technical Report arXiv :1806.07470, arXiv.

[Van Fraassen, 1980] Van Fraassen, B. C. (1980). *The Scientific Image*. Clarendon Library of Logic and Philosophy. Clarendon Press ; Oxford University Press, Oxford : New York.

[Van Looveren and Klaise, 2020] Van Looveren, A. and Klaise, J. (2020). Interpretable Counterfactual Explanations Guided by Prototypes. Technical Report arXiv :1907.02584, arXiv.

[Van Tran et al., 2018] Van Tran, D., Navarin, N., and Sperduti, A. (2018). On Filter Size in Graph Convolutional Networks. Technical Report arXiv :1811.10435, arXiv.

[Vapnik, 2000] Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*. Springer, New York, NY.

[Vaswani et al., 2023] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention Is All You Need.

[Veličković et al., 2018a] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018a). Graph Attention Networks. Technical Report arXiv :1710.10903, arXiv.

[Veličković et al., 2018b] Veličković, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., and Hjelm, R. D. (2018b). Deep Graph Infomax. Technical Report arXiv :1809.10341, arXiv.

[Vellido et al., 2012] Vellido, A., Martín-Guerrero, J., and Lisboa, P. (2012). Making machine learning models interpretable.

[Venkatasubramanian and Alfano, 2020] Venkatasubramanian, S. and Alfano, M. (2020). The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 284–293, Barcelona Spain. ACM.

[Verma et al., 2018] Verma, N., Boyer, E., and Verbeek, J. (2018). FeaStNet : Feature-Steered Graph Convolutions for 3D Shape Analysis. Technical Report arXiv :1706.05206, arXiv.

[Verma et al., 2023] Verma, S., Armgaan, B., Medya, S., and Ranu, S. (2023). Empowering Counterfactual Reasoning over Graph Neural Networks through Inductivity. Technical Report arXiv :2306.04835, arXiv.

[Verma et al., 2022] Verma, S., Boonsanong, V., Hoang, M., Hines, K. E., Dickerson, J. P., and Shah, C. (2022). Counterfactual Explanations and Algorithmic Recourses for Machine Learning : A Review. Technical Report arXiv :2010.10596, arXiv.

[Verma et al., 2021] Verma, S., Lahiri, A., Dickerson, J. P., and Lee, S.-I. (2021). Pitfalls of Explainable ML : An Industry Perspective. Technical Report arXiv :2106.07758, arXiv.

[Vermeire and Martens, 2020] Vermeire, T. and Martens, D. (2020). Explainable Image Classification with Evidence Counterfactual. Technical Report arXiv :2004.07511, arXiv.

[Veyrin-Forrer et al., ] Veyrin-Forrer, L., Kamal, A., Duffner, S., Plantevit, M., and Robardet, C. On GNN explanability with activation patterns.

[Veyrin-Forrer et al., 2022] Veyrin-Forrer, L., Kamal, A., Duffner, S., Plantevit, M., and Robardet, C. (2022). What Does My GNN Really Capture ? On Exploring Internal GNN Representations. In *International Joint Conference on Artificial Intelligence 2022*, Vienna, Austria.

[Vignac et al., 2020] Vignac, C., Loukas, A., and Frossard, P. (2020). Building powerful and equivariant graph neural networks with structural message-passing. In *Advances in Neural Information Processing Systems*, volume 33, pages 14143–14155. Curran Associates, Inc.

[Vinyals et al., 2016] Vinyals, O., Bengio, S., and Kudlur, M. (2016). Order Matters : Sequence to sequence for sets. Technical Report arXiv :1511.06391, arXiv.

[Vj and Sk, 2004] Vj, S. and Sk, H. (2004). Comparison of three methods for generating group statistical inferences from independent component analysis of functional magnetic resonance imaging data. *Journal of magnetic resonance imaging : JMRI*, 19(3).

[von Kügelgen et al., 2022] von Kügelgen, J., Karimi, A.-H., Bhatt, U., Valera, I., Weller, A., and Schölkopf, B. (2022). On the Fairness of Causal Algorithmic Recourse. Technical Report arXiv :2010.06529, arXiv.

[Vu, ] Vu, M. N. PGM-Explainer : Probabilistic Graphical Model Explanations for Graph Neural Networks.

[Vu and Thai, 2020] Vu, M. N. and Thai, M. T. (2020). PGM-Explainer : Probabilistic Graphical Model Explanations for Graph Neural Networks. Technical Report arXiv :2010.05788, arXiv.

[Wachter et al., 2017] Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual Explanations Without Opening the Black Box : Automated Decisions and the GDPR. *SSRN Electronic Journal*.

[Wagner et al., 2019] Wagner, J., Kohler, J. M., Gindele, T., Hetzel, L., Wiedemer, J. T., and Behnke, S. (2019). Interpretable and Fine-Grained Visual Explanations for Convolutional Neural Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9089–9099, Long Beach, CA, USA. IEEE.

[Waleffe et al., 2022] Waleffe, R., Mohoney, J., Rekatsinas, T., and Venkataraman, S. (2022). MariusGNN : Resource-Efficient Out-of-Core Training of Graph Neural Networks.

[Wan et al., 2022a] Wan, C., Kyrillidis, A., Li, Y., Kim, N. S., Wolfe, C. R., and Lin, Y. (2022a). PIPEGCN : EFFICIENT FULL-GRAPH TRAINING OF GRAPH CONVOLUTIONAL NETWORKS WITH PIPELINED FEATURE COMMUNICATION.

[Wan et al., 2022b] Wan, C., Li, Y., Li, A., Kim, N. S., and Lin, Y. (2022b). BNS-GCN : Efficient Full-Graph Training of Graph Convolutional Networks with Partition-Parallelism and Random Boundary Node Sampling.

[Wang, 2001] Wang, D. (2001). Unsupervised Learning : Foundations of Neural Computation. *AI Magazine*, 22(2) :101–101.

[Wang et al., 2016] Wang, D., Cui, P., and Zhu, W. (2016). Structural Deep Network Embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1225–1234, San Francisco California USA. ACM.

[Wang and Khosravi, ] Wang, E. and Khosravi, P. Towards Probabilistic Sufficient Explanations.

[Wang and Kong, 2022] Wang, F. and Kong, A. W.-K. (2022). Exploiting the Relationship Between Kendall's Rank Correlation and Cosine Similarity for Attribution Protection. Technical Report arXiv :2205.07279, arXiv.

[Wang and Kong, 2023] Wang, F. and Kong, A. W.-K. (2023). A Practical Upper Bound for the Worst-Case Attribution Deviations. Technical Report arXiv :2303.00340, arXiv.

[Wang et al., 2023a] Wang, G., Chuang, Y.-N., Du, M., Yang, F., Zhou, Q., Tripathi, P., Cai, X., and Hu, X. (2023a). Accelerating Shapley Explanation via Contributive Cooperator Selection. Technical Report arXiv :2206.08529, arXiv.

[Wang et al., 2021a] Wang, H., Lian, D., Zhang, Y., Qin, L., He, X., Lin, Y., and Lin, X. (2021a). Binarized Graph Neural Network. *World Wide Web-internet and Web Information Systems*, 24(3) :825–848.

[Wang et al., 2018a] Wang, H., Wang, J., Wang, J., Zhao, M., Zhang, W., Zhang, F., Xie, X., and Guo, M. (2018a). GraphGAN : Graph Representation Learning With Generative Adversarial Nets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

[Wang et al., 2020a] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., and Hu, X. (2020a). Score-CAM : Score-Weighted Visual Explanations for Convolutional Neural Networks.

[Wang et al., 2021b] Wang, H., Yin, H., Zhang, M., and Li, P. (2021b). Equivariant and Stable Positional Encoding for More Powerful Graph Neural Networks.

[Wang et al., 2021c] Wang, L., Yin, Q., Tian, C., Yang, J., Chen, R., Yu, W., Yao, Z., and Zhou, J. (2021c). FlexGraph : A flexible and efficient distributed framework for GNN training. In *Proceedings of the Sixteenth European Conference on Computer Systems*, pages 67–82, Online Event United Kingdom. ACM.

[Wang et al., 2020b] Wang, M., Zheng, D., Ye, Z., Gan, Q., Li, M., Song, X., Zhou, J., Ma, C., Yu, L., Gai, Y., Xiao, T., He, T., Karypis, G., Li, J., and Zhang, Z. (2020b). Deep Graph Library : A Graph-Centric, Highly-Performant Package for Graph Neural Networks.

[Wang and Vasconcelos, 2020] Wang, P. and Vasconcelos, N. (2020). SCOUT : Self-Aware Discriminant Counterfactual Explanations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8978–8987, Seattle, WA, USA. IEEE.

[Wang et al., 2020c] Wang, R., Wang, X., and Inouye, D. I. (2020c). Shapley Explanation Networks.

[Wang et al., 2019a] Wang, S., Zhou, T., and Bilmes, J. (2019a). Bias Also Matters : Bias Attribution for Deep Neural Network Explanation. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6659–6667. PMLR.

[Wang et al., a] Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., and MacNeille, P. A Bayesian Framework for Learning Rule Sets for Interpretable Classification.

[Wang et al., 2020d] Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., and Tang, J. (2020d). KEPLER : A Unified Model for Knowledge Embedding and Pre-trained Language Representation.

[Wang et al., 2021d] Wang, X., Ji, H., Shi, C., Wang, B., Cui, P., Yu, P., and Ye, Y. (2021d). Heterogeneous Graph Attention Network.

[Wang et al., 2021e] Wang, X., Ji, H., Shi, C., Wang, B., Cui, P., Yu, P., and Ye, Y. (2021e). Heterogeneous Graph Attention Network. Technical Report arXiv :1903.07293, arXiv.

[Wang and Shen, 2023] Wang, X. and Shen, H.-W. (2023). GNNInterpreter : A Probabilistic Generative Model-Level Explanation for Graph Neural Networks. Technical Report arXiv :2209.07924, arXiv.

[Wang et al., 2020e] Wang, X., Wu, Y., Zhang, A., He, X., and Chua, T.-s. (2020e). Causal Screening to Interpret Graph Neural Networks.

[Wang et al., b] Wang, X., Wu, Y.-X., Zhang, A., He, X., and Chua, T.-S. Towards Multi-Grained Explainability for Graph Neural Networks.

[Wang and Zhang, 2022] Wang, X. and Zhang, M. (2022). How Powerful are Spectral Graph Neural Networks. In *Proceedings of the 39th International Conference on Machine Learning*, pages 23341–23362. PMLR.

[Wang et al., 2021f] Wang, Y., Ding, Q., Wang, K., Liu, Y., Wu, X., Wang, J., Liu, Y., and Miao, C. (2021f). The Skyline of Counterfactual Explanations for Machine Learning Decision Models. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2030–2039, Virtual Event Queensland Australia. ACM.

[Wang et al., c] Wang, Y., Feng, B., Li, G., Li, S., Deng, L., Xie, Y., and Ding, Y. GNNAdvisor : An Adaptive and Efficient Runtime System for GNN Acceleration on GPUs.

[Wang et al., 2023b] Wang, Y., Huang, M., Deng, H., Li, W., Wu, Z., Tang, Y., and Liu, G. (2023b). Identification of vital chemical information via visualization of graph neural networks. *Briefings in Bioinformatics*, 24(1) :bbac577.

[Wang et al., 2019b] Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., and Solomon, J. M. (2019b). Dynamic Graph CNN for Learning on Point Clouds.

[Wang et al., 2022] Wang, Y., Yi, K., Liu, X., Wang, Y. G., and Jin, S. (2022). ACMP : Allen-Cahn Message Passing with Attractive and Repulsive Forces for Graph Neural Networks.

[Wang et al., 2020f] Wang, Y. G., Li, M., Ma, Z., Montufar, G., Zhuang, X., and Fan, Y. (2020f). Haar Graph Pooling. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9952–9962. PMLR.

[Wang et al., 2020g] Wang, Y. G., Li, M., Ma, Z., Montufar, G., Zhuang, X., and Fan, Y. (2020g). Haar Graph Pooling. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9952–9962. PMLR.

[Wang et al., 2020h] Wang, Z., Guan, Y., Sun, G., Niu, D., Wang, Y., Zheng, H., and Han, Y. (2020h). GNN-PIM : A Processing-in-Memory Architecture for Graph Neural Networks. In Dong, D., Gong, X., Li, C., Li, D., and Wu, J., editors, *Advanced Computer Architecture*, volume 1256, pages 73–86. Springer Singapore, Singapore.

[Wang et al., 2018b] Wang, Z., Lv, Q., Lan, X., and Zhang, Y. (2018b). Cross-lingual Knowledge Graph Alignment via Graph Convolutional Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 349–357, Brussels, Belgium. Association for Computational Linguistics.

[Wang et al., 2021g] Wang, Z., Ye, X., Wang, C., Cui, J., and Yu, P. S. (2021g). Network Embedding with Completely-imbalanced Labels. *IEEE Transactions on Knowledge and Data Engineering*, 33(11) :3634–3647.

[Wang et al., 2023c] Wang, Z. J., Wortman Vaughan, J., Caruana, R., and Chau, D. H. (2023c). GAM Coach : Towards Interactive and User-centered Algorithmic Recourse. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20, Hamburg Germany. ACM.

[Warmsley et al., 2022] Warmsley, D., Waagen, A., Xu, J., Liu, Z., and Tong, H. (2022). A Survey of Explainable Graph Neural Networks for Cyber Malware Analysis. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 2932–2939.

[Weber et al., 2019] Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., and Leiserson, C. E. (2019). Anti-Money Laundering in Bitcoin : Experimenting with Graph Convolutional Networks for Financial Forensics. Technical Report arXiv :1908.02591, arXiv.

[Wei et al., a] Wei, D., Dash, S., Gao, T., and Günlük, O. Generalized Linear Rule Models.

[Wei et al., b] Wei, D., Nair, R., Dhurandhar, A., Varshney, K. R., Daly, E. M., and Singh, M. On the Safety of Interpretable Machine Learning : A Maximum Deviation Approach.

[Wei et al., 2017] Wei, Y., Feng, J., Liang, X., Cheng, M.-M., Zhao, Y., and Yan, S. (2017). Object Region Mining with Adversarial Erasing : A Simple Classification to Semantic Segmentation Approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6488–6496, Honolulu, HI. IEEE.

[Weis et al., 2020] Weis, S., Patil, K. R., Hoffstaedter, F., Nostro, A., Yeo, B. T. T., and Eickhoff, S. B. (2020). Sex Classification by Resting State Brain Connectivity. *Cerebral Cortex*, 30(2) :824–835.

[Weisfeiler and Leman, ] Weisfeiler, B. Y. and Leman, A. A. THE REDUCTION OF A GRAPH TO CANONICAL FORM AND THE ALGEBRA WHICH APPEARS THEREIN.

[Wellawatte et al., 2021] Wellawatte, G. P., Seshadri, A., and White, A. D. (2021). Model agnostic generation of counterfactual explanations for molecules.

[Weller, 2019] Weller, A. (2019). Transparency : Motivations and Challenges. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R., editors, *Explainable AI : Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pages 23–40. Springer International Publishing, Cham.

[Wexler et al., 2019] Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viegas, F., and Wilson, J. (2019). The What-If Tool : Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1.

[White, 2020] White, A. (2020). Measurable Counterfactual Local Explanations for Any Classifier. *Santiago de Compostela*.

[White and d'Avila Garcez, 2019] White, A. and d'Avila Garcez, A. (2019). Measurable Counterfactual Local Explanations for Any Classifier. Technical Report arXiv :1908.03020, arXiv.

[Wijesinghe and Wang, 2021] Wijesinghe, A. and Wang, Q. (2021). A New Perspective on "How Graph Neural Networks Go Beyond Weisfeiler-Lehman ?".

[Wijesinghe and Wang, 2019] Wijesinghe, W. O. K. A. S. and Wang, Q. (2019). DFNets : Spectral CNNs for Graphs with Feedback-Looped Filters. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[Winn et al., 2005] Winn, J., Criminisi, A., and Minka, T. (2005). Object categorization by learned universal visual dictionary. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 1800–1807 Vol. 2, Beijing, China. IEEE.

[Wolpert and Macready, 1997] Wolpert, D. and Macready, W. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1) :67–82.

[Wrobel, ] Wrobel, S. ERIK.STRUMBELJ@FRI.UNI-LJ.SI IGOR.KONONENKO@FRI.UNI-LJ.SI.

[Wu et al., 2019a] Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., and Weinberger, K. (2019a). Simplifying Graph Convolutional Networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6861–6871. PMLR.

[Wu et al., 2022a] Wu, F., Wu, L., Li, S., Radev, D., and Huang, W. (2022a). Rethinking the Explanation of Graph Neural Network via Non-parametric Subgraph Matching.

[Wu et al., 2019b] Wu, F., Zhang, T., de Souza Jr., A. H., Fifty, C., Yu, T., and Weinberger, K. Q. (2019b). Simplifying Graph Convolutional Networks. Technical Report arXiv :1902.07153, arXiv.

[Wu et al., 2019c] Wu, F., Zhang, T., de Souza Jr., A. H., Fifty, C., Yu, T., and Weinberger, K. Q. (2019c). Simplifying Graph Convolutional Networks.

[Wu et al., 2021a] Wu, H., Chen, W., Xu, S., and Xu, B. (2021a). Counterfactual Supporting Facts Extraction for Explainable Medical Record Based Diagnosis with Graph Network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 1942–1955, Online. Association for Computational Linguistics.

[Wu et al., 2019d] Wu, S., Tang, Y., Zhu, Y., Wang, L., Xie, X., and Tan, T. (2019d). Session-based Recommendation with Graph Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01) :346–353.

[Wu et al., 2020] Wu, T., Ren, H., Li, P., and Leskovec, J. (2020). Graph Information Bottleneck. In *Advances in Neural Information Processing Systems*, volume 33, pages 20437–20448. Curran Associates, Inc.

[Wu et al., 2022b] Wu, Y., Chen, C., Che, J., and Pu, S. (2022b). FAM : Visual Explanations for the Feature Representations from Deep Convolutional Networks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10297–10306, New Orleans, LA, USA. IEEE.

[Wu et al., 2021b] Wu, Y., Ma, K., Cai, Z., Jin, T., Li, B., Zheng, C., Cheng, J., and Yu, F. (2021b). Seastar : Vertex-centric programming for graph neural networks. In *Proceedings of the Sixteenth European Conference on Computer Systems*, pages 359–375, Online Event United Kingdom. ACM.

[Wu et al., 2021c] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2021c). A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1) :4–24.

[Wu et al., 2017] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. (2017). MoleculeNet : A benchmark for molecular machine learning. *Chemical Science*, 9(2) :513–530.

[Wu et al., 2018] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. (2018). MoleculeNet : A Benchmark for Molecular Machine Learning. Technical Report arXiv :1703.00564, arXiv.

[Wu et al., 2023] Wu, Z., Wang, J., Du, H., Jiang, D., Kang, Y., Li, D., Pan, P., Deng, Y., Cao, D., Hsieh, C.-Y., and Hou, T. (2023). Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nature Communications*, 14(1) :2585.

[Xhonneux et al., 2020] Xhonneux, L.-P., Qu, M., and Tang, J. (2020). Continuous Graph Neural Networks. In *Proceedings of the 37th International Conference on Machine Learning*, pages 10432–10441. PMLR.

[Xia et al., 2022] Xia, W., Lai, M., Shan, C., Zhang, Y., Dai, X., Li, X., and Li, D. (2022). Explaining Temporal Graph Models through an Explorer-Navigator Framework.

[Xie and Grossman, 2018] Xie, T. and Grossman, J. C. (2018). Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters*, 120(14) :145301.

[Xie et al., 2022] Xie, Y., Katariya, S., Tang, X., Huang, E., Rao, N., Subbian, K., and Ji, S. (2022). Task-Agnostic Graph Explanations.

[Xie, ] Xie, Z. Graphiler : Optimizing Graph Neural Networks with Message Passing Data Flow Graph.

[Xinyi and Chen, 2018] Xinyi, Z. and Chen, L. (2018). Capsule Graph Neural Network.

[Xinyi and Chen, 2019] Xinyi, Z. and Chen, L. (2019). CAPSULE GRAPH NEURAL NETWORK.

[Xiong et al., 2020] Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., Li, Z., Luo, X., Chen, K., Jiang, H., and Zheng, M. (2020). Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *Journal of Medicinal Chemistry*, 63(16) :8749–8760.

[Xu et al., 2018] Xu, B., Shen, H., Cao, Q., Qiu, Y., and Cheng, X. (2018). Graph Wavelet Neural Network.

[Xu et al., 2019a] Xu, B., Shen, H., Cao, Q., Qiu, Y., and Cheng, X. (2019a). Graph Wavelet Neural Network.

[Xu et al., 2015] Xu, C., Li, C., Wu, H., Wu, Y., Hu, S., Zhu, Y., Zhang, W., Wang, L., Zhu, S., Liu, J., Zhang, Q., Yang, J., and Zhang, X. (2015). Gender Differences in Cerebral Regional Homogeneity of Adult Healthy Volunteers : A Resting-State fMRI Study. *BioMed Research International*, 2015 :1–8.

[Xu et al., 2021a] Xu, H., Sang, S., Yao, H., Herghelegiu, A. I., Lu, H., Yurkovich, J. T., and Yang, L. (2021a). APRILE : Exploring the Molecular Mechanisms of Drug Side Effects with Explainable Graph Neural Networks. Technical report, bioRxiv.

[Xu* et al., 2018] Xu*, K., Hu*, W., Leskovec, J., and Jegelka, S. (2018). How Powerful are Graph Neural Networks ?

[Xu et al., 2019b] Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2019b). How Powerful are Graph Neural Networks ? Technical Report arXiv :1810.00826, arXiv.

[Xu et al., 2018] Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K.-i., and Jegelka, S. (2018). Representation Learning on Graphs with Jumping Knowledge Networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5453–5462. PMLR.

[Xu et al., 2021b] Xu, K., Zhang, M., Jegelka, S., and Kawaguchi, K. (2021b). Optimization of Graph Neural Networks : Implicit Acceleration by Skip Connections and More Depth. In *Proceedings of the 38th International Conference on Machine Learning*, pages 11592–11602. PMLR.

[Xu et al., 2020] Xu, S., Venugopalan, S., and Sundararajan, M. (2020). Attribution in Scale and Space. Technical Report arXiv :2004.03383, arXiv.

[Xuanyuan et al., 2023] Xuanyuan, H., Barbiero, P., Georgiev, D., Magister, L. C., and Lió, P. (2023). Global Concept-Based Interpretability for Graph Neural Networks via Neuron Analysis. Technical Report arXiv :2208.10609, arXiv.

[Yan et al., 2020] Yan, M., Deng, L., Hu, X., Liang, L., Feng, Y., Ye, X., Zhang, Z., Fan, D., and Xie, Y. (2020). HyGCN : A GCN Accelerator with Hybrid Architecture. Technical Report arXiv :2001.02514, arXiv.

[Yan and Procaccia, 2021] Yan, T. and Procaccia, A. D. (2021). If You Like Shapley Then You'll Love the Core. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6) :5751–5759.

[Yanardag and Vishwanathan, 2015] Yanardag, P. and Vishwanathan, S. (2015). Deep Graph Kernels. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1365–1374, Sydney NSW Australia. ACM.

[Yang et al., 2015] Yang, B., Yih, W.-t., He, X., Gao, J., and Deng, L. (2015). Embedding Entities and Relations for Learning and Inference in Knowledge Bases. Technical Report arXiv :1412.6575, arXiv.

[Yang et al., 2023a] Yang, C., Huo, K.-B., Geng, L.-F., and Mei, K.-Z. (2023a). DRGN : A dynamically reconfigurable accelerator for graph neural networks. *Journal of Ambient Intelligence and Humanized Computing*, 14(7) :8985–9000.

[Yang et al., 2021a] Yang, C., Liu, J., and Shi, C. (2021a). Extract the Knowledge of Graph Neural Networks and Go Beyond it : An Effective Knowledge Distillation Framework.

[Yang et al., 2021b] Yang, C., Liu, J., and Shi, C. (2021b). Extract the Knowledge of Graph Neural Networks and Go Beyond it : An Effective Knowledge Distillation Framework. In *Proceedings of the Web Conference 2021*, pages 1227–1237, Ljubljana Slovenia. ACM.

[Yang et al., 2020] Yang, C., Wang, R., Yao, S., Liu, S., and Abdelzaher, T. (2020). Revisiting Over-smoothing in Deep GCNs.

[Yang et al., 2023b] Yang, C., Wu, Q., Wang, J., and Yan, J. (2023b). Graph Neural Networks are Inherently Good Generalizers : Insights by Bridging GNNs and MLPs. Technical Report arXiv :2212.09034, arXiv.

[Yang et al., 2021c] Yang, F., Alva, S. S., Chen, J., and Hu, X. (2021c). Model-Based Counterfactual Synthesizer for Interpretation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1964–1974.

[Yang et al., 2019] Yang, F., Du, M., and Hu, X. (2019). Evaluating Explanation Without Ground Truth in Interpretable Machine Learning. Technical Report arXiv :1907.06831, arXiv.

[Yang et al., 2018] Yang, J., Lu, J., Lee, S., Batra, D., and Parikh, D. (2018). Graph R-CNN for Scene Graph Generation. Technical Report arXiv :1808.00191, arXiv.

[Yang et al., 2022a] Yang, J., Tang, D., Song, X., Wang, L., Yin, Q., Chen, R., Yu, W., and Zhou, J. (2022a). GNNLab : A factored system for sample-based GNN training over GPUs. In *Proceedings of the Seventeenth European Conference on Computer Systems*, pages 417–434, Rennes France. ACM.

[Yang et al., 2021d] Yang, J., Zhao, P., Rong, Y., Yan, C., Li, C., Ma, H., and Huang, J. (2021d). Hierarchical Graph Capsule Network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12) :10603–10611.

[Yang and Kim, 2019] Yang, M. and Kim, B. (2019). Benchmarking Attribution Methods with Relative Feature Importance.

[Yang et al., 2022b] Yang, M., Shen, Y., Li, R., Qi, H., Zhang, Q., and Yin, B. (2022b). A New Perspective on the Effects of Spectrum in Graph Neural Networks. In *Proceedings of the 39th International Conference on Machine Learning*, pages 25261–25279. PMLR.

[Yang et al., 2022c] Yang, N., Kang, T., and Jung, K. (2022c). Deriving Explainable Discriminative Attributes Using Confusion About Counterfactual Class. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1730–1734.

[Yang et al., 2023c] Yang, Q., Ma, C., Zhang, Q., Gao, X., Zhang, C., and Zhang, X. (2023c). Interpretable Research Interest Shift Detection with Temporal Heterogeneous Graphs. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, WSDM 2023, Singapore, 27 February 2023 - 3 March 2023*.

[Yang et al., 2023d] Yang, R., Shi, J., Xiao, X., Yang, Y., Bhowmick, S. S., and Liu, J. (2023d). PANE : Scalable and effective attributed network embedding. *The VLDB Journal*.

[Yang et al., 2022d] Yang, S. C.-H., Folke, T., and Shafto, P. (2022d). A Psychological Theory of Explainability. Technical Report arXiv :2205.08452, arXiv.

[Yang et al., 2021e] Yang, Y., Liu, T., Wang, Y., Zhou, J., Gan, Q., Wei, Z., Zhang, Z., Huang, Z., and Wipf, D. (2021e). Graph Neural Networks Inspired by Classical Iterative Algorithms. In *Proceedings of the 38th International Conference on Machine Learning*, pages 11773–11783. PMLR.

[Yang et al., 2021f] Yang, Y., Qiu, J., Song, M., Tao, D., and Wang, X. (2021f). Distilling Knowledge from Graph Convolutional Networks.

[Yang et al., 2021g] Yang, Y., Qiu, J., Song, M., Tao, D., and Wang, X. (2021g). Distilling Knowledge from Graph Convolutional Networks.

[Yang et al., 2016] Yang, Z., Cohen, W. W., and Salakhutdinov, R. (2016). Revisiting Semi-Supervised Learning with Graph Embeddings. Technical Report arXiv :1603.08861, arXiv.

[Yao et al., 2018] Yao, L., Mao, C., and Luo, Y. (2018). Graph Convolutional Networks for Text Classification. Technical Report arXiv :1809.05679, arXiv.

[Ye et al., 2019] Ye, Z., Liu, K. S., Ma, T., Gao, J., and Chen, C. (2019). Curvature Graph Network.

[Yeh et al., 2019a] Yeh, C.-K., Hsieh, C.-Y., Suggala, A. S., Inouye, D. I., and Ravikumar, P. (2019a). On the (In)fidelity and Sensitivity for Explanations.

[Yeh et al., 2019b] Yeh, C.-K., Hsieh, C.-Y., Suggala, A. S., Inouye, D. I., and Ravikumar, P. (2019b). On the (In)fidelity and Sensitivity for Explanations. Technical Report arXiv :1901.09392, arXiv.

[Yeh et al., ] Yeh, C.-K., Kim, B., Arık, S. Ö., Li, C.-L., Pfister, T., and Ravikumar, P. On Completeness-aware Concept-Based Explanations in Deep Neural Networks.

[Yeh et al., 2018] Yeh, C.-K., Kim, J. S., Yen, I. E. H., and Ravikumar, P. (2018). Representer Point Selection for Explaining Deep Neural Networks.

[Yi et al., 2016a] Yi, L., Kim, V. G., Ceylan, D., Shen, I.-C., Yan, M., Su, H., Lu, C., Huang, Q., Sheffer, A., and Guibas, L. (2016a). A scalable active framework for region annotation in 3D shape collections. *ACM Transactions on Graphics*, 35(6) :1–12.

[Yi et al., 2016b] Yi, L., Su, H., Guo, X., and Guibas, L. (2016b). SyncSpecCNN : Synchronized Spectral CNN for 3D Shape Segmentation. Technical Report arXiv :1612.00606, arXiv.

[Yin et al., 2017] Yin, H., Benson, A. R., Leskovec, J., and Gleich, D. F. (2017). Local Higher-Order Graph Clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 555–564, Halifax NS Canada. ACM.

[Ying et al., 2021] Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., and Liu, T.-Y. (2021). Do Transformers Really Perform Badly for Graph Representation ?

[Ying et al., 2019a] Ying, R., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. (2019a). GNNExplainer : Generating Explanations for Graph Neural Networks.

[Ying et al., 2018a] Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L., and Leskovec, J. (2018a). Graph Convolutional Neural Networks for Web-Scale Recommender Systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 974–983.

[Ying et al., 2019b] Ying, R., You, J., Morris, C., Ren, X., Hamilton, W. L., and Leskovec, J. (2019b). Hierarchical Graph Representation Learning with Differentiable Pooling. Technical Report arXiv :1806.08804, arXiv.

[Ying et al., 2019c] Ying, Z., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. (2019c). GNNExplainer : Generating Explanations for Graph Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[Ying et al., 2018b] Ying, Z., You, J., Morris, C., Ren, X., Hamilton, W., and Leskovec, J. (2018b). Hierarchical Graph Representation Learning with Differentiable Pooling. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

[You et al., 2021] You, J., Ying, R., and Leskovec, J. (2021). Design Space for Graph Neural Networks. Technical Report arXiv :2011.08843, arXiv.

[You et al., 2018] You, J., Ying, R., Ren, X., Hamilton, W., and Leskovec, J. (2018). GraphRNN : Generating Realistic Graphs with Deep Auto-regressive Models. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5708–5717. PMLR.

[You et al., 2020] You, Y., Chen, T., Wang, Z., and Shen, Y. (2020). L2-GCN : Layer-Wise and Learned Efficient Training of Graph Convolutional Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2127–2135.

[Yu et al., 2018a] Yu, B., Yin, H., and Zhu, Z. (2018a). Spatio-Temporal Graph Convolutional Networks : A Deep Learning Framework for Traffic Forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 3634–3640.

[Yu et al., 2022] Yu, H., Wang, L., Wang, B., Liu, M., Yang, T., and Ji, S. (2022). GraphFM : Improving Large-Scale GNN Training via Feature Momentum.

[Yu et al., 2020] Yu, J., Xu, T., Rong, Y., Bian, Y., Huang, J., and He, R. (2020). Graph Information Bottleneck for Subgraph Recognition. Technical Report arXiv :2010.05563, arXiv.

[Yu et al., 2023] Yu, P., Xu, C., Bifet, A., and Read, J. (2023). Linear TreeShap.

[Yu et al., 2018b] Yu, W., Zheng, C., Cheng, W., Aggarwal, C. C., Song, D., Zong, B., Chen, H., and Wang, W. (2018b). Learning Deep Network Representations with Adversarially Regularized Autoencoders. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2663–2671, London United Kingdom. ACM.

[Yu et al., 2018c] Yu, W., Zheng, C., Cheng, W., Aggarwal, C. C., Song, D., Zong, B., Chen, H., and Wang, W. (2018c). Learning Deep Network Representations with Adversarially Regularized Autoencoders. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2663–2671, London United Kingdom. ACM.

[Yu and Gao, 2023] Yu, Z. and Gao, H. (2023). MotifExplainer : A Motif-based Graph Neural Network Explainer. Technical Report arXiv :2202.00519, arXiv.

[Yuan and Ji, 2019] Yuan, H. and Ji, S. (2019). StructPool : Structured Graph Pooling via Conditional Random Fields.

[Yuan et al., 2020] Yuan, H., Tang, J., Hu, X., and Ji, S. (2020). XGNN : Towards Model-Level Explanations of Graph Neural Networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 430–438, Virtual Event CA USA. ACM.

[Yuan et al., 2022] Yuan, H., Yu, H., Gui, S., and Ji, S. (2022). Explainability in Graph Neural Networks : A Taxonomic Survey. Technical Report arXiv :2012.15445, arXiv.

[Yuan et al., 2021] Yuan, H., Yu, H., Wang, J., Li, K., and Ji, S. (2021). On Explainability of Graph Neural Networks via Subgraph Explorations. Technical Report arXiv :2102.05152, arXiv.

[Yun et al., 2019] Yun, S., Jeong, M., Kim, R., Kang, J., and Kim, H. J. (2019). Graph Transformer Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[Yutao Hu, ] Yutao Hu, S. W. A Slice-level vulnerability detection and interpretation method based on graph neural network. *Journal of Software*.

[Zachary, 1977] Zachary, W. W. (1977). An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research*, 33(4) :452–473.

[Zang and Wang, 2020] Zang, C. and Wang, F. (2020). Neural Dynamics on Complex Networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 892–902, Virtual Event CA USA. ACM.

[Zech et al., 2018] Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs : A cross-sectional study. *PLOS Medicine*, 15(11) :e1002683.

[Zeiler and Fergus, 2013] Zeiler, M. D. and Fergus, R. (2013). Visualizing and Understanding Convolutional Networks. Technical Report arXiv :1311.2901, arXiv.

[Zeiler and Fergus, 2014a] Zeiler, M. D. and Fergus, R. (2014a). Visualizing and Understanding Convolutional Networks. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 818–833, Cham. Springer International Publishing.

[Zeiler and Fergus, 2014b] Zeiler, M. D. and Fergus, R. (2014b). Visualizing and Understanding Convolutional Networks. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 818–833, Cham. Springer International Publishing.

[Zemni et al., 2023] Zemni, M., Chen, M., Zablocki, É., Ben-Younes, H., Pérez, P., and Cord, M. (2023). OCTET : Object-aware Counterfactual Explanations. Technical Report arXiv :2211.12380, arXiv.

[Zeng and Prasanna, 2020a] Zeng, H. and Prasanna, V. (2020a). GraphACT : Accelerating GCN Training on CPU-FPGA Heterogeneous Platforms. In *Proceedings of the 2020 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pages 255–265.

[Zeng and Prasanna, 2020b] Zeng, H. and Prasanna, V. (2020b). GraphACT : Accelerating GCN Training on CPU-FPGA Heterogeneous Platforms. In *Proceedings*

*of the 2020 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pages 255–265.

[Zeng et al., 2022] Zeng, H., Zhang, M., Xia, Y., Srivastava, A., Malevich, A., Kannan, R., Prasanna, V., Jin, L., and Chen, R. (2022). Decoupling the Depth and Scope of Graph Neural Networks. Technical Report arXiv :2201.07858, arXiv.

[Zeng et al., 2019a] Zeng, H., Zhou, H., Srivastava, A., Kannan, R., and Prasanna, V. (2019a). Accurate, Efficient and Scalable Graph Embedding. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 462–471.

[Zeng et al., 2019b] Zeng, H., Zhou, H., Srivastava, A., Kannan, R., and Prasanna, V. (2019b). GraphSAINT : Graph Sampling Based Inductive Learning Method.

[Zeng et al., 2020a] Zeng, H., Zhou, H., Srivastava, A., Kannan, R., and Prasanna, V. (2020a). GraphSAINT : Graph Sampling Based Inductive Learning Method.

[Zeng et al., 2020b] Zeng, H., Zhou, H., Srivastava, A., Kannan, R., and Prasanna, V. (2020b). GraphSAINT : Graph Sampling Based Inductive Learning Method. Technical Report arXiv :1907.04931, arXiv.

[Zerilli et al., 2019] Zerilli, J., Knott, A., Maclaurin, J., and Gavaghan, C. (2019). Transparency in Algorithmic and Human Decision-Making : Is There a Double Standard ? *Philosophy & Technology*, 32(4) :661–683.

[Zhang et al., 2020a] Zhang, B., Zeng, H., and Prasanna, V. (2020a). Hardware Acceleration of Large Scale GCN Inference. In *2020 IEEE 31st International Conference on Application-specific Systems, Architectures and Processors (ASAP)*, pages 61–68.

[Zhang et al., 2022a] Zhang, B., Zheng, W., Zhou, J., and Lu, J. (2022a). Bort : Towards Explainable Neural Networks with Bounded Orthogonal Constraint.

[Zhang et al., 2019a] Zhang, C., Song, D., Huang, C., Swami, A., and Chawla, N. V. (2019a). Heterogeneous Graph Neural Network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 793–803, Anchorage AK USA. ACM.

[Zhang et al., 2020b] Zhang, D., Huang, X., Liu, Z., Hu, Z., Song, X., Ge, Z., Zhang, Z., Wang, L., Zhou, J., Shuang, Y., and Qi, Y. (2020b). AGL : A Scalable System for Industrial-purpose Graph Machine Learning.

[Zhang, 2022] Zhang, F. (2022). TGP : Explainable Temporal Graph Neural Networks for Personalized Recommendation.

[Zhang et al., 2022b] Zhang, H., Wu, B., Yuan, X., Pan, S., Tong, H., and Pei, J. (2022b). Trustworthy Graph Neural Networks : Aspects, Methods and Trends. Technical Report arXiv :2205.07424, arXiv.

[Zhang et al., 2021a] Zhang, H., Yu, Z., Dai, G., Huang, G., Ding, Y., Xie, Y., and Wang, Y. (2021a). Understanding GNN Computational Graph : A Coordinated Computation, IO, and Memory Perspective.

[Zhang et al., 2021b] Zhang, H., Yu, Z., Dai, G., Huang, G., Ding, Y., Xie, Y., and Wang, Y. (2021b). Understanding GNN Computational Graph : A Coordinated Computation, IO, and Memory Perspective.

[Zhang et al., 2018a] Zhang, J., Bargal, S. A., Lin, Z., Brandt, J., Shen, X., and Sclaroff, S. (2018a). Top-Down Neural Attention by Excitation Backprop. *International Journal of Computer Vision*, 126(10) :1084–1102.

[Zhang et al., 2016] Zhang, J., Lin, Z., Brandt, J., Shen, X., and Sclaroff, S. (2016). Top-down Neural Attention by Excitation Backprop. Technical Report arXiv :1608.00507, arXiv.

[Zhang et al., a] Zhang, J., Shi, X., Xie, J., Ma, H., King, I., and Yeung, D.-Y. GaAN : Gated Attention Networks for Learning on Large and Spatiotemporal Graphs.

[Zhang et al., 2018b] Zhang, J., Shi, X., Xie, J., Ma, H., King, I., and Yeung, D.-Y. (2018b). GaAN : Gated Attention Networks for Learning on Large and Spatiotemporal Graphs.

[Zhang et al., 2018c] Zhang, J., Shi, X., Xie, J., Ma, H., King, I., and Yeung, D.-Y. (2018c). GaAN : Gated Attention Networks for Learning on Large and Spatiotemporal Graphs. Technical Report arXiv :1803.07294, arXiv.

[Zhang et al., 2019b] Zhang, J., Wang, Y., Molino, P., Li, L., and Ebert, D. S. (2019b). Manifold : A Model-Agnostic Framework for Interpretation and Diagno-

sis of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, 25(1) :364–373.

[Zhang et al., 2021c] Zhang, L., Lai, Z., Li, S., Tang, Y., Liu, F., and Li, D. (2021c). 2PGraph : Accelerating GNN Training over Large Graphs on GPU Clusters. In *2021 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 103–113.

[Zhang et al., 2020c] Zhang, L., Wang, X., Li, H., Zhu, G., Shen, P., Li, P., Lu, X., Shah, S. A. A., and Bennamoun, M. (2020c). Structure-Feature based Graph Self-adaptive Pooling. In *Proceedings of The Web Conference 2020*, pages 3098–3104, Taipei Taiwan. ACM.

[Zhang and Chen, 2018] Zhang, M. and Chen, Y. (2018). Link Prediction Based on Graph Neural Networks. Technical Report arXiv :1802.09691, arXiv.

[Zhang and Chen, 2020] Zhang, M. and Chen, Y. (2020). Inductive Matrix Completion Based on Graph Neural Networks. Technical Report arXiv :1904.12058, arXiv.

[Zhang et al., 2018d] Zhang, M., Cui, Z., Neumann, M., and Chen, Y. (2018d). An End-to-End Deep Learning Architecture for Graph Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

[Zhang et al., 2018e] Zhang, M., Cui, Z., Neumann, M., and Chen, Y. (2018e). An End-to-End Deep Learning Architecture for Graph Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

[Zhang et al., 2018f] Zhang, Q., Wu, Y. N., and Zhu, S.-C. (2018f). Interpretable Convolutional Neural Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8827–8836, Salt Lake City, UT. IEEE.

[Zhang et al., b] Zhang, S., Liu, Y., Shah, N., and Sun, Y. GStarX : Explaining Graph Neural Networks with Structure-Aware Cooperative Games.

[Zhang et al., 2023] Zhang, S., Zhang, J., Song, X., Adeshina, S., Zheng, D., Faloutsos, C., and Sun, Y. (2023). PaGE-Link : Path-based Graph Neural Network Explanation for Heterogeneous Link Prediction. Technical Report arXiv :2302.12465, arXiv.

[Zhang and Lim, 2022] Zhang, W. and Lim, B. Y. (2022). Towards Relatable Explainable AI with the Perceptual Process. In *CHI Conference on Human Factors in Computing Systems*, pages 1–24.

[Zhang et al., 2019c] Zhang, W., Liu, H., Liu, Y., Zhou, J., and Xiong, H. (2019c). Semi-Supervised Hierarchical Recurrent Graph Neural Network for City-Wide Parking Availability Prediction.

[Zhang et al., 2022c] Zhang, W., Shen, Y., Lin, Z., Li, Y., Li, X., Ouyang, W., Tao, Y., Yang, Z., and Cui, B. (2022c). PaSca : A Graph Neural Architecture Search System under the Scalable Paradigm. In *Proceedings of the ACM Web Conference 2022*, pages 1817–1828, Virtual Event, Lyon France. ACM.

[Zhang et al., 2021d] Zhang, W., Yang, M., Sheng, Z., Li, Y., Ouyang, W., Tao, Y., Yang, Z., and Cui, B. (2021d). Node Dependent Local Smoothing for Scalable Graph Learning.

[Zhang et al., 2021e] Zhang, X., He, Y., Brugnone, N., Perlmutter, M., and Hirn, M. (2021e). MagNet : A Neural Network for Directed Graphs.

[Zhang et al., 2019d] Zhang, X., Li, Y., Shen, D., and Carin, L. (2019d). Diffusion Maps for Textual Network Embedding. Technical Report arXiv :1805.09906, arXiv.

[Zhang et al., c] Zhang, X., Solar-Lezama, A., and Singh, R. Interpreting Neural Network Judgments via Minimal, Stable, and Symbolic Corrections.

[Zhang et al., 2018g] Zhang, X., Solar-Lezama, A., and Singh, R. (2018g). Interpreting Neural Network Judgments via Minimal, Stable, and Symbolic Corrections. Technical Report arXiv :1802.07384, arXiv.

[Zhang et al., 2019e] Zhang, X., Tan, S., Koch, P., Lou, Y., Chajewska, U., and Caruana, R. (2019e). Axiomatic Interpretability for Multiclass Additive Models. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 226–234, Anchorage AK USA. ACM.

[Zhang and Zitnik, 2020] Zhang, X. and Zitnik, M. (2020). GNNGuard : Defending Graph Neural Networks against Adversarial Attacks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9263–9275. Curran Associates, Inc.

[Zhang et al., 2018h] Zhang, Y., Pal, S., Coates, M., and Üstebay, D. (2018h). Bayesian graph convolutional neural networks for semi-supervised classification. Technical Report arXiv :1811.11103, arXiv.

[Zhang et al., 2019f] Zhang, Y., Pal, S., Coates, M., and Ustebay, D. (2019f). Bayesian Graph Convolutional Neural Networks for Semi-Supervised Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01) :5829–5836.

[Zhang et al., 2021f] Zhang, Y., Regol, F., Pal, S., Khan, S., Ma, L., and Coates, M. (2021f). Detection and Defense of Topological Adversarial Attacks on Graphs. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 2989–2997. PMLR.

[Zhang et al., 2019g] Zhang, Y., Song, K., Sun, Y., Tan, S., and Udell, M. (2019g). "Why Should You Trust My Explanation ?" Understanding Uncertainty in LIME Explanations. Technical Report arXiv :1904.12991, arXiv.

[Zhang et al., 2021g] Zhang, Y., Wang, X., Shi, C., Liu, N., and Song, G. (2021g). Lorentzian Graph Convolutional Networks. In *Proceedings of the Web Conference 2021*, pages 1249–1261, Ljubljana Slovenia. ACM.

[Zhang et al., 2020d] Zhang, Z., Cui, P., and Zhu, W. (2020d). Deep Learning on Graphs : A Survey. Technical Report arXiv :1812.04202, arXiv.

[Zhang et al., 2022d] Zhang, Z., Cui, P., and Zhu, W. (2022d). Deep Learning on Graphs : A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(1) :249–270.

[Zhang et al., 2021h] Zhang, Z., Wang, X., and Zhu, W. (2021h). Automated Machine Learning on Graphs : A Survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 4704–4712.

[Zhao et al., 2021a] Zhao, H., Jiang, L., Jia, J., Torr, P., and Koltun, V. (2021a). Point Transformer. Technical Report arXiv :2012.09164, arXiv.

[Zhao et al., 2021b] Zhao, J., Dong, Y., Ding, M., Kharlamov, E., and Tang, J. (2021b). Adaptive Diffusion in Graph Neural Networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 23321–23333. Curran Associates, Inc.

[Zhao et al., 2020a] Zhao, J., Dong, Y., Tang, J., Ding, M., and Wang, K. (2020a). Generalizing Graph Convolutional Networks via Heat Kernel.

[Zhao and Akoglu, 2019] Zhao, L. and Akoglu, L. (2019). PairNorm : Tackling Oversmoothing in GNNs.

[Zhao and Akoglu, 2020] Zhao, L. and Akoglu, L. (2020). PairNorm : Tackling Oversmoothing in GNNs. Technical Report arXiv :1909.12223, arXiv.

[Zhao et al., 2022] Zhao, L., Jin, W., Akoglu, L., and Shah, N. (2022). From Stars to Subgraphs : Uplifting Any GNN with Local Structure Awareness. Technical Report arXiv :2110.03753, arXiv.

[Zhao and Hastie, 2021] Zhao, Q. and Hastie, T. (2021). Causal Interpretations of Black-Box Models. *Journal of Business & Economic Statistics*, 39(1) :272–281.

[Zhao, 2020] Zhao, Y. (2020). Fast Real-time Counterfactual Explanations. Technical Report arXiv :2007.05684, arXiv.

[Zhao et al., 2020b] Zhao, Y., Wang, D., Bates, D., Mullins, R., Jamnik, M., and Lio, P. (2020b). Learned Low Precision Graph Neural Networks.

[Zhao et al., 2020c] Zhao, Y., Wang, D., Gao, X., Mullins, R., Lio, P., and Jamnik, M. (2020c). Probabilistic Dual Network Architecture Search on Graphs.

[Zheng et al., 2022a] Zheng, C., Chen, H., Cheng, Y., Song, Z., Wu, Y., Li, C., Cheng, J., Yang, H., and Zhang, S. (2022a). ByteGNN : Efficient graph neural network training at large scale. *Proceedings of the VLDB Endowment*, 15(6) :1228–1242.

[Zheng et al., 2021a] Zheng, D., Ma, C., Wang, M., Zhou, J., Su, Q., Song, X., Gan, Q., Zhang, Z., and Karypis, G. (2021a). DistDGL : Distributed Graph Neural Network Training for Billion-Scale Graphs. Technical Report arXiv :2010.05337, arXiv.

[Zheng et al., 2022b] Zheng, D., Song, X., Yang, C., LaSalle, D., and Karypis, G. (2022b). Distributed Hybrid CPU and GPU training for Graph Neural Networks on Billion-Scale Graphs. Technical Report arXiv :2112.15345, arXiv.

[Zheng et al., 2021b] Zheng, X., Zhou, B., Gao, J., Wang, Y., Lió, P., Li, M., and Montufar, G. (2021b). How Framelets Enhance Graph Neural Networks. In

*Proceedings of the 38th International Conference on Machine Learning*, pages 12761–12771. PMLR.

[Zhirong Wu et al., 2015] Zhirong Wu, Song, S., Khosla, A., Fisher Yu, Linguang Zhang, Xiaoou Tang, and Xiao, J. (2015). 3D ShapeNets : A deep representation for volumetric shapes. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920, Boston, MA, USA. IEEE.

[Zhou et al., 2019] Zhou, B., Bau, D., Oliva, A., and Torralba, A. (2019). Comparing the Interpretability of Deep Networks via Network Dissection. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R., editors, *Explainable AI : Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pages 243–252. Springer International Publishing, Cham.

[Zhou et al., 2015] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2015). Learning Deep Features for Discriminative Localization. Technical Report arXiv :1512.04150, arXiv.

[Zhou et al., 2018] Zhou, B., Sun, Y., Bau, D., and Torralba, A. (2018). Interpretable Basis Decomposition for Visual Explanation. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, volume 11212, pages 122–138. Springer International Publishing, Cham.

[Zhou et al., 2022a] Zhou, H., He, L., Zhang, Y., Shen, L., and Chen, B. (2022a). Interpretable Graph Convolutional Network Of Multi-Modality Brain Imaging For Alzheimer's Disease Diagnosis. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5.

[Zhou et al., 2022b] Zhou, H., Zhang, Y., Chen, B. Y., Shen, L., and He, L. (2022b). Sparse Interpretation of Graph Convolutional Networks for Multi-modal Diagnosis of Alzheimer's Disease. In Wang, L., Dou, Q., Fletcher, P. T., Speidel, S., and Li, S., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, Lecture Notes in Computer Science, pages 469–478, Cham. Springer Nature Switzerland.

[Zhou et al., 2020a] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. (2020a). Graph neural networks : A review of methods and applications. *AI Open*, 1 :57–81.

[Zhou et al., 2021a] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. (2021a). Graph Neural Networks : A Review of Methods and Applications. Technical Report arXiv :1812.08434, arXiv.

[Zhou and Troyanskaya, 2015] Zhou, J. and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12(10) :931–934.

[Zhou et al., 2021b] Zhou, K., Dong, Y., Wang, K., Lee, W. S., Hooi, B., Xu, H., and Feng, J. (2021b). Understanding and Resolving Performance Degradation in Deep Graph Convolutional Networks. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2728–2737, Virtual Event Queensland Australia. ACM.

[Zhou et al., 2020b] Zhou, K., Huang, X., Li, Y., Zha, D., Chen, R., and Hu, X. (2020b). Towards Deeper Graph Neural Networks with Differentiable Group Normalization. In *Advances in Neural Information Processing Systems*, volume 33, pages 4917–4928. Curran Associates, Inc.

[Zhou et al., 2021c] Zhou, Y., Booth, S., Ribeiro, M. T., and Shah, J. (2021c). Do Feature Attribution Methods Correctly Attribute Features ? Technical Report arXiv :2104.14403, arXiv.

[Zhu et al., 2020a] Zhu, H., Feng, F., He, X., Wang, X., Li, Y., Zheng, K., and Zhang, Y. (2020a). Bilinear Graph Neural Network with Neighbor Interactions. volume 2, pages 1452–1458.

[Zhu and Koniusz, 2020] Zhu, H. and Koniusz, P. (2020). Simple Spectral Graph Convolution.

[Zhu et al., 2020b] Zhu, J., Yan, Y., Zhao, L., Heimann, M., Akoglu, L., and Koutra, D. (2020b). Beyond Homophily in Graph Neural Networks : Current Limitations and Effective Designs. In *Advances in Neural Information Processing Systems*, volume 33, pages 7793–7804. Curran Associates, Inc.

[Zhu et al., 2021a] Zhu, M., Wang, X., Shi, C., Ji, H., and Cui, P. (2021a). Interpreting and Unifying Graph Neural Networks with An Optimization Framework. In *Proceedings of the Web Conference 2021*, pages 1215–1226, Ljubljana Slovenia. ACM.

[Zhu et al., 2019] Zhu, R., Zhao, K., Yang, H., Lin, W., Zhou, C., Ai, B., Li, Y., and Zhou, J. (2019). AliGraph : A comprehensive graph neural network platform. *Proceedings of the VLDB Endowment*, 12(12) :2094–2105.

[Zhu and Ghahramani, 2003] Zhu, X. and Ghahramani, Z. (2003). Learning from Labeled and Unlabeled Data with Label Propagation.

[Zhu et al., ] Zhu, X., Lei, Z., Liu, X., Shi, H., and Li, S. Z. Face Alignment Across Large Poses : A 3D Solution. *Computer Vision and Pattern Recognition*, page 10.

[Zhu et al., 2018] Zhu, X., Singla, A., Zilles, S., and Rafferty, A. N. (2018). An Overview of Machine Teaching.

[Zhu et al., 2021b] Zhu, X., Xu, C., and Tao, D. (2021b). Where and What ? Examining Interpretable Disentangled Representations. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5857–5866, Nashville, TN, USA. IEEE.

[Zhu et al., 2022] Zhu, X., Zhang, Y., Zhang, Z., Guo, D., Li, Q., and Li, Z. (2022). Interpretability Evaluation of Botnet Detection Model based on Graph Neural Network. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6.

[Zhuang et al., 2019] Zhuang, J., Dvornek, N., Li, X., and Duncan, J. S. (2019). Ordinary differential equations on graph networks.

[Zintgraf et al., 2017] Zintgraf, L. M., Cohen, T. S., Adel, T., and Welling, M. (2017). Visualizing Deep Neural Network Decisions : Prediction Difference Analysis. Technical Report arXiv :1702.04595, arXiv.

[Zitnik and Leskovec, 2017] Zitnik, M. and Leskovec, J. (2017). Predicting multicellular function through multi-layer tissue networks. *Bioinformatics (Oxford, England)*, 33(14) :i190–i198.

[Zou et al., 2019] Zou, D., Hu, Z., Wang, Y., Jiang, S., Sun, Y., and Gu, Q. (2019). Layer-Dependent Importance Sampling for Training Deep and Large Graph Convolutional Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[Zou and Lerman, 2020] Zou, D. and Lerman, G. (2020). Graph convolutional neural networks via scattering. *Applied and Computational Harmonic Analysis*, 49(3) :1046–1074.

[Zügner and Günnemann, 2018] Zügner, D. and Günnemann, S. (2018). Adversarial Attacks on Graph Neural Networks via Meta Learning.

[Zügner and Günnemann, 2020] Zügner, D. and Günnemann, S. (2020). Certifiable Robustness of Graph Convolutional Networks under Structure Perturbations. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1656–1665, Virtual Event CA USA. ACM.

# Annexe A

# Moyens techniques mises en oeuvres pour la réalisation de ces travaux

## A.1 Moyens matériels et logiciels

Dans cette partie nous donnons les caractéristiques techniques matérielle et logicielle qui nous ont permit de mener nos travaux :

- nous avons utilisé principalement trois ordinateurs :
  - les spécifications du premier :
    - CPU : Intel(R) Core(TM) i7-10700 CPU @ 2.90GHz
    - GPU : NVIDIA Corporation Quadro P2200 5Go
    - RAM : 32Go
    - OS : Ubuntu 22.04.1 LTS
  - les spécifications du second :
    - CPU : Intel(R) Xeon(R) Silver 4208 CPU @ 2.10GHz
    - GPU : NVIDIA Corporation A100 40Go
    - RAM : 192Go
    - OS : Ubuntu 20.04.1 SMP
  - les spécifications du troisième :
    - CPU : Intel(R) Xeon(R) Silver 4208 CPU @ 2.10GHz
    - GPU : NVIDIA Corporation V100S 32Go
    - RAM : 384 Go
    - OS : Ubuntu 20.04.1 SMP

— nous avons utilisé le language de programmation Python (`https://www.python.org/`) dans sa version 3.0 à 3.10.

# Annexe B

# Compléments scientifiques relatifs aux travaux

## B.1 Choix pour l'orientation scientifique de nos travaux

Nous l'avons évoqué tout au long de ce manuscrit, le comportement d'un modèle d'apprentissage machine, profond ou non, varie selon tout un ensemble de degrés de liberté. Par exemple , la structure du modèle, les propriétés topologiques des espaces sur lequel il opère, les propriétés analytiques des fonctions qu'il utilise, leurs adaptations ou non au contexte du problème a résoudre, les représentations en machine des nombres calculés, les graines d'initialisation des tirages aléatoires en machine, les biais et variances éventuelles des données et leurs éventuels prétraitements, la méthode d'optimisation et ses propriétés, etc. À titre d'exemple, la détermination d'une méthode explicative locale et post hoc, le paradigme se prêtant le mieux aux mesures statistiques usuelles et donc à la possibilité de la valorisation dans la communauté, aurait pu ce faire sous l'angle d'une approche purement inférentielle, combinatoire, algébrique ou analytique ; en s'aidant des théories de l'information ou de l'apprentissage par renforcement, des modèles probabilistes graphiques ou encore de la théorie spectrale, etc. Des formulations pratiques sous l'angle de la coupe de graphe, de la segmentation par exemple, avec des métriques variées et originales. Toutes ces formulations sont surement isomorphes à un certain point mais font appel à des

outils et des communautés variées, avec elles aussi, des possibilités de valorisations variables. Guidés par nos intuitions, nos sensibilités personnelles, les attentes de la communauté pour la reconnaissance de nos travaux et celles que nous souhaitons lui communiquer, nous avons fait des choix et avons fixé certains des degrés de liberté mentionnés tout au long de ce document. Par ailleurs, nous devons garder à l'esprit que les résultats présentés dans les chapitres précédents peuvent varier selon les degrés de liberté que nous avons fixés.

## B.2 Étude de l'état de l'art détaillée

### B.2.1 Les méthodes explicatives pour le modèle de l'apprentissage machine

#### Méthodes explicatives à base de donnée contradictoires

**Étude de l'état de l'art** Dans cette partie, nous evoquons les prinpales études faite sur la littérature à propos de l'explicabilité des méthodes d'apprentissage machine profonde ou non, s'intéréssant aux approches dite de données contradictoire. Les voici : [Karimi et al., 2023, Mishra et al., 2023, Tesic and Hahn, 2022, Guidotti, 2022, Moreira et al., 2022, Spreitzer and Haned, , de Oliveira and Martens, 2021, Keane and Smyth, 2020, Verma et al., 2022, Byrne, 2019, Miller, 2021, Bodria et al., 2023, Madsen et al., 2023, Díaz-Rodríguez et al., 2023, Ali et al., 2023b, Dwivedi et al., 2023a, Nauta et al., 2023, Rudin et al., 2022, Verma et al., 2021, Bhatt et al., 2020, Molnar et al., 2020, Chalkiadakis, , Mittelstadt et al., 2019, Carvalho et al., 2019, Murdoch et al., 2019, Hohman et al., 2019, Miller, 2019a, Yang et al., 2019, Mueller et al., 2019, Adadi and Berrada, 2018, Došilović et al., 2018, Qin et al., 2018, Doshi-Velez and Kim, 2017b, Hoffman and Klein, 2017, Hoffman and Klein, 2017, Klein, 2018, Hoffman et al., 2018, Gevrey et al., 2003]

**Ouvrage de référence** [Dağlarli, 2020, Ancona et al., 2019, Arras et al., 2019, Fong and Vedaldi, 2019, Hansen and Rieger, 2019, Hong et al., 2019, Hu et al., 2019, Kindermans et al., 2019, Montavon, 2019, Montavon et al., 2019, Nguyen et al., 2019, Oh et al., 2019, Samek et al., 2019, Samek and Müller, 2019, Weller, 2019, Zhou et al.,

2019, Miller, 2018, Seifert et al., 2017]

**Méthode explicatives basé sur les données contradictoires** Ici nous évoquons les principales contributions répondant aux approches avec données contradictoires. Nous ne détaillerons pas les différents paradigmes explicatif des méthodes ci dessous. [Chen et al., 2023, Wang et al., 2023c, Zemni et al., 2023, Jiang et al., 2022, Sieger et al., 2023, Ma et al., a, Augustin et al., , Spooner et al., 2021, Artelt et al., 2021, Pawelczyk et al., 2022, Chen et al., 2022b, Dominguez-Olmedo et al., , Blanc et al., , Ko et al., , Serrurier et al., 2023, Yu and Gao, 2023, Jeanneret et al., 2022, Tolkachev et al., 2022, Kanamori et al., , Black et al., 2022, Asher et al., 2022, Khorram and Fuxin, 2022, Datta et al., 2022, Tan et al., 2022, von Kügelgen et al., 2022, Chen et al., 2022a, Robeer et al., 2021, Rodriguez et al., 2021, Ali et al., 2023a, Upadhyay et al., 2021, Pawelczyk and Bielawski, , Wang et al., 2021f, Korikov et al., 2021, Bajaj et al., 2022, Tavares et al., , Yang et al., 2021c, Luss et al., 2021, Smyth and Keane, 2021, Samoilescu et al., 2021, Schleich et al., 2021, Chapman-Rounds et al., 2021, Madaan et al., 2021, Kim et al., 2020, Mothilal et al., 2021, Sixt et al., 2020, Sauer and Geiger, 2021, Bica et al., 2020, Bertossi, 2020, Cheng et al., 2020a, Keane and Smyth, 2020, O'Shaughnessy et al., , Teney et al., 2020, Fu et al., 2020a, White, 2020, Sharma et al., 2020, Poyiadzi et al., 2020b, Freiesleben, 2022, Delaney et al., 2021, Aïvodji et al., 2020, Artelt and Hammer, 2020b, Barredo-Arrieta and Del Ser, 2020, Gomez et al., 2020, Aïvodji et al., 2020, Hashemi and Fathi, 2020, Nemirovsky et al., 2021, Vermeire and Martens, 2020, Fernández-Loría et al., 2021, von Kügelgen et al., 2022, Pawelczyk et al., 2020b, Singla et al., 2020, Kaushik et al., 2020, Pawelczyk et al., 2020a, Barocas et al., 2020, Mothilal et al., 2020, Venkatasubramanian and Alfano, 2020, Lucic et al., 2020, Artelt and Hammer, 2020a, Zhao, 2020, Downs et al., , Kanamori et al., 2020, Albini et al., 2020, Ghazimatin et al., 2020, Akula et al., 2020, Wang and Vasconcelos, 2020, Dandl et al., 2020, Chapman-Rounds et al., 2019, Fernández et al., 2020, Laugel et al., 2019, Kanehira et al., 2019, Laugel et al., 2020, Guidotti et al., 2019, Dhurandhar et al., 2019, Besserve et al., 2019, Lucic et al., 2021, Joshi et al., 2019, Rathi, 2019, Van Looveren and Klaise, 2020, Qin et al., 2019, Goyal et al., 2019, Ramakrishnan et al., 2020, White and d'Avila Garcez, 2019, Luss et al., 2021, Wexler et al., 2019, Ustun et al., 2019, Russell, 2019, Sokol and Flach, , Chang et al., 2019, Samangouei et al., 2018, van der Waa et al., 2018, Dhurandhar et al.,

2018, Zhang et al., 2018g, Laugel et al., 2018, Hendricks et al., 2018a, Hendricks et al., 2018b, Zhang et al., c, Laugel et al., 2017, Tolomei et al., 2017, Russell et al., , Chawla and Wang, 2017, Fernández-Loría et al., 2021, Aggarwal et al., 2010, Lim and Dey, 2009, Lewis, 1973]

**Méthode explicatives basé sur l'importance des descripteurs** Ici nous évoquons les principales contributions répondant aux approches type importance des descripteurs. Nous ne détaillerons pas les différents paradigmes explicatif des méthodes ci dessous. [Wang and Kong, 2023, Huang and Kong, 2022, Lin et al., 2022a, Pillai et al., 2022, Aflalo et al., 2022, Chen and Zhao, 2022, Wu et al., 2022b, Sammani et al., 2022, Khakzar et al., 2022, Arenas et al., , Wang and Kong, 2022, Yu et al., 2023, Schoch et al., , Amoukou and Brunel, , Amoukou et al., , Han et al., , Colin et al., , Parekh et al., 2022, Stalder et al., , Hu et al., , Levin et al., , Crabbé et al., , Dhurandhar et al., 2022, Kowal, , Wang et al., 2023a, Dasgupta et al., 2022, Ali et al., 2022, Gat et al., 2022, Yang et al., 2022d, Rong et al., 2022, Lundstrom et al., 2022, Macdonald et al., , Schnake et al., 2022, Yang et al., 2022c, Jethani et al., 2022, Arora et al., 2022, Fang and Choromanska, 2022, Sood and Craven, 2021, Jin et al., 2022, Zhou et al., 2021c, Shitole et al., , Kumar et al., , La Malfa et al., 2021, Marques-Silva et al., 2021, Crabbé and van der Schaar, 2021, Yuan et al., 2021, Lin et al., , Chang et al., 2021a, Li et al., 2020e, Petsiuk et al., 2021, Gu and Dong, 2021, Chefer et al., 2021, Ge et al., 2021, Lim et al., 2021b, Lee et al., , Mardaoui and Garreau, 2021, Covert and Lee, 2021, Yan and Procaccia, 2021, Nam et al., 2020, Sattarzadeh et al., 2020, Pillai and Pirsiavash, 2021, Nam et al., 2020, Schlichtkrull et al., 2020, Sahoo et al., 2020, Srinivas and Fleuret, 2020, Frye et al., 2020, Jeyakumar et al., , Frye et al., , Luo et al., , Vu, , Sturmfels et al., 2020, Rebuffi et al., 2020, Liu et al., , Xu et al., 2020, Goh et al., 2021, Hu et al., 2020a, Nam et al., 2019, Chen and Jordan, 2019, Setzu et al., 2021, Chalasani et al., 2020, Jin et al., 2020b, Janzing et al., , Wang et al., 2019a, Singla et al., 2019, Etmann et al., 2019, Ancona et al., , Srinivas and Fleuret, , Ghorbani et al., , Ying et al., 2019c, Kapishnikov et al., 2019, Fong et al., 2019, Chen et al., 2018b, Wagner et al., 2019, Pope et al., 2019a, Chen et al., a, Zhang et al., 2019g, Montavon, 2019, Sundararajan and Najmi, 2020, Alvarez Melis and Jaakkola, 2018, Nie et al., , Chen et al., 2018c, Seo et al., 2018, Zhou et al., 2018, Ming et al., 2018, Zhang et al., 2019b, Zhang et al., 2018a, Petsiuk, , Dabkowski

and Gal, , Montavon et al., 2017, Fong and Vedaldi, 2017, Singh and Lee, 2017, Bau et al., 2017, Zintgraf et al., 2017, Apley and Zhu, 2019, Mahendran and Vedaldi, 2016, Zhang et al., 2016, Kindermans et al., 2016, Datta et al., 2016, Bach et al., 2015a, Zhou and Troyanskaya, 2015, Štrumbelj and Kononenko, 2014, Maleki et al., 2014, Altmann et al., 2010, Wrobel, , Robnik-Šikonja and Kononenko, 2008, Gevrey et al., 2003]

**Méthode explicatives basé sur les données influantes** Ici nous évoquons les principales contributions répondant aux approches type données influantes. Nous ne détaillerons pas les différents paradigmes explicatif des méthodes ci dessous. [Koh and Liang, 2020b, Caruana et al., ]

**Méthode explicative basé modèle prédictif** Ici nous évoquons les principales contributions répondant au paradigme specifique basé sur le modèle prédictif. [Roy et al., , Jeong et al., , Li et al., 2020d, Bénard et al., 2021, Yuan et al., 2020, Hüyük et al., 2020, Rahman et al., , Lakkaraju et al., 2019, Chen et al., b, Dhurandhar et al., , Lipton, 2018, Ross et al., 2017, Frosst and Hinton, 2017, Huysmans et al., 2011, Janizek et al., 2020, Inouye et al., , Tsang et al., 2020, Zhao and Hastie, 2021, Krause et al., 2016, Friedman, 2001]

**Méthode explicatives localisé sur une partie du modèle prédictif** Ici nous évoquons les principales contributions répondant aux approches s'intéréssant à l'explication localisé sur une partie du modèle prédictif considéré. Nous ne détaillerons pas les différents paradigmes explicatif des méthodes ci dessous. [Durrani et al., 2020, Bau et al., 2019, Dalvi et al., 2018]

**Divers** Ici nous évoquons les contributions discutant moins formellement de l'explicabilité des modèle prédictif. Certaine contribution s'intéresse par exemple au lien entre l'explicabilité des modèles prédictifs et la psychologie humaine, etc. [Zhang and Lim, 2022, Yang et al., 2022d, Wei et al., b, Crabbé, , Li et al., , Shrotri et al., 2022, Paleja et al., , Yeh et al., , Ismail et al., , Gao et al., 2021a, Izza and Marques-Silva, 2021, Nair et al., 2021, Wang and Khosravi, , Yuan et al., 2021, Zhu et al., 2021b, Shen et al., 2021a, Hancox-Li, 2020, Crabbe et al., , Plumb et al., 2020, Cheng et al.,

2020b,Mueller et al., 2021,Bayani and Mitsch, 2022,Tekkesinoglu et al., 2020,Hoernle et al., 2020,Ghorbani et al., 2019d,Rudin, 2019b,Fisher et al., 2019,Doshi-Velez and Kim, 2017b,Wei et al., 2017,Lei et al., 2017,Martens and Provost, 2014]

**Methode explicatives basées sur les prototypes** Ici nous évoquons les principales contributions répondant aux approches basé prototypes. Nous ne détaillerons pas les différents paradigmes explicatif des méthodes ci dessous. [Ribeiro et al., 2018,Gurumoorthy et al., 2019,Kim et al., 2016b,Kim et al., ]

**Méthode d'évaluation des méthode explicative** Ici nous évoquons les principales contributions portant sur l'évaluation des méthodes explicatives. [Agarwal et al., 2023a,Hase and Bansal, 2022,Nauta et al., 2023,Yang et al., 2021f,Lin et al., 2020b,Hsieh et al., 2021,Tomsett et al., 2020,Hooker et al., ,Montavon et al., 2018]

**L'état de l'art des implémentation logicielle**

- AIF360
- AIX360
- Anchor
- Alibi
- Alibi-detect
- BlackBoxAuditing
- Brain2020
- Boruta-Shap
- casme
- Captum
- cnn-exposed
- ClusterShapley
- DALEX
- Deeplift
- DeepExplain
- Deep Visualization Toolbox
- dianna
- Eli5

— explabox
— explainx
— ExplainaBoard
— ExKMC
— Facet
— Grad-cam-Tensorflow
— GRACE
— Innvestigate
— imodels
— InterpretML
— interpret-community
— Integrated-Gradients
— Keras-grad-cam
— Keras-vis
— keract
— Lucid
— LIT
— Lime
— LOFO
— modelStudio
— M3d-Cam
— NeuroX
— neural-backed-decision-trees
— Outliertree
— InterpretDL
— polyjuice
— pytorch-cnn-visualizations
— pytorch-grad-cam
— PDPbox
— py-ciu
— PyCEbox
— path_explain

— rulefit
— rulematrix
— Saliency
— SHAP
— Shapley
— Skater
— TCAV
— skope-rules
— TensorWatch
— tf-explain
— Treeinterpreter
— torch-cam
— WeightWatcher
— What-if-tool
— XAI
— Xplique
— Alibi
— actionable-recourse
— CEML
— Dice
— ContrastiveExplanation
— actionable-recourse
— cf-feasibility
— Mace
— Strategic-Decisions
— Contrastive-Explanation-Method
— Alibi
— PDPbox
— PyCEbox

**Modèle prédictif directement interprétable** Ici nous évoquons les principales contributions répondant aux approches s'intéréssant à l'interprétabilité des méthodes directement interprétable. Nous ne détaillerons pas les différents paradigmes

explicatif des méthodes ci dessous. [Zhang et al., 2022a, Sarkar et al., 2022, Finkelstein et al., , Nguyen et al., , Souza et al., , Agarwal et al., 2022, Barbiero et al., 2022, Shen et al., 2021b, Moshkovitz et al., 2021, Han et al., 2021, Arik and Pfister, 2020, Wang et al., 2020c, Agarwal et al., 2021, Liang et al., 2020a, Wei et al., a, Zhang et al., 2019e, Zhang et al., 2018f, Dash et al., , Melis and Jaakkola, a, Melis and Jaakkola, b, Freitas and Freitas, , Vellido et al., 2012]

### B.2.2 État de l'art relatif aux réseaux de neurones sur graphes

**Étude de l'état de l'art** Dans cette partie, nous evoquons les principales études faites sur la littérature à propos des réseaux de neurones sur graphes. Les voici : [Wu et al., 2021c, Sun et al., 2022, Bronstein et al., 2017, Battaglia et al., 2018b, Lee et al., 2018b, Zhang et al., 2020d, Zhou et al., 2021a, Zhou et al., 2020a, Wu et al., 2021c, Zhang et al., 2022d, Abadal et al., 2022]

**L'état de l'art des implémentations logicielles** [Zhu et al., 2019, Liu et al., 2021a, Cen et al., 2023, Grattarola and Alippi, 2020, Wang et al., 2020b, Fey and Lenssen, 2019] et

- PyG: Graph Neural Network Library for PyTorch
- DGL: Python Package Built to Ease Deep Learning on Graph
- Graph Nets: Build Graph Nets in Tensorflow
- Euler: A Distributed Graph Deep Learning Framework
- StellarGraph: Machine Learning on Graphs
- Spektral: Graph Neural Networks with Keras and Tensorflow 2
- PGL: An Efficient and Flexible Graph Learning Framework Based on PaddlePaddle
- CogDL: An Extensive Toolkit for Deep Learning on Graphs
- DIG: A Turnkey Library for Diving into Graph Deep Learning Research
- Jraph: A Graph Neural Network Library in Jax
- Graph-Learn: An Industrial Graph Neural Network Framework
- DeepGNN: a Framework for Training Machine Learning Models on Large Scale Graph Data

**Théorie** Maintenant, nous allons évoquer les principales contributions spécifiquement dédiées à la théorie des réseaux de neurones sur graphes.

**Apprentissage machine sur graphes : les contributions fondatrices** Ces papiers sont les principales contributions s'intéressant aux méthodes d'apprentissage machine sur des données de type graphes : [Sperduti and Starita, 1997, Gori et al., 2005, Scarselli et al., 2009, Gallicchio and Micheli, 2010, Li et al., 2017, Dai et al., ]

**Opérations sur les signaux des graphes pour une bonne représentation des données**

- le passage de message : [Gilmer et al., 2017b, Vignac et al., 2020, Corso et al., 2020, Balcilar et al., 2021, Tailor et al., 2021b, Wang et al., 2022]
- la convolution : Ici nous donnons les principaux papiers s'intéressant aux opérations réalisables sur des signaux sur graphes. En particulier sur celle de la convolution et ses différentes formulations (spatiale et spectrale) :
  - Approche spectral : Fourier et ondelettes : [Bruna et al., 2014, Susnjara et al., 2015, Defferrard et al., 2017, Kipf and Welling, 2017, Levie et al., 2018, Wu et al., 2019b, Xu et al., 2019a, Gama et al., 2018, Bruna et al., 2013, Defferrard et al., 2016, Kipf and Welling, 2016a, Wu et al., 2019b, NT et al., 2021, Levie et al., 2019, Monti et al., 2018, Liao et al., 2018, Gasteiger et al., 2018, Gasteiger et al., 2019, Chen et al., 2020d, Bianchi et al., 2022a, Wijesinghe and Wang, 2019, Luan et al., 2019, Zhu et al., 2020a, Chen et al., 2020c, Sun et al., 2020a, Tong et al., 2020, Zhao et al., 2020a, Zhu and Koniusz, 2020, Chien et al., 2020, Chen et al., 2020b, Cong et al., 2021, Zhang et al., 2021e, He et al., 2021a, Isufi et al., 2022, Dong et al., 2021b, Ma et al., 2021a, Liu et al., 2021c, Zhao et al., 2021b, Jin et al., 2021, He et al., 2022a, Guo et al., 2022, Fu et al., 2022a, Yang et al., 2022b, Wang and Zhang, 2022, Li et al., 2022d, Singh and Chen, 2022, Donnat et al., 2018, Xu et al., 2018, Li et al., 2020c, Zheng et al., 2021b, Gama et al., 2018, Gama et al., 2019, Gao et al., 2019a, Zou and Lerman, 2020, Perlmutter et al., 2020, Min et al., 2020, Min et al., 2021]
  - approche spatiale : [Micheli, 2009, Duvenaud et al., , Atwood and Towsley, 2016, Gilmer et al., 2017b, Simonovsky and Komodakis, 2017, Monti et al.,

2016, Chang et al., , Van Tran et al., 2018, Gasteiger et al., 2022a, Atwood and Towsley, 2016, Hamilton et al., 2017, Chen et al., 2018a, Xu et al., 2018, Xu* et al., 2018, Morris et al., 2019, Chiang et al., 2019a, Pei et al., 2019, Liu et al., 2020b, Zhu et al., 2020b, Noutahi et al., 2019, Liu et al., 2021d, Bodnar et al., 2021, Valsesia et al., 2021, Bodnar et al., 2021, Ding et al., 2021, Zhang et al., 2021d, Wijesinghe and Wang, 2021, Bevilacqua et al., 2021, Wang et al., 2021b, Niepert et al., 2016, Gao et al., 2018]

— les plus utilisées par la communauté : [Kipf and Welling, 2017, Defferrard et al., 2017, Hamilton et al., 2018, Hamilton et al., 2018, Morris et al., 2021, Qasim et al., 2019, Li et al., 2017, Bresson and Laurent, 2018, Veličković et al., 2018a, Veličković et al., 2018a, Zhang et al., 2021a, Brody et al., 2022, Shi et al., 2021, Thekumparampil et al., 2018, Du et al., 2018, Xu et al., 2019b, Hu et al., 2020c, Bianchi et al., 2021, Wu et al., 2019b, Zhu and Koniusz, 2020, Gasteiger et al., 2022a, Duvenaud et al., 2015, Schlichtkrull et al., 2017, Schlichtkrull et al., 2017, Busbridge et al., 2019, Derr et al., 2018, Fey, 2019, Qi et al., 2017a, Qi et al., 2017b, Monti et al., 2016, Fey et al., 2018, Gilmer et al., 2017b, Xie and Grossman, 2018, Wang et al., 2019b, Wang et al., 2019b, Li et al., 2018c, Deng et al., 2018, Verma et al., 2018, Zhao et al., 2021a, Bai et al., 2020a, Ranjan et al., 2020, Corso et al., 2020, Chiang et al., 2019a, Li et al., 2020a, Chen et al., 2020d, Ma et al., 2021c, Weisfeiler and Leman, , Togninalli et al., 2019, Brockschmidt, 2020, Kim and Oh, 2020, Bo et al., 2021, Tailor et al., 2022a, Rozemberczki et al., 2021b, You et al., 2021, Hu et al., 2020e, Mo et al., 2021, Wang et al., 2021d, He et al., 2020, Shi et al., 2020, Rampášek et al., 2023, Gravina et al., 2022, Rossi et al., 2023]

— les opérations d'agrégation des signaux selon la théorie de l'apprentissage géométrique profond : [Hamilton et al., 2018, Xu et al., 2019b, Corso et al., 2020, Li et al., 2020a, Tailor et al., 2022a, Xu et al., 2019b, Corso et al., 2020, Tailor et al., 2022a, Li et al., 2020a, Bartunov et al., 2022, Buterez et al., 2022, Hoyt, 2023, Corso et al., 2020, Tailor et al., 2022a, Geometric, 2022, Corso et al., 2020, Tailor et al., 2022a, Li et al., 2020a, Li et al., 2020a, Hamilton et al., 2018, Buterez et al., 2022, Vinyals et al., 2016, Corso et al., 2020, Zhang et al.,

2018d, Baek et al., 2021, Li et al., 2019c, Bartunov et al., 2022, Buterez et al., 2022, Buterez et al., 2022, Buterez et al., 2022]

— les opérations principales pour la normalisation des signaux selon la théorie de l'apprentissage géométrique profond : [Ioffe and Szegedy, 2015, Ioffe and Szegedy, 2015, Ulyanov et al., 2017, Ba et al., 2016, Ba et al., 2016, Cai et al., 2021a, Dwivedi et al., 2022a, Zhao and Akoglu, 2020, Yang et al., 2020, Li et al., 2020a, Zhou et al., 2020b, Zhao and Akoglu, 2019, Zhou et al., 2021b, Zhou et al., 2020b, Cai et al., 2021a]

— les opérations principales pour la mutualisation des signaux selon la théorie de l'apprentissage géométrique profond :
[Gao and Ji, 2019a, Cangea et al., 2018, Knyazev et al., 2019, Lee et al., 2019a, Knyazev et al., 2019, Yanardag and Vishwanathan, 2015, Diehl, 2019, Ranjan et al., 2020, Ma et al., 2021c, Khasahmadi et al., 2020, Dhillon et al., 2007, Simonovsky and Komodakis, 2017, Qi et al., 2017b, Ying et al., 2019b, Ma et al., 2019b, Wang et al., 2020f, Lee et al., 2019b, Noutahi et al., 2020, Gra, , Wang et al., 2020g]

— les opérations principales pour la démutualisation des signaux selon la théorie de l'apprentissage géométrique profond : [Qi et al., 2017b]

— les opérations principales pour la mutualisation globale des signaux selon la théorie de l'apprentissage géométrique profond :
[Ying et al., 2019b, Bianchi et al., 2020, Tsitsulin et al., 2023]

— les opérations principales pour l'encodage positionnel des signaux selon la théorie de l'apprentissage géométrique profond : [Vaswani et al., 2023, Cong et al., 2022, Henderson et al., 2021]

— les opérations principales pour l'encodage globale des signaux selon la théorie de l'apprentissage géométrique profond : [Schlichtkrull et al., 2017, Zhang et al., 2018e, Ying et al., 2018b, Gao and Ji, 2019a, Lee et al., 2019a, Murphy et al., 2019, Yuan and Ji, 2019, Bianchi et al., 2020, Zhang et al., 2020c, Bianchi et al., 2022b, Li et al., 2020b, Baek et al., 2020, Gao et al., 2022, Cai and Wang, 2022, Frasca et al., 2020a, Kipf and Welling, 2017, Wang et al., 2021e, Zhao et al., 2022, Zhao et al., 2022, Gilmer et al., 2017b, Dwivedi et al., 2022a, Dwivedi et al., 2022b, Rossi et al., 2022, Azabou et al., 2023, Bazhenov et al.,

2023]

— les approche principales pour opérateurs attentionnels pour les réseaux de neurones sur graphes : [Veličković et al., 2018a, Zhang et al., 2018c, Liu et al., 2018b, Gao and Ji, 2019b, Veličković et al., 2018a, Hou et al., 2019, Abu-El-Haija et al., 2019, Zhang et al., a, Lee et al., 2018a, Gao and Ji, 2019a, Bo et al., 2021, He et al., 2021b, Kim and Oh, 2020, Brody et al., 2021]

— les approches bayésiennes avec les réseaux de neurones sur graphes : [Ng et al., , Zhang et al., 2018h, Zhang et al., 2019f, Stadler et al., 2021]

— les limites des réseaux de neurones sur graphes identifiées par la communauté : [Alon and Yahav, 2020, Topping et al., 2021]

**Optimisation calculatoire des réseaux de neurones sur graphes** Ici nous allons évoquer les principales contributions s'intéressant aux méthodes pour effectuer des calculs avec des réseaux de neurones plus rapides d'un point de vue théorique, logiciel et matériel :

— optimisation des méthodes d'apprentissages pour les réseaux de neurones sur graphes : [Hamilton et al., , Chen et al., 2018d, Huang et al., 2018, Chen et al., 2018a, Chiang et al., 2019a]

— optimisation des architectures de réseau de neurones sur graphes permettant un passage à l'échelle efficace d'un point de vue calculatoire : [Chen et al., 2020b, Wu et al., 2019c, Frasca et al., 2020b, Tailor et al., 2022b, Li et al., 2022b, Xu et al., 2018, Dua, , Gao and Ji, 2019a, Abu-El-Haija et al., 2019]

— méthode pour la recherche d'architectures optimales pour les réseaux de neurones sur graphes : [Gao et al., 2019b, Zhao et al., 2020c, Zhang et al., 2021h]

— méthode pour le traitement efficace de grands graphes et méthodes d'échantillonnage efficace sur grands graphes : [You et al., 2020, Chiang et al., 2019b, Zeng et al., 2020a, Bojchevski et al., 2020, Fey et al., 2021a, Markowitz et al., 2021, Rong et al., 2019, Chen et al., 2018a, Chen et al., 2018d, Huang et al., 2018, Zou et al., 2019, Hasanzadeh et al., 2020, Zeng et al., 2019b, Cong et al., 2020, Hamilton et al., 2018, Hu et al., 2020e, Chiang et al., 2019a, Chiang et al., 2019a, Zeng et al., 2020b, Zeng et al.,

2022, Hamilton et al., 2018, Hamilton et al., 2018, Hu et al., 2020e, Chiang et al., 2019a,Chiang et al., 2019a,Zeng et al., 2019b,Zeng et al., 2022,Hamilton et al., 2018]

— méthode de compression des données adaptées aux réseaux de neurones sur graphes :

[Zhao et al., 2020b, Tailor et al., 2021a, Bahri et al., 2021]

— méthode de compression des réseaux de neurones sur graphes permettant la distillation de connaissance :

[Chen et al., 2021b, Yang et al., 2021g, Yang et al., 2021a, Deng and Zhang, 2021, Joshi et al., 2022, Huang et al., 2022a, Liu et al., 2022c, Yang et al., 2021g, Wang et al., 2021a, Feng et al., 2020a]

— méthode matérielle pour le calcul efficace avec des réseaux de neurones sur graphes :

[Zeng et al., 2019a,Liang et al., 2020c,Zeng and Prasanna, 2020a,Chen et al., 2020e, Abadal et al., 2021]

— méthode de projection pour le calcul efficace avec des réseaux de neurones sur graphes : [Zhang et al., 2021b, Fu et al., 2022b, Huang et al., 2020a, Tian et al., 2020a, Fu et al., 2022c, Chen et al., 2020h, Tian et al., 2020b, Rahman et al., 2021]

— méthode de compilations pour réseaux de neurones sur graphes :

[Xie, , Wu et al., 2021b, Hu et al., 2020d]

— méthode pour le calcul distribué pour réseaux de neurones sur graphes :

[Peng et al., 2022, Wan et al., 2022b, Zhang et al., 2022c, Wan et al., 2022a, Ramezani et al., 2022, Zheng et al., 2022b, Md et al., 2021, Chakaravarthy et al., 2021, Zhang et al., 2021c, Gandhi and Iyer, , Thorpe et al., , Wang et al., 2021c, Cai et al., 2021b, Tripathy et al., 2020, Liu et al., 2020a, Zheng et al., 2021a, Jia et al., , Zhang et al., 2020b, Ma et al., b]

— méthode d'apprentissage pour le calcul distribué sur réseaux de neurones sur graphes : [Waleffe et al., 2022, Zheng et al., 2022a, Yu et al., 2022, Fey et al., 2021b, Wang et al., c, Gong et al., 2022, Li et al., 2022e, Chen et al., 2020f, Zeng and Prasanna, 2020b]

— méthode pour le traitement distribué des données de type graphes pour ré-

seaux de neurones sur graphes :
[Kaler et al., 2022, Yang et al., 2022a, Dong et al., 2021a, Huang et al., 2021, Lin et al., 2020c, Deutsch et al., 2019]

— méthode d'inférence rapide pour les réseaux de neurones sur graphes :
[Yang et al., 2023a, Sohrabizadeh et al., 2022, Huang et al., 2022c, Huang et al., 2022c, Chen et al., 2021a, He et al., 2021c, Geng et al., 2021, Li et al., 2021b, Wang et al., 2020h, Zhang et al., 2020a, Auten et al., 2020, Geng et al., 2020, Kiningham et al., 2020, Yan et al., 2020]

— méthode de prétraitement pour un apprentissage optimale des réseaux de neurones sur graphes :
[Hu* et al., 2019, Qiu et al., 2020]

— méthode de mesure pour les réseaux de neurones sur graphes : [Li et al., 2018a, Xu et al., 2019b, Morris et al., 2021, Li et al., 2018a, Chen et al., 2020a, Oono and Suzuki, 2019, Hou et al., 2019, Chen et al., 2020d, Liu et al., 2021c]

— méthode de stress (c-à-d attaques) pour les réseaux de neurones sur graphes :
[Zügner and Günnemann, 2018, Zügner and Günnemann, 2020, Zügner and Günnemann, 2018, Bojchevski and Günnemann, 2019, Zhang and Zitnik, 2020, Zhang et al., 2021f, Chang et al., 2021b]

**Divers** Ici nous listons des contributions avec des contextes applicatifs très particuliers, par exemple certaines contributions s'intéressent à la génération de graphe par apprentissage adversariale, d'autre s'intéressent aux réseaux de neurones sur graphes adaptés aux graphes ayant une composante temporelle dans leur définition ou bien encore pour le traitement de graphes hétérogènes, etc. Ces contributions sont plus marginales pour la communauté puisque leurs cas d'usage sont assez restreints : [Schlichtkrull et al., 2017, Derr et al., 2018, Ma et al., 2018b, Liao et al., 2019, Feng et al., 2019, Hassani and Khasahmadi, 2020, Zhu et al., 2021a, Loukas, 2019, Yang et al., 2021b, Sun et al., 2021c, Hu et al., 2020b, Sun et al., 2020b, Zhu et al., 2020a, Li et al., 2021a, Fey et al., 2021c, Yang et al., 2021e, Dasoulas et al., 2021, Balcilar et al., 2020, Yang et al., 2021g, Errica et al., 2019, Ye et al., 2019, Wu et al., 2020, Lukovnikov and Fischer, 2021, Xu et al., 2021b, Luo et al., 2021, Ma et al., 2019a, Feng et al., 2020b, Sato et al., 2019, Chen et al., 2020g, Brockschmidt, 2020, Geerts and Reutter, 2021, Seo et al., 2016, Trivedi et al., 2017, Li et al.,

2018d, Yu et al., 2018a, Guo et al., 2019, Geng et al., 2019, Li et al., 2019a, Zhang et al., 2019c, Yun et al., 2019, Hu et al., 2020e, Kreuzer et al., 2021, Ying et al., 2021, Frey et al., 2021, Kuang et al., 2021, Bastos et al., 2022, Bo et al., 2022, Wang et al., 2018a, Bojchevski et al., 2018, Chami et al., 2019, Liu et al., 2019, Gulcehre et al., 2018, Bachmann et al., 2020, Lou et al., 2020, Zhang et al., 2021g, Yang et al., 2021g, Sun et al., 2021a, Sun et al., 2021b, Xinyi and Chen, 2018, Yang et al., 2021d, Zhang et al., 2019a, Zhang et al., 2019a, Fu et al., 2020b, Li et al., 2018b, Xinyi and Chen, 2019, Bacciu et al., 2019]

**Application** Ici nous listons les principaux cas d'usage des réseaux de neurones sur graphes :

- dans le domaine de la vision par ordinateur : [Qi et al., 2017c, Yi et al., 2016b, Santoro et al., 2017, Johnson et al., 2018, Qi et al., 2017a, Chen et al., 2018e, Landrieu and Simonovsky, 2018, Narasimhan et al., 2018, Liang et al., , Yang et al., 2018, Qi et al., 2018, Guo et al., 2018, Te et al., 2018, Wang et al., 2019b]
- dans le domaine du le traitement du langage naturel : [Zhang et al., 2019d, Beck et al., 2018, Wang et al., 2018b, Liu et al., 2018a, Marcheggiani et al., 2018, Yao et al., 2018]
- dans le domaine de l'analyse du web : [Nguyen and Grishman, 2018, Rahimi et al., 2018, Zügner and Günnemann, 2018, Qiu et al., 2018]
- dans le domaine des systèmes de recommandation : [Ying et al., 2018a, Wu et al., 2019d]
- dans le domaine de la santé : [Choi et al., 2017, Liang et al., 2020b, Shang et al., 2019]
- dans le domaine de la physique : [Battaglia et al., 2016, Hoshen, 2018]
- dans le domaine de la chimie : [Kearnes et al., 2016, Fout et al., ]
- dans le domaine du traitement numérique des réseaux sociaux : [Wang et al., 2016, Cao et al., 2016a, Kipf and Welling, 2016b, Zhang and Chen, 2018, Shaham et al., 2018, Tu et al., 2018, Yu et al., 2018b, Pan et al., 2018a]
- dans le domaine de la résolution d'équations aux dérivées partielles : [Zhuang et al., 2019, Xhonneux et al., 2020, Zang and Wang, 2020, Belbute-Peres et al., 2020, Anandkumar et al., 2020, Li et al., 2020f, Iakovlev et al., 2020, Cham-

berlain et al., 2021]

**Les réseaux de neurones sur graphe standard de la littérature** Parmis tous les réseaux de neurones sur graphes concevables, certains ont de meilleures propriétés que le réseau de neurones sur graphes moyen, voici les plus connues de ces modèles :

**Modèles de réseaux de neurones sur graphes usuels** [Kipf and Welling, 2017, Hamilton et al., 2018, Xu et al., 2019b, Veličković et al., 2018a, Brody et al., 2022, Corso et al., 2020, Wang et al., 2019b, Xu et al., 2018, Battaglia et al., 2018b, Grover and Leskovec, 2016, Veličković et al., 2018b, Kipf and Welling, 2016b, Kipf and Welling, 2016b, Kipf and Welling, 2016b, Pan et al., 2019, Pan et al., 2019, Derr et al., 2018, Jin et al., 2020a, Gao and Ji, 2019a, Schütt et al., 2017, Gasteiger et al., 2022c, Gasteiger et al., 2022b, Dong et al., 2017, Li et al., 2019b, Li et al., 2020a, Rossi et al., 2020, Zhu and Ghahramani, 2003, Huang et al., 2020b, Xiong et al., 2020, Wang et al., 2021g, Lim et al., 2021a, He et al., 2020, Shi et al., 2021, Li et al., 2022c, Park et al., 2021, Yang et al., 2023b, Bordes et al., , Trouillon et al., 2016, Yang et al., 2015, Sun et al., 2019]

**Génération de graphes et ensemble de données usuel** Ici nous évoquons les principales méthodes pour la génération de graphes et de principaux ensembles de données utilisés dans la littérature :

- méthodes de génération de graphes : [Johnson, 2017, De Cao and Kipf, 2022, Bojchevski et al., 2018, You et al., 2018, Ma et al., c, Cao et al., 2016b, Pan et al., 2018b, Yu et al., 2018c, You et al., 2018, Simonovsky and Komodakis, 2018, Ma et al., 2018a, Salha et al., 2019]
- les ensembles de données usuels :
  - les ensembles de données usuels composés de graphes homogènes : [Zachary, 1977, Dwivedi et al., 2022a, Yang et al., 2016, Carlson et al., 2010, Bojchevski and Günnemann, 2018, Shchur et al., 2019, Shchur et al., 2019, Zitnik and Leskovec, 2017, Hamilton et al., 2018, Zeng et al., 2020b, Zeng et al., 2020b, Zeng et al., 2020b, Zeng et al., 2020b, Wu et al., 2018, Wu et al., 2018, Sterling and Irwin, 2015, Gómez-Bombarelli et al., 2018, Dwi-

vedi et al., 2022a, Sorkun et al., 2019, Wu et al., 2018, Schlichtkrull et al., 2017, Schlichtkrull et al., 2017, Bai et al., 2020b, Yang et al., 2023d, Monti et al., 2016, Bogo et al., 2014, Bogo et al., 2017, Yi et al., 2016a, Zhirong Wu et al., 2015, Ranjan et al., 2018, Cosmo et al., 2016, Guerrero et al., 2018, Armeni et al., 2016, Pareja et al., 2019, Cong et al., 2023, Jin et al., 2020a, Jin et al., 2020a, Cho et al., 2013, Bourdev and Malik, 2009, Ham et al., 2016, Bordes et al., 2013, Dettmers et al., 2018, Bordes et al., 2013, Mernyei and Cangea, 2022, Pei et al., 2019, Rozemberczki et al., 2021a, Platonov et al., 2023, Pei et al., 2019, Dou et al., 2021, Rozemberczki et al., 2021a, Rozemberczki et al., 2021a, Rozemberczki and Sarkar, 2020, Rozemberczki and Sarkar, 2020, Rozemberczki et al., 2019, Rozemberczki et al., 2021a, Ribeiro et al., 2017, Dwivedi et al., 2023b, Freitas et al., , Adamic and Glance, 2005, Yin et al., 2017, Lim et al., 2021a, Weber et al., 2019, Weber et al., 2019, Huang et al., 2023, Choudhury et al., 2020, Bonnet et al., 2023, Kumar et al., 2019, Wang et al., 2020d]

— les ensembles de données usuels composés de graphes hétérogènes : [Sun et al., 2017, Dong et al., 2017, Hu et al., 2021, Fu et al., 2020b, Lv et al., 2021, Zhang and Chen, 2020, He et al., 2020]

— les ensembles de données synthétiques : [Kim and Oh, 2020, Abu-El-Haija et al., 2019, Ying et al., 2019c, Baldassarre and Azizpour, 2019a, Luo et al., 2020, Azzolin et al., 2023, Ying et al., 2019c]

### B.2.3 État de l'art relatif aux méthodes explicatives pour les réseaux de neurones sur graphes

[Li et al., 2023b] constatent que les réseaux de neurones sur graphes ont démontré une amélioration significative des performances de prédiction sur les données de type graphes. Dans le même temps, les prédictions faites par ces modèles sont souvent difficiles à interpréter. À cet égard, de nombreux efforts ont été déployés pour expliquer les mécanismes de prédiction de ces modèles à partir de perspectives telles que GNNExplainer, XGNN et PGExplainer. Bien que de tels travaux présentent des cadres standards pour interpréter les réseaux de neurones sur graphe, une revue sur les réseaux de neurones sur graphe explicables n'est pas disponible. Dans cette en-

quête, ils présentent un examen complet des méthodes explicatives développées pour les réseaux de neurones sur graphe. Ils se concentrent sur les réseaux de neurones à graphes explicables et les catégorisent en fonction de l'utilisation de méthodes explicables. Ils fournissent en outre les mesures de performance communes pour les explications des réseaux de neurones sur graphe et soulignent plusieurs orientations de recherche futures.

[Yuan et al., 2022] constatent que les méthodes d'apprentissage profond atteignent des performances toujours croissantes sur de nombreuses tâches d'intelligence artificielle. Une limitation majeure des modèles profonds est qu'elles ne se prêtent pas à l'explicabilité. Cette limitation peut être enlevée en développant des méthodes post-hoc pour expliquer les prédictions, donnant ainsi naissance au domaine d'explicabilité. Récemment, l'explicabilité des modèles profonds sur des images et des textes a réalisé des progrès significatifs. Dans le domaine des données de type graphes, les réseaux de neurones sur graphes et leur explicabilité connaissent des développements rapides. Cependant, il n'existe ni un traitement unifié des méthodes d'explicabilité pour les réseaux de neurones sur graphe, ni un référentiel standard et un banc d'essai pour leurs évaluations. Dans cette enquête, ils fournissent une vue unifiée et taxonomique des méthodes d'explicabilité pour réseaux de neurones sur graphe actuelles. Leurs traitements unifiés et taxonomiques de ce sujet mettent en lumière les points communs et les différences des méthodes existantes et préparent le terrain pour de futurs développements méthodologiques. Pour faciliter les évaluations, ils génèrent un ensemble d'ensembles de données de référence spécifiquement pour l'explicabilité des réseaux de neurones sur graphe. Ils résument les ensembles de données et les mesures actuelles pour évaluer l'explicabilité des réseaux de neurones sur graphe. Dans l'ensemble, ce travail fournit un traitement méthodologique unifié de l'explicabilité des réseaux de neurones sur graphe et un banc d'essai standardisé pour les évaluations.

[Zhang et al., 2022b] constatent que les réseaux de neurones sur graphes sont apparus comme une série de méthodes d'apprentissage sur graphe performantes pour divers scénarios du monde réel, allant des applications quotidiennes telles que les systèmes de recommandation et la réponse aux questions technologiques telles que la découverte de médicaments dans les sciences de la vie et la simulation des

corps en astrophysique. Cependant, l'exécution des tâches n'est pas la seule exigence des réseaux de neurones sur graphe. Les réseaux de neurones sur graphe axés sur les performances ont montré des effets négatifs potentiels tels que la vulnérabilité aux attaques contrefactuelles, une discrimination inexplicable à l'encontre des groupes défavorisés ou une consommation excessive de ressources dans les environnements informatiques . Pour éviter ces dommages involontaires, il est nécessaire de créer des réseaux de neurones sur graphe performants et dignes de confiance. À cette fin, ils proposent une feuille de route complète pour créer des réseaux de neurones sur graphe fiables du point de vue des différentes technologies informatiques impliquées. Dans cette enquête, ils introduisent des concepts de base et résument de manière exhaustive les efforts existants pour des réseaux de neurones sur graphe fiables sous six aspects, notamment la robustesse, l'explicabilité, la confidentialité, l'équité, la responsabilité et le bien-être environnemental. De plus, ils mettent en évidence les relations croisées complexes entre les six aspects ci-dessus des réseaux de neurones sur graphe dignes de confiance. Enfin, ils présentent un aperçu complet des tendances visant à faciliter la recherche et l'industrialisation de réseaux de neurones sur graphe dignes de confiance.

[Warmsley et al., 2022] constatent que les activités malveillantes en matière de cybersécurité sont devenues de plus en plus inquiétantes pour les particuliers comme pour les entreprises. Bien que les méthodes d'apprentissage automatique telles que les réseaux de neurones sur graphes se soient révélées efficaces dans la tâche de détection des logiciels malveillants, leurs résultats sont souvent difficiles à comprendre. Des méthodes de détection de logiciels malveillants explicables sont nécessaires pour identifier automatiquement les programmes malveillants et présenter les résultats aux analystes de logiciels malveillants d'une manière interprétable par l'Homme. Dans cette enquête, ils décrivent un certain nombre de méthodes d'explicabilité pour réseaux de neurones sur graphe et comparent leurs performances sur un ensemble de données de détection de logiciels malveillants réels. Plus précisément, ils ont formulé le problème de détection comme un problème de classification de graphes sur les graphes de flux de contrôle (CFG) des logiciels malveillants. Ils constatent que les méthodes basées sur le gradient surpassent les méthodes basées sur les perturbations en termes de dépenses de calcul et de performances sur des

métriques spécifiques aux méthodes explicatives (par exemple, fidélité et parcimonie). Leurs résultats fournissent des informations sur la conception de nouveaux modèles basés sur réseaux de neurones sur graphe pour la détection et l'attribution de cybermalwares.

[Longa et al., 2023] constatent que suite à une première avancée rapide dans l'apprentissage basé sur les graphes, les réseaux de neurones sur graphes ont atteint un certain niveau de performance dans de nombreux domaines scientifiques, ce qui a suscité le besoin de méthodes permettant de comprendre leur processus de décision. Les méthodes explicatives pour réseaux de neurones sur graphe ont commencé à émerger ces dernières années, avec une multitude de méthodes à la fois nouvelles ou adaptées d'autres domaines. Pour trier cette pléthore de méthodes, plusieurs études ont comparé les performances de différentes méthodes explicatives par rapport à différentes mesures. Cependant, ces travaux antérieurs ne tentent pas de comprendre pourquoi différentes architectures de réseaux de neurones sur graphe sont plus ou moins explicables, ni quelle méthode explicative devrait être préférée dans un contexte donné. Dans cette enquête, ils comblent ces lacunes en concevant une étude expérimentale standard, qui teste dix méthodes explicatives sur huit architectures, sur six ensembles de données de classification de graphes et de noeuds. Avec leurs résultats, ils fournissent des informations clés sur le choix et l'applicabilité des méthodes explicatives pour réseaux de neurones sur graphe, ils isolent les composants clés qui les rendent utilisables et efficaces et fournissent des recommandations sur la manière d'éviter les pièges d'interprétation courants. Ils concluent en mettant en évidence les questions ouvertes et les orientations de recherches futures possibles.

### B.2.4 Principales contributions

[Yuan et al., 2022] propose une méthode unifiée pour l'évaluation des méthodes explicatives pour les réseaux de neurones sur graphes. Il propose un protocole standard et des ensembles de données adaptés pour cette évaluation.

[Ying et al., 2019c] propose une première méthode explicative pour les réseaux de neurones sur graphes, agnostique au modèle, locale et post-hoc. Il propose GNNExplainer, une méthode adaptée à n'importe quel réseau de neurones sur graphes.

Elle fournit ses explications en cherchant des sous-graphes conservant les propriétés de classification du modèle prédictif via une optimisation de l'information mutuelle entre le graphe initial et le sous-graphe considéré.

[Pope et al., 2019b] ont adapté, aux réseaux de neurones sur graphes, les méthodes explicatives déjà existantes dans la littérature de l'explicabilité des modèles prédictifs tels que Saliency ou encore Class Activation Mapping et leurs dérivés.

[Yuan et al., 2020] propose d'expliquer les réseaux de neurones sur graphe en entraînant un générateur de graphes afin que les graphes générés maximisent la qualité de prédiction du modèle prédictif. Il formule la génération de graphes comme une tâche d'apprentissage par renforcement, où pour chaque étape, le générateur de graphes prédit comment ajouter une arête dans le graphe courant. Le générateur de graphe est formé via une méthode de gradient de politique basée sur les informations provenant des réseaux de neurones sur graphe considérés. De plus, il intègre plusieurs règles sur graphes pour encourager la validité de ces derniers.

[Sanchez-Lengeling et al., ] évalue les méthodes de la littérature sous l'angle de nouvelle métrique d'évaluation telles que la précision, la stabilité, la fidélité et la consistance de la méthode explicative.

[Vu and Thai, 2020] propose PGM-Explainer, une méthode explicative agnostique au modèle prédictif basée sur les modèles probabilistes graphiques, pour les réseaux de neurones sur graphes. Étant donné une prédiction à expliquer, PGM-Explainer identifie les composants du graphe cruciaux et génère une explication sous la forme d'un modèle probabiliste graphique se rapprochant de cette prédiction. De plus, PGM-Explainer est capable de démontrer les dépendances des descripteurs expliquées sous forme de probabilités conditionnelles.

[Baldassarre et al., 2020] propose d'utiliser les méthodes explicatives pour les réseaux de neurones sur graphes dans le contexte de la vision par ordinateur et notamment pour des problèmes de reconnaissance d'objet.

[Lu and Li, 2020] propose une architecture de réseaux de neurones sur graphes permettant une certaine explicabilité et de résoudre le problème de détection d'information fausse.

[Yuan et al., 2021] proposent une nouvelle méthode, connue sous le nom de SubgraphX, pour expliquer les réseaux de neurones sur graphe en identifiant des

sous-graphes importants. Étant donné un modèle de réseau de neurones sur graphe entraîné et un graphe, SubgraphX explique ses prédictions en explorant efficacement différents sous-graphes avec la recherche arborescente de Monte Carlo. Pour rendre la recherche arborescente plus efficace, ils proposent d'utiliser les valeurs de Shapley comme mesure de l'importance des sous-graphes, qui peuvent également capturer les interactions entre différents sous-graphes. Pour accélérer les calculs, ils proposent des schémas d'approximation efficaces pour calculer les valeurs de Shapley pour les données de type graphe. Leur travail représente une première tentative d'expliquer les réseaux de neurones sur graphe via l'identification explicite et directe des sous-graphes.

[Laird et al., 2023] proposent une méthode explicative pour les réseaux de neurones sur graphe appelé eXplainable Insight (XInsight) qui génère une distribution d'explications de modèle à l'aide de GFlowNets [Bengio et al., 2021]. Étant donné que les GFlowNets génèrent des objets avec des probabilités proportionnelles à une récompense, XInsight peut générer un ensemble diversifié d'explications, par rapport aux méthodes précédentes qui n'apprennent que l'échantillon de récompense maximale. Ils démontrent que XInsight en générant des explications pour les réseaux de neurones sur graphe formés sur deux tâches de classification de graphes : classifier les composés mutagènes avec l'ensemble de données MUTAG et classifier les graphes acycliques avec un ensemble de données synthétiques en open source. Ils montrent l'utilité des explications de XInsight en analysant les composés générés à l'aide de la modélisation QSAR, constatent que XInsight génère des composés qui se regroupent par lipophilie, une corrélation connue pour la mutagénicité. Leurs résultats montrent que XInsight génère une distribution d'explications qui révèle les relations sous-jacentes démontrées par le modèle. Ils soulignent également l'importance de générer un ensemble diversifié d'explications, car cela permet de découvrir des relations cachées dans le modèle et fournit des conseils précieux pour une analyse plus approfondie.

[Verma et al., 2023] proposent une méthode explicative basée sur un raisonnement contrefactuel. L'objectif principal est d'apporter des modifications minimes au graphe d'un réseau de neurones sur graphe afin de modifier sa prédiction. Bien que plusieurs algorithmes aient été proposés pour des explications contrefactuelles

des réseaux de neurones sur graphe, la plupart d'entre eux présentent deux inconvénients principaux. Premièrement, ils considèrent uniquement les suppressions d' arêtes comme des perturbations. Deuxièmement, les modèles d'explication contrefactuels sont transductifs, ce qui signifie qu'ils ne se généralisent pas à certaines données. Dans cette étude, ils introduisent un algorithme inductif appelé INDUCE, qui surmonte ces limitations. En menant des expériences approfondies sur plusieurs ensembles de données, ils démontrent que l'incorporation d'ajouts d'arêtes conduit à de meilleurs résultats contrefactuels par rapport aux méthodes existantes. De plus, la méthode de modélisation inductive permet à INDUCE de prédire directement les perturbations contrefactuelles sans nécessiter de formation spécifique à une instance. Cela entraîne des améliorations significatives de la vitesse de calcul par rapport aux méthodes de base et permet une analyse contrefactuelle évolutive pour les réseaux de neurones sur graphe.

[Li et al., 2023a] proposent une nouvelle méthode : MSInterpreter. MSInterpreter fournit un schéma de sélection de passage de messages (MSScheme) pour sélectionner les chemins critiques pour les agrégations de messages des réseaux de neurones sur graphe, qui vise à atteindre l'auto-explication plutôt que des explications post-hoc. En détail, le MSScheme élaboré est conçu pour calculer les facteurs de pondération des chemins d'agrégation de messages en considérant la structure canonique et les composants d'intégration de noeuds, où la base de structure vise les facteurs de pondération parmi les sous-structures induites par les noeuds ; d'autre part, la base d'intégration de noeuds se concentre sur les facteurs de poids via des plongements de noeuds obtenus par réseaux de neurones sur graphe monocouche.

[Bonabi Mobaraki and Khan, 2023] proposent un nouveau cadre de travail pour les méthodes explicatives pour réseaux de neurones sur graphes. Plusieurs méthodes d'explicabilité pour les réseaux de neurones sur graphe ont été proposées récemment. Cependant, comme elles n'ont pas été comparées de manière approfondie les unes par rapport aux autres, leurs compromis et leur efficacité dans le contexte des réseaux de neurones sur graphe sous-jacents et des applications ne sont pas clairs. Pour soutenir davantage les recherches dans ce domaine, ils ont développé un outil interactif de bout en bout, nommé gInterpreter, en réimplémentant 15 méthodes d'explicabilité pour réseaux de neurones sur graphe récentes sous un protocole commun avec un cer-

tain nombre de réseaux de neurones sur graphe utilisés pour différentes tâches. Cette contribution présente gInterpreter avec un profilage interactif des performances de 15 méthodes récentes d'explicabilité pour réseaux de neurones sur graphe, visant à expliquer les raisonnements complexes des méthodes profondes sur des données de type graphes.

[Prado-Romero et al., 2023] proposent une méthode explicative de type contrefactuelle, fortement inspirée du domaine image, appelée CounteRGAN. Ils considèrent alors les graphes comme des images en nuage de gris pour encoder les adjacences de ces derniers.

[Agarwal et al., 2023b] proposent une méthode de travail pour évaluer les méthodes explicatives de réseaux de neurones sur graphes. Ils introduisent un générateur de données type graphes synthétiques, ShapeGGen, qui peut générer une variété d'ensembles de données de référence (par exemple, différentes tailles de graphes, distributions de degrés, graphes homophiles ou hétérophiles) accompagnés d'explications de vérité terrain. La flexibilité nécessaire pour générer divers ensembles de données synthétiques et les explications correspondantes de la vérité terrain permet à ShapeGGen d'imiter les données dans divers domaines du monde réel. Ils incluent ShapeGGen et plusieurs ensembles de données de type graphes du monde réel dans une bibliothèque d'explicabilité pour réseaux de neurones sur graphe, GraphXAI. En plus des ensembles de données de type graphes synthétiques et réels avec des explications de la vérité terrain, GraphXAI fournit des chargeurs de données, des fonctions de traitement de données, des visualiseurs, des implémentations de modèle de réseaux de neurones sur graphe et des métriques d'évaluation pour comparer les méthodes d'explicabilité pour réseaux de neurones sur graphe.

[Homberg et al., 2023] introduisent une nouvelle méthode pour interpréter les prédictions des réseaux de neurones sur graphes basées sur les valeurs de Myerson issues de la théorie des jeux coopératifs. Les valeurs de Myerson sont étroitement liées aux valeurs de Shapley et fournissent ainsi une méthode d'explicabilité similaire aux valeurs SHAP. Ils ont développé une méthode pour des applications dans la découverte de médicaments, mais elle peut être utilisée avec n'importe quel graphe. En utilisant les réseaux de neurones sur graphes comme un jeu de coalition et le graphe interprété comme structure de coopération, les valeurs de Myerson déterminent la

valeur de chaque noeud du graphe. La valeur de tous les noeuds du graphe s'ajoute à la valeur prédite du modèle, permettant une interprétation simple et intuitive de la prédiction. Pour interpréter les prédictions sur les graphes moléculaires, ils montrent des explications visuelles sur les structures moléculaires à l'aide de deux ensembles de données moléculaires.

[Wu et al., 2023] proposent une nouvelle méthode explicative pour les réseaux de neurones sur graphes adaptée au domaine de la chimie. La plupart des méthodes explicatives existantes pour les réseaux de neurones sur graphe en chimie se concentrent sur l'attribution des prédictions du modèle à des noeuds, arêtes ou fragments individuels qui ne sont pas nécessairement dérivés d'une segmentation chimiquement significative des molécules. Pour relever ce défi, il propose une méthode appelée explication du masque de sous-structure (SME). SME est basé sur des méthodes de segmentation moléculaire bien établies et fournit une interprétation qui correspond à la compréhension des chimistes. Il applique le SME pour élucider comment les réseaux de neurones sur graphe apprennent à prédire la solubilité aqueuse, la génotoxicité, la cardiotoxicité et la perméation de la barrière hémato-encéphalique pour les petites molécules. SME fournit une interprétation cohérente avec la compréhension des chimistes, les alerte en cas de performances peu fiables et les guides dans l'optimisation structurelle des propriétés cibles. Par conséquent, il pense que SME permet aux chimistes d'exploiter en toute confiance la relation structure-activité (SAR) à partir de réseaux de neurones sur graphe fiables grâce à une inspection transparente de la manière dont les réseaux de neurones sur graphe captent des signaux utiles lors de l'apprentissage à partir de données.

[Liu et al., 2023] proposent d'exploiter une méthode déjà existante dans la littérature des méthodes explicatives pour les réseaux de neurones sur graphes dans le contexte médical. Le principe du goulot d'étranglement de l'information (IB) a été étendu aux réseaux de neurones sur graphe pour identifier un sous-graphe compact le plus informatif pour les étiquettes de classe, ce qui améliore considérablement l'explicabilité de la décision. Cependant, les modèles GIB (Graph Information Bottleneck) existants sont soit instables pendant la formation (en raison de la difficulté d'estimation de l'information mutuelle), soit se concentrent uniquement sur un type particulier de graphe (par exemple, les réseaux cérébraux) qui souffrent d'une

mauvaise généralisation aux ensembles de données de type graphes généraux. avec différentes tailles de graphes. Dans ce travail, ils étendent le goulot d'étranglement de l'information cérébrale (BrainIB) récemment développé aux graphes en utilisant l'information mutuelle d'ordre $\alpha$ de Renyi basées sur une matrice pour stabiliser l'entraînement ; et en concevant une nouvelle stratégie de masque pour traiter différentes tailles de graphes, de sorte que la nouvelle méthode puisse également être utilisée pour les réseaux sociaux, les molécules, etc.

[Bui et al., 2023] proposent une nouvelle méthode explicative pour les réseaux de neurones sur graphe, générale et rapide nommée SCALE. SCALE forme plusieurs apprenants spécialisés pour expliquer les réseaux de neurones sur graphe, car selon eux, il est compliqué de créer une méthode explicative puissante pour examiner les attributions des interactions dans les graphes. En formation, un modèle de réseau de neurones sur graphe en boîte noire guide les apprenants sur la base d'un paradigme de distillation des connaissances en ligne. Pendant la phase d'explication, les explications des prédictions sont générées par plusieurs méthodes explicatives. Le masquage des arêtes et les procédures de marche aléatoire avec redémarrage sont implémentés pour fournir des explications structurelles pour les prédictions au niveau du graphe et du noeud. Un module d'attribution de descripteurs fournit des résumés globaux et des contributions aux descripteurs au niveau de l'instance.

[Wang and Shen, 2023] proposent une méthode d'explication agnostique au modèle prédictif pour différents réseaux de neurones sur graphe qui suivent le schéma de transmission de messages. Leur méthode graphInterpreter apprend une distribution de graphes génératifs probabilistes qui produit le modèle de graphe le plus discriminant que le réseaux de neurones sur graphe tente de détecter lors d'une certaine prédiction en optimisant une nouvelle fonction objectif spécifiquement conçue pour l'explication des réseaux de neurones sur graphe au niveau du modèle.

[Azzolin et al., 2022] proposent GLGExplainer (Global Logic-based GNN Explainer), le premier Global Explainer capable de générer des explications sous forme de combinaisons booléennes arbitraires de concepts graphes appris. GLGExplainer est une architecture entièrement différentiable qui prend des explications locales comme entrées et les combine dans une formule logique sur des concepts graphes, représentés sous forme de groupes d'explications locales. Contrairement aux solutions

existantes, GLGExplainer fournit des explications globales précises et interprétables par l'Homme, parfaitement alignées avec les explications de la vérité terrain (sur des données synthétiques) ou correspondants aux connaissances du domaine existant (sur des données du monde réel). Les formules extraites sont fidèles aux prédictions du modèle, au point de fournir un aperçu de certaines règles parfois incorrectes, apprises par le modèle.

[Xia et al., 2022] proposent T-GNNExplainer pour l'explication du modèle prédictif adaptée au graphe spatio-temporel. Plus précisément, ils considèrent un graphe temporel constitué d'une séquence d'événements temporels. Étant donné un événement cible, leur tâche consiste à trouver un sous-ensemble d'événements survenus précédemment qui conduisent à la prédiction du modèle. Pour gérer ce problème d'optimisation combinatoire, T-GNNExplainer comprend un explorateur pour trouver les sous-ensembles d'événements avec la recherche arborescente de Monte Carlo (MCTS) et un navigateur qui apprend les corrélations entre les événements et permet de réduire l'espace de recherche. En particulier, le navigateur est formé au préalable puis intégré à l'explorateur pour accélérer la recherche et obtenir de meilleurs résultats. GraphExplainer est le premier une méthode explicative adaptée aux modèles de graphes temporels.

[Li et al., 2022f] proposent une nouvelle méthode explicative pour réseaux de neurones sur graphes. La littérature existante se concentre principalement sur la sélection d'un sous-graphe, via l'optimisation combinatoire, pour fournir des explications fidèles. Cependant, la taille exponentielle de l'ensemble des sous-graphes candidats limite l'applicabilité des méthodes aux réseaux de neurones sur graphe à grande échelle. Il améliore cela grâce à une méthode différente : en proposant une structure générative pour réseau de neurones sur graphe basée sur GFlowNets (GFlowExplainer), ils transforment le problème d'optimisation en un problème génératif étape par étape. GFlowExplainer vise à apprendre une politique qui génère une distribution de sous-graphes pour laquelle la probabilité d'un sous-graphe est proportionnelle à sa récompense. L'méthode proposée élimine l'influence de la séquence de noeuds et ne nécessite donc aucune stratégie de pré-entraînement. Il propose également une nouvelle matrice de sommets coupés pour explorer efficacement les états parents de la structure GFlowNets, rendant ainsi la méthode utilisable à grande

échelle.

[Liu et al., 2022a] proposent un paramétrage des distributions prédites par les réseaux de neurones sur graphe en utilisant l'attribution axiomatique, où les distributions se trouvent sur une variété de faible dimension dans un espace d'intégration de haute dimension. Ils exploitent le point de vue géométrique différentiel pour modéliser l'évolution distributionnelle sous forme de courbes sur la variété. Ils reparamètrent les familles de courbes sur la variété et conçoivent un problème d'optimisation convexe pour trouver une courbe unique qui se formule de manière concise de l'évolution distributionnelle pour l'interprétation humaine.

[Bui et al., 2022] introduisent INGREX, une méthode explicative interactive pour les réseaux de neurones sur graphe conçue pour aider les utilisateurs à comprendre les prédictions du modèle. Leur cadre de travail est implémenté sur la base de plusieurs algorithmes d'explication et de bibliothèques existantes.

[Fan et al., 2022] ont fait le constat suivant : les réseaux de neurones graphes ont amélioré les performances de nombreuses tâches liées aux graphes. Malgré leur grand succès, des études récentes ont montré que les réseaux de neurones sur graphe sont très vulnérables aux attaques adverses, où les adversaires peuvent induire en erreur les prédictions des réseaux de neurones sur graphe en modifiant les graphes. D'autre part, l'explication des réseaux de neurones sur graphe (GNNExplainer) permet une meilleure compréhension d'un modèle de réseau de neurones sur graphe entraîné en générant un petit sous-graphe et les caractéristiques les plus influentes pour sa prédiction. Dans cette contribution, ils effectuent des études empiriques sur les arêtes pour valider que GNNExplainer peut agir comme un outil d'inspection et avoir le potentiel de détecter les perturbations contrefactuelles des graphes. Cette découverte les motivent à lancer une nouvelle étude sur le problème : est-ce qu'un réseau de neurones sur graphe et ses explications peuvent être attaqués conjointement en modifiant des graphes avec des désirs malveillants ? Il est difficile de répondre à cette question puisque les objectifs des attaques contrefactuelles et GNNExplainer se contredisent essentiellement. Dans ce travail, il apporte une réponse confirmative à cette question en proposant un nouveau cadre d'attaque (GEAttack), qui peut attaquer à la fois un modèle de réseau de neurones sur graphe et ses explications en exploitant simultanément leurs vulnérabilités. Des expériences approfondies sur

deux méthodes explicatives (GNNExplainer et PGExplainer) sous divers ensembles de données du monde réel démontrent l'efficacité de la méthode proposée.

[Zhang et al., 2023] proposent une méthode explicative pour réseaux de neurones sur graphe adaptée aux tâches de prédiction de liens qui sont souvent écartées dans la littérature. Étant donné que les méthodes explicatives pour réseau de neurones sur graphes existantes ne concernent que les tâches au niveau des noeuds/graphes, ils proposent une méthode explicative pour la prédiction de liens hétérogènes (PaGE-Link) qui génère des explications avec explicabilité de connexion, bénéficie de l'évolutivité du modèle et gère l'hétérogénéité des graphes. Qualitativement, PaGE-Link peut générer des explications sous forme de chemins reliant une paire de noeuds, qui capturent naturellement les connexions entre les deux noeuds et les transfèrent facilement vers des explications interprétables par l'Homme.

[Yang et al., 2023c] ont fait le constat suivant : les chercheurs se consacrent aux problèmes de recherche qui les intéressent et ont souvent des intérêts de recherche en évolution dans leur carrière universitaire. L'étude de la détection des changements d'intérêt en recherche peut aider à trouver des faits pertinents pour les parcours de formation scientifique, les tendances en matière de financement scientifique et la découverte de connaissances. Les méthodes existantes définissent des structures de graphes spécifiques tels que des réseaux d'auteurs-conférences-sujets et des réseaux de co-citations pour détecter les changements d'intérêt de recherche. Soit ils ignorent le facteur temporel, soit ils passent à côté d'informations hétérogènes caractérisant les activités académiques. Plus importants encore, les résultats de détection ne contiennent pas d'interprétations de la manière dont les intérêts de recherche évoluent au fil du temps, réduisant ainsi la crédibilité du modèle. Pour résoudre ces problèmes, ils proposent un nouveau modèle de détection de changement d'intérêt de recherche interprétable avec des graphes hétérogènes temporels. Ils construisent d'arête des graphes hétérogènes temporels pour représenter les intérêts de recherche des auteurs cibles. Pour rendre la détection interprétable, ils conçoivent un réseau de neurones profond pour paramétrer le processus de génération d'interprétation sur les résultats prédits sous la forme d'un sous-graphe pondéré. De plus, pour améliorer le processus de formation, ils proposent une stratégie d'échantillonnage de données négatives sémantique pour générer des graphes de décalage auxiliaires non intéressants

sous forme d'échantillons contrastés.

[Fang et al., 2023] s'intéressent à l'explicabilité des réseaux de neurones sur graphe au niveau des caractéristiques d'entrée et des neurones. Ils proposent une méthode explicative agnostique, CooperativGNNExplanation (CGE) pour générer simultanément le sous-graphe explicatif et le sous-réseau. , qui expliquent conjointement comment le modèle de réseau de neurones sur graphe est arrivé à sa prédiction. Plus précisément, ils initialisent les scores d'importance des caractéristiques des arêtes d'entrée et des neurones cachés avec des réseaux de masquage. Ensuite, ils recalculent de manière itérative les scores d'importance, affinant le sous-graphe et le sous-réseau saillants en supprimant les caractéristiques et les neurones à faible score à chaque itération. Grâce à un tel apprentissage coopératif, CGE génère non seulement des explications fidèles et concises, mais montre également comment les informations importantes circulent en activant et en désactivant les neurones.

[Int, ] proposent un réseau de neurones sur graphes à noyau moléculaire (MolKGNN) pour l'apprentissage de la représentation moléculaire, qui présente l'invariance de conformation, la chiralité et l'explicabilité. Pour MolKGNN, ils construisent cela par une convolution de graphe moléculaire pour capturer le modèle chimique en comparant la similarité de l'atome avec les noyaux moléculaires de la vérité terrain. De plus, ils propagent le score de similarité pour capturer le modèle chimique.

[Yutao Hu, ] proposent une méthode de détection et d'interprétation de vulnérabilité au niveau des découpes de graphe selon le modèle prédictif. La méthode est écrite en C/C++ et extrait des parties de graphes pour réduire l'interférence des informations redondantes dans les échantillons ; deuxièmement, un modèle de réseau de neurones sur graphe est utilisé pour intégrer les tranches afin d'obtenir leurs représentations vectorielles afin de préserver les informations structurelles et les caractéristiques de vulnérabilité du code source ; puis les représentations vectorielles des tranches sont introduites dans le modèle de détection de vulnérabilité à des fins d'entraînement et de prédiction ; enfin, le modèle de détection de vulnérabilité entraîné et les tranches de vulnérabilité à expliquer sont introduits dans l'interpréteur de vulnérabilité pour obtenir les lignes spécifiques du code de vulnérabilité.

[Gao et al., 2023] proposent un réseau de neurones sur graphes spatio-temporels explicables pour la détection de colmatage altérant l'efficacité des toupies creusant

des tunnels sous-terrain. La méthode permet de juger du niveau de risque de colmatage pour chaque anneau de la toupie. Les graphes sont transformés à partir de données de surveillance de séries chronologiques multidimensionnelles originales, qui incluent des informations spatio-temporelles, dont les risques sont qualifiés d'étiquettes par les experts. Un cas pratique de tunneling est appliqué pour discuter des résultats du modèle et les données d'une autre ligne de tunneling sont appliquées pour valider la robustesse. Le processus de mise en commun hiérarchique et le mécanisme d'attention mettent en évidence les noeuds importants de chaque graphe en tant qu'explicabilité spatio-temporelle. Des caractéristiques de colmatage typiques peuvent être observées dans ces noeuds importants. Avec des indices de réseau complexes, des explications plus détaillées sont obtenues.

[Tian et al., 2023] proposent un nouveau modèle de réseaux de neurones sur graphe appelé réseau de graphes focalisés itérativement (IFGN), qui peut identifier progressivement les atomes/groupes clés de la molécule qui sont étroitement liés aux propriétés prédites par le mécanisme de focalisation en plusieurs étapes. Dans le même temps, la combinaison du mécanisme de mise au point en plusieurs étapes avec la visualisation peut également générer des interprétations en plusieurs étapes, permettant ainsi d'acquérir une compréhension approfondie des comportements prédictifs du modèle. Pour les huit ensembles de données étudiés, le modèle IFGN a obtenu de bonnes performances de prédiction, ce qui indique que le mécanisme de focalisation en plusieurs étapes proposé peut également améliorer les performances du modèle, en plus d'augmenter l'explicabilité du modèle construit.

[Wang et al., 2023b] est un travail plutôt à dominante applicative . Ils ont adopté plusieurs méthodes d'attribution avancées pour différents modèles de réseaux de neurones sur graphes et ils les ont appliquées pour expliquer les tâches de prédiction des propriétés de plusieurs molécules médicamenteuses, permettant l'identification et la visualisation d'informations chimiques vitales dans les réseaux.

[Chatzianastasis et al., 2023] introduisent un nouveau modèle de réseaux de neurones sur graphes multicouches explicables (EMGNN) pour identifier les gènes du cancer en exploitant plusieurs réseaux d'interaction génique et des données multi-omiques pan-cancer ?. Contrairement à l'apprentissage sur graphe conventionnel sur un seul réseau biologique, EMGNN utilise un réseau neuronal graphe multicouche

pour apprendre de plusieurs réseaux biologiques afin de prédire avec précision les gènes du cancer. Leur méthode surpasse systématiquement toutes les méthodes existantes, avec une amélioration moyenne de 7,15% de l'aire sous la courbe de précision-rappel (AUC) par rapport à la méthode actuelle. Il est important de noter qu EMGNN a intégré plusieurs graphes pour hiérarchiser les gènes de cancer nouvellement prédits avec des prédictions contrefactuelles provenant de réseaux biologiques uniques. Pour chaque prédiction, EMGNN a fourni des informations biologiques précieuses via à la fois des explications sur l'importance des caractéristiques au niveau du modèle et une analyse de l'enrichissement des ensembles de gènes au niveau moléculaire.

[Zhang et al., b] font le constat suivant : la valeur de Shapley issue de la théorie des jeux a été proposée comme méthode privilégiée pour calculer l'importance des caractéristiques dans les prédictions de modèles sur les images, le texte, les données tabulaires et, récemment, les réseaux de neurones graphes. Dans ce travail, ils revisitent la pertinence de la valeur de Shapley pour l'explication des réseaux de neurones sur graphes, où la tâche consiste à identifier le sous-graphe et les noeuds constitutifs les plus importants pour les prédictions du réseau de neurones sur graphes. Ils affirment que la valeur de Shapley n'est pas un choix idéal pour les données de type graphe car elle n'est par définition pas sensible à la structure. Ils proposent une méthode d'explication basée sur la structure du graphe (GStarX) pour exploiter les informations critiques sur la structure du graphe afin d'améliorer l'explication. Plus précisément, ils définissent une fonction de notation basée sur une nouvelle valeur sensible à la structure issue de la théorie des jeux coopératifs proposée par Hamiache et Navarro (HN). Lorsqu'elle est utilisée pour évaluer l'importance des noeuds, la valeur HN utilise des structures graphes pour attribuer le surplus de coopération entre les noeuds voisins, ressemblant à la transmission de messages dans les réseaux de neurones sur graphes, de sorte que les scores d'importance des noeuds reflètent non seulement l'importance des caractéristiques du noeud, mais également les rôles structurels des noeuds.

[Xie et al., 2022] proposent une méthode explicative pour réseaux de neurones sur graphe indépendante des contextes (TAGE) agnostiques aux modèles prédictifs et formée sous l'autosupervision. TAGE permet l'explication des réseaux de neu-

rones sur graphe et permet une explication efficace des modèles multitâches. Leurs expériences montrent que TAGE peut considérablement accélérer l'efficacité de l'explication en utilisant le même modèle pour expliquer les prédictions de plusieurs tâches tout en obtenant une qualité d'explication correcte.

[Feng et al., 2022b] proposent DEGREE (Decomposition based Explanation for GRaph nEural nEtworks) pour fournir une explication fidèle des prédictions des réseaux de neurones sur graphe. En décomposant le mécanisme de génération et d'agrégation d'informations des réseaux de neurones sur graphe, DEGREE permet de suivre les contributions de composants spécifiques du graphe à la prédiction finale. Sur cette base, ils conçoivent en outre un algorithme d'interprétation au niveau du sous-graphe pour révéler des interactions complexes entre les noeuds du graphe qui étaient négligées par les méthodes précédentes.

[Cucala et al., 2022] proposent une nouvelle famille de réseaux de neurones sur graphe qui peuvent être entraînés efficacement, mais où toutes les prédictions peuvent être expliquées symboliquement sous forme d'inférences logiques dans Datalog, un formalisme basé sur des règles connues. En particulier, ils montrent comment coder un graphe de connaissances dans un graphe avec des vecteurs de caractéristiques numériques, traiter ce graphe à l'aide d'un réseau de neurones sur graphe et décoder le résultat dans un graphe de connaissances de sortie. Il utilise une nouvelle classe de réseaux de neurones sur graphe monotones (MGNN) pour garantir que ce processus est équivalent à une série d'applications d'un ensemble de règles Datalog. Ils montrent également que, étant donné un MGNN arbitraire, ils peuvent automatiquement extraire des règles qui caractérisent complètement la transformation.

[Rao et al., 2021b] proposent xFraud, un modèle de réseau de neurones sur graphes construit pour la prédiction de transactions frauduleuses, explicables, qui est principalement composé d'un détecteur et d'une méthode explicative. Le détecteur xFraud peut prédire de manière efficace et efficiente la légitimité des transactions entrantes. Plus précisément, il utilise un réseau neuronal graphe hétérogène pour apprendre des représentations expressives à partir des entités informatives typées de manière hétérogène dans les journaux de transactions. L'une méthode explicative de xFraud peut générer des explications significatives et compréhensibles par l'Homme à partir de graphes pour faciliter les processus ultérieurs dans l'unité commerciale.

Dans leurs expériences avec xFraud sur des réseaux de transactions réels comptant jusqu'à 1,1 milliard de noeuds et 3,7 milliards d'arêtes, xFraud est capable de surpasser divers modèles de base dans de nombreuses mesures d'évaluation tout en restant évolutif dans des environnements distribués.

[Shin et al., 2022] proposent Prototype-bAsed GNN-Explainer (PAGE), une nouvelle méthode explicative qui explique ce que le modèle sous-jacent a appris en fournissant des informations humainement interprétables, tels que des prototypes interprétables. Plus précisément, leur méthode effectue un clustering sur l'espace d'intégration du modèle de réseau de neurones sur graphe sous-jacent ; extrais les intégrations dans chaque cluster ; et découvre des prototypes, qui servent d'explications de modèle, en estimant le sous-graphe commun maximum (MCS) à partir des plongements extraits.

[Feng et al., 2022a] ont fait le constat suivant : les noyaux de graphes sont historiquement la méthode la plus utilisée pour les tâches de classification de graphes. Cependant, ces méthodes souffrent de performances limitées en raison des caractéristiques combinatoires artisanales des graphes. Ces dernières années, les réseaux de neurones sur graphes sont devenus les approches clés pour des tâches liées aux graphes en raison de leurs performances supérieures. La plupart des réseaux de neurones sur graphe sont basés sur les frameworks MPNN (Message Passing Neural Network). Cependant, des études récentes montrent que les MPNN ne peuvent pas dépasser la puissance de l'algorithme de Weisfeiler-Lehman (WL) dans le test d'isomorphisme des graphes. Pour répondre aux limites des méthodes existantes de noyau de graphe et de réseaux de neurones sur graphe, ils proposent dans cette contribution un nouveau cadre pour les réseaux de neurones sur graphe, appelé Kernel Graph Neural Networks (KerGNN), qui intègre les noyaux de graphe dans le processus de transmission de messages des réseaux de neurones sur graphe. Inspirés par les filtres de convolution des réseaux de neurones convolutifs (CNN), les KerGNN adoptent des graphes cachés entraînables en tant que filtres graphes qui sont combinés avec des sous-graphes pour mettre à jour les intégrations de noeuds à l'aide de noyaux de graphes. De plus, ils montrent que les MPNN peuvent être considérés comme des cas particuliers de KerGNN. Ils appliquent les KerGNN à plusieurs tâches liées aux graphes et utilisons la validation croisée pour effectuer des

comparaisons équitables avec des références. Ils montrent que leur méthode atteint des performances compétitives par rapport aux méthodes existantes, démontrant le potentiel d'augmenter la capacité de représentation des réseaux de neurones sur graphe. Ils montrent également que les filtres graphes entraînés dans les KerGNN peuvent révéler les structures graphes locales de l'ensemble de données, ce qui améliore considérablement l'explicabilité du modèle par rapport aux modèles de réseaux de neurones sur graphe conventionnels.

[Liu et al., 2022b] font le constat suivant : les graphes sont largement présents dans l'analyse des réseaux sociaux et les schémas électroniques, où les réseaux de neurones sur graphes constituent le modèle de choix pour leurs représentations . Les réseaux de neurones sur graphe peuvent être biaisés en raison d'attributs sensibles et de la topologie du réseau. Avec les travaux existants qui apprennent une représentation équitable des noeuds ou une matrice d'adjacence, obtenir une forte garantie d'équité de groupe tout en préservant l'exactitude des prédictions reste un défi, le compromis équité-précision restant obscur pour les décideurs humains. Ils définissent et analysent les arêtes, pour mesurer l'équité de celles-ci afin d'optimiser la matrice d'adjacence sans nuire de manière significative à la précision des prédictions. Pour comprendre la nuance du compromis équité-précision, ils proposent en outre des méthodes explicatives macroscopiques et microscopiques pour révéler les compromis et l'espace que l'on peut exploiter. La méthode d'explication macroscopique est basée sur un échantillonnage stratifié et une programmation linéaire pour expliquer de manière déterministe la dynamique de l'équité du groupe et la précision des prédictions. En descendant jusqu'au niveau microscopique, ils proposent une explication basée sur le chemin qui révèle comment la topologie du réseau conduit au compromis.

[Prado-Romero and Stilo, 2022] présentent GRETEL, une méthode unifiée pour développer et tester des méthodes GCE dans plusieurs contextes. GRETEL est une méthode d'évaluation hautement extensible qui favorise la science ouverte et la reproductibilité de l'évaluation en fournissant un ensemble de mécanismes bien définis à intégrer et à gérer facilement : des ensembles de données réels et synthétiques, des modèles d'apprentissage machine, des méthodes explicatives et les mesures d'évaluation.

[Saha et al., 2022] ont fait le constat suivant : les réseaux de neurones sur graphes apprennent les représentations de noeuds en agrégeant le vecteur de caractéristiques d'un noeud avec ses voisins. Ils fonctionnent bien dans une variété de tâches graphes. Cependant, pour améliorer la fiabilité et la crédibilité de ces modèles lors de leur utilisation dans des scénarios critiques, il est essentiel d'examiner les mécanismes de prise de décision de ces modèles plutôt que de les traiter comme des boîtes noires. Leur méthode centrée sur le modèle donne un aperçu du type d'informations apprises par les réseaux de neurones sur graphe sur les voisinages des noeuds au cours de la tâche de classification des noeuds. Ils proposent un générateur de quartier comme une méthode explicative qui génère des quartiers optimaux pour maximiser une prédiction de classe particulière du réseau de neurones sur graphe entraîné. Ils formulent la génération de quartier comme un problème d'apprentissage par renforcement et utilisent une méthode de gradient de politique pour former leur générateur en utilisant le classificateur de noeuds basé sur les réseaux de neurones sur graphe construit en amont. Leur méthode fournit des explications intelligibles des mécanismes d'apprentissage des réseaux de neurones sur graphe sur des jeux de données synthétiques et réels et met même en évidence certaines lacunes de ces modèles.

[Veyrin-Forrer et al., 2022] proposent une méthode explicative qui isole les caractéristiques internes du modèle prédictif, cachées dans les couches du modèle, qui sont automatiquement identifiées par le réseau de neurones sur graphe pour classer les graphes. Ils montrent que cette méthode permet de connaître les parties des graphes utilisés par les réseaux de neurones sur graphe avec beaucoup moins de biais que les méthodes de l'état de l'art et donc d'apporter de la confiance dans le processus de décision.

[Kamal et al., 2023] ont fait le constat suivant : des travaux récents ont proposé d'expliquer les réseaux de neurones sur graphe à l'aide de règles d'activation. Les règles d'activation permettent de capturer des configurations spécifiques dans l'espace d'intégration d'une couche donnée qui sont discriminantes pour la décision des réseaux de neurones sur graphes. Ces règles détectent également les descripteurs cachés des graphes d'entrée. Cela nécessite d'associer ces règles à des graphes représentatifs. Dans cette contribution, ils proposent d'une part une analyse d'algo-

rithmes heuristiques pour extraire les règles d'activation, et d'autre part l'utilisation de distances de graphes issu du transport optimale pour associer chaque règle au graphe le plus spécifique qui les déclenche.

[Zhou et al., 2022b] proposent un modèle prédictif basé sur les réseaux de neurones sur graphes convolutifs, interprétable et clairsemé, dénommé (SGCN) pour l'identification et la classification de la maladie d'Alzheimer (MA) à l'aide de données d'imagerie cérébrale avec de multiples modalités. SGCN applique un mécanisme d'attention avec parcimonie pour identifier la structure de sous-graphe la plus discriminante et les caractéristiques de noeud importantes pour la détection de la MA. Le modèle apprend les probabilités d'importance faible pour chaque caractéristique et arrête de noeud avec l'entropie et la régularisation de l'information mutuelle. Ils ont ensuite utilisé ces informations pour trouver les signatures des régions d'intérêt (ROI) et mettre l'accent sur les connexions du réseau cérébral spécifiques à la maladie en détectant la différence significative de connecteurs entre les régions des groupes de contrôle sain (HC) et celles des groupes types AM. Ils ont évalué le SGCN sur la base de données ADNI avec des données d'imagerie provenant de trois modalités, dont VBM-MRI, FDG-PET et AV45-PET, et ont observé que les probabilités importantes apprises sont efficaces pour l'identification de l'état de la maladie et la faible explicabilité des données spécifiques à la maladie. Les ROIs saillantes détectées et les connexions entre les réseaux les plus discriminants sont interprétées par leur méthode. Ils montrent une forte correspondance avec les preuves de neuro-imagerie antérieures associées à la MA.

[He et al., 2022b] proposent Illuminati, une méthode explicative pour les applications de cybersécurité utilisant des réseaux de neurones sur graphes. Grâce à un graphe et à un modèle de réseau de neurones sur graphe pré-entraîné, Illuminati est capable d'identifier les noeuds, arrêtes et attributs importants qui contribuent à la prédiction sans nécessiter aucune connaissance préalable des réseaux de neurones sur graphes. Ils évaluent Illuminati dans deux applications de cybersécurité, à savoir la détection de vulnérabilités de code et la détection de vulnérabilités de contrats intelligents. Les expériences montrent qu'Illuminati obtient des résultats d'explication plus précis que les méthodes , en particulier, 87,6 % des sous-graphes identifiés par Illuminati sont capables de conserver leur prédiction originale, soit une amélioration

de 10,3 % par rapport aux autres à 77,3 %. De plus, l'explication des Illuminati peut être facilement comprise par les experts du domaine, suggérant l'utilité significative pour le développement d'applications de cybersécurité.

[Zhu et al., 2022] proposent une méthode pour évaluer la fiabilité des modèles de détection de botnets basés sur réseaux de neurones sur graphe, appelée BD-GNNExplainer. Concrètement, BD-GNNExplainer extrait les données qui contribuent le plus à la décision de réseaux de neurones sur graphe en réduisant la perte entre les résultats de classification générés par le sous-graphe sélectionné en entrée du modèle de réseaux de neurones sur graphe et les résultats générés par l'ensemble du graphe en entrée. Ils calculent la pertinence entre les données basées sur le modèle et les données informatives pour quantifier un score exprimant l'explicabilité. Pour les modèles de réseaux de neurones sur graphe de structure différente, ces scores leur indiquent lesquels sont les plus fiables et deviendront finalement une base essentielle pour l'optimisation du modèle.

[Li et al., 2022a] proposent Shapley Explainer, qui fournit des scores d'importance raisonnable aux noeuds d'entrée d'un réseau de neurones sur graphe dans un coût de calcul approprié, fournissant ainsi une interprétation valide et raisonnable du réseau neuronal graphe sur un réseau défini par logiciel. La méthode proposée dérive le classement d'importance des noeuds topologiques en combinant les valeurs Shapley avec une matrice de masque discret souple. Ils appliquent Shapley Explainer au modèle RouteNet, un modèle de réseau de neurones sur graphe qui fournit des prédictions intelligentes des mesures de performances du réseau SDN. Les résultats expérimentaux montrent que Shapley Explainer peut fournir des interprétations efficaces pour RouteNet. Il vérifie également que le modèle RouteNet peut apprendre correctement la relation entre les descripteurs, ce qui peut fournir une meilleure compréhension du processus de prédiction de RouteNet, favorisant ainsi l'application des systèmes SDN basés sur réseau de neurones sur graphe dans la pratique de l'ingénierie.

[He et al., 2022c] font le constat suivant : les réseaux de neurones à graphes temporels (TGNN) ont été largement utilisés pour modéliser des tâches liées aux graphes évoluant dans le temps en raison de leur capacité à capturer à la fois la dépendance topologique des graphes et la dynamique temporelle non linéaire. L'ex-

plication des TGNN est d'une importance vitale pour un modèle transparent et fiable. Cependant, la structure topologique complexe et la dépendance temporelle rendent l'explication des modèles TGNN très difficile. Dans cette contribution, ils proposent une nouvelle méthode explicative pour les modèles TGNN. Étant donné une série chronologique sur un graphe à expliquer, le cadre peut identifier les explications dominantes sous la forme d'un modèle graphe probabiliste sur une période donnée. Des études de cas sur le domaine des transports démontrent que la méthode proposée peut découvrir des structures de dépendance dynamiques dans un réseau routier pendant une période donnée.

[Han and Zhao, 2022] font le constat suivant : le calcul de réservoir graphe (GraphRC) attire de plus en plus l'attention en raison de sa grande efficacité de formation. Cependant, étant donné que GraphRC est développé sans connaissance de son mécanisme interne, son déploiement dans la pratique ne peut pas être entièrement fiable. Bien qu'il existe certaines méthodes existantes qui peuvent être étendues pour interpréter GraphRC, le rôle spécifique joué par chaque neurone (c'est-à-dire le noeud réservoir) de GraphRC est beaucoup moins exploré. Pour résoudre ce problème, la propriété de mémoire latente à court terme de chaque noeud réservoir de GraphRC est caractérisée qualitativement pour dévoiler son rôle dans la prédiction du signal graphe, permettant ainsi un GraphRC interprétable. Plus précisément, ils déduisent d'arête l'équivalence entre le GraphRC et le calcul de réservoir (RC) conventionnel. Ensuite, les propriétés de mémoire sous-jacentes du GraphRC et de ses noeuds réservoir peuvent être caractérisées en théorie par l'accessibilité multisource parmi les noeuds réservoir dans le RC transformé. De plus, les modèles temporels distincts cachés dans les noeuds de réservoir sont identifiés, puis un mécanisme d'attention basé sur les modèles temporels identifiés est déployé dans GraphRC pour améliorer ses performances. De plus, l'efficacité de l'explicabilité de GraphRC et de GraphRC amélioré est vérifiée sur le système dynamique spatio-temporel Lorenz-96. Les résultats expérimentaux du système chaotique spatio-temporel Lorenz-96 et de trois ensembles de données de trafic réels démontrent que le GraphRC amélioré est supérieur au GraphRC original et peut atteindre des performances de prédiction comparables aux modèles de base , mais avec beaucoup moins de coût pour la formation.

[Spinelli et al., 2022] étudient le degré d'explicabilité des réseaux de neurones

sur graphes . Les méthodes explicatives existantes fonctionnent en trouvant des sous-graphes globaux/locaux pour expliquer une prédiction, mais ils sont appliqués une fois qu'un réseau de neurones sur graphe a déjà été formé. Ici, ils proposent un une méthode explicative pour améliorer le niveau d'explicabilité d'un réseau de neurones sur graphe directement au moment de la formation, en orientant la procédure d'optimisation vers des minima qui permettent aux méthodes explicatives post hoc d'obtenir de meilleurs résultats, sans sacrifier la précision globale du réseau de neurones sur graphes. Leur cadre (appelé MATE, MetA-Train to Explain) entraîne conjointement un modèle pour résoudre la tâche d'origine, par exemple la classification des noeuds, et pour fournir des sorties facilement traitables pour les algorithmes qui expliquent les décisions du modèle d'une manière conviviale. En particulier, ils méta-entraînent les paramètres du modèle pour minimiser rapidement l'erreur d'un GNNExplainer au niveau de l'instance formée à la volée sur des noeuds échantillonnés aléatoirement. La représentation interne finale repose sur un ensemble de descripteurs qui peuvent être « mieux » comprises par un algorithme d'explication, par exemple une autre instance de GNNExplainer. Leur méthode indépendante du modèle peut améliorer les explications produites pour différentes architectures de réseau de neurones sur graphe et utiliser n'importe quelle une méthode explicative basée sur un exemple. Des expériences sur des ensembles de données synthétiques et réelles pour la classification des noeuds et des graphes montrent qu'ils peuvent produire des modèles systématiquement plus faciles à expliquer par différents algorithmes. De plus, cette augmentation de l'explicabilité n'a aucun coût pour la précision du modèle.

[Bacciu and Numeroso, 2022] présentent une nouvelle méthode explicative locale spécifiquement adaptée aux données de type graphes et aux réseaux de neurones sur graphes. Leur méthode exploite l'apprentissage par renforcement pour générer des perturbations locales significatives du graphe, dont ils cherchent une interprétation de la prédiction. Ces points de données perturbés sont obtenus en optimisant un score multi-objectif prenant en compte les similitudes tant au niveau structurel qu'au niveau des sorties du modèle profond. De cette manière, ils sont en mesure de remplir un ensemble d'échantillons voisins informatifs pour le graphe de requête, qui est ensuite utilisé pour adapter localement un modèle interprétable pour le compor-

tement prédictif du réseau profond à la prédiction du graphe de requête. Ils montrent l'efficacité de la méthode explicative proposée par une analyse qualitative sur deux ensembles de données chimiques, TOX21 et ESOL et par des résultats quantitatifs sur un ensemble de données de référence pour les explications, CYCLIQ.

[Veyrin-Forrer et al., ] proposent d'explorer les modèles d'activation dans les couches cachées pour comprendre comment les réseaux de neurones sur graphes perçoivent le monde. Le problème n'est pas de découvrir des modèles d'activation qui sont individuellement hautement discriminants pour une sortie du modèle. Au lieu de cela, le défi consiste à fournir un petit ensemble de modèles couvrant tous les graphes d'entrée. À cette fin, ils introduisent le domaine des modèles d'activation subjectifs. Ils définissent un algorithme efficace et fondé sur des principes pour énumérer les modèles d'activations dans chaque couche cachée. La méthode proposée pour quantifier l'intérêt de ces modèles est ancrée dans la théorie de l'information et est capable de prendre en compte les connaissances de base sur les données de type graphes. Les modèles d'activation peuvent ensuite être redécrits grâce à des langages de motifs impliquant des descripteurs interprétables. Ils montrent que les modèles d'activation fournissent des informations sur les caractéristiques utilisées par le réseau de neurones sur graphe pour classer les graphes. Cela permet notamment d'identifier les descripteurs cachés construits par le réseau de neurones sur graphe à travers ses différentes couches. En outre, ces modèles peuvent ensuite être utilisés pour expliquer les décisions du réseau de neurones sur graphes. Les expériences sur des ensembles de données synthétiques et réels montrent des performances très compétitives, avec une amélioration de la fidélité allant jusqu'à 200 % dans l'explication de la classification des graphes par rapport aux méthodes de l'état de l'art.

[Li et al., 2022g] ont construit un nouveau modèle de réseaux de neurones sur graphe transparent et explicable en distillant les connaissances à partir de modèles de « boîte noire » pré-entraînés. Plus précisément, en préservant simultanément la fidélité aux comportements du modèle d'origine et en optimisant la perte de prédiction, un réseau neuronal graphe peu profond avec des poids de « contribution » explicites entre deux noeuds est formé. Ensuite, une stratégie de sélection de voisins est construite sur ces pondérations explicites pour garantir des niveaux élevés

de performance et d'explicabilité. Pour évaluer le cadre proposé, leur méthode est intégrée à quatre modèles : GCN, GAT, GraphSAGE et AM-GCN. Les résultats expérimentaux sur trois ensembles de données du monde réel montrent l'efficacité du cadre proposé.

[Li et al., 2022h] font le constat suivant : avec le déploiement rapide de méthodes basées sur les réseaux de neurones sur graphes dans un large éventail d'applications telles que la prédiction de liens, la détection de communautés et la classification de noeuds, l'explicabilité des réseaux de neurones sur graphe devient un élément indispensable pour une prise de décision prédictive et fiable. Pour atteindre cet objectif, certains travaux récents se concentrent sur la conception de modèles de réseaux de neurones sur graphe explicables tels que GNNExplainer, PGExplainer et Gem. Ces méthodes explicatives pour réseaux de neurones sur graphe ont montré des performances remarquables dans l'explication des résultats prédictifs des réseaux de neurones sur graphe. Malgré leur succès, la robustesse de ces méthodes explicatives est moins explorée en termes de vulnérabilités des méthodes explicatives pour réseaux de neurones sur graphe. Les perturbations telles que les attaques contrefactuelles peuvent conduire à des explications inexactes et par conséquent provoquer des catastrophes. Ainsi, dans cette contribution, ils font le premier pas et ils s'efforcent d'explorer la robustesse des méthodes explicatives pour réseaux de neurones sur graphe. Pour être précis, ils définissent d'arête deux scénarios d'attaque contrefactuelles : un adversaire agressif et un adversaire conservateur pour contaminer les structures de graphes. Ils étudient ensuite les impacts des graphes empoisonnés sur l'explicabilité de trois méthodes explicatives pour réseaux de neurones sur graphe courants avec trois métriques d'évaluation standard : Fidelity, Fidelity+, Fidelity -, et la parcimonie. Ils mènent des expériences sur des ensembles de données synthétiques et réels et ils se concentrent sur deux tâches d'exploration de graphes populaires : la classification de noeuds et la classification de graphes. Leurs résultats empiriques suggèrent que les méthodes explicatives pour les réseaux de neurones sur graphes ne sont généralement pas robustes aux attaques adverses provoquées par les bruits structurels des graphes.

[Ragno et al., 2022] font le constat suivant : les réseaux de neurones sur graphes se sont révélés être un outil clé pour traiter de nombreux problèmes et domaines

tels que la chimie, le traitement du langage naturel et les réseaux sociaux. Bien que la structure des couches soit simple, il est difficile d'identifier les modèles appris par le réseau neuronal graphe. Plusieurs travaux proposent des méthodes post-hoc pour expliquer les prédictions de graphes, mais peu d'entre eux tentent de générer des modèles interprétables. A l'inverse, le thème des modèles interprétables est très étudié en reconnaissance d'images. Compte tenu de la similitude entre les domaines d'image et de graphe, ils analysent l'adaptabilité des réseaux de neurones basés sur des prototypes pour la classification des graphes et des noeuds. En particulier, ils étudient l'utilisation de deux réseaux interprétables, ProtoPNet et TesNet, dans le domaine des graphes. Ils montrent que les réseaux adaptés parviennent à atteindre des scores de précision meilleurs ou plus élevés que leurs modèles de boîte noire respectifs et des performances comparables avec des modèles auto-explicables . Montrant comment extraire les explications ProtoPNet et TesNet sur les réseaux de neurones sur graphes, ils étudient plus en détail comment obtenir des explications globales et locales pour les modèles formés. Ils évaluent ensuite les explications des modèles interprétables en les comparant avec des méthodes post-hoc et des modèles auto-explicables. Leurs résultats montrent que l'application de TesNet et ProtoPNet au domaine des graphes produit des prédictions qualitatives tout en améliorant leur fiabilité et leur transparence.

[Pereira et al., 2022] proposent une méthode explicative simple pour les réseaux de neurones sur graphes. Ils s'inspirent de travaux récents montrant que les réseaux de neurones sur graphe peuvent souvent être simplifiés sans aucun impact sur les performances, ils proposent de distiller les réseaux de neurones sur graphe en modèles plus simples (linéaires) et d'expliquer ces derniers à la place. Après distillation, ils extraient les explications en résolvant un problème d'optimisation convexe qui identifie les noeuds les plus pertinents pour une prédiction donnée au niveau du noeud.

[Lv et al., 2022] proposent Glocal-Explainer pour générer des explications au niveau du modèle prédictif, qui traitent les informations locales des sous-structures dans le graphe pour rechercher une explicabilité globale. Plus précisément, ils étudient la fidélité et la généralité de chaque explication candidate. Dans la littérature, la fidélité et l'infidélité sont largement considérées comme mesurant la fidélité, mais

les deux mesures peuvent ne pas s'aligner l'une sur l'autre et n'ont encore été intégrées ensemble dans aucune méthode explicative. Au contraire, la généralité, qui mesure combien d'instances partagent la même structure d'explication, n'est pas encore explorée en raison du coût de calcul lié à l'exploration fréquente de sous-graphes. Ils introduisent une méthode d'exploration de sous-graphes adaptée pour mesurer la généralité ainsi que la fidélité lors de la génération de candidats explicatifs. De plus, ils définissent formellement le problème de génération d'explications de GLocal et le lien avec les problèmes classiques de recouvrement de graphes. Un algorithme glouton est utilisé pour trouver la solution.

[Zhou et al., 2022a] proposent une nouvelle méthode explicative pour les réseaux de neurones convolutifs sur graphe pour l'identification et la classification de la maladie d'Alzheimer (MA) à l'aide de données d'imagerie cérébrale multimodales. Plus précisément, ils ont étendu la méthode Grad-CAM (Gradient Class Activation Mapping) pour quantifier les caractéristiques les plus discriminantes identifiées par réseaux de neurones sur graphes convolutifs à partir des modèles de connectivité cérébrale. Ils les ont ensuite utilisés pour trouver la signature des régions d'intérêt (ROI) signatures en détectant la différence de caractéristiques entre les régions des groupes de contrôle sain (HC), de déficience cognitive légère (MCI) et de MA. Ils ont mené les expériences sur la base de données ADNI avec des données d'imagerie provenant de trois modalités, dont VBM-MRI, FDG-PET et AV45-PET, et ils ont montré que les descripteurs de retour sur investissement apprises par leur méthode étaient efficaces pour améliorer les performances de prédiction des scores cliniques et l'identification de l'état de la maladie. Ils ont également identifié avec succès des biomarqueurs associés à la MA et au MCI.

[Pfeifer et al., 2022] présentent une nouvelle méthode pour la détection de sous-réseaux de maladies via des réseaux de neurones sur graphe explicables. Dans cette méthode, chaque patient est représenté par la topologie d'un réseau protéine-protéine (PPI) et les noeuds sont enrichis de données moléculaires multimodales, telles que l'expression des gènes et la méthylation de l'ADN. Par conséquent, leur nouvelle modification de réseaux de neurones sur GraphExplainer pour des explications à l'échelle du modèle peut détecter des sous-réseaux potentiels de maladies, ce qui présente une grande pertinence pratique.

[Orlichenko et al., 2022] proposent une nouvelle méthode explicative pour les réseaux de neurones sur graphes dans la problématique de l'IRMf. Ils utilisent PGI-GCN pour prédire l'âge des enfants et des jeunes adultes sur la base des données IRMf multi-paradigmes de l'ensemble de données de la Philadelphia Neurodevelopmental Cohort (PNC). Ils montrent que PGI-GCN a une capacité prédictive supérieure par rapport à un modèle profond plus simple qui utilise la connectivité fonctionnelle plus le sexe sans le graphe au niveau de la population. Un masque apprenable identifie 3 différences importantes de connectivité intra-réseau (récupération de la mémoire, attention dorsale et sous-corticale) et 3 différences inter-réseaux importantes (visuelle-cérébelleuse, attention visuelle-dorsale et sous-corticale-cérébelleuse) entre les enfants et les jeunes adultes.

[Ji et al., 2022] proposent une nouvelle méthode d'explicabilité post-hoc basée sur la fidélité locale, appelée TraP2, qui peut générer une explication haute fidélité. Considérant qu'à la fois la structure du graphe pertinente et les caractéristiques importantes à l'intérieur de chaque noeud doivent être mises en évidence, une architecture à trois couches dans TraP2 est conçue : i) le domaine d'interprétation est défini à l'avance par la couche de traduction ; ii) les comportements prédictifs locaux des réseaux de neurones sur graphe expliqués sont sondés et surveillés par la couche de perturbation, dans laquelle de multiples perturbations pour la structure du graphe et le niveau des caractéristiques sont effectuées dans le domaine d'interprétation ; et iii) des explications très fidèles sont générées en ajustant la limite de décision locale des réseaux de neurones sur graphe expliqués via la couche.

[Herath et al., 2022] proposent une nouvelle méthode explicative adaptée aux applications en sécurité informatique, et utilise les réseaux de neurones sur graphes pour traiter les logiciels malveillants sous la forme de graphe de flux de contrôle (CFG) et classifier ces logiciels malveillants. Ils proposent CFGExplainer, un modèle basé sur l'apprentissage profond pour interpréter les résultats de classification des logiciels malveillants orientés pour les réseaux de neurones sur graphes. CFGExplainer identifie un sous-graphe du malware CFG qui contribue le plus à la classification et donne un aperçu de l'importance des noeuds qu'il contient. À leur connaissance, CFGExplainer est la première méthode expliquant la classification des logiciels malveillants basée sur les réseaux de neurones sur graphes.

[Park, 2023] font le constat suivant : l'apprentissage avec des données structurées sous forme de graphes, telles que les réseaux sociaux, biologiques et financiers, nécessite des représentations efficaces de faible dimension pour gérer leurs interactions vastes et complexes. Récemment, avec les progrès des réseaux de neurones et des algorithmes d'intégration, de nombreuses méthodes non supervisées ont été proposées pour de nombreuses tâches avec des résultats prometteurs ; cependant, peu de recherches ont été menées sur l'interprétation des représentations non supervisées et, plus particulièrement, sur la compréhension des parties des noeuds voisins qui contribuent à la représentation d'un noeud. Pour atténuer ce problème, ils proposent une méthode statistique pour interpréter les représentations apprises. De nombreux travaux existants, conçus pour les modèles de présentation de noeuds supervisés, calculent la différence entre les scores de prédiction après avoir perturbé les arrêtes d'un noeud d'explication candidat ; cependant, leur cadre proposé exploite un test statistique basé sur la prédiction conforme (CP) pour vérifier l'importance du noeud candidat dans chaque représentation de noeud. Lors de leur évaluation, le cadre proposé a été vérifié dans de nombreux contextes expérimentaux et a présenté des résultats prometteurs par rapport à ceux des méthodes de référence récentes.

[Bonet et al., 2022] proposent une nouvelle méthode explicative basée sur la sélection de noeuds kernelisée tenant compte de la topologie des graphes et ils appliquent leur méthode à des problématique liées à la pollution atmosphérique. Grâce à leur modèle , ils sont capables de capturer efficacement la topologie des graphe et, pour certain noeud du graphe, d'en déduire ses noeuds les plus pertinents. De plus, ils proposent un nouveau noeud topologiquement intégré pour chaque noeud, capturant sous une forme vectorielle les parcours du graphe par rapport à tous les autres noeuds du graphe.

[Lin et al., 2022b] font le constat suivant : les tâches de manipulation, comme charger un lave-vaisselle, peuvent être vues comme une séquence de contraintes spatiales et de relations entre différents objets. Leur objectif est de découvrir ces règles à partir de démonstrations en posant la manipulation comme un problème de classification sur un graphe, dont les noeuds représentent des entités pertinentes pour la tâche comme des objets et des objectifs, et de présenter une architecture des réseaux de neurones sur graphe pour résoudre ce problème à partir de démonstrations. Dans

leurs expériences, un seul réseau de neurone sur graphes formé à l'aide de l'apprentissage par imitation (IL) sur 20 démonstrations d'experts peut résoudre les tâches d'empilement de blocs, de réarrangement et de chargement du lave-vaisselle. Une fois que le modèle prédictif a appris la structure spatiale, elle peut se généraliser à un plus grand nombre d'objets, de configurations d'objectifs et passer de la simulation au monde réel. Ces expériences montrent que l'IL basé sur les graphes peut résoudre des problèmes complexes de manipulation à long terme sans nécessiter de descriptions détaillées des tâches.

[Teufel et al., 2022] proposent un nouveau réseau de neurones sur graphe basé sur l'attention avec multi-explications (MEGAN). Leur modèle comporte plusieurs canaux d'explication, qui peuvent être choisis indépendamment des spécifications de la tâche. Ils valident leur modèle sur un ensemble de données synthétiques pour la régression, où leur modèle produit des explications monocanal avec une qualité similaire à GNNExplainer. De plus, ils démontrent les avantages des explications multicanaux sur un ensemble de données synthétiques et deux ensembles de données réelles : la prédiction de la solubilité dans l'eau des graphes moléculaires et la classification des sentiments des critiques de films. Ils constatent que leur modèle produit des explications conformes à l'intuition humaine.

[Zhang, 2022] font le constat suivant : la majorité des algorithmes de récupération d'éléments dans les systèmes de recommandation structurés typiques de type « récupération-rang-reclassement » peuvent être séparés en trois catégories : les recommandations latents profonds, séquentiels et basés sur des graphes, qui collectent respectivement des signaux de filtrage collaboratif, séquentiels et homogènes. Cependant, il existe un chevauchement conceptuel entre les recommandations séquentielles et graphes sur les éléments ayant déjà interagi avec un utilisateur. Cela déclenche l'idée selon laquelle les signaux séquentiels, de filtrage collaboratif et homogènes peuvent être inclus dans une structure de données formatée par un graphe temporel, et les algorithmes d'apprentissage séquentiel et sur graphe, peuvent être résumés comme un encodeur de graphe temporel. Dans cet contribution, ils proposent Temporal-Graph-Plugin, un encodeur de graphe temporel explicable pour compléter les algorithmes latents profonds avec un message de voisinage temporel agrégé de k-hop via un module d'attention locale.

[Wu et al., 2022a] ils proposent une nouvelle méthode explicative permettant d'obtenir correspondance entre sous graphe et sous graphe explicatifs, nommé MatchExplainer. MatchExplainer couple le graphe cible avec d'autres instances homologues et identifie la sous-structure commune la plus cruciale en minimisant la distance basée sur les noeuds correspondants entre elles. Après cela, un classement de graphes externe est suivi pour sélectionner la sous-structure la plus informative parmi tous les sous-graphes candidats. Ainsi, MatchExplainer est entièrement non paramétrique. De plus, les méthodes actuelles d'échantillonnage de graphes ou de suppression de noeuds souffrent généralement du problème d'échantillonnage des faux positifs. Pour résoudre ce problème, ils corrigent la partie la plus informative du graphe et effectuent des augmentations de graphe sur le reste de la partie moins informative, ce procédé est appellé MatchDrop.

[Fang et al., 2022] étudie le rôle de la régularisation dans l'explicabilité des réseaux de neurones sur graphe du point de vue de la théorie de l'information. Leurs principales conclusions sont les suivantes : la régularisation recherche essentiellement l'équilibre entre deux phases, son coefficient optimal est proportionnel à la rareté des explications, les méthodes existantes impliquent un effet de régularisation implicite du mécanisme stochastique et ses effets contrefactuelles sur deux phases sont responsables du problème de non-distribution (OOD) dans l'explicabilité post-hoc. Sur la base de ces résultats, ils proposent deux méthodes d'optimisation courantes, qui peuvent renforcer les performances des méthodes explicatives actuelles via des schémas de régularisation adaptative à la parcimonie et résistants à l'OOD.

[Serra and Niepert, 2023] inspirés par le paradigme apprendre à expliquer (L2X), ils proposent L2XGNN, une méthode explicative pour les réseaux de neurones sur graphe explicables qui fournit des explications fidèles dès leurs conceptions. L2XGNN apprend un mécanisme de sélection de sous-graphes explicatifs (motifs) qui sont exclusivement utilisés dans les opérations de transmission de messages des réseaux de neurones sur graphe. L2XGNN est capable de sélectionner, pour chaque graphe, un sous-graphe avec des propriétés spécifiques telles que le fait d'être clairsemé et connecté. Imposer de telles contraintes aux motifs conduit souvent à des explications plus interprétables et plus efficaces.

[Xuanyuan et al., 2023] font le constat suivant : les méthodes d'explicabilité

actuelles sont généralement locales et traitent les réseaux de neurones sur graphe comme des boîtes noires. Ils ne regardent pas à l'intérieur du modèle, ce qui inhibe la confiance humaine dans le modèle et ses explications. Motivés par la capacité des neurones à détecter des concepts sémantiques de haut niveau dans les modèles de vision, ils effectuent une nouvelle analyse du comportement des neurones du réseau de neurones sur graphes pour répondre aux questions sur l'explicabilité du réseau de neurones sur graphe et proposent de nouvelles mesures pour évaluer l'explicabilité des neurones du réseau de neurones sur graphes. Ils proposent une nouvelle méthode pour produire des explications globales des réseaux de neurones sur graphes en utilisant des concepts au niveau des neurones afin de permettre aux chercheurs d'avoir une vue de haut niveau sur le modèle. Plus précisément, il s'agit du premier travail montrant que les neurones réseau de neurones sur graphe agissent comme des détecteurs de concepts et sont fortement alignés sur les concepts formulés sous forme de compositions logiques de degrés de noeuds et de propriétés de voisinage ; ils évaluent quantitativement l'importance des concepts détectés et identifient un compromis entre la durée de la formation et l'explicabilité au niveau des neurones ; ils démontrent que leur méthode d'explicabilité globale présente des avantages par rapport à l'état de l'art actuel : ils peuvent démêler l'explication en concepts individuels interprétables soutenus par des descriptions logiques, ce qui réduit le risque de biais et améliore la compréhension du modèle prédictif.

[Jiang and Li, 2022] font le constat suivant : l'attaque par porte dérobée est un algorithme d'attaque puissant pour le modèle d'apprentissage profond. Récemment, la vulnérabilité de réseaux de neurones sur graphe aux attaques par porte dérobée a été prouvée, notamment dans le cadre de tâches de classification de graphes. Dans cet contribution, ils proposent la première méthode de détection et de défense de porte dérobée sur réseaux de neurones sur graphe. La plupart des attaques par porte dérobée dépendent de l'injection d'un déclencheur petit mais influent dans l'échantillon propre. Pour les données de type graphes, les attaques de porte dérobée actuelles se concentrent sur la manipulation de la structure du graphe pour injecter le déclencheur. Ils constatent qu'il existe des différences apparentes entre les échantillons bénins et les échantillons malveillants dans certaines mesures d'évaluation telles que la fidélité et l'infidélité. Après avoir identifié l'échantillon malveillant,

l'explicabilité du modèle de réseaux de neurones sur graphe peut aider à capturer le sous-graphe le plus significatif qui est probablement le déclencheur dans un graphe de type cheval de Troie.

[Shumovskaia et al., 2022] font le constat suivant : les algorithmes d'apprentissage social fournissent des modèles de formation d'opinions sur les réseaux sociaux résultant de raisonnements locaux et d'échanges entre agents. Les interactions se produisent sur une topologie graphe sous-jacente, qui décrit le flux d'informations entre les agents. Pour tenir compte des conditions de dérive dans l'environnement, ce travail adopte une stratégie d'apprentissage social adaptatif, capable de suivre les variations des statistiques de signaux sous-jacentes. Compte tenu des observations de l'évolution des explications au fil du temps, ils visent à déduire la topologie du graphe sous-jacente, à découvrir les influences par paires entre les agents et à identifier les trajectoires d'informations significatives dans le réseau.

[Shan et al., ] propose une nouvelle méthode explicative pour réseau de neurones sur graphes. Étant donné un modèle de réseau de neurones sur graphe entraîné, une méthode explicative pour réseau de neurones sur graphe vise à identifier un sous-graphe le plus influent pour interpréter la prédiction d'une instance (par exemple, un noeud ou un graphe), qui est essentiellement un problème d'optimisation combinatoire sur graphe. Les travaux existants résolvent ce problème par relaxation continue ou heuristiques basées sur la recherche de singularité. Mais ils souffrent de problèmes clés tels que la violation de la transmission des messages et des heuristiques élaborées à la main, conduisant à une explicabilité de qualité inférieure. Pour résoudre ces problèmes, ils proposent une méthode explicative pour réseau de neurones sur graphe par apprentissage par renforcement , RG-Explainer, qui se compose de trois composants principaux : la sélection du point de départ, la génération itérative de graphes et l'apprentissage des critères d'arrêt. RG-Explainer construit un sous-graphe explicatif connecté en ajoutant séquentiellement des noeuds à partir de la limite du graphe actuellement généré, ce qui est cohérent avec le schéma de transmission de messages. De plus, ils conçoivent un localisateur de graines d'échantillonnage efficace pour sélectionner le point de départ et apprenent les critères d'arrêt pour générer des explications supérieures.

[Wang et al., b] font le constat suivant : lorsqu'un réseau de neurones sur graphe

fait une prédiction, on peut se poser la question suivante : « Quelle fraction du graphe d'entrée a la plus d'influence sur la décision du modèle ? Produire une réponse nécessite de comprendre le fonctionnement interne du modèle en général et de mettre l'accent sur les informations sur la décision pour l'instance concernée. Néanmoins, la plupart des méthodes actuelles se concentrent uniquement sur un seul aspect : l'explicabilité locale, qui explique chaque instance indépendamment, ne présente donc guère les modèles de classe ; et l'explicabilité globale, qui caractérise les modèles dans leur ensemble. Cette dichotomie limite considérablement la flexibilité et l'efficacité des une méthode explicatives. Un paradigme performant vers une explicabilité multi-grains fait jusqu'à présent défaut et constitue donc un objectif de leur travail. Dans ce travail, ils utilisent l'idée des explications multi-grains, via la phase de préformation et de mise au point. Plus précisément, la phase de pré-formation prend en compte la contrastivité entre les différentes classes, de manière à mettre en évidence les caractéristiques de chaque classe d'un point de vue global ; ensuite, la phase de mise au point adapte les explications au contexte local.

[Yu et al., 2020] font le constat suivant : Compte tenu du graphe et de son étiquette/propriété, plusieurs problèmes clés de l'apprentissage des graphes, tels que la recherche de sous-graphes interprétables. Ces sous-graphes doivent être aussi informatifs que possible, tout en contenant des structures les moins redondantes et les moins bruyantes. Cette problématique est étroitement liée au principe bien connu du goulot d'étranglement de l'information (IB), qui a cependant moins été étudié pour les données de type graphes et pour les réseaux de neurones sur graphes. Dans cet contribution, ils proposent une nouvelle méthode Graph Information Bottleneck (GIB) permettant de résoudre le problème de reconnaissance de sous-graphes sous l'angle du goulot d'étranglement. Cela permet de reconnaître le sous-graphe à la fois informatif et compressif, nommé sous-graphe IB. Cependant, l'objectif GIB est difficile à optimiser, principalement en raison de la difficulté à optimiser l'information mutuelle. Pour cela, ils proposent : un objectif GIB basé sur un estimateur d'information mutuelle pour les données de type graphes ; un schéma d'optimisation à deux niveaux pour maximiser l'objectif GIB ; une perte de connectivité pour stabiliser le processus d'optimisation.

[Faber et al., 2021] propose des nouvelles métriques d'évaluation des méthodes

explicatives spécifiques aux réseaux de neurones sur graphes.

[Jaume et al., 2021] développe une nouvelle méthode explicative pour les réseaux de neurones sur graphes , adapté à la biologie et plus particulièrement au praticien s'intéressant aux maladies humaines. Dans ce travail, ils arrêtent ce problème en adoptant un traitement de graphes basé sur des entités biologiques et des une méthode explicatives de graphes permettant des explications accessibles aux pathologistes. Dans ce contexte, un défi majeur consiste à discerner des explications significatives, en particulier de manière standardisée et quantifiable. À cette fin, ils proposent un ensemble de nouvelles métriques quantitatives basées sur des statistiques de séparabilité des classes utilisant des concepts pathologiquement mesurables pour caractériser les une méthode explicatives de graphes. Ils utilisent les métriques proposées pour évaluer trois types d'une méthode explicatives de graphes, à savoir les méthodes de propagation de pertinence par couche, de saillance basée sur le gradient et d'élagage de graphes, afin d'expliquer les représentations de graphes cellulaires pour les sous-types du cancer du sein. Les métriques proposées sont également applicables dans d'autres domaines en utilisant des concepts spécifiques au domaine.

[Wu et al., 2021a] font le constat suivant : fournir une explication fiable du diagnostique clinique basée sur le dossier médical électronique (DME) est fondamental pour l'application de l'intelligence artificielle dans le domaine médical. Les méthodes actuelles traitent principalement le DME comme une séquence de texte et fournissent des explications basées sur une base de connaissances médicales précises, spécifiques à la maladie et difficiles à obtenir pour les experts dans la réalité. Il propose dans cette contribution une méthode d'extraction de graphe multi-granularités contrefactuels (CMGE) pour extraire les information du DME sans bases de connaissances externes. Plus précisément, ils structurent la séquence d'EMR dans un réseau de neurones sur graphes hiérarchiques, puis obtiennent la relation causale entre les caractéristiques multi-granularités et les résultats du diagnostic grâce à une intervention contrefactuelle sur le graphe. Les caractéristiques ayant le lien causal le plus fort avec les résultats fournissent un support interprétatif au diagnostique.

[Schnake et al., 2022] propose une nouvelle méthode explicative et montre que les réseaux de neurones sur graphe peuvent en fait être naturellement expliqués en

utilisant des développements d'ordre supérieur, c'est-à-dire en identifiant des groupes d'arêtes qui contribuent conjointement à la prédiction. En pratique, ils constatent que de telles explications peuvent être extraites à l'aide d'un schéma d'attribution imbriqué, dans lequel des méthodes existantes telles que la propagation de pertinence par couche (LRP) peuvent être appliquées à chaque étape. La sortie est une collection de parcours dans le graphe d'entrée pertinents pour la prédiction. Leur nouvelle méthode d'explication Graph-LRP, est applicable à un large éventail de réseaux de neurones sur graphes et permet d'extraire des informations pratiquement pertinentes sur l'analyse des sentiments des données textuelles, les relations structure-propriété en chimie quantique et la classification des images.

[Chereda et al., 2021] ont étendu la méthode LRP de manière à la rendre disponible pour les modèles prédictifs de type Graph-CNN. Ils ont testé son applicabilité sur un vaste ensemble de données sur le cancer du sein. Il s'agit d'une nouvelle méthode pour expliquer les décisions prises par les Graph-CNN. Ils démontrent une vérification de l'intégrité du GLRP développé sur un ensemble de données de chiffres manuscrits, puis applique la méthode aux données d'expression génique. Ils montrent que le GLRP fournit des sous-réseaux moléculaires spécifiques au patient qui concordent largement avec les connaissances cliniques et identifient les facteurs communs ainsi que nouveaux et potentiellement médicamenteux de la progression tumorale.

[Jiménez-Luna et al., 2021] propose une étude visant à améliorer la transparence de la modélisation pour la conception moléculaire rationnelle en appliquant la méthode GradCAM pour les réseaux de neurones sur graphes. Des modèles prédicitifs ont été formés pour prédire la liaison aux protéines plasmatiques, l'inhibition du canal hERG, la perméabilité passive et l'inhibition du cytochrome P450. La méthodologie proposée a mis en évidence les caractéristiques moléculaires et les éléments structurels qui sont en accord avec les motifs pharmacophores connus ; a correctement identifié leurs propriétés et a fournir un aperçu des interactions non spécifiques.

[Xu et al., 2021a] proposent APRILE, une nouvelle méthode explicative pour réseaux de neurones sur graphes permettant d'explorer les mécanismes moléculaires sous-jacents aux effets secondaires de la polypharmacie. À partir d'une liste d'effets secondaires et des paires de médicaments qui les provoquent, APRILE identifie

un ensemble de protéines (cibles ou non des médicaments) et les termes d'ontologie génétique (GO) associés et les effets secondaires associés. Grâce à APRILE, ils ont généré des explications pour 843 318 (appris) et 93 966 (nouveaux) événements associés à des effets secondaires et à des médicaments, couvrant 861 effets secondaires (472 maladies, 485 symptômes et 9 troubles mentaux) et 20 catégories de maladies. Ils montrent que leurs deux nouvelles mesures - l'utilisation des informations pharmacogénomiques et l'utilisation des informations sur les interactions protéine-protéine - fournissent des estimations quantitatives de la complexité des mécanismes moléculaires. Les explications étaient significativement cohérentes avec les associations maladie-gène pour 232/239 (97 %) effets secondaires. En outre, APRILE a généré de nouvelles connaissances sur les mécanismes moléculaires de quatre catégories diverses d'effets indésirables des médicaments : les infections, les maladies métaboliques, les maladies gastro-intestinales et les troubles mentaux, y compris les effets secondaires paradoxaux. Ils démontrent la viabilité de la méthode par la découverte effective de mécanismes d'effets secondaires de la polypharmacie.

[Kasanishi et al., 2021] propose une adaptation des méthodes explicatives pour réseaux de neurones sur graphe déjà existantes pour les problématiques de classification d'arêtes.

[Gui et al., 2023] font le constat suivant : alors que la plupart des méthodes actuelles se concentrent sur l'explication des noeuds, des arêtes ou des caractéristiques des graphes, ils soutiennent que, en tant que mécanisme fonctionnel inhérent aux réseaux de neurones sur graphes, les flux de messages sont plus naturels pour effectuer l'explicabilité. À cette fin, ils proposent ici une nouvelle méthode explicative, FlowX, permettant d'expliquer les réseaux de neurones sur graphe en identifiant les flux de messages importants. Pour quantifier l'importance des flux, ils proposent de suivre la philosophie des valeurs de Shapley issue de la théorie des jeux coopératifs. Pour surmonter la complexité du calcul des contributions marginales de toutes les coalitions, ils proposent un schéma d'approximation pour calculer des valeurs de Shapley. Ils proposent ensuite un algorithme d'apprentissage pour entraîner les scores de flux et améliorer l'explicabilité.

[Cho et al., 2021] proposent une nouvelle méthode explicative pour améliorer la qualité d'explication des tâches de classification de noeuds qui peut être appliquée

de manière post hoc grâce à l'agrégation d'explications auxiliaires provenant de noeuds voisins importants, nommée SEEN. SEEN ne nécessite pas de modification du graphe et peut être utilisée dans divers contexte applicatifs. Les expériences sur la correspondance des noeuds participant à des motifs importants à partir d'un graphe donné montrent une grande amélioration de la précision des explications, jusqu'à 12,71 %, et démontrent la corrélation entre les explications auxiliaires et la précision améliorée des explications grâce à l'exploitation de leurs liens.

[Sun et al., 2021d] proposent une nouvelle méthode explicative polyvalente, permettant d'acquérir un masque qui indique les perturbations topologiques importantes des graphes. Ils proposent aussi un système de visualisation interactif (GNNViz) qui peut remplir plusieurs objectifs : préserver, promouvoir ou attaquer les prédictions de réseaux de neurones sur graphe.

[Rathee et al., 2021] proposent une nouvelle méthode explicative appelée Kedge pour fragmenter explicitement le graphe en supprimant les noeuds inutiles. Ils se sont basés sur une méthode de parciminification utilisant la distribution Hard Kumaraswamy qui peut être utilisée avec n'importe quels réseaux de neurones sur graphe. Kedge apprend les masques de arrête de manière modulaire via une optimisation par gradient. Enfin, ils montrent aussi que Kedge contrecarre efficacement les phénomènes de lissage excessif présent dans les réseaux de neurones sur graphe profonds en maintenant de bonnes performances sur les tâches d'apprentissage avec l'augmentation des couches des réseaux de neurones sur graphe.

[Kazi et al., 2021] s'intéressent a l'explicabilité des réseaux de neurones sur graphes dans le domaine médical. Dans ce contexte, ils proposent une nouvelle méthode explicative qui permet d'obtenir la pertinence clinique des caractéristiques d'entrée pour la tâche ; utilise l'explication pour améliorer les performances du modèle prédictif et apprendre un nouveau graphe qui peut être utilisé pour interpréter le comportement de la cohorte de patients. Dans un scénario clinique, un tel modèle peut aider les experts cliniques à prendre de meilleures décisions en matière de diagnostic et de planification du traitement. La principale nouveauté réside dans le module d'attention interprétable (IAM), qui opère directement sur des descripteurs multimodales. Leur IAM apprend l'attention portée à chaque descripteur en fonction des pertes uniques spécifiques à l'explicabilité.

[Ghorbani et al., ] s'intéressent à la classification des noeuds et des graphes via les réseaux de neurones sur graphe. La méthode s'appelle GCExplainer. GCExplainer est une méthode non supervisée pour la découverte et l'extraction post-hoc d'explications globales basées sur des concepts pour les réseaux de neurones sur graphe, qui met l'humain dans la boucle pour le calcul de l'explication.

[Magister et al., 2021] étendent le domaine d'applicabilité de la méthode GCExplainer.

[Loveland et al., 2021] propose une analyse théorique des méthodes explicatives pour réseaux de neurones sur graphes appliquée à la chimie.

[BK et al., 2021] proposent une nouvelle méthode d'évaluation automatique pour les explications fournies par les méthodes explicatives pour les réseaux de neurones sur graphe.

[Rao et al., 2021a] s'intéresse à la vérité terrain concernant l'évaluation des méthodes explicatives pour les réseaux de neurones sur graphes. En effet, ils indiquent que celles-ci reposent en fin de compte sur des jugements subjectifs des humains, de sorte que la qualité de l'interprétation du modèle est difficile à évaluer objectivement et quantitativement. Dans cette contribution, ils construisent trois niveaux d'ensembles de données de référence pour évaluer quantitativement les méthodes explicatives de l'état de l'art pour les modèles de réseaux de neurones sur graphe.

[Chan et al., 2022] font le constat suivant : l'augmentation de l'utilisation des modèles linguistiques pré-entraînés avec des graphes de connaissances (KG) a permis d'améliorer le traitement de diverses tâches d'apprentissage. Cependant, pour une instance de tâche donnée, le KG, ou certaines parties du KG, peuvent ne pas être utiles. Bien que les modèles augmentés par KG utilisent souvent l'attention pour se concentrer sur des composants KG spécifiques, le KG est toujours utilisé et le mécanisme d'attention n'est jamais explicitement enseigné quels composants KG doivent être utilisés. Parallèlement, les méthodes de saillance peuvent mesurer dans quelle mesure un descripteur KG (par exemple, graphe, noeud, chemin) influence le modèle pour effectuer la prédiction correcte, expliquant ainsi quelles descripteurs KG sont utiles. Cet contribution explore comment les explications par saillance peuvent être utilisées pour améliorer les performances des modèles augmentés par KG. Ils proposent de créer des explications par saillance grossières, (c-à-d le KG est-il utile ?) et

fines (c-à-d Quels noeuds/chemins dans le KG sont utiles ?). Deuxièmement, pour motiver la supervision basée sur la saillance, ils analysent les modèles augmentés par Oracle KG qui utilisent directement les explications de saillance comme entrées supplémentaires pour guider leur attention. Troisièmement, ils proposent SalKG, une méthode permettant aux modèles augmentés par KG d'apprendre à partir d'explications de saillance grossière et/ou fine. Compte tenu des explications de saillance créées à partir de l'ensemble d'entraînement d'une tâche, SalKG entraîne conjointement le modèle pour prédire les explications, puis résout la tâche en s'occupant des caractéristiques de KG mises en évidence par les explications prédites.

[Gao et al., 2021b] ont fait le constat suivant : de nombreuses méthodes sont proposées pour expliquer les prédictions des réseaux de neurones sur graphes, en se concentrant sur « comment générer des explications ». Cependant, les questions de recherche telles que « si les explications du réseau de neurones sur graphe sont inexactes », « que se passe-t-il si les explications sont inexactes » et « comment ajuster le modèle pour générer des explications plus précises » n'ont pas été suffisamment explorées. Pour répondre aux questions ci-dessus, cette contribution propose une méthode explicative supervisée pour les réseau de neurones sur graphe (GNES) pour apprendre de manière adaptative à expliquer plus correctement les réseaux de neurones sur graphes. Plus précisément, la méthode proposée optimise conjointement à la fois la prédiction du modèle et l'explication du modèle en appliquant à la fois la régularisation sur le graphe entier et une faible supervision des explications du modèle. Pour la régularisation des graphes, ils proposent une formulation d'explication unifiée pour les explications au niveau des noeuds et au niveau des arrêtes en renforçant la cohérence entre elles. Les méthodes explicatives au niveau des noeuds et des arrêtes qu'ils proposent sont également génériques.

[Ma et al., 2021b] font le constat suivant : les données relationnelles sont universelles dans la société d'aujourd'hui, comme les réseaux sociaux, les réseaux de citations et les molécules. Ils peuvent être transformés en données de type graphes constituées de noeuds et d'arêtes. Par exemple, la police antidrogue s'appuie sur les données sociales pour identifier les toxicomanes cachés. Cependant, lors du déploiement des réseaux de neurones sur graphes, la prise en charge par les policiers nécessite toujours des méthodes explicatives. Parce que les réseaux de neurones sur

graphe intègrent les informations des noeuds et les transmettent grâce à la matrice d'adjacence. Cependant, les méthodes explicatives courantes utilisées dans d'autres modèles ne sont pas directement applicables aux réseaux de neurone sur graphes. Et ces méthodes d'explication dans les CNN ne sont pas non plus applicables compte tenu de la variété topologique du graphe. Afin d'expliquer leur modèle de réseau de neurones sur graphe sur les réseaux sociaux liés à l'abus de drogues, ils ont proposé une méthode pour prétraiter les données de type graphes, entraîner le modèle de réseau de neurones sur graphe et appliquer leur méthode d'information mutuelle maximale (MMI) pour aider à expliquer la prédiction faite par le réseau de neurones sur graphes. Ils évaluent les performances du modèle d'explication avec la cohérence et la contrastivité des métriques d'évaluation. Le modèle d'explication fonctionne de manière impressionnante à la fois dans les données du monde réel et dans deux données synthétiques par rapport à d'autres références. Les résultats montrent également que leur modèle peut imiter l'explication réseau de neurones sur graphe faite par l'Homme dans la classification des noeuds.

[Himmelhuber et al., ] s'intérésse aux biais de confirmation dans le cadre des méthodes explicatives pour réseaux de neurones sur graphe. En effet il indique que un certain nombre de méthodes, telles que la génération de masques d'importance, ont été développées pour fournir un aperçu du processus de prise de décision de ces réseaux de neurones sur graphes. Il s'agit de premières étapes importantes sur la voie de la modélisation de l'explicabilité, mais laisser l'interprétation de ces explications aux analystes humains peut s'avérer problématique puisque les humains s'appuient naturellement sur leurs connaissances de base et donc aussi sur leurs préjugés sur les données et leur domaine. Pour surmonter ce problème, ils introduisent une méthode conceptuelle en suggérant l'extraction de règles d'explication au niveau du modèle via une méthode d'apprentissage standard en boîte blanche à partir des masques d'importance générés.

[Holzinger et al., 2021] indiquent que les réseaux de neurones sont des modèles prédictifs intéressants pour les problématiques de fusions de données et notamment les données multimodales. Ils effectuent une étude pour les réseaux de neurones sur graphes.

[Peach et al., 2020] font le constat suivant : les réseaux sont largement utilisés

comme modèles mathématiques de systèmes complexes dans de nombreuses disciplines scientifiques, non seulement en biologie et en médecine, mais également en sciences sociales, en physique, en informatique et en ingénierie. Des décennies de travaux ont produit un vaste corpus de recherches caractérisant les propriétés topologiques, combinatoires, statistiques et spectrales des graphes. Chaque propriété de graphe peut être considérée comme un descripteur qui capture les caractéristiques importantes (et parfois se chevauchant) d'un réseau. Dans l'analyse de graphes du monde réel, il est crucial d'intégrer systématiquement un grand nombre de caractéristiques graphes diverses afin de caractériser et de classer les réseaux, ainsi que de faciliter la découverte scientifique basée sur les réseaux. Ils présentent HCGA, une nouvelle méthode explicative permettant l'analyse comparative d'ensembles de données de type graphe qui calcule plusieurs milliers de caractéristiques de graphe à partir d'un modèle prédictif donné. HCGA propose également une suite d'outils d'apprentissage statistique et d'analyse de données pour l'identification et la sélection automatisées de caractéristiques importantes et interprétables qui sous-tendent la caractérisation des ensembles de données de type graphe.

[Faber et al., 2020] propose une nouvelle méthode explicative Distribution Compliant Explanation (DCE). Cette contribution présente une nouvelle méthode explicative contrastive pour réseaux de neurones sur graphe (CoGE) .

[Jaume et al., 2020] L'explicabilité des méthodes d'apprentissage automatique en pathologie numérique (DP) est d'une grande importance pour faciliter leur large adoption dans les cliniques. Récemment, des réseaux de neurones sur graphes codant des entités biologiques pertinentes ont été utilisées pour représenter et évaluer les images biologiques. Un tel changement de paradigme, passant d'une analyse par pixel à une analyse par entité, offre davantage de contrôle sur la représentation des concepts. Dans cet contribution, ils introduisent un une méthode explicative post-hoc pour dériver des explications compactes par instance mettant l'accent sur les entités importantes dans le graphe. Bien que ils se concentrent sur leurs analyses de leurs cellules et les interactions cellulaires dans le cancer du sein, la méthode explicative proposée est suffisamment générique pour être étendue à d'autres contextes applicatifs. Les analyses qualitatives et quantitatives démontrent l'efficacité de la méthode explicative pour générer des explications complètes et compactes.

[Saueressig et al., 2021] utilise les réseaux de neurones sur graphe pour la détection automatique de tumeur cérébrale.

[Lin et al., 2020a] proposent GISST. GIIST combine un mécanisme d'attention et une régularisation de parcimonie pour produire un sous-ensemble important de descripteurs de sous-graphe et de noeud lié à toute tâche basée sur un graphe. Grâce à une seule couche d'auto-attention, un modèle GISST apprend une probabilité d'importance pour chaque caractéristique de noeud et chaque arête dans le graphe. En incluant ces probabilités d'importance dans la fonction de perte du modèle, les probabilités sont optimisées de bout en bout et liées aux performances spécifiques à la tâche. De plus, GISST répartit ces probabilités d'importance avec l'entropie et la régularisation $l_1$ pour réduire le bruit dans la topologie du graphe et les caractéristiques des noeuds. leurs modèles GISST atteignent une précision supérieure en termes de caractéristiques de noeuds et d'explication des arrêtes dans les ensembles de données synthétiques, par rapport aux méthodes d'interprétation de l'état de l'art. De plus, leurs modèles GISST sont capables d'identifier une structure graphe importante dans des ensembles de données du monde réel. Ils démontrent en théorie que l'importance des caractéristiques des arrêtes et plusieurs types de arrêtes peuvent être pris en compte en les incorporant dans le calcul de probabilité des arrêtes du GISST. En prenant en compte conjointement la topologie, les descripteurs de noeud et les descripteurs de périphérie, GISST fournit intrinsèquement des interprétations simples et pertinentes pour tous les modèles et tâches.

[Wang et al., 2020e] propose une méthode explicative basée sur l'explicabilité causale dans les réseaux de neurones sur graphes, elle s'appelle Causal Screening. Il sélectionne progressivement une caractéristique d'arrête du graphe avec une attribution causale importante, qui est formulée comme l'effet causal individuel sur la prédiction du modèle. En tant que méthode agnostique au modèle, Causal Screening peut être utilisé pour générer des explications fidèles et concises pour n'importe quel modèle de réseau de neurone sur graphes.

[Hu et al., 2020a] proposent une nouvelle méthode d'explication basée sur la propagation de pertinence par couche pour les réseaux de neurones sur graphes convolutifs, à savoir GCN-LRP. Cette méthode est basée sur la méthode de [Binder et al., 2016]

[Wellawatte et al., 2021] proposent une nouvelle méthode explicative basée sur la méthode des données contrefactuelles. Leur méthode est basée sur la méthode « Superfast Traversal, Optimization, Novelty, Exploration and Discovery » [Nigam et al., 2021] qui est efficace et indépendant du modèle pour générer des données contrefactuelles.

[Schlichtkrull et al., 2020] proposent une nouvelle méthode explicative post-hoc pour interpréter les prédictions des réseau de neurones sur graphe qui identifie les arêtes inutiles. Étant donné un modèle de réseau de neurones sur graphe entraîné, ils apprennent un classificateur simple qui, pour chaque arête de chaque couche, prédit si cette arête peut être supprimée.

[Funke et al., 2020] se concentre sur l'explication ou l'interprétation de la logique sous-jacente à une prédiction donnée de réseaux neuronaux sur graphe pour la tâche de classification des noeuds. Les méthodes existantes pour interpréter les réseaux de neurones sur graphe tentent de trouver des sous-ensembles de descripteurs et de noeuds importants en apprenant un masque continu. Leur objectif est de trouver des masques discrets qui sont sans doute plus interprétables tout en minimisant l'écart attendu par rapport à la prédiction du modèle sous-jacent.

[Huang et al., 2020c] propose GraphLIME, une méthode explicative locale pour les réseaux de neurones sur graphes utilisant le lasso du critère d'indépendance de Hilbert-Schmidt (HSIC), qui est une méthode de sélection de caractéristiques non linéaire. GraphLIME est une méthode explicative générique de modèle de réseaux de neurones sur graphe qui apprend un modèle interprétable non linéaire localement dans le sous-graphe du noeud expliqué. Plus précisément, pour expliquer un noeud, ils génèrent un modèle interprétable non linéaire à partir de son voisinage $L$-hop, puis calculent les K caractéristiques les plus représentatives comme explications de sa prédiction à l'aide du HSIC Lasso.

[Liu et al., 2020c] font le constat suivant : les graphes de connaissances biomédicales permettent une méthode informatique du raisonnement sur les systèmes biologiques. La nature des données biologiques conduit à une structure graphe qui diffère de celles généralement rencontrées dans les ensembles de données d'analyse comparative. Pour comprendre les implications que cela peut avoir sur les performances des algorithmes de raisonnement, ils mènent une étude empirique basée sur

la tâche réelle de réutilisation de médicaments. Ils formulent cette tâche comme un problème de prédiction de liens où les composés et les maladies correspondent à des entités dans un graphe de connaissances. Pour surmonter les faiblesses apparentes des algorithmes existants, ils proposent une nouvelle méthode, PoLo, qui combine des parcours guidés par des politiques basées sur l'apprentissage par renforcement et des règles logiques. Ces règles sont intégrées à l'algorithme en utilisant une nouvelle fonction de récompense. Ils appliquent leur méthode à Hetionet, qui intègre des informations biomédicales provenant de 29 bases de données bioinformatiques de premier plan. leurs expériences montrent que leur méthode surpasse plusieurs méthodes pour la prédiction de liens tout en offrant une explicabilité.

### B.2.5 Bibliothèques logicielles

— pour le calcul d'explication : [Geometric, 2023, Liu et al., 2021b, Holdijk et al., 2021, Fey and Lenssen, 2019]

— pour la visualisation d'explication : [Ott et al., 2022]

# Annexe C

# Détails de certains calculs

## C.1 Décomposition de l'erreur quadratique moyenne selon le biais et la variance du modèle prédictif

Pour alléger les notations, nous noterons dans la suite simplement le nom des variables aléatoires en indice des espérances. Il faut alors développer la quantité suivante :

$$\mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{Y}[(f(X)-Y)^2|X]]]$$

On a donc :

$$\mathop{\mathbb{E}}_{Y}[(f(X)-Y)^2|X] = \mathop{\mathbb{E}}_{Y}[((f(X)-\mathop{\mathbb{E}}_{D}[f(X)])+(\mathop{\mathbb{E}}_{D}[f(X)]-Y))^2|X] \tag{C.1}$$

$$= \mathop{\mathbb{E}}_{Y}[(f(X)-\mathop{\mathbb{E}}_{D}[f(X)])^2|X] + \mathop{\mathbb{E}}_{Y}[(\mathop{\mathbb{E}}_{D}[f(X)]-Y)^2|X]$$

$$+ 2\mathop{\mathbb{E}}_{Y}[(f(X)-\mathop{\mathbb{E}}_{D}[f(X)])(\mathop{\mathbb{E}}_{D}[f(X)]-Y)|X] \tag{C.2}$$

or

$$\mathop{\mathbb{E}}_{Y}[(f(X)-\mathop{\mathbb{E}}_{D}[f(X)])(\mathop{\mathbb{E}}_{D}[f(X)]-Y)|X] = \mathop{\mathbb{E}}_{Y}[f(X)\mathop{\mathbb{E}}_{D}[f(X)]|X] - \mathop{\mathbb{E}}_{Y}[f(X)Y|X] - \mathop{\mathbb{E}}_{Y}[(\mathop{\mathbb{E}}_{D}[f(X)])^2|X]$$

$$+ \mathop{\mathbb{E}}_{Y}[\mathop{\mathbb{E}}_{D}[f(X)]Y|X] \tag{C.3}$$

$$= f(X)\mathop{\mathbb{E}}_{D}[f(X)] - f(X)\mathop{\mathbb{E}}_{Y}[Y|X] - (\mathop{\mathbb{E}}_{D}[f(X)])^2$$

$$+ \mathop{\mathbb{E}}_{D}[f(X)]\mathop{\mathbb{E}}_{Y}[Y|X] \tag{C.4}$$

ce qui implique que :

$$\mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[f(X)\mathop{\mathbb{E}}_{D}[f(X)] - f(X)\mathop{\mathbb{E}}_{Y}[Y|X] - (\mathop{\mathbb{E}}_{D}[f(X)])^2 + \mathop{\mathbb{E}}_{D}[f(X)]\mathop{\mathbb{E}}_{Y}[Y|X]]] \tag{C.5}$$

$$= \mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{D}[f(X)\mathop{\mathbb{E}}_{D}[f(X)] - f(X)\mathop{\mathbb{E}}_{Y}[Y|X] - (\mathop{\mathbb{E}}_{D}[f(X)])^2 + \mathop{\mathbb{E}}_{D}[f(X)]\mathop{\mathbb{E}}_{Y}[Y|X]]] \tag{C.6}$$

$$= \mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{D}[f(X)\mathop{\mathbb{E}}_{D}[f(X)]] - \mathop{\mathbb{E}}_{D}[f(X)\mathop{\mathbb{E}}_{Y}[Y|X]] - \mathop{\mathbb{E}}_{D}[(\mathop{\mathbb{E}}_{D}[f(X)])^2] + \mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{D}[f(X)]\mathop{\mathbb{E}}_{Y}[Y|X]]] \tag{C.7}$$

$$= \mathop{\mathbb{E}}_{X}[(\mathop{\mathbb{E}}_{D}[f(X)])^2 - \mathop{\mathbb{E}}_{D}[f(X)\mathop{\mathbb{E}}_{Y}[Y|X]] - \mathop{\mathbb{E}}_{D}[(\mathop{\mathbb{E}}_{D}[f(X)])^2] + \mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{D}[f(X)]\mathop{\mathbb{E}}_{Y}[Y|X]]] \tag{C.8}$$

$$= \mathop{\mathbb{E}}_{X}[(\mathop{\mathbb{E}}_{D}[f(X)])^2 - \mathop{\mathbb{E}}_{Y}[Y|X]\mathop{\mathbb{E}}_{D}[f(X)] - (\mathop{\mathbb{E}}_{D}[f(X)])^2 + \mathop{\mathbb{E}}_{Y}[Y|X]\mathop{\mathbb{E}}_{D}[f(X)]]] \tag{C.9}$$

$$= 0 \tag{C.10}$$

d'où :

$$\mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{Y}[(f(X) - Y)^2|X]]] = \mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{Y}[(f(X) - \mathop{\mathbb{E}}_{D}[f(X)])^2|X] + \mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{Y}[(\mathop{\mathbb{E}}_{D}[f(X)] - Y)^2|X]]] \tag{C.11}$$

$$= \mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[(f(X) - \mathop{\mathbb{E}}_{D}[f(X)])^2|X]] + \mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{Y}[(\mathop{\mathbb{E}}_{D}[f(X)] - Y)^2|X]]] \tag{C.12}$$

cependant :

$$\mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{Y}[(\mathop{\mathbb{E}}_{D}[f(X)] - Y)^2|X]]] = \mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{Y}[(\mathop{\mathbb{E}}_{D}[f(X)] - \mathop{\mathbb{E}}_{Y}[Y|X] + \mathop{\mathbb{E}}_{Y}[Y|X] - Y)^2|X]]] \tag{C.13}$$

$$\begin{aligned} &= \mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{Y}[(\mathop{\mathbb{E}}_{D}[f(X)] - \mathop{\mathbb{E}}_{Y}[Y|X])^2 + (\mathop{\mathbb{E}}_{Y}[Y|X] - Y)^2 \\ &\qquad + 2(\mathop{\mathbb{E}}_{D}[f(X)] - \mathop{\mathbb{E}}_{Y}[Y|X])(\mathop{\mathbb{E}}_{Y}[Y|X] - Y)|X]]] \end{aligned} \tag{C.14}$$

$$\begin{aligned} &= \mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{Y}[(\mathop{\mathbb{E}}_{D}[f(X)] - \mathop{\mathbb{E}}_{Y}[Y|X])^2|X]]] + \mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{Y}[(\mathop{\mathbb{E}}_{Y}[Y|X] - Y)^2|X]] \\ &\qquad + 2\mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{Y}[(\mathop{\mathbb{E}}_{D}[f(X)] - \mathop{\mathbb{E}}_{Y}[Y|X])(\mathop{\mathbb{E}}_{Y}[Y|X] - Y)|X]]] \end{aligned} \tag{C.15}$$

or :

$$\mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{Y}[(\mathop{\mathbb{E}}_{D}[f(X)] - \mathop{\mathbb{E}}_{Y}[Y|X])(\mathop{\mathbb{E}}_{Y}[Y|X] - Y)|X]]] = \mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{Y}[\mathop{\mathbb{E}}_{D}[f(X)]\mathop{\mathbb{E}}_{Y}[Y|X] - Y\mathop{\mathbb{E}}_{D}[f(X)]$$

$$- (\mathop{\mathbb{E}}_{Y}[Y|X])^2 + Y\mathop{\mathbb{E}}_{Y}[Y|X]]|X]] \tag{C.16}$$

$$= \mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{D}[f(X)]\mathop{\mathbb{E}}_{Y}[Y|X] - \mathop{\mathbb{E}}_{Y}[Y|X]\mathop{\mathbb{E}}_{D}[f(X)] - (\mathop{\mathbb{E}}_{Y}[Y|X])^2 + \mathop{\mathbb{E}}_{Y}[Y|X]\mathop{\mathbb{E}}_{Y}[Y|X]]|X]] \tag{C.17}$$

$$= 0 \tag{C.18}$$

Donc

$$\mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{Y}[(\mathop{\mathbb{E}}_{D}[f(X)]-Y)^2|X]]] = \mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{Y}[(\mathop{\mathbb{E}}_{D}[f(X)]-\mathop{\mathbb{E}}_{Y}[Y|X])^2|X]]]+\mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{Y}[(\mathop{\mathbb{E}}_{Y}[Y|X]-Y)^2|X]] \tag{C.19}$$

C'est-à-dire :

$$\mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{Y}[(f(X)-Y)^2|X]]] = \mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{Y}[(f(X)-\mathop{\mathbb{E}}_{D}[f(X)])^2|X] + \mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{Y}[(\mathop{\mathbb{E}}_{D}[f(X)]-\mathop{\mathbb{E}}_{Y}[Y|X])^2|X]]] + \mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{Y}[(\mathop{\mathbb{E}}_{Y}[Y|X]-Y)^2|X]] \tag{C.20}$$

$$= \underbrace{\mathop{\mathbb{E}}_{D}[\mathop{\mathbb{E}}_{X}[(f(X)-\mathop{\mathbb{E}}_{D}[f(X)])^2|X]}_{\text{Variance de } f} + \underbrace{\mathop{\mathbb{E}}_{X}[(\mathop{\mathbb{E}}_{D}[f(X)]-\mathop{\mathbb{E}}_{Y}[Y|X])^2|X]]}_{\text{Biais}^2} + \underbrace{\mathop{\mathbb{E}}_{X}[\mathop{\mathbb{E}}_{Y}[(\mathop{\mathbb{E}}_{Y}[Y|X]-Y)^2|X]]}_{\text{Erreur irréductible ne dépendant que des données}} \tag{C.21}$$

## C.2 Solution pour la maximisation de la vraisemblance dans le cas de la régression linéaire

Notre problème est alors équivalent à résoudre :

$$a^\star, b^\star = \arg\max_{a,b} e^{-\frac{1}{2}\sum_i^N \epsilon_i^2} \tag{C.22}$$

Car $\sqrt{2\pi}^{-N}$ est constant par rapport à $a$ et $b$. Or résoudre

$$\max_{a,b} e^{-\frac{1}{2}\sum_i^N \epsilon_i^2}$$

est équivalent à

$$\max_{a,b} -\frac{1}{2}\sum_i^N \epsilon_i^2$$

par stricte croissance de la fonction exponentielle. De plus, résoudre

$$\max_{a,b} -\frac{1}{2}\sum_{i}^{N}\epsilon_i^2$$

est équivalent à

$$\min_{a,b}\frac{1}{2}\sum_{i}^{N}\epsilon_i^2$$

par stricte décroissance de $x \mapsto -x$ sur $\mathbb{R}$. Enfin, cela est équivalent à :

$$\min_{a,b}\sum_{i}^{N}\epsilon_i^2$$

c'est à dire :

$$\min_{a,b}\sum_{i}^{N}\epsilon_i^2 = \min_{a,b}\sum_{i}^{N}(aX_i + b - y_i)^2 = \min_{a,b}\sum_{i}^{N}(f(X_i) - y_i)^2 \tag{C.23}$$

Pour minimiser cette quantité par rapport à $a$ et $b$, il faut alors remarquer que la fonction :

$$\begin{array}{rccl} m: & \mathbb{R}^2 & \to & \mathbb{R} \\ & (a,b) & \mapsto & \frac{1}{N}\sum_i^N (aX_i + b - y_i)^2 \end{array}$$

Est différentiable sur tout $\mathbb{R}^2$ en tant que somme de fonction différentiable sur $\mathbb{R}^2$, de plus par convexité de la fonction à variable réelle $x \mapsto x^2$, $m$ est convexe sur $\mathbb{R}^2$. Ainsi, $m$ admet un unique point critique (qui est son minimum) au point où son gradient est nul, c'est à dire où ses dérivées partielles s'annulent.

Or d'une part :

$$\frac{\delta m}{\delta b} = 2\frac{1}{N}\sum_{i}^{N}(aX_i + b - y_i) \tag{C.24}$$

ce qui implique que résoudre $\frac{\delta m}{\delta b} = 0$ par rapport à $b$, on a :

$$b = -a\frac{1}{N}\sum_{i}^{N}X_i + \frac{1}{N}\sum_{i}^{N}y_i \tag{C.25}$$

Donc dans le plan $\{(a,b) \in \mathbb{R}^2, b = -a\frac{1}{N}\sum_i^N X_i + \frac{1}{N}\sum_i^N y_i\}$ , $m$ prend la valeur :

$$m(a, -a\frac{1}{N}\sum_{i}^{N}X_i + \frac{1}{N}\sum_{i}^{N}y_i) = \sum_{i}^{N}(aX_i - a\frac{1}{N}\sum_{i}^{N}X_i + \frac{1}{N}\sum_{i}^{N}y_i - y_i)^2 \tag{C.26}$$

$$= \sum_{i}^{N}a(X_i - \frac{1}{N}\sum_{i}^{N}X_i) - (y_i - \frac{1}{N}\sum_{i}^{N}y_i))^2 \tag{C.27}$$

$$= \sum_i^N a^2(X_i - \frac{1}{N}\sum_i^N X_i)^2 - 2a(X_i - \frac{1}{N}\sum_i^N X_i)(y_i - \frac{1}{N}\sum_i^N y_i) \tag{C.28}$$

$$+(y_i - \frac{1}{N}\sum_i^N y_i)^2 \tag{C.29}$$

donc, dans le plan susmentionné

$$\frac{\delta m}{\delta a} = \sum_i^N 2a(X_i - \frac{1}{N}\sum_i^N X_i)^2 - 2(X_i - \frac{1}{N}\sum_i^N X_i)(y_i - \frac{1}{N}\sum_i^N y_i) \tag{C.30}$$

donc résoudre :

$$\frac{\delta m}{\delta a} = 0 \tag{C.31}$$

$$\sum_i^N 2a(X_i - \frac{1}{N}\sum_i^N X_i)^2 - 2(X_i - \frac{1}{N}\sum_i^N X_i)(y_i - \frac{1}{N}\sum_i^N y_i) = 0 \tag{C.32}$$

$$\tag{C.33}$$

d'où :

$$a = \frac{\sum_i^N (X_i - \frac{1}{N}\sum_i^N X_i)(y_i - \frac{1}{N}\sum_i^N y_i)}{\sum_i^N (X_i - \frac{1}{N}\sum_i^N X_i)^2} \tag{C.34}$$

donc pour conclure :

$$a^\star = \frac{\sum_i^N (X_i - \frac{1}{N}\sum_i^N X_i)(y_i - \frac{1}{N}\sum_i^N y_i)}{\sum_i^N (X_i - \frac{1}{N}\sum_i^N X_i)^2} \tag{C.35}$$

$$b^\star = -a^\star \frac{1}{N}\sum_i^N X_i + \frac{1}{N}\sum_i^N y_i \tag{C.36}$$

Et nous retrouvons pour $a^\star$, par multiplication au numérateur et dénominateur par $\frac{1}{N}$ , l'expression de $a^\star$ exprimée en fonction de la covariance de la variable aléatoire des étiquettes et des observations et de la variance de celle représentant les observations.