UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1

ÉCOLE DOCTORALE N° 605
Biologie Santé
Spécialité : *Génétique, Génomique et bioinformatique*

Par
Maël CONAN

Approche prédictive pour évaluer la génotoxicité des contaminants de l'environnement

Thèse présentée et soutenue à Rennes, le 23 mars 2021
Unité de recherche :
Institut de Recherche en Santé, Environnement et Travail (IRSET). Equipe Dymec (Univ Rennes 1, INSERM, EHESP).
Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA). Equipe Dyliss (Univ Rennes 1, INRIA, CNRS).

Rapporteurs avant soutenance :

Sabine PERES	Maitresse de conférence, Université Paris Saclay
Fabien JOURDAN	Directeur de recherche, INRAE, Toulouse

Composition du Jury :

Président :	Cédric LHOUSSAINE	Professeur, Université de Lille
Examinateurs :	Karine AUDOUZE	Maitresse de conférence, Université de Paris
	Fabien JOURDAN	Directeur de recherche, INRAE, Toulouse
	Nathalie THERET	Directrice de recherche, Université de Rennes
	Sabine PERES	Maitresse de conférence, Université Paris Saclay
Dir. de thèse :	Sophie LANGOUET	Directrice de recherche, INSERM, Rennes
Co-dir. de thèse :	Anne SIEGEL	Directrice de recherche, CNRS, Rennes

Sommaire

TABLE DES FIGURES

LISTE DES TABLEAUX

PRÉAMBULE

Le foie joue un rôle majeur dans le maintien de l'homéostasie métabolique. Il est notamment responsable du métabolisme, de la distribution, de la détoxification et de l'excrétion des substances chimiques auxquelles nous sommes quotidiennement exposés. Il peut être agressé par différents composés exogènes ce qui peut mener entre autres à des maladies chroniques hépatiques. Celles-ci sont de plus en plus une cause majeure de morbidité et de mortalité, responsables de plus de 2 millions de morts dans le monde chaque année. L'étude des conséquences de l'exposition chronique à des composés exogènes, ou xénobiotiques, est donc un domaine de recherche essentiel en santé publique.
Parmi les xénobiotiques, notre équipe s'intéresse à une catégorie de composés préoccupants, les amines hétérocycliques aromatiques (AHAs). Celles-ci sont catégorisées comme possiblement et probablement cancérigènes par l'IARC mais on ne dispose que de peu de données du devenir de ces contaminants chez l'homme. Parmi les 30 AHAs aujourd'hui identifiées, l'activation métabolique de seulement trois a été caractérisée chez l'homme en utilisant des hépatocytes humains primaires : AαC, MeIQx et PhIP. Identifier des biomarqueurs d'expositions dérivés des AHAs permettrait de réaliser des études épidémiologiques dont nous avons besoin pour clarifier le potentiel génotoxique des AHAs. Pour cela, les méthodes de prédictions *in silico* sont essentielles car elles permettent d'orienter la recherche de ces biomarqueurs d'expositions et de prédire le potentiel génotoxique de nombreux composés.

Dans ce contexte, ce travail de thèse propose une approche de prédiction de la bioactivation des AHAs permettant à la fois de prédire l'activation métabolique, c'est-à-dire les réactions que peuvent subir les AHA et les métabolites dérivés, mais aussi de prédire la réactivité des métabolites vis-à-vis de l'ADN et la probabilité que les métabolites soient produits. Cette nouvelle approche se base sur la prédiction de cartes du métabolisme des xénobiotiques à travers l'utilisation de règles de biotransformations. Ces cartes sont ensuite enrichies par des annotations et notamment des annotations provenant d'outils prédicteurs des sites du métabolisme. Ces annotations, utilisées avec un formalisme de réseau bayésien, permettent de calculer un score de probabilité de production permettant d'évaluer les chances de synthétiser un métabolite suivant une condition enzymatique précise.

Cette approche a été appliquée et validée en utilisant une molécule bien décrite dans la littérature, la caféine. Les résultats que nous avons obtenus ont permis de montrer la capacité

de notre score à discriminer les métabolites expérimentalement identifiés des autres. Cela nous a permis d'utiliser ce score comme un outil de filtration des cartes prédites du métabolisme. Ensuite, nous avons pu déterminer le rôle de différentes enzymes du métabolisme des xénobiotiques sur la production de métabolites réactifs vis-à-vis de l'ADN en appliquant notre approche aux AHAs. Nous avons également pu déterminer les conditions, en termes de disponibilité des enzymes, favorisant la production de métabolites réactifs à l'ADN et donc d'adduits à l'ADN.

Cette nouvelle approche pourra être appliquée à de nombreux contaminants de l'environnement d'intérêt comme les perturbateurs endocriniens et les mycotoxines pour lesquels des données manquent chez l'homme. De plus, notre approche sera particulièrement utile pour déterminer l'effet des conditions physio-pathologiques sur la synthèse de métabolites d'intérêts comme les métabolites réactifs vis-à-vis de l'ADN ou encore les métabolites fixant des récepteurs membranaires ou nucléaires spécifiques.

Chapitre 1

Introduction

Ce premier chapitre permet de décrire le contexte biologique qui borne ce travail de thèse. Ce travail s'articule autour de la prédiction et la caractérisation de la génotoxicité des contaminants de l'environnement. Nous allons, dans un premier temps, décrire le contexte biologique de ces recherches. Puis nous introduirons l'objet de nos travaux c'est-à-dire les Amines Hétérocycliques Aromatiques (AHA). Puis nous présenterons l'ensemble des outils et ressources disponibles concernant la prédiction et la caractérisation de la génotoxicité des xénobiotiques. Enfin nous décrirons notre projet de recherche, notre problématique et les méthodes, outils et ressources dont nous disposons pour répondre à notre problématique. Enfin nous aborderons la contribution de ce travail de recherche.

1.1 Contexte biologique

Notre projet s'intègre dans l'étude du devenir des contaminants de l'environnement dans le foie humain qu'il soit sain ou pathologique. En effet la recherche autour des effets des contaminants de l'environnement au niveau du foie ou des effets des pathologies du foie sur le métabolisme de ces contaminants est un vrai enjeu de santé publique ; le nombre de maladies hépatiques lié à l'environnement est en augmentation constante depuis plus d'une décennie sans que les causes soient totalement identifiées [50].

1.1.1 Les maladies chroniques hépatiques

Les maladies chroniques hépatiques sont une cause majeure de morbidité et de mortalité [27]. Elles sont à l'origine de plus de 2 millions de morts dans le monde chaque année et sont en constante augmentation [50]. Ces maladies sont devenues un enjeu de santé publique et économique important [71].

La plupart des maladies chroniques hépatiques sont associées au développement d'une fibrose qui témoigne des agressions répétées du foie. Elle se caractérise par une accumulation de protéines de la matrice extracellulaire. Elle induit une déformation de l'architecture hépatique en formant une cicatrice fibreuse. Le stade ultime de la fibrose est la cirrhose qui se caractérise par un dépôt de tissu conjonctif, de la régénération, des troubles vasculaires et une nécrose des hépatocytes.

90% des carcinomes hépatocellulaires (CHC), le cancer primaire du foie, se développe sur foie cirrhotique [4].

Les causes des maladies chroniques hépatiques sont diverses comme le virus de l'hépatite B (HBV), de l'hépatite C (HCV), l'alcool (à l'origine d'alcool liver disease (ALD)) ou l'obésité (à l'origine du non alcoholic fatty liver disease (NAFLD)). La cause principale de maladies chroniques hépatiques est le virus de l'hépatite B mais l'incidence de ce virus diminue tandis que celui des NAFLD augmente ainsi que les contaminants de l'environnement [50].

1.1.2 Les contaminants de l'environnement

Le foie est un organe qui maintient l'homéostasie métabolique. Il est responsable du métabolisme, de la distribution et de l'excrétion des substances chimiques exogènes ce qui fait qu'il est considéré comme le premier organe d'exposition [3]. Les contaminants de l'environnement sont considérés comme des facteurs majeurs du développement des NAFLD et donc des maladies chroniques hépatiques. L'étude du devenir des contaminants de l'environnement, ou xénobiotiques, par leur métabolisme et leur bioactivation et de leurs effets est une question cruciale permettant de comprendre de nombreuses pathologies. Parmi les questions de recherche actuelles un enjeu est de comprendre l'étude des effets de l'exposition chronique à de faibles doses de xénobiotiques ainsi que les effets des mélanges de contaminants dits "effets cocktails".

Notre équipe s'intéresse plus particulièrement à une famille de contaminants préoccupants, les AHAs. Ces AHAs sont des produits de pyrolyse de la viande ou du poisson ; on les retrouve également dans la fumée de cigarette ou les gaz d'échappement [73]. Elles ont été catégorisées comme cancérogènes possibles (2B) et probables par l'International Agency for Research on Cancer (IARC) [59, 31]. En effet elles sont mutagènes chez la bactérie et cancérogènes chez l'animal. Le manque de données chez l'homme explique la classification de l'IARC. Il est donc nécessaire d'identifier des marqueurs d'exposition de ces AHAs afin de permettre d'étudier l'exposition de la population à ces molécules. Ces marqueurs d'exposition peuvent être des métabolites et/ou des adduits à l'ADN. À ce jour, 30 AHAs ont été identifiées ; elles sont décrites dans l'annexe 1. L'identification des métabolites et des adduits à l'ADN ont été décrits,dans des hépatocytes humains primaires, pour trois de ces trente AHAs [41, 40, 52, 6]. Afin d'explorer l'activation métabolique des 30 AHAs et la capacité des métabolites à former des adduits à l'ADN, des méthodes de prédiction et de modélisation du métabolisme ont été développées [19] afin de générer des cartes prédictives du métabolisme.

1.2 État de l'art de la prédiction du métabolisme et de la réactivité

Dans ce contexte, ma problématique durant ma thèse a été de déterminer, à partir de cartes du métabolisme des AHAs, l'influence des enzymes du métabolisme des xénobiotiques sur la formation des adduits à l'ADN dérivés de ces AHA. Pour cela nous allons introduire plusieurs notions. Tout d'abord, nous allons rappeler les connaissances disponibles sur les AHAs et le métabolisme connu des AHAs. Ensuite, nous allons nous intéresser aux méthodes capables de prédire le métabolisme, de prédire l'influence des enzymes et de prédire la réactivité. Enfin, nous ferons l'état des ressources dont nous disposons qui permettra de situer notre projet de recherche.

1.2.1 Les Amines Hétérocycliques Aromatiques et leur métabolisme

Trente AHAs ont été identifiées à ce jour. Le détail de leurs noms et structures chimiques est répertorié en annexe 1. Elles peuvent être réparties en deux classes en fonction de la réaction responsable de leur formation. Il y a d'abord les AHA pyrolytiques qui sont formées par une réaction de pyrolyse des acides aminés à une température pouvant atteindre 250°C. Un exemple d'amine de ce groupe est le 2-Amino-9H-pyrido[2,3-*b*]indole (AαC). Les aminoimidazoarènes correspondent à la deuxième classe d'AHAs qui sont produites par la réaction de Maillard entre un hexose et des acides aminés à une température supérieure à 150°C. La 2-amino-3,8--diméthylimidazo[4,5-textitf]quinoxaline (MeIQx), la 2-amino-1-méthyl-6-phénylimidazo[4,5-*b*]-pyridine (PhIP) et la 2-amino-3-méthylimidazo[4,5-*f*]quinoléine (IQ) appartiennent à cette classe avec respectivement un noyau de quinoxaline, de pyridine et de quinoléine. Parmi les trente AHAs identifiées, la bioactivation de trois a été étudiée dans des hépatocytes humains primaires, il s'agit de AαC, PhIP et MeIQx [41, 40, 52, 6]. Ces travaux ont permis de caractériser les métabolites et les adduits à l'ADN dérivés de ces trois AHAs. Ils ont également permis d'identifier les enzymes impliquées dans la formation de ces métabolites. Les tables 1.1, 1.2 et 1.3 décrivent les métabolites dérivés de ces trois AHAs ainsi que leur réactivité vis-à-vis de l'ADN [8, 73, 6].

Métabolite	Formule SMILES	Structure 2D	Réactif vis-à-vis de l'ADN
AαC	Nc1ccc2c(n1)[nH]c1ccccc12		NON
AαC-3-O-Gluc	Nc1nc2[nH]c3ccccc3c2cc1OC1OC-(C(=O)O)C(O)C(O)C1O		NON
AαC-3-OH	Nc1nc2[nH]c3ccccc3c2cc1O		NON
AαC-3-O-SO3H	Nc1nc2[nH]c3ccccc3c2cc1OS(=O)-(=O)O		NON
AαC-6-O-Gluc	Nc1ccc2c(n1)[nH]c1ccc(OC3OC(C-(=O)O)C(O)C(O)C3O)cc12		NON
AαC-6-OH	Nc1ccc2c(n1)[nH]c1ccc(O)cc12		NON
AαC-6-O-SO3H	Nc1ccc2c(n1)[nH]c1ccc(OS(=O)-(=O)O)cc12		NON
AαC-HN2-O-Gluc	O=C(O)C1OC(ONc2ccc3c(n2)-[nH]c2ccccc23)C(O)C(O)C1O		OUI
AαC-HN2-OH	ONc1ccc2c(n1)[nH]c1ccccc12		OUI
AαC-N2-Gluc	O=C(O)C1OC(Nc2ccc3c(n2)[nH]-c2ccccc23)C(O)C(O)C1O		NON
N-Acetoxy-AαC	CC(=O)ONC1=NC2=C(C=C1)-C1=CC=CC=C1N2		OUI
N-Sulfonyloxy-AαC	OS(=O)(=O)ONC1=NC2=C-(C=C1)C1=CC=CC=C1N2		OUI

TABLE 1.1 – **Table descriptive des métabolites dérivés de AαC** Chaque métabolite est associé à une formule SMILES qui permet de décrire la structure 2D des composés. Cette structure apparaît dans la colonne *Structure 2D*. De plus la colonne *Réactif vis-à-vis de l'ADN* précise si ce métabolite est décrit dans la littérature comme étant réactif vis-à-vis de l'ADN [8, 73, 6]

L'activation métabolique des AHAs Un grand nombre de contaminants sont des procarcinogènes, c'est-à-dire qu'ils nécessitent d'être métabolisés pour devenir toxiques ou génotoxiques en altérant l'ADN. C'est le cas des AHAs qui doivent subir plusieurs étapes de biotransformation pour être élimnier et c'est au cours de cette biotransformation que des métabolites intermédiaires très réactifs peuvent se fixer de façon covalente à l'ADN et former des adduits. Le métabolisme des AHAs a été étudié chez l'animal et chez l'homme et les principales étapes sont décrites par la figure 1.1.

L'activation métabolique des AHA se décompose en deux phases de biotransformation biochimique formant des métabolites et une phase d'excrétion. Tout d'abord le xénobiotique est

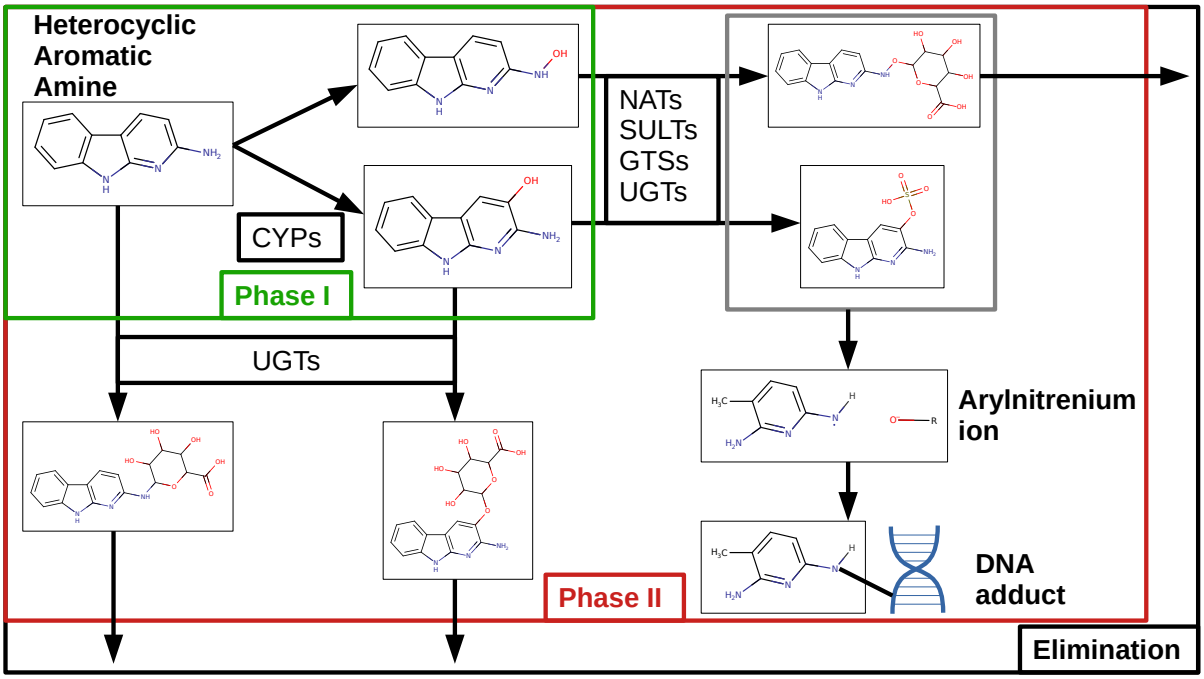


FIGURE 1.1 – **Description du métabolisme des AHAs.** L'activation métabolique des AHA se déroule en deux phases : une première phase d'oxydation nommée "phase I" (délimitée en vert), cette phase est principalement catalysée par les enzymes du métabolisme de phase I : les cytochromes P450 (CYPs). Le métabolite oxydé ainsi formé subit ensuite une réaction de conjugaison dite "phase II" (délimitée en rouge). Elle est catalysée par les enzymes de phase II que sont les UDP-glucuronyl transferases (UGTs), les N-acetyl transférases (NATs), les Sulfotransférases (SULTs) ou les Glutathion-S-transferase (GSTs) Le métabolite conjugué peut alors soit être excrété, c'est la phase III du métabolisme des xénobiotique (délimitée en noir). Dans certain cas le métabolite conjugué subit un clivage hétérolytique et forme un ion arylnitrenium qui est réactif vis-à-vis de l'ADN et peut conduire à la formation d'un adduit à l'ADN.

Métabolite	Formule SMILES	Structure 2D	Réactif vis-à-vis de l'ADN
MeIQx	Cc1cnc2ccc3c(nc(N)n3C)c2n1		NON
7-oxo-MeIQx	Cc1nc2c(ccc3c2nc(N)n3C)nc1O		NON
8-CH2OH-IQx	Cn1c(N)nc2c3nc(CO)cnc3ccc21		NON
HON-MeIQx	Cc1cnc2ccc3c(nc(NO)n3C)c2n1		OUI
HON-MeIQx-N2-Gluc	Cc1cnc2ccc3c(nc(N(O)C4OC(C-(=O)O)C(O)C(O)C4O)n3C)c2n1		NON
IQx-8-COOH	Cn1c(N)nc2c3nc(C(=O)O)-cnc3ccc21		NON
MeIQx-N2-Gluc	Cc1cnc2ccc3c(nc(NC4OC(C(=O)-O)C(O)C(O)C4O)n3C)c2n1		NON
MeIQx-N2-SO3H	Cc1cnc2ccc3c(nc(NS(=O)(=O)O)-n3C)c2n1		NON
N-Acetoxy-MeIQx	CN1C(NOC(C)=O)=NC2=C1C-=CC1=C2N=C(C)C=N1		OUI
N-desmethyl-7-oxo-MeIQx	CC1=NC2=C(NC1=O)C=CC1-=C2N=C(N)N1		NON
N-Sulfonyloxy-MeIQx	CN1C(NOS(O)(=O)=O)=NC2-=C1C=CC1=C2N=C(C)C=N1		OUI

TABLE 1.2 – **Table descriptive des métabolites dérivés de MeIQx** Chaque métabolite est associé à une formule SMILES qui permet de décrire la structure 2D des composés. Cette structure apparaît dans la colonne *Structure 2D*. De plus la colonne *Réactif vis-à-vis de l'ADN* précise si ce métabolite est décrit dans la littérature comme étant réactif à l'ADN [8, 73, 6]

transformé par une première phase dite d'oxydation ou phase I. Cette phase est catalysée par les cytochromes P450 (CYPs) et notamment le CYP1A2 chez l'homme [72, 1, 41, 40]. On décrit les enzymes qui catalysent les réactions d'oxydation de cette phase comme les enzymes de phase I.

Les métabolites obtenus sont ensuite pris en charge par la phase II et sont conjugués. Ces réactions de conjugaison sont catalysées par différentes enzymes, dites de phase II et de conjugaison, que sont les UDP-glucuronyl transférases (UGTs), les N-acetyls transférases (NATs), les Sulfotransférases (SULTs) et les Glutathion-S-transférases (GSTs). Une fois conjugué, le métabolite peut soit être éliminé, c'est-à-dire être excrété par la phase III du métabolisme, soit être bioactivé. Dans ce cas, le métabolite conjugué est instable et peut subir un clivage hétérolytique. Cette réaction produit alors un ion arylnitrenium qui est particulièrement réactif vis-à-vis de l'ADN. Dans ce cas le métabolite peut alors potentiellement former un adduit à l'ADN qui entraîne

Métabolite	Formule SMILES	Structure 2D	Réactif vis-à-vis de l'ADN
PhIP	Cn1c(N)nc2ncc(-c3ccccc3)cc21		NON
4'HO-PhIP	Cn1c(N)nc2ncc(-c3ccc(O)cc3)cc21		NON
4'-Ogluc-PhIP	Cn1c(N)nc2ncc(-c3ccc(OC4OC(C-(=O)O)C(O)C(O)C4O)cc3)cc21		NON
4'-OSO3H-PhIP	Cn1c(N)nc2ncc(-c3ccc(OS(=O)-(=O)O)cc3)cc21		NON
HON-PhIP	Cn1c(NO)nc2ncc(-c3ccccc3)cc21		OUI
HON-PhIP-N2-Gluc	Cn1c(N(O)C2OC(C(=O)O)C(O)C-(O)C2O)nc2ncc(-c3ccccc3)cc21		NON
PhIP-N2-Gluc	Cn1c(NC2OC(C(=O)O)C(O)C(O)-C2O)nc2ncc(-c3ccccc3)cc21		NON
PhIP-N3-Gluc	Cn1c(N)n(C2OC(C(=O)O)C(O)C-(O)C2O)c2ncc(-c3ccccc3)cc21		NON
N-Acetoxy-PhIP	CC(=O)ONC1=NC2=C(N1C)-C=C(C=N2)C3=CC=CC=C3		OUI
N-Sulfonyloxy-PhIP	CN1C2=C(N=CC(=C2)C3=CC=-CC=C3)N=C1NOS(=O)(=O)O		OUI

TABLE 1.3 – **Table descriptive des métabolites dérivés de PhIP** Chaque métabolite est associé à une formule SMILES qui permet de décrire la structure 2D des composés. Cette structure apparaît dans la colonne *Structure 2D*. De plus la colonne *Réactif vis-à-vis de l'ADN* précise si ce métabolite est décrit dans la littérature comme étant réactif à l'ADN [8, 73, 6]

une l'altération de l'ADN. Si cette altération n'est pas corrigée par les systèmes de réparation de l'ADN, cela peut alors être à l'origine de mutations ; première étape du cancer. Ces connaissances ont été à la base de travaux précédents au laboratoire permettant la reconstruction de cartes du métabolisme des AHAs [19].

1.2.2 Méthodes de prédiction du métabolisme et de la réactivité

Notre objectif et de définir une méthode de prédiction du métabolisme des xénobiotiques permettant d'interroger le rôle des enzymes dans la formation des adduits à l'ADN dérivés des métabolites des AHAs. Pour cela, nous allons présenter les outils qui permettent de prédire le métabolisme, les outils permettant de prédire la réactivité à l'ADN et les outils qui permettent d'obtenir des informations sur le rôle des enzymes.

1.2.2.1 Les outils de prédiction du métabolisme

Les outils permettant la prédiction du métabolisme fonctionnent tous sur un même principe initial, celui de l'application de règles de biotransformations [15, 45, 39, 78, 49, 25, 21, 60]. Concrètement une règle de biotransformation décrit une réaction biochimique avec une structure chimique d'entrée et une structure chimique de sortie. Une réaction biochimique consomme des molécules et en produit d'autres. Cependant lorsque les molécules sont complexes, seulement une partie de la structure chimique est impactée par la réaction. La spécificité d'une règle de biotransformation par rapport à la complète description de la réaction chimique est qu'elle ne décrit que les parties des molécules qui varient au cours de la réaction. Ainsi la règle de biotransformation est plus générale que la description totale d'une réaction biochimique. Les outils de prédiction du métabolisme utilisent un ensemble de ces règles pour prédire des métabolites. Ils parcourent ces règles et la structure chimique de la molécule dont on prédit le métabolisme. Pour chaque règle, l'outil recherche des sous-structures, il s'agit d'une structure qui est complètement intégrée dans la structure chimique de la molécule. Lorsque la sous-structure correspond à la structure d'entrée de la règle de biotransformation, la règle est appliquée et permet de générer une nouvelle structure chimique, c'est un métabolite. Ensuite, l'outil recommence à rechercher des sous-structures. Une autre méthode très similaire consiste à utiliser des règles de biotransformations non plus par des structures chimiques d'entrée et de sortie mais par des empreintes moléculaires. L'idée est de définir l'empreinte d'une structure chimique d'entrée en détaillant chaque atome, leur position, les atomes voisins ou encore les liaisons que l'atome forme. Ensuite, lors du parcours des sous-structures de la molécule, on définit également l'empreinte de la sous-structure. Puis, en comparant les empreintes par un système de score on peut alors définir la proximité des structures. Si cette proximité est considérée comme suffisamment grande alors la règle est appliquée et un nouveau métabolite est formé [10].

Les outils peuvent être accessibles à tous, ou bien sous licence commerciale. Ils diffèrent les uns des autres de part l'ensemble des règles de biotransformation qu'ils utilisent. Ces règles de biotransformation, que ces outils utilisent, sont classées en deux catégories : les règles dites expertes et les règles provenant de l'apprentissage. La première catégorie de règles décrit les règles qui proviennent de l'analyse de la littérature et de l'expertise des auteurs de l'outil de prédiction. Les auteurs recherchent alors dans la littérature, les réactions biochimiques d'intérêt et les traduisent ensuite sous la forme de règles de biotransformation. Les règles peuvent également être issues de l'exploration automatique de bases de données de réactions biochimiques comme BioVia, DrugBank, PharmGKB ou XMETDB [76, 75, 68]. Ces bases de données sont explorées à travers des outils d'apprentissage automatique. Ces outils extraient des règles de biotransformation à partir d'une réaction chimique ou d'un ensemble de réactions chimiques en comparant la structure de la molécule consommée et de la molécule produite. Les outils de prédiction les plus

cités comme Meteor Nexus ou ADMET Predictor [45, 32] utilisent tous deux des règles issues de l'exploration des bases de données mais également expertes. C'est une tendance générale qui rejoint l'idée que les outils de prédiction du métabolisme nécessitent une coopération entre un système d'apprentissage capable d'explorer une grande base de données et des connaissances expertes [38]. Encore récemment, l'outil BioTransformer, publié en 2019, qui est un outil de prédiction du métabolisme a montré cette tendance. Il utilise une base de données de réactions biochimiques MetXBioDB. Cette base de données permet de retrouver pour chaque réaction la règle de biotransformation définissant cette réaction. La base de données est construite à partir d'export des bases de données de réactions DrugBank, PharmGKB, XMETDB, et SuperCYP [76, 75, 68, 57] mais elle intègre également plus de 100 réactions, et donc règles de transformation, extraites de la littérature. L'utilisation d'un grand nombre de règles de biotransformation conduit inévitablement à la prédiction d'un grand nombre de métabolites [7]. Les métabolites prédits par ces outils sont généralement représentés sous la forme de carte du métabolisme. Il s'agit d'un graphe qui représente les métabolites ainsi que les réactions qui les relient. La figure 2.1 présentée dans le chapitre 2 décrit une telle carte.

1.2.2.2 Les outils de prédiction du site du métabolisme

D'autres outils de prédictions sont complémentaires des outils de prédictions des métabolites, il s'agit des outils de prédictions du site du métabolisme ou SOM. Un SOM désigne un atome ou des atomes d'une molécule qui sont soit catalysés par une enzyme, soit réactifs par rapport à un ensemble de réactions données. Par exemple, dans une réaction qui ajoute un groupe à une molécule, il peut s'agir de l'atome qui forme une liaison avec le groupe ajouté. Les outils qui prédisent des SOMs, prédisent en fait la réactivité des atomes par rapport à un groupe de réactions données. Les groupes de réactions sont formés afin de représenter les réactions qui sont catalysées par une enzyme donnée. Ainsi, prédire un SOM c'est prédire la réactivité des atomes par rapport à des réactions catalysées par une enzyme donnée. Les outils prédicteurs de SOMs associent à chaque atome d'une molécule un score qui décrit la probabilité que l'atome soit un SOM. Il existe différentes stratégies pour prédire des SOMs [36].
Tout d'abord, il existe des méthodes qui utilisent des descripteurs moléculaires. Ce sont des caractéristiques que chaque atome d'une molécule possède. Il peut s'agir de sa nature, du nombre de liaisons que l'atome forme, du type de liaison, de la nature des atomes auquel il est lié ou de la charge de l'atome. Un descripteur particulier est l'énergie nécessaire à l'abstraction d'un hydrogène de l'atome analysé. Cette caractéristique est prédictive de la réactivité des atomes par rapport aux CYPs [66]. Des outils QMBO, CypScore, SMARTCyp ou MetaSite [28, 2, 63, 64, 15, 14] utilisent ce type de descripteur spécifique. Plutôt que d'utiliser ce seul descripteur, une autre stratégie est d'utiliser un vaste ensemble de descripteurs moléculaires et des méthodes d'apprentissage automatique. L'idée est tout d'abord de définir un ensemble de réactions, caractérisées

par le fait qu'elles sont catalysées par une enzyme ou une famille d'enzymes. Ensuite on repère dans cet ensemble de réactions le site du métabolisme donc l'atome de la molécule consommée qui réagit. On décrit chaque SOM à partir d'un ensemble de descripteurs moléculaires, puis on utilise l'apprentissage automatique pour inférer, à partir de ces descripteurs, des règles de calculs. Ces règles permettent de calculer un score, à partir de la valeur des descripteurs, qui décrit la probabilité d'être un SOM. À partir d'un ensemble de descripteurs, caractéristiques d'un atome, on peut alors calculer ce score et prédire si un atome est un SOM ou non. L'utilisation des méthodes d'apprentissages se retrouvent dans des outils comme Way2Drug SOMP, FAst MEtabolizer (FAME) ou XenoSite Metabolism 1.0 [61, 37, 65, 29, 30]. Enfin une dernière stratégie est d'utiliser des méthodes de recherche de similarités entre des structures de ligands à une enzyme et celle du composé dont on cherche à prédire les SOMs en complément d'une approche de docking [43, 9, 54].

Les outils prédicteurs de SOMs se concentrent principalement sur la prédiction de SOMs par rapport aux enzymes du métabolisme des xénobiotiques et principalement autour des CYPs, en effet prédire la réactivité des molécules par rapport aux enzymes de ce métabolisme est une question cruciale dans le développement de médicaments notamment pour prédire une potentielle toxicité.

1.2.2.3 Les outils de prédictions de la réactivité vis-à-vis de l'ADN

Nous nous intéressons à la question de la formation des adduits à l'ADN dérivés des AHAs. Il existe un ensemble d'outils de prédictions qui permettent non pas de prédire la formation d'adduits mais la réactivité d'une molécule vis-à-vis de l'ADN. Cette prédiction reste tout de même relative à la question de la formation des adduits.

Il existe différents outils de prédictions de cette réactivité. Tout d'abord des outils permettant la prédiction de la toxicité d'une molécule. Ces prédictions portent sur différentes toxicités et notamment la génotoxicité et donc la réactivité vis-à-vis de l'ADN. Tout d'abord, il existe des outils basés sur la recherche de structures spécifiques [46, 33, 55]. Ces structures sont des structures dont on sait, expérimentalement, qu'elles lient l'ADN. Retrouver une telle structure dans une molécule la classe donc comme réactive vis-à-vis de l'ADN. Une autre façon de déterminer si un composé peut réagir avec l'ADN est d'utiliser les relations structure-toxicité. C'est la méthode des modèles QSAR (Quantitative Structure-Activity Relationship), ces modèles utilisent les descripteurs moléculaires des composés. Cette fois les descripteurs décrivent la molécule et non plus un atome comme pour la prédiction des SOMs. La génotoxicité est déterminée par une fonction mathématique appliquée au descripteur d'un composé[77, 46, 33, 62]. Plus récemment, des outils ont utilisé les mêmes principes que pour la prédiction des SOMs par machine learning. XenoSite Reactivity v1 [29, 30] utilise donc la même méthode que la prédiction des SOMs,

c'est-à-dire qu'il utilise les descripteurs moléculaires des atomes. Cette fois l'outil ne prédit plus la probabilité qu'un atome soit l'atome de la réaction, par rapport à un groupe de réactions, catalysées par la même enzyme mais plutôt la probabilité qu'un atome soit l'atome liant l'ADN, par rapport à un groupe de réactions qui seraient des réactions de formation d'adduit à l'ADN. C'est la prédiction de sites de réactivité ou SORs. Comme pour la prédiction des SOMs, l'outil détaille les descripteurs moléculaires des atomes identifiés comme liant l'ADN, dans des réactions de formation d'adduit à l'ADN. Ces réactions proviennent de bases de données de réactions. Une fois les atomes décrits, une méthode de réseau de neurones permet d'inférer des règles de calculs permettant de calculer, à partir des descripteurs moléculaires, la probabilité que l'atome soit un SOR.

1.2.3 Les ressources disponibles

Notre problématique est de prédire le rôle des enzymes du métabolisme des xénobiotiques et les conditions de formation des adduits à l'ADN dérivés des AHAs. Pour cela, nous devons donc être en mesure de prédire à la fois le métabolisme et la réactivité à l'ADN des métabolites. Nous devons pouvoir prédire un effet des enzymes sur les réactions prédites afin de déterminer des conditions de formation des adduits à l'ADN. Pour cela nous avons présenté une série de méthodes de prédictions. Nous allons présenter ici les applications de cette méthode à des problématiques similaires.

1.2.3.1 Prédiction de carte du métabolisme des AHAs et de la formation d'adduits à l'ADN

Tout d'abord, l'ensemble de ces méthodes a été appliqué aux 30 AHAs [19]. Cette première étude de Delannée et. al propose un workflow permettant de prédire le métabolisme des AHAs et leur réactivité vis-à-vis de l'ADN. La figure 1.2 provient de la thèse à l'origine de cette étude et détaille le pipeline [20].

Ce pipeline a permis de générer les cartes du métabolisme de 30 AHAs [18] et de décrire les métabolites réactifs vis-à-vis de l'ADN. La prédiction du métabolisme utilise l'outil Metaprint2D-React, aujourd'hui indisponible. Cet outil basait la prédiction du métabolisme sur des règles de biotransformation d'empreintes moléculaires et non de stricte structure chimique. L'outil utilisait une méthode provenant de Metaprint2D encore disponible qui est un outil prédicteur de sites du métabolisme spécialisé sur les CYPs. Ensuite le pipeline utilise les prédicteurs de SOMs WaytoDrugs SOMP et XenoSite Metabolism pour filtrer certains métabolites. L'idée est d'identifier les scores SOMs de la réaction, c'est-à-dire le score associé à l'atome de la réaction, et si celui-ci est supérieur ou égal à un seuil alors la réaction est conservée. Sinon la réaction est éliminée et l'ensemble des métabolites qui ne sont plus liés par une réaction à l'AHA d'origine sont

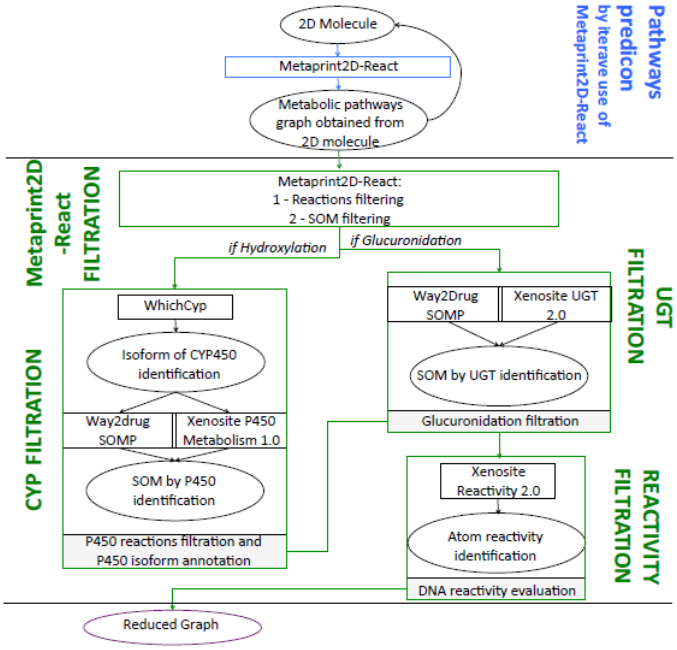


FIGURE 1.2 – **Pipeline pour la prédiction des voies métaboliques des xénobiotiques et l'identification des métabolites réactifs à l'ADN** Ce pipeline provient de la thèse à l'origine de l'étude Delannée et al. 2019 [20]. Il décrit un pipeline en deux parties permettant de générer des cartes du métabolisme des AHAs. La première partie est une étape de prédiction des métabolites en utilisant Metaprint2D-React (en bleu) comme outil de prédiction du métabolisme. Ensuite vient une partie de filtration des résultats de la première partie à partir de différents outils de prédictions de sites du métabolisme (en vert). Les prédictions sont également filtrées à travers les résultats de XenoSite Reactivitity afin d'orienter la construction des cartes du métabolisme vers la formation d'adduits à l'ADN.

également éliminés. Les seuils utilisés ont été déterminés en utilisant un ensemble de composés chimiques analogues aux AHAs et dont le métabolisme est connu. Ces seuils sont les seuils à partir desquels aucun métabolite expérimentalement identifié n'est filtré.

Ensuite ce pipeline permet d'identifier les métabolites réactifs vis-à-vis de l'ADN. Ces métabolites sont définis comme les métabolites ayant un score de réactivité supérieur ou égal à 0,85. Ce score provient de l'outil XenoSite Reactivity et il est le score le plus fort parmi les scores de l'ensemble des atomes du métabolite. Le seuil de 0,85 a été déterminé à partir d'un jeu de données test de molécules connues pour former des adduits à l'ADN.

Les cartes du métabolisme, prédites par ce pipeline, ont été utilisées afin de déterminer un ratio : celui du nombre de métabolites réactifs vis-à-vis de l'ADN divisé par le nombre total de métabolites. Ce ratio a ensuite servi à désigner les AHAs les plus à même de former des adduits à l'ADN. Par rapport à notre projet, ce travail détaille à la fois le métabolisme des AHAs et la réactivité des métabolites. Le rôle des enzymes est pris en compte mais seulement en tant qu'outil de filtration des prédictions et ne permet pas ensuite de déterminer un rôle des enzymes dans la formation de tel ou tel métabolite réactif. Ensuite, ce travail se concentre sur les métabolites réactifs vis-à-vis de l'ADN et occulte donc une partie de la carte du métabolisme lorsque certains métabolites sont filtrés sur la base du score de réactivité. En effet, pour concentrer ce travail sur la prédiction des potentiels adduits à l'ADN, lorsqu'un métabolite est associé à un score de réactivité plus faible que la molécule dont il est dérivé, il est éliminé ce qui réduit la carte aux seules voies formant des adduits à l'ADN.

1.2.3.2 Prédiction de site du métabolisme associés à la N-Dealkylation et application à la terbinafine

Une autre étude utilise des outils de prédiction du site du métabolisme d'une manière inédite [17]. Cette étude de Dang et al. relate l'utilisation de l'algorithme des outils prédicteurs de SOMs XenoSite [47, 79] à l'application de la prédiction de SOMs spécifiques des réactions de N-dealkylation. La prédiction de ces SOMs a été appliquée à la terbinafine (TBF) et différents dérivés de TBF. La figure 1.3 est extraite de leur étude [17] et présente la carte du métabolisme du TBF auquel la prédiction de SOMs a été appliquée.

Cette étude a montré une utilisation intéressante des scores SOMs. En effet ceux-ci permettent d'évaluer les réactions. Pour cela, on associe à la réaction le score SOM de l'atome de la réaction. Ce score associé à la réaction devient la probabilité que la réaction se produise. En utilisant l'ensemble des scores SOMs associés à une suite de réactions, on peut déterminer la probabilité de cette voie métabolique. Dans la figure 1.3 il s'agit des voies métaboliques A,B,C et D. Chacune est associée à une probabilité et les voies ont pu être comparées entre elles. De

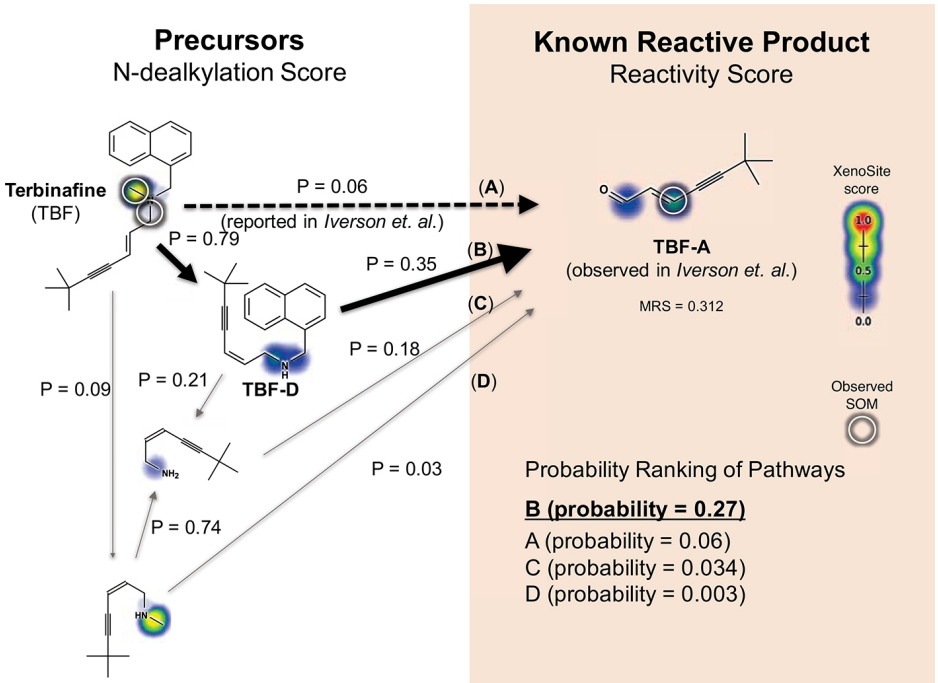


FIGURE 1.3 – **Figure extraite de la publication de Dang et al. 2018** [17] Cette figure décrit une carte du métabolisme de la terbinafine (TBF). La prédiction des sites du métabolisme a permis d'évaluer les réactions (score sur les réactions). Cette information a permis d'évaluer les probabilités de différentes voies métaboliques et de montrer que la voie tranformant le TBF en desmethyl terbinafine (TBF-D) puis en 7,7-dimethylhept-2-ene--4-ynal (TBF-A) était une voie plus probable selon cette prédiction qu'une transformation de TBF directement vers TBF-A.

plus la voie TBF vers TBF-D puis vers TBF-A, prédite grâce à cette utilisation des SOMs, est cohérente avec la littérature qui rapporte la production de TBF-D en plus de la production de TBF-A. Ce résultat démontre qu'utiliser les scores SOMs comme la probabilité qu'une réaction se produise a un sens biologique et peut permettre l'exploration des voies métaboliques d'une carte du métabolisme.

1.3 Contribution

1.3.1 Problématique de recherche

Nous rappelons que notre problématique est de déterminer, pour les AHAs, les influences des enzymes sur la formation des adduits à l'ADN. L'état de l'art des outils de prédictions du métabolisme, de la réactivité vis-à-vis de l'ADN et des sites du métabolisme nous a permis de relever deux études les plus proches de notre problématique. Tout d'abord, l'étude de Delannée et al. 2019 [19] nous a permis de montrer qu'il est possible de prédire le métabolisme et la réactivité vis-à-vis de l'ADN des AHAs en utilisant à la fois des outils prédicteurs du métabolisme, de la réactivité et prédicteurs de SOMs. Cependant une limite de ce travail est qu'il ne permet pas d'aborder l'influence ou le rôle des enzymes sur la formation des métabolites ou seulement des métabolites réactifs. En effet les cartes de cette étude [18] ne relatent que les structures des métabolites, les réactions reliant les métabolites ainsi que le nom de chaque réaction et la réactivité des métabolites par rapport à l'ADN. Il est donc impossible que le pipeline de cette étude puisse nous informer sur l'implication des enzymes dans la formation des adduits à l'ADN. Ensuite l'étude de Dang et al. 2018 nous a montré une utilisation particulière des outils prédicteurs de SOMs. Les scores de ces outils ne servent plus uniquement à identifier des atomes des molécules susceptibles d'être réactifs mais permettent de comparer des voies métaboliques au sein d'une carte. La limite principale de cette méthode par rapport à notre problématique est qu'elle ne permet pas de prédire le rôle des enzymes dans la formation des métabolites mais plutôt leurs rôles dans les voies métaboliques. Nous proposons donc dans cette thèse une nouvelle méthodologie permettant de répondre à la question du probable rôle des enzymes dans la formation des métabolites réactifs vis-à-vis de l'ADN dérivés des AHAs.

1.3.2 Un nouveau pipeline pour répondre à notre problématique

Notre contribution est un nouveau pipeline de prédiction et d'enrichissement de cartes du métabolisme. Ce pipeline est spécialisé dans la prédiction du métabolisme des xénobiotiques de par l'utilisation d'outils prédicteurs de SOM associés aux enzymes de ce métabolisme. Les cartes enrichies permettent de prédire les conditions favorisant la formation des métabolites réactifs vis-à-vis de l'ADN dérivés des AHAs.

Pour cela, nous utilisons des cartes du métabolisme des AHAs. Notre subtilité par rapport à Delannée et al. 2019 sera d'annoter ces cartes et notamment les réactions avec un score SOM. Ce score est annoté de la même manière que dans l'étude de Dang et al. 2018. Cependant, il ne servira pas à comparer la probabilité des voies métaboliques, mais à calculer la probabilité de production des métabolites. Pour cela nous utiliserons les propriétés des réseaux Bayésiens. Notre méthode consiste premièrement à prédire des cartes du métabolisme. Ensuite ces cartes sont enrichies à travers différentes annotations. Puis ces annotations nous permettent de transformer la carte du métabolisme en réseau Bayésien. Cela permet *in fine* de calculer la probabilité de production des métabolites. Cette probabilité constitue la dernière annotation des cartes enrichies. L'ensemble du pipeline regroupant ces étapes est détaillé dans le chapitre 2.

1.3.3 Applications de notre pipeline

Notre pipeline a été appliqué à 7 molécules. Tout d'abord la caféine pour valider notre approche puis à 6 AHAs d'intérêts que nous avons sélectionné.

1.3.3.1 Application à la caféine

Une fois notre pipeline développé nous l'avons appliqué à la caféine. Cette application est détaillée dans le chapitre 3. Nous avons choisi d'appliquer notre pipeline à la caféine dans le but de le valider. Nous avons choisi la caféine car c'est un xénobiotique et que son métabolisme chez l'homme implique des enzymes communes à celles responsables de l'activation métabolique des AHAs. De plus, la caféine présente un métabolisme spécifique. En effet certains métabolites peuvent être produits par différentes voies métaboliques et notre méthode peut prendre en compte la multiplicité des voies métaboliques. L'analyse de la simple prédiction du métabolisme de la caféine nous à permis de constater plusieurs résultats importants. Tout d'abord la grande majorité des métabolites de la caféine, identifiés expérimentalement, a été retrouvée. Ces résultats démontrent que l'outil SyGMa permet bien de prédire le métabolisme des xénobotiques. Ensuite nous avons pu mettre en lumière certaines lacunes dans les prédictions de SyGMa autour des réactions catalysées par les NATs. Enfin nous avons pu montrer l'intérêt de notre score de probabilité de production. En effet, celui-ci a permis de discriminer les métabolites de la caféine expérimentalement identifiés, des métabolites inconnus. Nous avons donc proposé l'utilisation de ce score comme outil de filtration des cartes du métabolisme obtenues par des outils de prédictions du métabolisme.

1.3.3.2 Application aux AHAs

Après avoir évalué notre pipeline, nous l'avons appliqué aux AHAs. Cette application est décrite dans le chapitre 4. Nous avons sélectionné six AHAs d'intérêts dont AαC, MeIQx et

PhIP qui sont les trois AHAs dont l'activation métabolique et la formation des adduist à l'ADN a été décrit dans des hépatocytes humains primaires. Tout d'abord l'analyse des métabolites prédits par notre pipeline nous a permis de confirmer que la majorité des métabolites dérivés des AHAs identifiés expérimentalement est prédite par SyGMa. Ensuite nous avons pu confirmer des lacunes de SyGMa : d'une part sur la prédiction de certaines réactions précises et d'autre part sur la prédiction des réactions associées aux N-acetyl transferases (comme nous l'avions constaté lors de l'application du pipeline à la caféine). Nous avons également pu utiliser le score de probabilité de production comme un outil de filtration. La filtration des cartes du métabolisme des AHAs a permis de mettre en lumière une quantité importante de métabolites prédits mais peu soutenus par les outils de prédictions de SOMs. Par la suite nous avons utilisé notre score de probabilité de production comme un outil permettant de répondre à notre problématique. En effet ce score permet de comparer la production de deux métabolites. On peut donc comparer deux métabolites entre eux mais également un métabolite avec lui-même dans des conditions physiopathologiques différentes. En effet nous montrerons l'influence des conditions physiopathologiques sur le calcul du score de probabilité de production lors de l'application du pipeline aux AHAs. Nous avons pu obtenir, à partir de ces comparaisons de conditions, des signatures enzymatiques qui favorisent la production de métabolites réactifs à l'ADN. Ce résultat à permis de montrer l'importance des enzymes CYP1A2 (isoforme de CYP), UGTs et SULTs dans la production des métabolites réactifs. Nous avons pu également montrer la spécificité de l'enzyme CYP3A4 dans la production des métabolites réactifs dérivés des AHAs proches de MeIQx.

1.3.4 Conclusion

L'ensemble de ce travail permet de proposer un nouveau pipeline de prédiction du métabolisme. Celui-ci sera prochainement amélioré, notamment afin de combler les lacunes identifiées dans les prédictions de SyGMa. Ce pipeline sera également appliqué aux 24 AHAs non sélectionnés dans nos travaux. Les résultats obtenus pourront être analysés à travers les données d'expressions de carcinomes hépatocellulaires issues des bases de données TCGA ou GTEx [48, 44].

Chapitre 2

Construction de cartes métaboliques enrichies. Probabilités de production.

Dans le contexte des maladies chroniques hépatiques, on s'intéresse à prédire l'activation métabolique d'une famille de contaminants de l'environnement classés comme probablement ou possiblement cancérigènes[59, 31, 26] : les (AHAs). Notre objectif est de prédire à la fois, les métabolites produits par chaque AHA, mais aussi la réactivité de ces métabolites vis-à-vis de l'ADN et les conditions physiologiques qui permettent la production du plus grand nombre de ces métabolites potentiellement réactifs. Il existe différents outils permettant de prédire l'activation métabolique, la réactivité vis-à-vis de l'ADN ou encore les sites du métabolisme (SOMs) qui décrivent les atomes réactifs vis-à-vis de certaines enzymes.

Une étude précédente de Delannée et al. 2019 [19], propose une utilisation combinée d'outils de prédictions du métabolisme, de prédictions de SOMS et de prédiction de la réactivité vis-à-vis de l'ADN pour établir des cartes du métabolisme des AHAs. Ces cartes[18] répertorient plusieurs informations comme la réaction reliant deux métabolites ou encore la réactivité du métabolite. Cependant ces cartes ne permettent pas d'évaluer la crédibilité d'un métabolite prédit par rapport à un autre, elles ne permettent pas non plus d'étudier l'influence de conditions physio-pathologiques sur la production des métabolites. Une autre étude de Dang et al. 2018[17] propose également une utilisation des prédictions des sites du métabolisme mais cette fois pour pondérer les réactions et évaluer les voies métaboliques menant au même métabolite. Cette pondération représente la chance que la réaction a de se produire.

Dans ce contexte, ce travail de thèse a consisté en l'élaboration d'un pipeline de prédiction du métabolisme des xénobiotiques. Ce pipeline est basé 1) d'une part, par la représentation du métabolisme sous la forme d'une carte du métabolisme annotée, inspirée de l'étude de Delannée et al. 2019 et 2) d'autre part par une nouvelle méthode, évaluant la *probabilité qu'un métabolite soit produit*, dans le cadre du métabolisme des xénobiotiques, à travers un score. Cette méthode

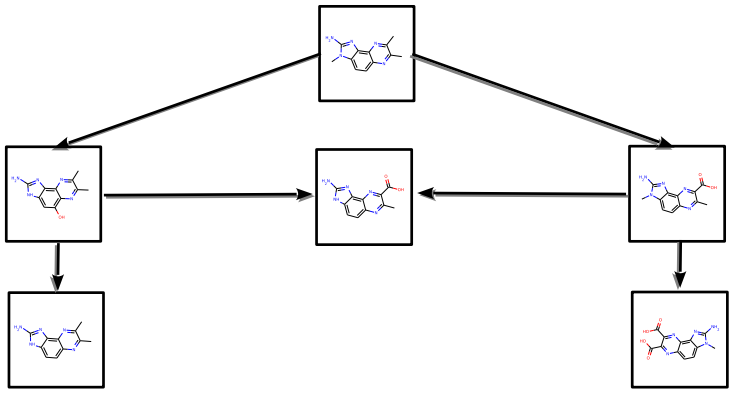


FIGURE 2.1 – **Représentation d'une carte du métabolisme** Celle-ci relie les différentes molécules, ou métabolites, par des réactions. Les métabolites sont représentés par leur structure 2D et sont encadrés. Les réactions sont représentées par des flèches.

consiste à modéliser un réseau Bayésien à partir de la carte annotée du métabolisme et d'utiliser les propriétés de ce genre de réseaux pour calculer le score de *probabilité de production*. Cette probabilité permet de comparer les métabolites d'une même carte entre eux et de les organiser du plus probable au moins probable.

Le pipeline de prédiction du métabolisme que nous proposons est divisé en trois étapes. La première étape du pipeline est la prédiction des métabolites à partir du composé d'intérêt et d'un outil de prédiction du métabolisme formant une *carte du métabolisme*. Une carte du métabolisme est un graphe, c'est-à-dire un ensemble de noeuds et d'arêtes. La figure2.1 décrit une telle carte. Les métabolites sont les noeuds du graphe et les réactions sont les arêtes dirigées partant du métabolite qui réagit et pointant vers le produit. Ensuite, vient une étape d'annotation de la carte du métabolisme. Ces annotations permettent la conversion de la carte métabolique en une structure de réseau Bayésien. La dernière étape est le calcul du score de confiance en utilisant les propriétés des réseaux Bayésiens. Ce pipeline a été appliqué dans un premier temps à la caféine utilisée preuve de concept et validation de la méthode et ensuite aux 30 AHAs connues.

2.1 Reconstruction d'une carte du métabolisme annotée

Ces cartes du métabolisme annotées sont nécessaires à la production par la suite d'un réseau Bayésien et donc du calcul du score de probabilité de production. Pour obtenir une telle carte il faut d'abord prédire une carte du métabolisme à partir d'un outil de prédiction du métabolisme et de la molécule dont on veut prédire le métabolisme sous la forme d'une formule SMILES. Cette formule est une chaîne de caractère permettant de retrouver la structure en deux dimensions d'une molécule. Les formules SMILES que nous avons utilisé proviennent de la base de

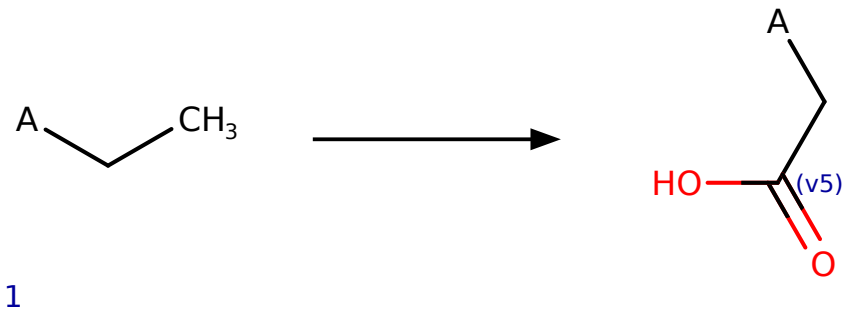


FIGURE 2.2 – **Exemple d'une réaction de carboxylation**. Représentation visuel d'une réaction de carboxylation d'un groupe éthyle, obtenue à l'aide de l'outil de visualisation MarvinView [11] à partir de la formule SMIRKS : [* :1][CH2][CH3]>>[* :1][CH2][CH](=[O])[OH]

données PubChem[35]. Une fois la carte du métabolisme obtenue nous l'annotons à partir de différents outils. Les métabolites de la carte sont annotés par un *identifiant*, la *formule SMILES* du métabolite, un *score de réactivité à l'ADN* et pour finir le *score de probabilité de production* du métabolite. Cette dernière annotation n'est obtenue qu'après avoir utilisé le modèle Bayésien pour le calculer ; les cartes annotées utilisées par ce modèle ne contiennent donc pas cette première annotation. Les réactions de la carte du métabolisme sont elles annotées par un *identifiant*, un nom de réaction ou *SMIRKS label*, une *famille d'enzymess* catalysant la réaction, un *numéro d'atome de la réaction*, un *rang*, un *score SOM* et une *enzyme*. Cette dernière annotation est plus précise que l'annotation *famille d'enzymes*. L'ensemble des annotations seront décrites lorsque l'on détaillera l'annotation des cartes du métabolisme.

2.1.1 Production de la carte du métabolisme

Notre pipeline commence par la production d'une carte du métabolisme. Pour cela il intègre et utilise un outil de prédiction du métabolisme qui prédit à la fois l'ensemble des métabolites et les réactions les produisant. L'ensemble de ces outils se base sur l'application de règles de biotransformations. Ces règles décrivent une structure chimique particulière qui est transformée en une seconde structure chimique particulière.

Par exemple une réaction de carboxylation d'un groupe éthyle est décrite par la transformation du groupe éthyle en un groupe carboxyle lié à un carbone. Elle peut être visuellement représentée ou encodée sous la forme d'une formule SMIRKS. Une formule SMIRKS décrit la réaction en

liant une structure chimique d'entrée et une structure chimique de sortie, séparées par le symbole ». Une représentation visuelle de la réaction précédente est celle de la figure2.2. On peut voir dans la représentation de cette réaction le symbole "*A*", correspondant aux caractères "*[* :1]*" de la formule SMIRKS, qui signifie qu'une autre structure chimique peut être liée à la structure de l'éthyle. De ce fait, la description de la réaction est généralisée autant par la représentation visuelle que par la formule SMIRKS.

Un ensemble de réactions ainsi encodées constitue un ensemble de règles de biotransformations. Un outil de prédiction du métabolisme recherche alors, pour chaque règle qu'il intègre, l'ensemble des structures chimiques incluses dans la structure de la molécule d'intérêt et correspondant à la structure d'entrée de la règle. À chaque fois qu'une telle structure est identifiée la règle est appliquée et un nouveau métabolite est généré. Les outils de prédictions du métabolisme diffèrent cependant de par la manière dont les règles sont constituées.

Ainsi certaines règles sont dites expertes, c'est-à-dire qu'elles proviennent de l'analyse de la littérature des auteurs de l'outil de prédiction. Les règles peuvent également être issues de l'exploration de bases de données de réactions biochimiques comme the Metabolite Database (Accerlys BioVia), DrugBank, PharmGKB ou XMETDB. Ces bases de données sont explorées à travers des outils de machine learning qui extraient, à partir des réactions chimiques, des règles de biotransformations. Il existe aujourd'hui une tendance à baser les outils de prédictions du métabolisme à la fois sur des règles de biotransformations expertes et issues de l'exploration de bases de données. Par exemple, pour le récent outil BioTransformer, plus de 100 réactions, et donc règles de transformation, ont été extraites de la littérature et ajoutées à la base de données MetXBioDB. Celle-ci est également constituée d'exports des bases de données de réactions DrugBank, PharmGKB, XMETDB, et SuperCYP.

Nous avons choisi d'utiliser l'outil de prédiction du métabolisme SyGMa (Systematic Generation of potential Metabolites)[60]. Cet outil existe sous la forme d'un packet python et permet la prédiction de métabolites produits par des réactions associées au métabolisme des xénobiotiques. SyGMa propose 176 règles de biotransformations SMIRKS divisées en deux groupes. Le premier groupe, composé de 149 règles est nommé *phase I* et répertorie un ensemble de règles associées aux réactions ayant lieux lors de la phase I du métabolisme des xénobiotiques, décrite dans la figure 1.1. Le second groupe, *phase II*, décrit 27 règles de biotransformations associées à la phase II du métabolisme des xénobiotiques. Ces règles sont issues de l'exploration de la base de données Metabolite Database maintenant connue sous le nom Accerlys BioVia, de Dassault System. Ces règles ne sont pas les seules que SyGMa prenne en compte, en effet il est possible d'ajouter un ensemble de règles, en formule SMIRKS, que SyGMa peut prendre en compte pour réaliser une prédiction du métabolisme. Dans notre pipeline, SyGMa est intégré en tant que package python, il prend en compte trois paramètres pour générer une carte du métabolisme :

— La molécule dont on cherche à prédire le métabolisme sous la forme d'une formule SMILES. Cette molécule est nommée *composé original* par la suite dans la carte du métabolisme.
— L'ensemble des règles de biotransformations à utiliser sous format SMIRKS. Ces règles peuvent être présentes directement dans l'outil sous les noms *phase I* et *phaseII* ou elles peuvent être des règles complémentaires ajoutées par l'utilisateur.
— Le nombre d'itérations de SyGMa. Ce paramètre définit le nombre d'itération de SyGMa. Par exemple si ce paramètre est de 2 alors des métabolites du composé original sont produits à partir des règles. Puis des métabolites de ces métabolites sont produits lors d'une seconde itération. À chaque itération, de nouveaux métabolites sont prédits à partir des métabolites prédits lors de l'itération précédente. Ce paramètre permet de définir le nombre de *rang* différents qui annoteront les réactions. Si une réaction est prédite lors de la première itération elle est de rang 1. Si elle est prédite à la seconde elle est de rang 2 et ainsi de suite.

Dans son application pour notre pipeline nous avons choisi d'utiliser les règles SMIRKS des groupes phase I et phase II et un rang de 2. Cela permet de reproduire le métabolisme dérivé des xénobiotiques, décrit dans la figure1.1, qui comporte une première réaction d'oxydation dite de phase I puis une seconde réaction de conjugaison dite de phase II. Les résultats de la prédiction du métabolisme de SyGMa sont alors retranscrits sous la forme d'une carte du métabolisme.

2.1.2 Annotation de la carte du métabolisme

Le score de probabilité de production est calculé pour chaque métabolite à partir d'un ensemble d'annotations permettant de générer la structure d'un réseau Bayésien à partir d'une carte du métabolisme. Les annotations nécessaires lors de cette étape sont les annotations *enzymes* et *SOM score* des réactions. Cependant il est nécessaire de passer par des annotations intermédiaires avant de produire ces annotations. Enfin l'annotation *score de réactivité à l'ADN* est une annotation facultative mais qui nous sera utile en vue d'explorer les voies de formation des adduits à l'ADN dans le chapitre 4 lorsque le pipeline sera appliqué aux AHAs. La figure 2.3 résume l'ensemble du workflow permettant d'annoter les métabolites et les réactions d'une carte du métabolisme. Un résumé des annotations sur les métabolites et les réactions est représenté dans la figure 2.4. Cet exemple est repris en fin de chapitre avec des exemples d'annotations réelles sur la même carte du métabolisme dans la figure 2.5.

2.1.2.1 Annotation des métabolites

Les métabolites sont annotés par quatres informations : un identifiant, une formule SMILES qui permet de retrouver la structure du métabolite, un score de probabilité de production et un

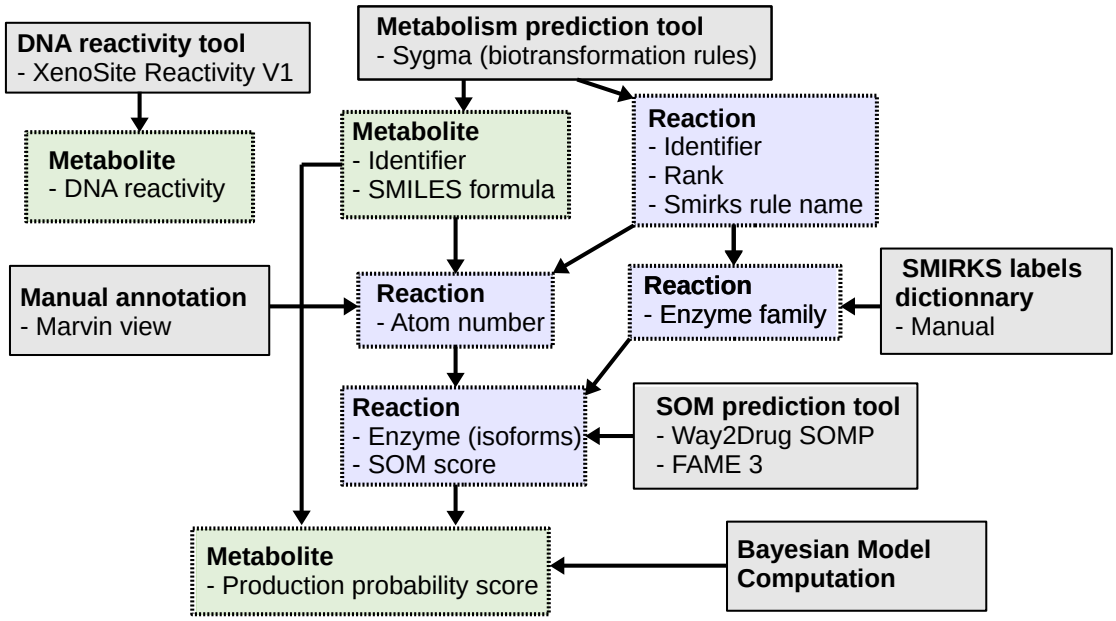


FIGURE 2.3 – **Pipeline de production des annotations des métabolites et des réactions.** Les annotations sont décrites dans les cadres verts (métabolites) et violets (réactions). Ces annotations sont obtenues par différents outils et procédés qui sont représentés par les cadres gris. Le "SMIRKS labels dictionary" est détaillé dans l'Annexe 2.

score de réactivité vis-à-vis de l'ADN qui est la probabilité que le métabolite soit réactif vis-à-vis de l'ADN.

Annotation de l'identifiant et de la formule SMILES Ces annotations proviennent directement de SyGMa. La formule SMILES est générée par le packet RDKIT[58] déjà intégré à SyGMa. L'identifiant quant à lui est un nombre qui est généré lors de la lecture des résultats de SyGMa pour former la carte du métabolisme. Chaque métabolite possède un identifiant unique.

Annotation du score de probabilité de production Ce score est le score que produit le réseau Bayésien produit à partir de la carte annotée du métabolisme. Cette annotation est donc réalisée à la fin du pipeline.

Annotation du score de réactivité Pour estimer la capacité de former des adduits à l'ADN dérivés des AHAs, nous avons cherché à évaluer la capacité des métabolites à former des adduits à l'ADN. Plusieurs outils permettent de prédire cette réactivité. Une première méthode est de rechercher des structures chimiques spécifiques à l'intérieur des structures des métabolites[46, 33, 55]. Ces structures spécifiques proviennent de structures de métabolites connus comme étant réactifs vis-à-vis de l'ADN. Une autre méthode peut être d'utiliser un score QSAR (Quantitative

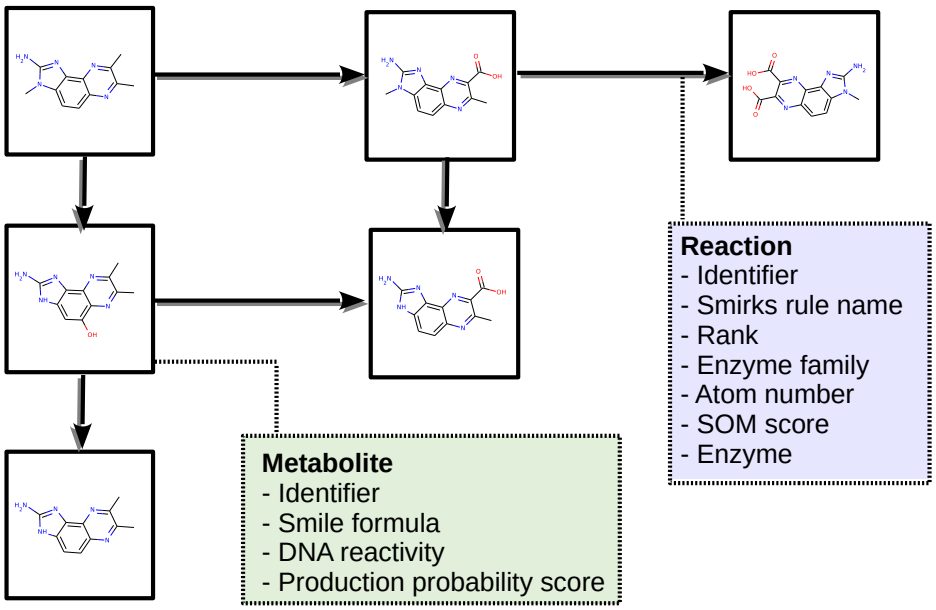


FIGURE 2.4 – **Annotations des cartes du métabolisme.** Les métabolites sont représentés par leur structure 2D dans des carrés noirs, ce sont les noeuds du graphe. Les flèches représentent les arêtes dirigées et donc les réactions. Les rectangles de couleur listent les annotations des métabolites (vert) et les annotations des réactions (violet).

Structure-Toxicity Relationship) qui est issue de l'analyse de *descripteurs moléculaires* de la molécule dont on cherche à évaluer la réactivité[46, 33, 62]. Ces descripteurs sont des éléments caractéristiques de la molécule comme l'affinité entre le ligand et son site de liaison, la lipophilie, les propriétés électroniques et stériques ou certaines caractéristiques structurelles [74]. Enfin une dernière méthode est d'évaluer la réactivité à l'ADN de chaque atome de la molécule. Pour cela on utilise des outils basés sur des méthodes d'apprentissage de type deep learning[29, 30]. Ces méthodes prédisent des sites de réactivités (SORs) en utilisant les descripteurs moléculaires caractérisant chaque atome d'une molécule. Ces outils extraient de bases de données de réactions, les réactions de fixation à l'ADN. Par la suite, ils identifient les atomes qui se sont liés à la réaction et génèrent les descripteurs moléculaires de ces atomes. Enfin, cet ensemble de descripteurs est analysé par deep learning permettant de faire émerger un ensemble de règles à appliquer aux différents descripteurs pour établir un score de réactivité. Une fois ces règles définies il ne reste plus qu'à les appliquer aux descripteurs de chaque atome d'une molécule. Chaque atome est donc associé à un score de réactivité, plus il est proche de 1 et plus l'outil prédit que l'atome est réactif. Nous avons utilisé XenoSite Reactivity v1 pour annoter nos métabolites. Puisque c'est l'ensemble du métabolite qui est annoté par un score de réactivité c'est le score le plus fort de chaque atome du métabolite qui annote le métabolite.

2.1.2.2 Annotation des réactions

Les réactions sont associées à sept annotations. Tout d'abord, comme les métabolites, elles sont associées à un identifiant unique. Elles sont également associées à un nom de réaction ou SMIRKS label. Les règles SMIRKS permettant la prédiction des métabolites sont toutes identifiées par un label, c'est le label de la règle SMIRKS qui a permis de générer la réaction qui est annotée. Ensuite la réaction est annotée par un rang qui décrit si la réaction se produit entre le composé d'origine et un métabolite ou si elle se produit entre un métabolite et un second métabolite. Cette annotation est relative au paramètre *rang* de SyGMa. La réaction est annotée manuellement par deux informations : 1° l'information *enzyme family* qui identifie la famille d'enzymes catalysant la réaction et 2) l'information *numéro d'atome de la réaction* qui identifie l'atome qui a réagi entre la molécule d'entrée de la réaction et la molécule produite. Enfin ces deux annotations sont utilisées pour annoter les informations *SOM score* et *enzymes* qui sont les annotations dont a besoin le modèle Bayésien.

Annotation de l'identifiant, du SMIRKS label et du rang Comme pour les métabolites ces annotations proviennent directement de SyGMa. L'identifiant est directement généré lors de la lecture des résultats de SyGMa pour générer la carte du métabolisme. Le SMIRKS label est directement relié à chaque réaction prédite par SyGMa et est extrait des résultats. Quant au rang, il est défini par l'analyse de la carte obtenue par SyGMa ; si la réaction a pour molécule d'entrée le composé d'origine elle est de rang 1, sinon elle est de rang 2. Le rang permettra plus tard d'établir l'annotation *enzyme* et *SOM score*.

Annotation de la famille d'enzymes Cette annotation identifie la famille d'enzymes du métabolisme des xénobiotiques qui peuvent catalyser cette réaction. Pour obtenir cette annotation nous utilisons un dictionnaire "*SMIRKS labels to enzymes*", décrit dans l'Annexe 2. Ce dictionnaire a été réalisé manuellement à partir de la description du métabolisme des xénobiotiques décrit dans la figure1.1. Il s'agit d'un ensemble d'équivalences qui, à chaque SMIRKS labels, associe une famille d'enzymess. Les SMIRKS labels associés aux règles SMIRKS du groupe *phase I* de SyGMa sont associés aux CYPs qui sont les enzymes qui catalysent les réactions de phase I chez les AHAs. Parmi les 27 SMIRKS labels du groupe des enzymes de phase II plusieurs groupes ont été définis. Un groupe de 13 SMIRKs labels a été associé aux UGTs, un autre de 5 SMIRKS labels a été associé aux NATs, 6 SMIRKS labels ont été associés au SULTs et 1 SMIRKS label a été associé au GSTs. Les 2 derniers SMIRKS labels contenus dans le groupe phase II n'ont été associés à aucune enzyme de la phase II du métabolisme des xénobiotiques. Les réactions ont donc été annotées suivant ce dictionnaire, dans le cas où la réaction est annotée par un des SMIRKS labels associé à aucune enzyme, les réactions sont éliminées de la carte du métabolisme annotée car elles ne permettent plus l'application de la méthode Bayésienne. Lorsqu'une réac-

tion est éliminée, elle entraîne l'élimination de l'ensemble des métabolites et réactions isolées, c'est-à-dire qui ne sont plus reliés au composé d'origine.

Annotation de l'atome de la réaction Chaque réaction est également caractérisée par un numéro d'atome de la réaction, c'est-à-dire l'index de l'atome de la molécule d'entrée qui réagit lorsque la réaction se produit. Cette annotation est obtenue manuellement à partir de la comparaison des structures 2D des molécules d'entrée et de sortie de chaque réaction. Pour cela nous avons utilisé l'outil MarvinView[11] qui permet la représentation d'une structure chimique en deux dimensions à partir d'une formule SMILES. Une fois l'atome identifié visuellement nous avons extrait son numéro d'index qui peut être affiché par MarvinView. Nous avons choisi l'index issu d'une numérotation normalisée suivant les règles de l'International Union of Pure and Applied Chemistry (IUPAC)[22]. En effet cet index est le même que celui que l'on retrouve dans les résultats d'outils de prédictions de site du métabolisme qui permettront de générer l'annotation *enzyme* et *SOM score*.

Annotation du SOM score et de l'enzyme Ces deux annotations sont les annotations permettant le calcul par le modèle Bayésien d'un score de probabilité de production pour chaque métabolite.

L'annotation SOM score est une annotation inspirée de l'étude de Dang et al. 2018 [17] qui utilisait les résultats de XenoSite Dealkylation, un outil prédisant les sites du métabolisme (SOMs), comme probabilité que la réaction se produise. Cela a permis à cette équipe de comparer la probabilité de chaque voie métabolique partant de la Terbinafine (TBF) et menant à la production de 7,7-dimethylhept-2-ene-4-ynal (TBF-A). Dans la même idée nous annotons nos réactions avec un score issu d'outils de prédictions de SOMs.

Les outils prédicteurs de SOMs sont utilisés afin de déterminer, pour chaque atome d'une molécule, un score représentant la probabilité que cet atome soit un site du métabolisme. Un site du métabolisme est ici l'atome qui réagit pour un ensemble de réactions données. Par exemple, dans le cas de la réaction montrée en exemple de la figure2.2 c'est le carbone à l'extrémité de la molécule de droite qui est l'atome qui réagit en passant d'un groupe méthyel à un groupe carboxyle. Ce carbone est un site du métabolisme.

Il existe différentes stratégies pour déterminer un score évaluant la capacité d'un atome à être un site du métabolisme. Tout d'abord certains outils [28, 2, 63, 64, 15, 14] utilisent le paramètre d'énergie nécessaire à la réalisation de la réaction d'abstraction d'un atome d'hydrogène. Ce paramètre est un bon indicateur de SOM pour des réactions catalysées par les CYPs. D'autres outils comme Way2Drug SOMP, FAst MEtabolizer (FAME) ou XenoSite Metabolism 1.0 [61, 37, 65] utilisent des descripteurs moléculaires. C'est un ensemble de paramètres décrivant un

atome comme le type d'atome (carbone, azote, oxygène...), la charge de l'atome, les atomes voisins ou le nombre de liaisons formées par l'atome. De la même manière que la prédiction de SORs par XenoSite Reactivity v1 [29, 30] ces outils isolent les atomes étant site du métabolisme d'un ensemble de réactions. Puis en décrivant ces atomes par des descripteurs moléculaires et en utilisant des méthodes d'apprentissage automatique, ces outils permettent de définir des règles de calcul d'un score prédictif des SOMs du sous ensemble de réactions considérées. Ce score représente la probabilité de l'atome d'être un SOM. En choisissant un ensemble de réactions catalysées par une isoforme d'enzyme précise ou par une famille enzymatique, le score SOM devient la probabilité que l'atome soit catalysé par l'isoforme ou la famille d'enzymes considérée. Enfin une dernière stratégie est d'utiliser des méthodes de docking associées aux similarités entre la structure du ligand et la structure de la molécule d'intérêt.

Nous avons choisi d'utiliser les outils FAME 3 [65] et Way2Drug SOMP[61] afin de couvrir un maximum d'enzymes de phases I et II du métabolisme des xénobiotiques. FAME 3 a permis de déterminer les scores SOMs associés aux enzymes CYPs, UGTs, NATs, SULTs et GSTs. Way2Drug SOMP a permis de déterminer les scores SOM des UGTs mais également des différentes isoformes de CYPs : CYP1A2, CYP2C19, CYP2C9, CYP2D6 et CYP3A4.

L'annotation par un score SOM des réactions est déterminée par : le numéro de l'atome de la réaction, l'annotation de famille d'enzymes et le rang. Tout d'abord le rang influence le choix de l'outil de prédiction des SOMs. Comme deux outils différents sont utilisés, pour couvrir un maximum d'enzymes du métabolisme des xénobiotiques, il n'est pas possible d'établir un score prenant en compte les résultats des deux outils pour les CYPs et UGTs. Nous avons fait le choix d'annoter les réactions de rang 1 par Way2Drug et les réactions de rang 2 par FAME 3. De cette manière, les réactions de rang 1 sont annotées avec plus de finesse pour les isoformes des CYPs. À l'inverse les réactions de rang 2 sont mieux annotées pour les réactions catalysées par les SULTs, NATs et GSTs.

Nous avons fait l'hypothèse que les réactions de rang 1 seront majoritairement des réactions associées aux enzymes de phase I tandis que la majorité des réactions associées aux enzymes de phase 2 seraient de rang 2. Cette hypothèse se base sur la description du métabolisme des xénobiotiques commençant par une réaction associée aux enzymes de phase I dite phase d'oxydation puis suivi d'une réaction de phase II dite de conjugaison avant une phase III consistant en l'élimination du xénobiotique oxydé et conjugué comme décrit dans la figure 1.1. Ainsi pour les réactions majoritairement associées aux enzymes de phase I du métabolisme des xénobiotiques nous proposons une fine prédiction des SOMs des CYPs avec un score SOM pour chaque isoforme proposé par Way2Drug SOMP. FAME 3 quant à lui permet de prendre en compte les

réactions associées aux NATs, SULTs et GSTs de rang 2 et donc d'être plus exhaustif que les scores SOMs UGTs que propose Way2Drug SOMP.

Ainsi pour chaque réaction de la carte on détermine l'outil à utiliser grâce au rang de la réaction, si l'annotation *famille d'enzymess* de la réaction propose une famille que l'outil peut prédire alors on applique la prédiction des SOMs à partir de la formule SMILES du métabolite d'entrée de la réaction. L'atome de la réaction permet d'identifier le score SOM qui doit être sélectionné. Le score SOM de l'atome de la réaction annote alors la réaction.

L'annotation *enzyme* est déterminée en même temps que le score SOM. Soit la famille d'enzymes de la réaction est UGTs, NATs, SULTs ou GSTs et dans ce cas l'annotation est la même que pour l'annotation *famille d'enzymes*. Soit la famille d'enzymes est la famille des CYPs. Dans ce cas, si la réaction annotée est de rang 1 alors elle est divisée en cinq réactions. Chacune est annotée par une isoforme de CYPs différente que propose Way2Drug SOMP. L'annotation SOM score change également car Way2Drug propose un score SOM pour chaque isoforme. Pour les réactions de rang 2 elles sont divisées et annotées par l'ensemble des isoformes des CYPs annotant les réactions de rang 1. En effet le score de FAME 3 n'étant pas exhaustif on considère que l'ensemble des isoformes se valent mais seules les isoformes qui annotent les réactions de rang 1, sont ajoutées.

L'ensemble des réactions qui n'ont pas pu être annotées par Way2Drug ou FAME 3 sont éliminées de la carte du métabolisme annotée. Comme pour l'annotation *famile d'enzymes* l'ensemble des réactions et métabolites isolés suite à cette élimination sont également éliminés.

Une fois l'ensemble de ces annotations appliquées à la carte du métabolisme, la carte est suffisamment enrichie pour permettre le calcul des scores de probabilité de production des métabolites. Un exemple concret des valeurs des annotations est représenté dans la figure 2.5.

2.2 Calcul du score de probabilité de production à partir de propriétés des réseaux Bayésiens

Le score de probabilité de production est un score permettant de représenter la probabilité qu'un métabolite, d'une carte du métabolisme, soit produit. Ce score est construit à partir des annotations des cartes du métabolisme précédentes et du formalisme réseaux Bayésien [34]. Le calcul de ce score est influencé par ce que nous définirons comme des *contextes enzymatiques* et peut permettre de comparer l'influence des enzymes sur ce score de probabilité de production.

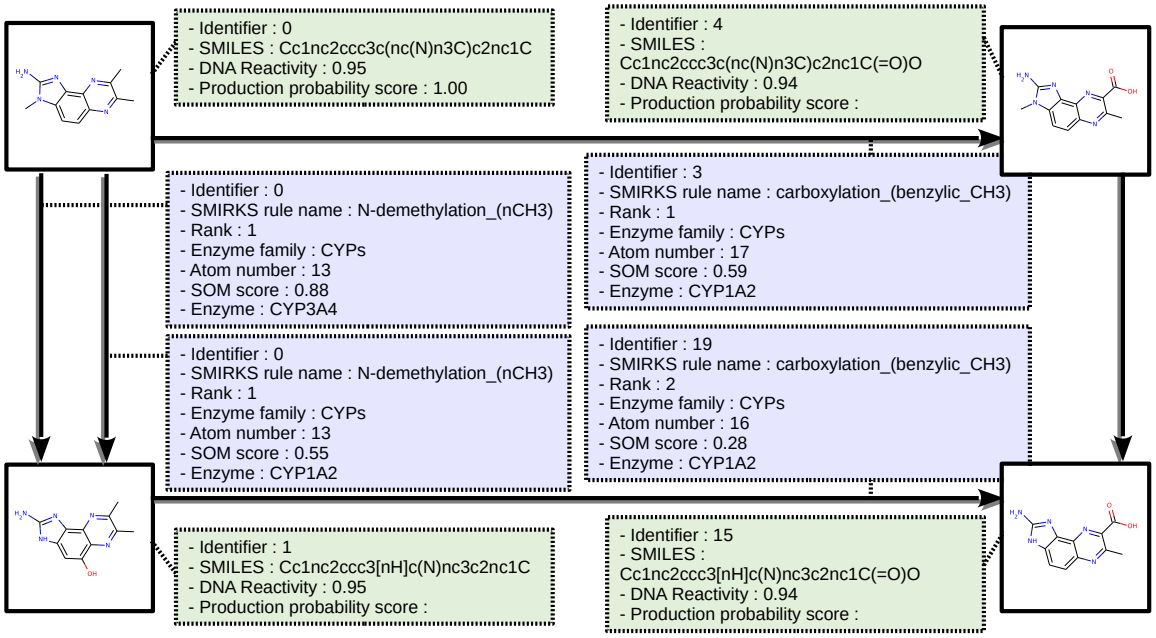


FIGURE 2.5 – **Exemple d'une carte du métabolisme annotée.** Cet exemple de carte du métabolisme annotée est un extrait de la carte de la figure 2.4, elle reprend les mêmes codes et donne des exemples concrets des valeurs que peuvent prendre les annotations.

2.2.1 Les réseaux Bayésiens

Les réseaux Bayésiens [42, 34]sont une modélisation mathématique des connaissances sous la forme d'un modèle graphique probabiliste. Ce modèle représente les connaissances sous la forme de variables, influencées par d'autres variables ce qui se traduit par une arête orientée de la variable influente vers la variable influencée. Ces réseaux permettent le calcul des probabilités des valeurs des variables à l'aide de tableaux de probabilités conditionnelles. L'exemple classique des réseaux Bayésien est celui du réseau *Asia* proposé par Lauritzen et Spiegelhalter en 1988 [42], ce réseau est présenté dans la figure2.6.

Il décrit les relations entre différentes variables ou *évènements* dans le contexte d'un patient présentant une maladie pulmonaire. Il est construit afin de prédire si cette maladie pulmonaire est plus probablement une bronchite, la tuberculose ou un cancer du poumon. Chaque variable est associée à deux valeurs *yes* ou *no*, et chaque valeur est associée à une certaine probabilité de se réaliser. Modifier cette probabilité, par exemple en répondant *yes* à la question *Visit to Asia ?*, influence la probabilité des valeurs des autres variables [5]. Dans ce cas cette variable influence directement l'évènement *Tuberculosis ?* en augmentant la probabilité de la valeur *yes* de cette variable. Cela traduit le fait qu'il est plus probable qu'un patient revenant d'Asie présente une tuberculose plutôt qu'un patient qui n'en revient pas.

Les probabilités dans les réseaux Bayésien sont des probabilités dites conditionnelles, c'est-

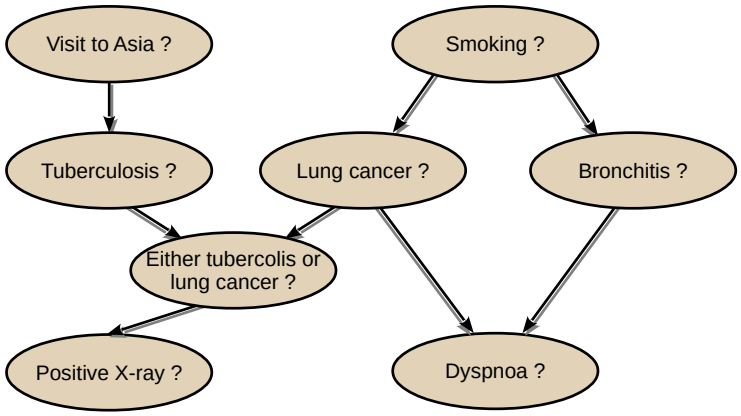


FIGURE 2.6 – **Représentation du réseau bayésien Asia** Ce réseau décrit des variables, représentées par des ovales, et les relations entre ces variables représentées par des flèches. Le réseau original a été proposé par Lauritzen et Spiegelhalter en 1988 [42] et décrit les influence de certaines variables sur d'autre dans un contexte de maladie pulmonaire.

Probabilité des valeurs de la variable *Tuberculosis ?*	Asia	
	yes	no
P(yes)	0,05	0,01
P(no)	0,95	0,99

TABLE 2.1 – **Table des probabilités conditionnelles de la variable *Tuberculosis ?*.** Cette table répertorie la probabilité que la variable *Tuberculosis ?* prenne la valeur *yes* ou *no* en fonction de la valeur de la variable *Visit to Asia ?* dans le réseau bayésien Asia.

à-dire la probabilité que la variable A prenne la valeur Y en connaissant la valeur d'autres variables influençant directement A. Dans l'exemple présenté la probabilité *yes* à l'évènement *Tuberculosis ?* est conditionnée par l'évènement *Visit to Asia ?*. Cela se traduit par une probabilité différente de la valeur *yes* pour l'évènement *Tuberculosis ?* selon si la valeur de l'évènement *Visit to Asia ?* est *yes* ou *no*. Dans un réseau Bayésien ces influences entre variables sont écrites sous la forme de table de probabilités conditionnelles. La table 2.1 décrit les probabilités conditionnelles de la variable *Tuberculosis ?* en fonction des valeurs que prendrait *Visit to Asia ?*. Ainsi la probabilité que *Tuberculosis ?*, prenne la valeur *yes*, $P(yes)$, lorsque la valeur de *Visit to Asia ?* est *yes* est de 0,05 tandis que cette probabilité est de 0,01 lorsque la valeur de *Visit to Asia ?* est *no*. Ces tables peuvent être plus complexes en fonction du nombre de variables impactant la probabilité de la variable dont on écrit la table.

L'intérêt d'une telle représentation dans le cas du réseau *Asia* est de permettre, en établissant les valeurs de certains évènements, d'établir un diagnostique pour le patient. Pour cela le réseau peut calculer, à partir des tables de probabilités conditionnelles et de la valeur de certaines variables, la probabilité que la variable *Tuberculosis* prenne la valeur *yes*.

2.2.2 Application des Modèles Bayésiens aux cartes du métabolisme des xénobiotiques

La strucutre du réseau Asia, figure 2.6, rappelle celle des cartes du métabolisme, figure 2.1. L'intérêt d'appliquer une représentation en réseau Bayésien aux cartes du métabolisme est de calculer la probabilité de chaque variable, ici les métabolites. Pour calculer la probabilité des noeuds il est nécessaire d'avoir des probabilités conditionnelles. Nous proposons de représenter les cartes du métabolisme annotées sous la forme de réseaux Bayésiens où les métabolites seraient des variables pouvant prendre la valeur *produit* ou *non produit*. De cette manière, calculer la probabilité d'un noeud revient à calculer la probabilité qu'un métabolite soit produit. Cette probabilité, c'est le score de *probabilité de production* que nous avons présenté précédemment.

2.2.2.1 Compatibilité des structures des cartes du métabolisme et des réseaux Bayésien

Pour calculer ce score il est d'abord nécessaire de transformer la carte du métabolisme en réseau Bayésien.

Tout d'abord un réseau Bayésien est un *directed acyclic graph* ou DAG, c'est-à-dire un graph où les arêtes sont orientées et ne forment pas de cycles. Ici les cartes prédites par SyGMa sont orientées, puisque les métabolites proviennent de réactions. Une réaction consomme un métabolite et produit un autre métabolite, elle est donc orientée. Les réactions sont orientées donc l'ensemble de la carte est également orientée. De plus, aucun cycle ne devrait être formé du fait du paramètre rang qui a pour valeur 2. Il reste le risque qu'une réaction renverse une

précédente réaction, par exemple une méthylation suivie par une déméthylation ce qui reviendrait à revenir au métabolite initial. Cependant, ce cas de figure ne s'est présenté sur aucune des 31 cartes que nous avons générées.

Ensuite pour construire un tel réseau il faut des variables, ou évènements, et des probabilités conditionnelles associées. Les métabolites peuvent être considérés comme les variables prenant la valeur *produit* ou *non produit*. Les probabilités conditionnelles peuvent être les scores SOMs annotant les réactions. En effet jusqu'à présent ce score représente la probabilité que la réaction se réalise donc la probabilité que le métabolite de sortie de la réaction soit produit en sachant que le métabolite d'entrée est également produit.

La structure des cartes du métabolisme est donc compatible avec la structure des réseaux Bayésiens. Transformer ces cartes en réseaux Bayésiens permet le calcul du score de probabilité de production de chaque métabolite.

2.2.2.2 Construction du réseau Bayésien à partir de la carte du métabolisme annotée

Tout d'abord, pour chaque métabolite de la carte une variable du modèle Bayésien a été construite pouvant prendre la valeur *produit/non produit*. Les tables des probabilités conditionnelles sont produites à partir des scores SOM annotant les réactions. Elles déterminent que la probabilité qu'une variable métabolite prenne la valeur *produit*, sachant que le métabolite précédent est lui même produit. Cependant un problème se pose, il est impossible de déterminer l'effet que deux réactions produisant le même métabolite peuvent avoir l'une sur l'autre. En effet, dans le cas des réactions qui ont été multipliées suite à l'annotation des isoformes de CYPs, il existe plusieurs réactions qui utilisent le même métabolite d'entrée. Cela rend ces réactions interdépendantes et le fait que l'une se réalise peut avoir un impact sur la réalisation de l'autre. De même, il peut exister des réactions qui produisent le même métabolite mais agissent sur les mêmes enzymes ce qui peut entraîner un effet d'une réaction sur une autre et cela même si le métabolite d'entrée est différent. Cet ensemble de cas où il est impossible de déterminer l'influence des réactions les unes envers les autres se traduit par l'incapacité à remplir la probabilité de certaines conditions précises dans la table de probabilité conditionnelle.

La figure 2.7 montre un exemple de carte du métabolisme tel que l'on peut l'observer dans les 31 cartes générées. Dans le cas du métabolite F, il est produit à travers les réactions consommant soit C soit D. On simplifie cet exemple dans un premier temps en ne considérant que deux réactions, celles annotées par E3. Lorsque l'on doit reconstruire la table des probabilités conditionnelles, décrite Table 2.2 on se retrouve dans l'impossibilité de remplir la probabilité que F soit produit, dans le cas où C et D sont produits. Et cela parce que l'enzyme catalysant ces réactions est la même.

Une possibilité pour résoudre ces cas impossibles est de rendre indépendantes les réactions

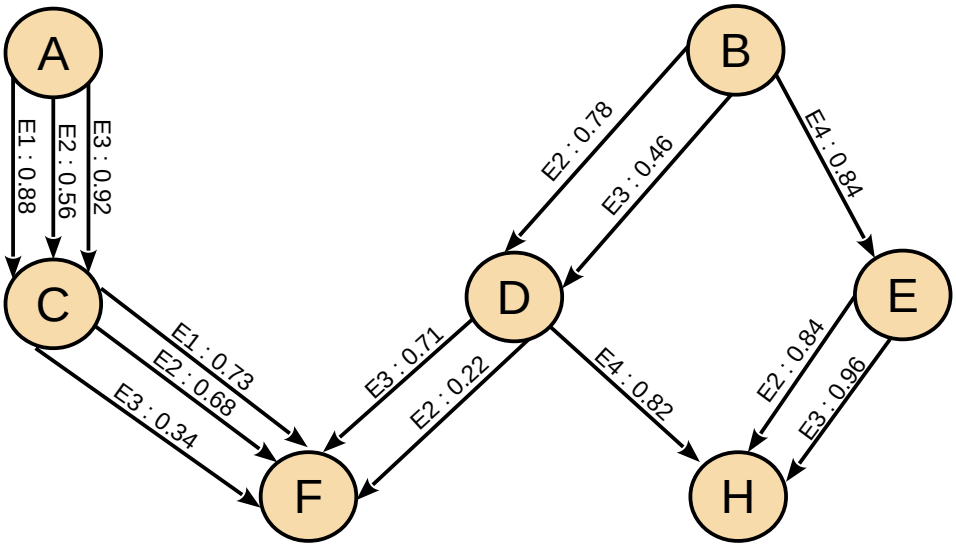


FIGURE 2.7 – **Exemple d'une carte du métabolisme annotée**. Cette carte est un exemple de carte du métabolisme où les métabolites sont identifiés par des lettres et les SOM score annotant chaque réaction sont identifiés sur les flèches représentant les réactions.

		D	
		Produit	Non Produit
C	Produit	? ?	0,34
	Non Produit	0,71	0,00

TABLE 2.2 – **Probabilités conditionnelles de la variable F de la figure 2.7** Cette variable est associée au métabolite F issue de la même figure et peut prendre la valeur *produit* ou *non produit*. Les probabilité de la table décrive la probabilité que la variable F prenne la valeur *produit*.

		D	
		Produit	Non Produit
C	Produit	0,92	0,73
	Non Produit	0,71	0,00

TABLE 2.3 – **Table des probabilités conditionnelles de la variable F, complétée à partir des SOMs scores de deux réactions indépendantes.** Cette variable est associée au métabolite F issue de la figure 2.7 et peut prendre la valeur *produit* ou *non produit*. Les probabilité de la table décrive la probabilité que la variable F prenne la valeur *produit* et on pu être complétée en utilisant les scores SOM annotant des réactions indépendantes.

lorsque l'on remplit la table des probabilités conditionnelles.

Calcul des probabilités conditionnelles simple Une modélisation du métabolisme a été définit afin de rendre indépendantes les réactions générant des cas insolubles. Si les réactions sont indépendantes alors on peut remplir la table de probabilités conditionnelles comme la table 2.3 qui prend en compte deux réactions indépendantes, celle annotée par E3 et consommant D et celle annotée par E1 et consommant C. La probabilité que F soit produit avec C *produit* et D *non produit* devient $P(F|C,\ D) = P(F|C)$. Dans cette formule représente la valeur *non produit* des variables, ici C peut être traduit par *C est produit* tandis que D peut être traduit par *D est non produit*. Le symbole | représente la condition, $P(F|C,\ D)$ se lit donc *probabilité de produire F sachant que C est produit et D est non produit*. De même la probabilité de $P(F|D,\ C)$ est la même que $P(F|D)$ du fait de l'indépendance de la production de F via C et D.

Calcul de la probabilité conditionnelle la plus complexe La probabilité la plus complexe à résoudre est la probabilité $P(F|C, D)$, qui est donc la probabilité d'obtenir F en sachant que C et D sont produits. Cette probabilité ne peut pas être résumée au produit de $P(F|C)$ et $P(F|D)$ car ce produit représente la probabilité d'obtenir F à la fois par C et par D ce qui est différent d'obtenir F via C, D ou les 2.

Résolution du calcul de la probabilité à travers l'exemple des pièces de monnaie Pour résoudre ce problème on peut temporairement remplacer chaque évènement, ainsi C serait un lancé d'une pièce de monnaie équilibrée et D serait un lancé d'une autre pièce de monnaie équilibrée. Chaque évènement est associé à deux valeurs *la pièce a été lancée*. F serait alors l'évènement *obtenir pile*, $P(F|C, D)$ est alors la probabilité d'obtenir pile en ayant deux lancés indépendants de pièces, C et D. La probabilité d'obtenir pile, $P(F|C, D)$ est la somme des probabilités des cas de figure où au moins un pile est obtenu. Cette probabilité est la même que $1 - P(\ F|C, D)$ c'est-à-dire 1 - la probabilité contraire de l'évènement *obtenir pile* c'est-à-dire ne jamais obtenir pile. En utilisant cet exemple on peut à présent résoudre $P(F|C, D)$ dans le cas des métabolites. Cette probabilité est la même que $1 - P(\ F|C, D)$ c'est-à-dire 1 moins la

probabilité que F soit *non produit*. Cette probabilité peut être traduite par $1 - P(\ F|C, D) = 1 - [P(\ F|C) * P(\ F|D)]$ de plus $P(\ F|C) = 1 - P(F|C)$ et $P(\ F|D) = 1 - P(F|D)$. Or on connaît les valeurs de $P(F|C)$ et de $P(F|D)$, il s'agit du score SOM annotant les réactions de C vers F et de D vers F respectivement. On retrouve alors que $P(F|C, D) = 1 - 0,27 * 0,29 = 0,9217$. On retrouve bien la valeur de la table 2.3.
Rendre les réactions indépendantes les unes avec les autres dans la carte du métabolisme permet donc d'établir pour chaque métabolite une table des probabilités conditionnelles et de construire un modèle Bayésien.

Rendre les réactions indépendantes Rendre les réactions indépendantes implique dans le cas de l'exemple de la figure 2.7 que les réactions multiples consommant et produisant les mêmes métabolites, comme les réactions de C vers F, doivent être réduites à une unique réaction. Ainsi les réactions deviennent indépendantes en ce qui concerne le métabolite consommé. En revanche il reste les réactions liées par les enzymes qui les catalysent. Il faut donc que des réactions produisant le même métabolite ne soient pas catalysées par les mêmes enzymes.

Règles de réductions pour construire un modèle Bayésien Nous proposons une série de règles permettant de réduire la carte du métabolisme afin de rendre indépendantes les réactions et de permettre la production d'un réseau Bayésien à partir d'un carte du métabolisme annotée. Cette série de règles se base sur l'hypothèse que les scores SOM représentent l'affinité du métabolite et de l'enzyme pour une réaction donnée. Lorsqu'il faut choisir une réaction parmi plusieurs lors de la réduction on choisit de maximiser cette affinité, c'est-à-dire le score SOM. L'application de ces règles à un exemple de carte du métabolisme est illustré dans la figure 2.8.

La première règle permet de réduire les multiples réactions consommant et produisant les mêmes métabolites. Ces réactions ne diffèrent que par les annotations *enzyme* et *SOM score*. Notre méthode conserve, sur cet ensemble de réactions uniquement la réaction associée au score SOM le plus fort, les autres réactions sont éliminées. La figure 2.8B correspond à ce cas de figure, des trois réactions provenant de A celle qui est retenue est la réaction associée à l'enzyme E3 et au SOM score $0,92$ qui est le score le plus fort.

La seconde règle concerne le cas des réactions qui produisent le même métabolite. Ici les réactions proviennent de différents métabolites. Notre méthode repère, parmi les SOMs scores de l'ensemble des réactions, le SOM score le plus fort. La réaction associée à ce score est alors retenue et l'ensemble des réactions provenant du même métabolite est éliminé. C'est le cas de la figure 2.8C pour les réactions de C vers F et de D vers F. Le score SOM le plus fort est $0,77$ et est associé à la réaction de D vers F annotée par l'enzyme E1. L'autre réaction de D vers F est éliminée. Ensuite l'ensemble des autres réactions qui sont annotées par la même isoforme d'enzyme ou enzyme sont également éliminées. Ici la réaction C vers F annotée par E1 est donc éliminée. De cette manière l'ensemble des autres réactions deviennent indépendantes de

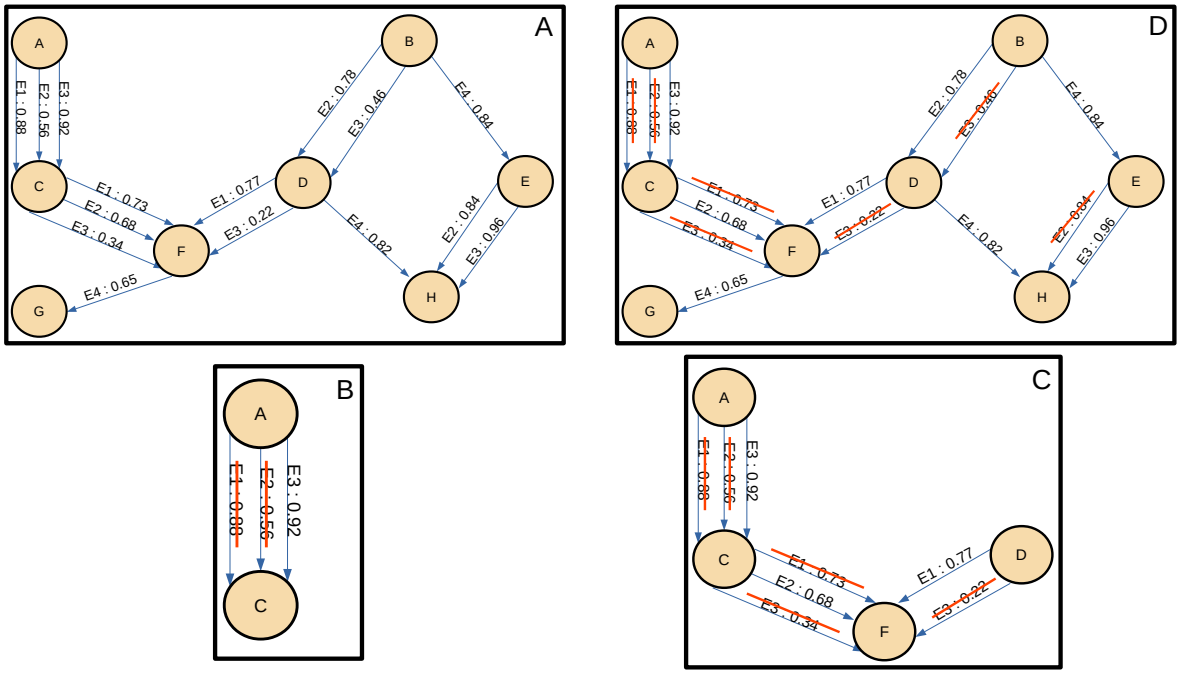


FIGURE 2.8 – **Application des règles de réduction des cartes du métabolisme annotées en vue de construire un modèle Bayésien** Cette figure montre la réduction d'une carte du métabolisme annotée (A) dont nous avons fait apparaître l'annotation de l'enzyme, ici E1, E2, E3 et E4, et du SOM score. La figure (B) illustre l'application de la première règle de réduction qui dit qu'une seule réaction peut consommer et produire les mêmes métabolites. Il y a une réaction consommant A et produisant C qui a été divisée en trois réactions, une pour chaque enzyme pouvant annoter la réaction. La réaction associée au score SOM le plus haut (0,68) a été sélectionnée. La figure (C) illustre l'application de la seconde règle de réduction qui concerne les réactions produisant le même métabolite mais consommant un métabolite différent. Dans cet exemple c'est le cas des réactions de C vers F et D vers F. Ici le score SOM maximal est 0,77 associé à une réaction de D vers F et annotée par l'enzyme E1. Cette réaction est retenue comme la seule réaction de D vers F, la réaction annotée par E3 de D vers F est éliminée. Ensuite les autres réactions annotées par E1 sont éliminées, ici c'est la réaction de C vers F annotée par E1 qui est éliminée. Il ne reste que deux réactions de C vers F, la première règle est appliquée et c'est la réaction annotée par E2 avec un score SOM de 0,68 qui est retenue. La figure (D) résume la réduction complète de la carte de la figure (A).

la première réaction sélectionnée. Une fois tout ce processus effectué, le processus recommence jusqu'à ce que chaque réaction produisant le même métabolite soit indépendante des autres. Lorsqu'il ne reste qu'un seul groupe de réactions à réduire qui proviennent toutes du même métabolite alors c'est la première règle qui s'applique. Dans notre exemple il ne reste plus qu'à réduire les réactions provenant de C et allant vers F restantes. Le score SOM le plus fort entre les deux réactions restantes est 0,68 et est associé à l'enzyme E2. C'est la réaction avec ces deux annotations qui est retenue. La figure 2.8D montre la carte du métabolisme complètement réduite et pouvant être utilisée comme un modèle Bayésien.

Une fois que la carte a été réduite, les scores SOM ont été interprétés comme des probabilités conditionnelles. Ainsi la table de probabilités conditionnelles de chaque métabolite a pu être

calculée.

Calcul du score de probabilité de production des métabolites. Le score de probabilité de production des métabolites est la probabilité que la variable associée à ce métabolite, dans le modèle Bayésien, prenne la valeur *produit*. Les propriétés du réseau Bayésien nous permettent de calculer une telle probabilité, il s'agit de la somme des probabilités jointes où le métabolite prend la valeur *produit*[34].

Pour décrire ce que sont les probabilités jointes nous allons prendre un exemple d'un petit réseau, celui de la figure2.8C. Ce réseau décrit 4 métabolites qui peuvent être *produit* ou non. On peut alors imaginer une combinaison de valeurs *produit* ou non pour chaque métabolite. Par exemple il se peut que A, C et D soient produits mais pas F. Les probabilités conditionnelles permettent de calculer la probabilité d'obtenir cette série de valeurs. Il s'agit de $P(A, C, D,\ F)$ que l'on peut lire *probabilité que A, C et D soient produits et que F soit non produit.* l'application des formules de Bayes permet de décomposer cette probabilité en $P(A, C, D,\ F) = P(A) * P(C|A) * P(D) * P(\ F|C, D)$. L'ensemble de ces probabilités peuvent être calculées à travers les tables des probabilités conditionnelles. La probabilité de A et D est paramétrée à 1.0 car ce sont des variables qui ne sont pas produites par une réaction. Dans les cartes que nous avons générées, un seul métabolite possède ce critère, il s'agit du composé d'origine dont la probabilité d'être produit est donc fixée à 1. Pour les autres probabilités on se réfère aux tables de probabilités conditionnelles. Ainsi $P(A) * P(C|A) * P(D) * P(\ F|C, D) = 1,0 * 0,92 * 1,0 * [(1 - P(F|C)) * (1 - P(F|D))]$. On obtient alors : $P(A, C, D,\ F) = 0,92 * [(1 - 0,68) * (1 - 0,77)] = 0,067712..$ Cette probabilité est la probabilité que A, C et D soient produits mais pas F. Chaque combinaison des valeurs *produit/non produit* peut être associée à une probabilité. La probabilité globale de F est donc la somme des probabilités des combinaisons de valeurs *produit/non produit* où F est produit. Dans cet exemple $P(F) = P(A, C, D, F) + P(A, C,\ D, F) + P(A,\ C, D, F) + P(A,\ C,\ D, F) + P(\ A, C, D, F) + P(\ A, C,\ D, F) + P(\ A,\ C, D, F) + P(\ A,\ C,\ D, F)$. Les probabilités que A et D soient produits sont paramétrées à 1,0 donc l'ensemble des combinaisons avec A ou D ont une probabilité de 0.0. Donc $P(F) = P(A, C, D, F) + P(A,\ C, D, F)$ c'est-à-dire $P(F) = 0,629952 + 0,2464 = 0,8776352$. Dans cet exemple le score de probabilité de production de F est donc de $0,8776352$ qui peut être arrondi à $0,88$.

Le nombre de combinaisons des valeurs *produit/non produit* peut vite devenir grand avec un réseau de grande taille cependant on peut démontrer qu'il n'est pas nécessaire de prendre en compte un ensemble de combinaisons exhaustives mais seulement un sous ensemble par métabolite. En effet si l'on reprend l'exemple de la figure 2.8C en y ajoutant le noeud G de la figure 2.8C et que l'on cherche à calculer le score de probabilité de production de F. Alors on sait depuis l'exemple précédent que $P(F) = P(A, C, D, F, G) + P(A,\ C, D, F, G) + P(A, C, D, F,\ G) + P(A,\ C, D, F,\ G)$. En décomposant ces probabilités on obtient $P(F) = P(A, C, D, F) * P(G|F) +$

$P(A,\ C, D, F) * P(G|F) + P(A, C, D, F) * P(\ G|F) + P(A,\ C, D, F) * P(\ G|F)$. On peut alors factoriser par $P(G|F)$ et $P(\ G|F)$. $P(F) = P(G|F) * [P(A, C, D, F) + P(A,\ C, D, F)] + P(\ G|F) * [P(A, C, D, F) + P(A,\ C, D, F)]$. En factorisant encore avec $P(A, C, D, F) + P(A,\ C, D, F)$ on obtient $P(F) = [P(A, C, D, F) + P(A,\ C, D, F)] * [P(G|F) + P(\ G|F)]$. Or $P(G|F) + P(\ G|F) = 1$ donc $P(F) = P(A, C, D, F) + P(A,\ C, D, F)$ ce qui est le même calcul que lorsque le réseau ne contenait pas G. Il est donc possible, pour chaque métabolite, de réduire le nombre de combinaisons dont il faut calculer la probabilité avant de pouvoir calculer le score de probabilité de production. Ce score peut être calculé même sur des cartes du métabolisme de grande taille comme les 31 cartes du métabolisme que nous avons utilisées.

2.3 Conclusion

Ce pipeline de prédiction du métabolisme a été développé afin de prédire le métabolisme des xénobiotiques, donc des AHAs, et de prédire la réactivité à l'ADN de ces métabolites. Il a également été développé afin de répondre à notre problématique qui est de déterminer le rôle des enzymes du métabolisme des xénobiotiques dans la production des métabolites réactifs à l'ADN. Pour cela on s'intéresse à prédire l'influence des conditions physiopathologiques sur la production des métabolites. On cherche à déterminer les conditions favorisant la production des métabolites réactifs à l'ADN.

2.3.1 Valeurs ajoutées

Notre pipeline répond aux limites que nous avions identifié lors de l'état de l'art des outils de prédictions du métabolisme.

Tout d'abord, par rapport à Delannée et al. 2019 [19], notre pipeline utilise l'outil de prédiction SyGMa et non pas Metaprint2D-React. SyGMa prédit des métabolites issus de réactions associées aux phases I et II du métabolisme des xénobiotiques. Nous avons choisi SyGMa car c'est un outil implémentable sous la forme d'un packet python et dont les prédictions peuvent être modifiées par l'ajout de nouvelles règles de biotransformations. Nous avons également choisi cet outil car contrairement à Metaprint2D-React celui-ci ne peut pas devenir inaccessible, ce qui permet la reproductibilité de notre protocole. En effet Metaprint2D-React était un outil uniquement accessible par une plateforme en ligne et hébergé par l'université de Cambridge. Suite à une décision de Dassault System, à qui appartient la base de données BioVia, qui était utilisée par Metaprint2D-React, l'outil en ligne a dû fermer. Le code de l'outil ainsi que le format de la base de données qu'ils utilisaient ne sont pas accessibles il est donc impossible de régénérer Metaprint2D-React tel qu'il a été utilisé par Delannée et al. Notre priorité a donc été d'utiliser au maximum des outils intégrables localement et modulables. Cela afin de permettre d'une part

l'amélioration du pipeline, et d'autre part de pouvoir reproduire notre pipeline. Ensuite notre pipeline enrichit les cartes par des annotations. Les métabolites prédits sont annotés par deux éléments essentiels. Tout d'abord le *score de réactivité à l'ADN*, il décrit la réactivité du métabolite vis-à-vis de l'ADN de la même manière que Delannée et al. Ensuite le *score de probabilité de production*, qui décrit la probabilité qu'un métabolite soit *produit*. Ce score permet de comparer les métabolites entre eux, de les ranger et même de comparer un même métabolite produit dans deux conditions différentes, ce qui n'était pas possible à partir du pipeline de Delannée et al. Il nous est ainsi possible de rechercher les conditions qui favorisent la formation des métabolites les plus réactifs à l'ADN et donc de déterminer les conditions physiopathologiques les plus favorables à la formation d'adduits à l'ADN.

Pour calculer ce *score de probabilité de production* notre pipeline s'inspire de l'étude de Dang et al. 2018 [17]. En effet notre pipeline utilise les annotations des réactions par un score SOM. Ceux-ci proviennent de plusieurs outils : WayToDrugs SOMP [61] et FAME 3 [65]. Si plusieurs outils sont utilisés c'est parce-que chacun permet de prédire des SOMs pour différentes enzymes du métabolisme des xénobiotiques. C'est là qu'intervient notre première subtilité par rapport à Dang et al. puisque l'on considère que les scores SOMs annotant les réactions ne sont pas la probabilité que la réaction se produise mais plutôt la probabilité que la réaction se produise en étant catalysée par une enzyme donnée. La probabilité est donc conditionnée par la présence de l'enzyme associée au score SOM ce qui permet de rendre la réalisation de la réaction dépendante à la présence ou disponibilité de l'enzyme. Cette notion permet alors de questionner le rôle des enzymes dans les probabilités que notre pipeline calcule. Ces probabilités sont utilisées par la modélisation d'un réseau Bayésien qui permet de calculer la probabilité que les métabolites soient produits. C'est notre seconde subtilité puisque l'on peut alors comparer les probabilités de voies métaboliques mais aussi les probabilités que le métabolite soit produit. Cela permet de comparer la production des métabolites dans différentes conditions enzymatiques, celles-ci relatent la disponibilité des enzymes du métabolisme des xénobiotiques. Ce sont ces comparaisons qui permettent de déterminer les conditions les plus favorables à la formation d'adduits à l'ADN.

2.3.2 Application du pipeline

Nous avons appliqué ce pipeline à 7 molécules. Dans un premier temps nous l'avons appliqué à la caféine (cf. chapitre 3) afin d'évaluer notre méthode et le score de probabilité de production. Ensuite nous avons appliqué le pipeline à six AHAs d'intérêts (cf. chapitre 4).

Les résultats obtenus suite à la prédiction du métabolisme de la caféine permettent de proposer une utilisation du *score de probabilité de production* en tant qu'outil de filtration des métabolites. Lors de l'application du pipeline aux AHAs nous proposons de calculer ce score

dans différentes conditions physiopathologiques. Nous proposons alors une utilisation du *score de probabilité de production* comme un critère permettant de comparer, pour un même métabolite, deux conditions physiopathologiques. Les effets de ces conditions sur le calcul et la valeur du score sont détaillés dans le chapitre 4 où nous introduirons les concepts de *contexte enzymatique* et de *signature enzymatique optimale*.

Chapitre 3

Application du pipeline de prédiction à la caféine et évaluation du score de probabilité de production

Un pipeline de prédiction du métabolisme et d'évaluation de la probabilité d'être produit pour chaque métabolites, a été développé et décrit dans le chapitre 2. Ce pipeline a été développé dans le but de prédire le métabolisme des xénobiotiques autrement dit des composés dont le métabolisme est dirigé par les enzymes de phase I et II du métabolisme des xénobiotiques. Pour évaluer notre pipeline de prédiction du métabolisme et de calcul d'un *score de probabilité de production*, nous avons utilisé le pipeline sur la caféine. En effet cette molécule présente plusieurs caractéristiques qui la rendent idéale pour évaluer notre pipeline de reconstruction de carte du métabolisme.

Afin d'évaluer notre pipeline nous avons établi une carte du métabolisme de la caféine. Elle a été produite manuellement par l'analyse de la littérature décrivant le métabolisme de la caféine chez l'homme [67, 70, 56, 13, 53]. Cette carte sera annotée par les enzymes catalysant les différentes réactions. Elle permettra de comparer les prédictions que propose SyGMa à partir de la caféine aux métabolites expérimentalement identifiés chez l'homme. Le score de probabilité de production associé aux métabolites déjà identifiés expérimentalement permettra de discuter l'intérêt du score de probabilité de production que propose notre méthode. Enfin nous proposerons une carte filtrée issue de l'application d'un seuil sur ces scores.

3.1 Production de la carte du métabolisme de la caféine issue des connaissances

Tout d'abord la caféine, (1,3,7 trimethylxanthine ou plus simplement 137x), est la substance psychoactive la plus consommée dans le monde [24] ce qui en fait un objet d'étude très important. Tout comme les AHAs, c'est un xénobitique. Son métabolisme est donc également soumis

aux enzymes du métabolisme des xénoiotiques. Ensuite, le métabolisme de la caféine est bien décrit chez l'homme [67, 13, 53]. Plus précisément ce métabolisme partage avec le métabolisme des AHAs des enzymes comme les CYPs, associées à la phase I du métabolisme des xénobiotiques, et les NATs, associées à la phase II. L'isoforme de CYP CYP1A2 est la principale enzyme du métabolisme de la caféine [67] et intervient également dans le métabolisme des AHAs et la production d'adduits à l'ADN [73, 52, 6]. Le métabolisme de la caféine implique également les isoformes CYP3A4, CYP2E1 et CYP2D6. Pour ce qui est de l'implication des NATs c'est l'isoforme NAT2 qui est impliquée dans la production de métabolites de la caféine, cette enzyme est également retrouvée dans le métabolisme des AHAs et notamment dans la production d'adduits à l'ADN. L'ensemble de ces critères font de ce composé un bon candidat pour évaluer notre approche de prédiction du métabolisme ainsi que notre modélisation basée sur les scores SOM associées aux enzymes de phase I et de phase II du métabolisme des xénobiotiques.

Nous avons identifié dans la littérature 15 métabolites de la caféine. Nous avons également identifié les voies principales de ce métabolisme, au nombre de quatre, ainsi que les principales enzymes impliquées. 13 des 15 métabolites identifiés sont obtenus par des réactions de déméthylation ou d'oxydation catalysées par les cythochromes P450 (CYPs), enzymes de phase I du métabolisme des xénobiotiques. Parmi les isoformes de CYPs, le CYP1A2 est responsable de la majorité de ces transformations[67] les CYPs 3A4, 2E1 et 2D6 jouent également un rôle dans certaines de ces transformations. Les 2 métabolites restants sont obtenus par des réactions de N-acetylation, dirigés par l'isoforme des NATs, NAT2 [67, 13], enzyme de phase II du métabolisme des xénobiotiques. Plusieurs métabolites de la caféine sont produits à partir de différents métabolites intermédiaires. Cela implique que ces métabolites peuvent être produits par des voies métaboliques distinctes. On retrouve également des voies métaboliques entrecroisées dans les cartes prédites par notre pipeline.

3.1.1 Les métabolites de la caféine

Parmi les 15 métabolites identifiés dans le métabolisme de la caféine, 5 sont des dérivés directement issus de la transformation de la caféine. Les 10 métabolites restants sont des dérivés de ces 5 métabolites. Ces 5 métabolites sont la paraxanthine (1,7-dimethylxanthine ou 17X), la theophylline (3,7-dimethylxanthine ou 37X) et la theobromine (1,3-dimethylxanthine ou 13X), le 1,3,7-trimethyluric acid (137U) et le 6-amino-5-(N-formylmethylamino)-1,3-dimethyluracil (137-DAU ou 137-TAU ou ADMU). Parmis ces 5 métabolites, 3 sont prédominants et représentent plus de 94% du métabolisme de la caféine, il s'agit de 17X, 37X et 13X qui représentent respectivement 79,6%, 10,8% et 3,7% de la totalité du métabolisme de la caféine. La formation de 137U, 137-TAU ou l'absence de transformation de la caféine représente les moins de 6% restant de ce métabolisme [67].

3.1.1.1 Métabolisme de la paraxanthine

La paraxanthine (17X) est le principal métabolite directement issu de la caféine[70, 56]. Ce métabolite est issu d'une réaction de démethylation largement catalysé par le CYP1A2, et dans une moindre mesure par le CYP1A1, CYP2E1 et CYP2D6 lorsque la concentration en caféine est suffisamment grande[67]. Ce métabolite peut être soit directement excrété en vue de son élimination, soit de nouveau métabolisé. Nous avons dénombré 4 métabolites de la paraxanthine :

— Le 1-methylxanthine (1MX), c'est un métabolite issu d'une déméthylation de 17X dirigée par le CYP1A2 [67].

— Le 5-acetylamino-6-formylamino-3-methyluracil (AFMU), il est produit lors de la déméthylation de 17X vers 1MX. Lors de cette déméthylation un métabolite transitoire est obtenu, soit la réaction se poursuit et 1MX est obtenu, soit ce métabolite de transition est recruté par la N-actelytransferase NAT2 [13]. Dans ce cas c'est AFMU qui est produit.

— Le 7-methylxanthine (7MX), c'est aussi un métabolite issu d'une déméthylation de 17X dirigée de nouveau par le CYP1A2[67]. 7MX et 1MX diffère de par le groupe méthyl qui a été éliminé lors de la déméthylation.

— le 1,7-dimethyluric acid (17U), il est obtenue via une réaction d'hydroxylation du 17X. Cette réaction est principalement dirigée par le CYP2A6 mais peut aussi être catalysée par le CYP1A2 [67].

Les métabolites 1MX et 7MX peuvent encore être métabolisés par une xanthine déhydrogénase, ou par le CYP1A2 dans le cas de 1MX, et produire respectivement le 1-methyluric acid (1MU) et le 7-methyluric acid (7MU) [67].

3.1.1.2 Métabolisme de la théophylline

La théophylline (37X) est formée par une réaction de démethylation de la caféine dirigée majoritairement par le CYP1A2 ou, dans une moindre mesure, par CYP2E1 et CYP2D6 [67]. Elle peut être excrétée, ce qui représente 10% de son métabolisme, mais elle est plus généralement métabolisée par les CYPs. Nous avons dénombré 3 métabolites de la théophylline :

— Le 7MX, il s'agit du même dérivé que celui obtenu par la déméthylation du 17X. Ici c'est les CYP1A2 et CYP2E1 qui dirigent cette déméthylation de 37X vers 7MX. [67]

— Le 3,7-dimethyluric acid (37U), il est obtenu par l'hydroxylation de 37X. Cette hydroxylation est également dirigée par le CYP1A2 et CYP2E1 [67].

— Le 3MX, il est obtenu par une déméthylation de 37X. Les enzymes dirigeant cette réaction sont les CYPs CYP1A2, CYP2E1 et CYP2A6 [67]. Tout comme 1MX et 7MX, 3MX peut être métabolisé par une xanthine déhydrogénase [67].

3.1.1.3 Métabolisme de la theobromine

La theobromine (13X) est le dernier métabolite de la caféine dont nous avons identifié des dérivés. Elle est issue d'une démethylation, celle-ci est catalysée principalement par le CYP1A2 et dans une moindre mesure par le CYP2D6 et le CYP2E1 [67]. Ce métabolite, comme les deux précédents, peut être excrété mais est principalement métabolisé. Nous avons dénombré 3 métabolites de la théophylline :

— Le 1,3-dimethyluric acid (13U). Il représente 50% du métabolisme de 13X. Il est obtenu par une hydroxylation du 13X dirigé par le CYP1A2, le CYP2E1 et le CYP3A4.
— Le 1MX, il s'agit du même métabolite que pour le 17X. Il est produit par une hydroxylation de 13X dirigée par le CYP1A2.
— Le 3MX, il s'agit du même métabolite que pour le 37X. Il est produit par une hydroxylation de 13X dirigée par le CYP1A2.

Les deux autres métabolites de la caféine, le 137U et le ADMU ne sont associés à aucun dérivés identifiés. 137U est obtenu par une hydroxylation de la caféine catalysée le CYP3A4 majoritairement mais aussi le CYP1A2, CYP2E1, CYP2C8 et CYP2C9[67]. Le 137-TAU est obtenu de la même manière que pour AFMU, dérivé de la paraxanthine, via une N-actetyl transferase[67]. Ces deux voies sont cependant très minoritaires. En effet elles représentent, avec la non transformation de la caféine, moins de 6% du métabolisme total de la caféine [67].

3.1.2 La carte du métabolisme de la caféine

L'ensemble des informations que nous avons décrites sur le métabolisme de la caféine et de ses dérivés a été compilé dans une carte du métabolisme annotée que nous représentons dans la figure 3.1.

3.2 Évaluation du pipeline de prédiction du métabolisme

Le pipeline est évalué à différents niveaux. Tout d'abord les métabolites de la carte de la figure 3.1 seront comparés aux métabolites proposés par SyGMa [60]. Cela permet de classer les métabolites issus de la prédiction en deux groupes. Les métabolites identifiés expérimentalement et les métabolites considérés comme inconnus. Nous pouvons ainsi déterminer la proportion des métabolites inconnus, comparer les métabolites des deux groupes ou rechercher des similarités de structures entre les métabolites inconnus. Ensuite les connaissances sur la quantification des métabolites directement dérivés de la caféine permettra de discuter le score de probabilité de production obtenu par notre modèle Bayésien. Enfin, nous proposons d'utiliser le score de probabilité de production comme outil de filtration, nous évaluerons donc la carte prédite du

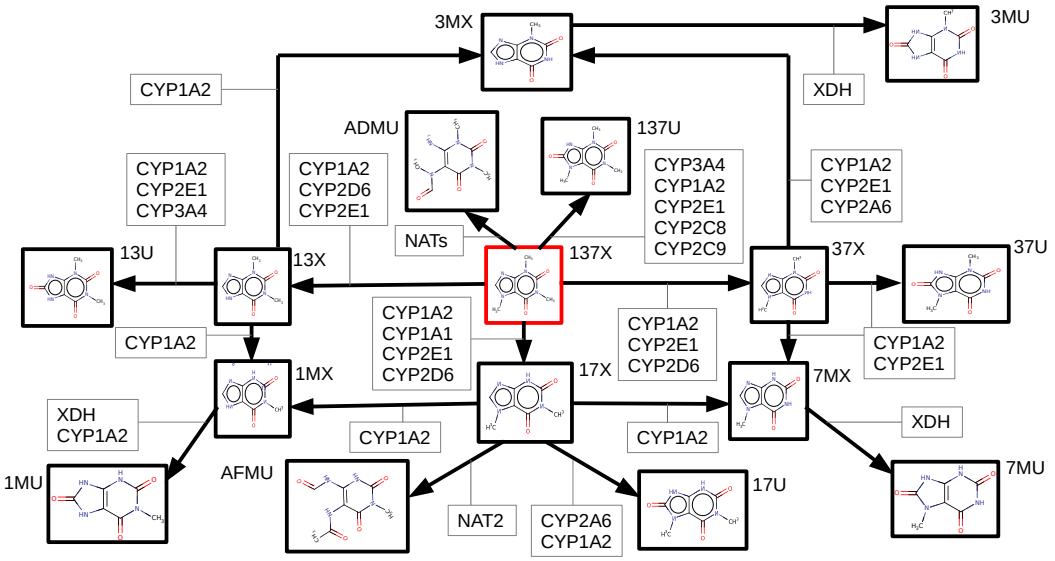


FIGURE 3.1 – **Représentation schématique des connaissances sur le métabolisme de la caféine.** Les enzymes catalysant les réactions sont représentées dans des cadres gris. XDH = xanthine déhydrogénase ; NATs = N-acetyl transférases nous n'avons pas pu identifier dans la littérature la ou les isoformes associées à cette annotation. Nous faisons l'hypothèse qu'il s'agit de NAT2 car il s'agit d'une réaction similaire à celle générant AFMU

métabolisme final, c'est-à-dire la carte obtenue après une filtration sur les scores de probabilité de production des métabolites à partir d'un seuil.

3.2.1 Evaluation de la prédiction des métabolites par SyGMa

La partie du pipeline prédisant une carte du métabolisme est décrite dans le chapitre 2. En résumé SyGMa est un outil qui permet de prédire des métabolites à partir de règles de biotransformations et d'une structure chimique. Il nécessite deux paramètres pour effectuer une prédiction :

— Un ensemble de règles de biotransformations décrites par des formules SMIRKS. Ces formules permettent de décrire des réactions métaboliques. Ici l'ensemble de règles décrit les biotransformations prises en compte pour la prédiction. Notre pipeline utilise les deux ensembles de règles SMIRKS déjà intégrés au sein de SyGMa nommés *phase I* et *phase II*. (cf. chapitre 2 et Annexe 2.

— Un nombre d'itérations. Ce nombre précise combien de fois SyGMa doit itérer les prédictions. À chaque nouvelle itération SyGMa utilise les métabolites prédits par l'itération précédente et applique une prédiction de nouveaux métabolites. Ce paramètre détermine le nombre maximal de réactions séparant, dans une carte, un métabolite et le composé d'origine. Dans notre pipeline il est paramétré à 2 afin d'imiter le métabolisme des xéno-

biotiques. Ce métabolisme se décompose en deux phases de transformation décrites dans la figure 1.1. La première phase est catalysée par les enzymes de phase I, majoritairement des CYPs, puis la seconde phase est catalysée par les enzymes de phase II, c'est-à-dire UGTs, NATs, SULTs et GSTs. Ce métabolisme comporte une phase III qui est une phase de transport permettant l'élimination du xénobiotique. Avec un nombre d'itérations de SyGMa égal à deux, les réactions obtenues dans la prédiction seront différenciées par un rang. Tout d'abord les réactions prédites lors de la première itération seront les réactions de rang 1. Ensuite les réactions prédites lors de la seconde itération seront les réactions de rang 2.

Ces deux paramètres forment ensemble un *scénario* qui peut ensuite être appliqué à n'importe quelles structures chimiques. Cette structure à partir de laquelle on commence la prédiction des métabolites est décrite par une formule SMILES. C'est la description d'une molécule sous la forme d'une ligne de texte. Il s'agit donc d'un encodage de la structure chimique qui peut être utilisé par différents outils. Dans le cas de la prédiction du métabolisme de la caféine avec SyGMa, nous avons utilisé la formule SMILES : CN1C=NC2=C1C(=O)N(C(=O)N2C) qui est extraite de la base de données PubChem[35]. Cette première molécule est donc à l'origine de toutes les prédictions et sera alors mise en évidence dans les cartes du métabolisme. Nous avons nommé cette molécule le *composé original*.

La figure 3.2 décrit la carte du métabolisme de la caféine résultant de la prédiction de SyGMa. Elle contient 23 métabolites reliés par 31 réactions. Le composé repéré en rouge est le composé original, la caféine. En comparant cette carte à la carte figure3.1 décrivant le métabolisme connu de la caféine, on constate que 11 des 16 métabolites connus sont retrouvés et sont représentés en vert. Enfin les 12 métabolites bleus sont les métabolites qui sont à notre connaissance non identifiés dans le métabolisme de la caféine et donc inconnus.

Parmi les 5 métabolites identifiés expérimentalement (ADMU, AFMU, 1MU, 3MU et 7MU) mais non retrouvés, 3 (1MU, 3MU et 7MU) sont issus d'une succession de trois réactions dans la figure 3.1. SyGMa a été paramétré pour itérer deux fois. Il est donc normal que ces métabolites ne soient pas retrouvés par le pipeline qui ne prédit que deux réactions consécutives. Les 2 autres métabolites non prédits sont AFMU et ADMU. Ce sont les seuls métabolites de la caféine associés à une réactions catalysée par les NATs. Afin de comprendre pourquoi ces métabolites ne sont pas obtenus suite à la prédiction de SyGMa nous avons analysé les règles SMIRKS associées aux NATs. L'association entre les règles SMIRKS et les enzymes catalysant ces réactions est décrite en Annexe 2. Il s'avère que les structures chimiques variables de 137X vers ADMU et de 17X vers AFMU ne sont pas décrites dans les règles SMIRKS proposées par SyGMA. Cela explique l'absence de prédiction de ces métabolites.

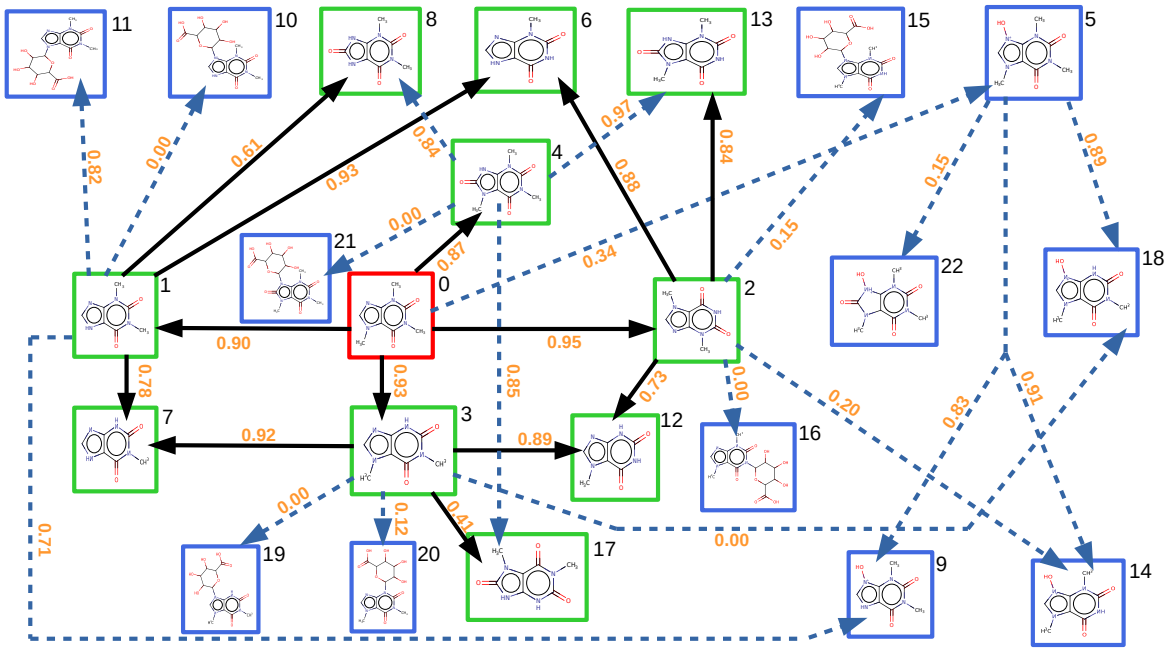


FIGURE 3.2 – **Carte du métabolisme de la caféine prédite en utilisant SyGMa** Les 23 métabolites sont représentés par leur structure 2D. Parmi eux, la caféine est représentée dans un cadre rouge. Les 10 métabolites identifiés expérimentalement sont représentés dans un cadre vert et les 12 métabolites inconnus dans un cadre bleu. Les 31 réactions sont représentées par des flèches. Parmi elles, les 13 flèches noires représentent les réactions qui sont identiques à la carte du métabolisme de la caféine de la figure 3.1. Les 18 autres réactions sont décrites par des flèches bleues hachurées. Cette carte présente également les scores SOM annotant les réactions en orange gras. Ces scores sont des scores provenant de Way2Drugs SOMP.

Parmi les réactions prédites, 16 consomment et produisent un métabolite identifié expérimentalement ou la caféine elle même. 13 de ces réactions ont été identifiées dans la figure 3.1 et sont représentées par des flèches noires. SyGMa prédit donc 3 nouvelles réactions, chacune transformant 137U (noeud 4) en 17U (noeud 17), 13U (noeud 8) ou 37U (noeud 13) via une réaction de déméthylation. La déméthylation est une réaction que l'on retrouve beaucoup dans le métabolisme de la caféine. C'est la réaction qui permet la transformation de 137 X vers 13X, 17X et 37X mais également la réaction qui produit 1MX, 3MX et 7MX à partir de 13X, 17X et 37X. Ces nouvelles réactions sont donc crédibles au vu du métabolisme connu de la caféine. Ce résultat suggère la capacité de 137U à être une autre voie de formation de ces métabolites, sa faible formation dans le métabolisme de la caféine (moins de 4%) ainsi que de sa rapide élimination rend le métabolite instable et pourrait expliquer l'absence de démonstration de ces réactions dans la littérature.

SyGMa retrouve 11 des 13 métabolites que l'on pouvait s'attendre à retrouver en prenant en compte la contrainte de 2 itérations ce qui rend ces prédictions consistantes avec la littérature. On note cependant l'absence de prédictions des réactions associées aux NATs et donc l'absence

des métabolites associés. SyGMa prédit 12 nouveaux métabolites, cela est en accord avec la littérature [7] qui constate un grand nombre de prédictions lors d'utilisation d'outils de prédictions du métabolisme.

Parmi les métabolites inconnus prédits, le noeud 22 (SMILES : CN1C(=O)[NH+](O)c2c1c(=O)-n(C)c(=O)n2C ; Nom IUAPC : 9-hydroxy-1,3,7-trimethyl-2,6,8-trioxo-2,3,6,7,8,9-hexahydro-1H-purin-9-ium), est issu d'une réaction qui agit sur la structure aromatique de la molécule du noeud 5. Celle-ci est perdue sur le noeud 22 et c'est une réaction atypique et plutôt improbable selon nos connaissances.
Ces prédictions sont donc encourageantes et permettent d'établir que SyGMa est un bon outil pour couvrir la prédiction du métabolisme des xénobiotiques. Elles justifient également l'emploi de méthodes additionnelles permettant la filtration ou une meilleure exploitation des résultats de prédictions comme c'est le cas pour l'exploration des résultats d'autres outils de prédictions du métabolisme [19].

3.2.2 Annotation et filtration de la carte du métabolisme de la caféine

La structure spéciale du noeud 22 justifie l'emploi de méthodes permettant d'estimer la confiance que l'on peut avoir sur chaque prédiction. Nous proposons d'utiliser notre score de probabilité de production en ce sens.

Tout d'abord rappelons les conditions de calcul de ce score. Lorsque le pipeline utilise la carte annotée du métabolisme pour calculer le score de probabilité de production, il prend en compte les annotations *score SOM* des réactions. Les outils Way2Drugs SOMP [61] et FAME 3 [65] sont les outils de prédictions de sites du métabolisme (SOM) que nous utilisons pour annoter les réactions. Ces outils associent à chaque atome d'une molécule un score. Ce score représente la probabilité de l'atome d'être un atome catalysé par une enzyme. Cette prédiction est réalisée, sur chaque atome, en prenant en compte un ensemble de critères qui décrit l'atome. Cet ensemble est aussi nommé ensemble de descripteurs moléculaires. C'est en partant des descripteurs des atomes dont on sait qu'il réagissent, à partir de réactions connues, que ces outils établissent des règles de calculs pour obtenir le score SOM de l'atome. Ce score représente donc la probabilité que l'atome soit effectivement un SOM. Nous utilisons ce score pour annoter les réactions en utilisant le score SOM correspondant à l'*atome de la réaction*. Cet atome est celui qui a réagi dans une réaction donnée. Le score SOM représente alors la chance que l'atome soit effectivement catalysé par l'enzyme qui catalyse *a priori* la réaction.
Pour choisir lequel des outils est utilisé pour prédire le SOM qui annotera la réaction notre pipeline fait une hypothèse. Celle qu'une majorité des réactions de rang 1, c'est-à-dire issue de la première itération de SyGMa et consommant le composé d'origine, serait en grande partie des

réactions associées aux enzymes de phase I. Les réactions de rang 2 quant à elles contiendraient une majorité des réactions associées eux enzymes de phase II. Cette hypothèse a été construite afin d'appliquer notre pipeline au métabolisme des xénobiotiques. Elle permet de sélectionner les outils de prédictions de site du métabolisme (SOM) annotant les réactions en fonction du rang. Les résultats de prédictions de SyGMa sur la caféine confortent cette hypothèse. 5 sur 5 réactions de rang 1 sont associées à des enzymes de phase I. Tandis que l'ensemble des réactions associées aux enzymes de phase II, toutes des glucuronidations, sont de rang 2. Ce résultat conforte donc notre hypothèse et justifie bien l'annotation des réactions de rang 1 par Way2Drugs SOMP, qui annote jusqu'à cinq isoformes de CYPs, et l'annotation des réactions de rang 2 par FAME 3, qui annote plus d'enzymes de phase II.

3.2.3 Classification des métabolites via le score de probabilité de production calculé par un modèle Bayésien

Après l'annotation des réactions de la carte prédite du métabolisme, le pipeline calcule un score de probabilité de production pour chaque métabolite en utilisant une approche semblable à celle des réseaux Bayésiens. Ce calcul repose notamment sur les annotations *enzyme* et *SOM score* et est décrit dans le chapitre 2. En résumé, le pipeline réduit d'abord la carte du métabolisme annotée en ne prenant plus en compte qu'une réaction parmi les multiples réactions consommant le même métabolite et produisant un autre même métabolite. Ensuite le pipeline utilise les scores SOM comme des probabilités. Par exemple le score annotant la réaction du métabolite A vers le métabolite B décrit la probabilité que B soit produit à condition que A soit lui-même produit. *In fine* c'est en calculant un ensemble de probabilités jointes à partir de ces probabilités conditionnelles que le pipeline propose une probabilité de production pour chaque métabolite.

3.2.3.1 Étude de la distribution des scores sur les métabolites

Après le calcul du score de probabilité de production sur l'ensemble des métabolites prédits par SyGMa, nous avons observé la distribution de ce score sur les métabolites afin d'évaluer la pertinence de ce score à l'aune des connaissances établies sur le métabolisme de la caféine. Nous avons représenté cette distribution sous la forme d'un histogramme, figure3.3 ordonné du score le plus grand au score le plus faible. Les barres vertes sont associées aux métabolites identifiés expérimentalement. Les barres bleues sont elles associées aux métabolites inconnus.

On constate que les 11 métabolites avec le score le plus grand sont les 11 métabolites identifiés expérimentalement. Les deux métabolites avec le score le plus fort après les métabolites identifiés expérimentalement sont le noeud 9 (SMILES : Cn1c(=O)c2[nH]c[n+](O)c2n(C)c1=O ; Nom IUPAC : 9-hydroxy-1,3-dimethyl-2,6-dioxo-2,3,6,7-tetrahydro-1H-purin-9-ium) et le noeud 11

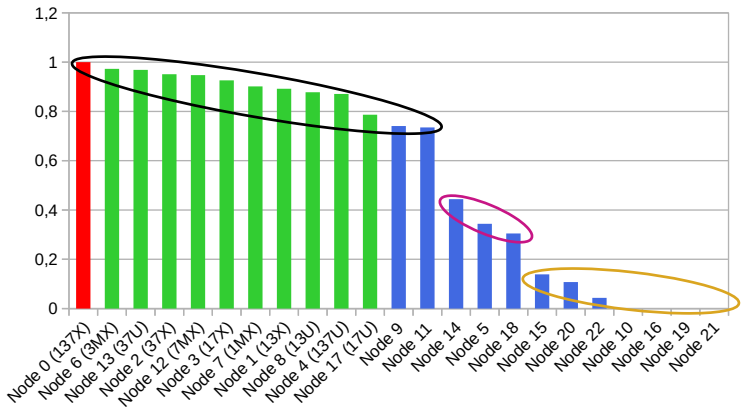


FIGURE 3.3 – **Distribution des scores de probabilités de production des métabolites de la carte de la caféine**. Le score de probabilité de production de chaque métabolite a été calculé à partir d'un réseau Bayésien dont les probabilités conditionnelles des variables sont extraites des score SOM annotant les réactions. Les barres rouges, vertes et bleues sont associées respectivements à la caféine, les métabolites expérimentalement identifiés de la caféine et les métabolites inconnus dans le métabolisme de la caféine. Les métabolites sont regroupés en trois groupes selon la valeur de leur score de probabilité de production (ellipses noire, violette et jaune).

(Cn1c(=O)c2c(ncn2C2OC(C(=O)O)C(O)C(O)C2O)n(C)c1=O ; 6-(1,3-dimethyl-2,6-dioxo-2,3,-6,7-tetrahydro-1H-purin-7-yl)-3,4,5-trihydroxyoxane-2-carboxylic acid). Ils présentent un score du même ordre de grandeur que ceux des métabolites identifiés expérimentalement, c'est-à-dire supérieur à 0,70. Ces 13 métabolites forment un premier groupe de métabolites avec un fort score de probabilité de production, ils sont identifiés par une ellipse noire dans la figure 3.3. Ensuite, les métabolites associés aux noeuds 14, 5 et 6 présentent un score éloigné du premier groupe mais suffisamment grand pour former un groupe de score intermédiaire, c'est-à-dire supérieur à 0,20 mais inférieur à 0,70, ce groupe est identifié par une ellipse violette. Enfin les noeuds 15, 20, 22, 10, 16, 19 and 21 constituent le dernier groupe avec un score de confiance très faible (inférieur à 0.20) ou nul. Ce groupe est identifié par une ellipse jaune.

Dans ce dernier groupe des scores faibles, certains des métabolites sont associés à un score de confiance nul. Ce sont les métabolites des noeuds 10, 16, 19 et 21. Un score de probabilité de production nul est surprenant car il implique que le métabolite n'a aucune chance d'être produit. Un score nul est permis par notre modèle dans un seul cas, celui des réactions annotées par un score de 0,0. Dans le cas de l'annotation des réactions par les SOMs scores produits par Way2Drugs SOMP, le score qui vas annoter la réaction peut être négatif. Way2Drug peut proposer des scores SOM négatifs, ils représentent alors la chance que l'atome ne soit pas un SOM. Au vu de l'implication des SOMs dans notre pipeline, nous avons décidé qu'un score négatif pouvait être remplacé par la valeur 0,0. En effet, utiliser des probabilités conditionnelles négatives n'avait pas de sens. Pour autant, la réaction a été annotée et n'a pas été éliminée par le pipeline. Dans le cas de l'annotation des réactions par les SOMs scores produits par FAME

3, FAME 3 peut proposer des probabilités de 0,0. La réaction est alors annotée mais n'a, selon ce score, aucune chance de se produire. Pour autant l'annotation ayant eu lieu, la réaction n'a pas été éliminée du pipeline.
Ainsi, si l'ensemble des voies métaboliques, menant à un métabolite, contiennent au moins une réaction annotée par un score nul, alors la probabilité que le métabolite existe devient nulle car la probabilité d'emprunter chacune des voies est nulle. Dans le cas de la production des métabolites de la caféine c'est le cas du noeud 10, où la réaction du noeud 1 vers le noeud 10 est annotée par un score de 0,0. Autrement dit, elle n'est soutenue par aucuns CYPs.

La distribution de la valeur de notre score de probabilité de production met en évidence que ce score permet de différencier les métabolites expérimentalement identifiés des autres métabolites. En effet, les métabolites avec le plus fort score sont les métabolites identifiés expérimentalement ainsi que les métabolites des noeuds 9 et 11. Les scores des autres métabolites sont bien distincts de ce groupe. Notre score semble bien représenter la chance d'être produit pour un métabolite car ce sont les métabolites effectivement produits que l'on retrouve avec les plus forts scores.

Nous avons alors comparé les scores de probabilité de production et la quantité produite de chaque métabolite de la caféine. Ces quantités ont été décrites lorsque le métabolisme de la caféine a été décrit. On sait que la paraxanthine (17X) est le métabolite majoritaire parmi les métabolites directement issus de la caféine. Elle représente 79,6% du métabolisme de la caféine. Viennent ensuite la theophylline (37X) avec 10,8% du métabolisme de la caféine et la theobromine (13X) avec 3,7% du métabolisme. En comparant les scores de probabilité de production, à partir de la figure 3.3, de ces trois métabolites on constate que 37X a un score plus grand que celui de 17X. 17X a quant a lui un score plus grand que 13X. Alors que des trois métabolites c'est 17X qui est de loin le plus produit il est associé à un score plus faible que celui du 37X. Nous démontrons par cet exemple que notre score n'est pas relatif à la quantité de métabolites produits. Il est plutôt relatif à la confiance que l'on accorde à la production d'un métabolite donné. On peut traduire ce score par la probabilité d'être produit au moins une fois.

3.2.4 Filtration de la carte du métabolisme

Le score de probabilité de production est un paramètre permettant de discriminer les métabolites. Nous proposons d'utiliser ce score afin de filtrer les grands nombres de prédictions obtenues par les différents outils de prédictions du métabolisme. La filtration élimine les métabolites avec un score de confiance inférieur à un seuil. Ensuite les réactions est les métabolites isolés, c'est-à-dire qui ne sont plus reliés au composé d'origine, ici la caféine, sont également éliminés. Dans le cas de la caféine nous avons défini un seuil de 0,70 afin de ne conserver que les métabolites associés à un fort score de probabilité de production.

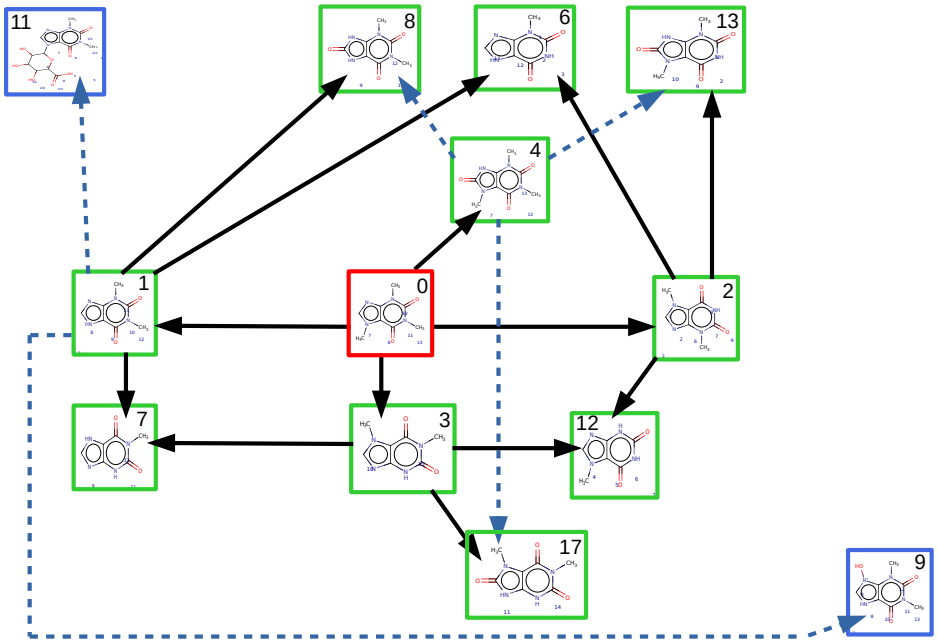


FIGURE 3.4 – **Carte du métabolisme de la caféine prédite par SyGMa et filtrée à partir du score de probabilité de production.** Cette carte est la carte obtenue après une filtration des métabolites de la carte précédente et cela afin de conserver uniquement les métabolites avec un fort score de probabilité de production. Le seuil de filtration utilisé est de 0,70 ce qui signifie que les métabolites associés à un score strictement inférieur sont éliminés. La carte contient 13 métabolites et 17 réactions. Deux métabolites, inconnus dans le métabolisme de la caféine, (noeuds 9 et 11), sont prédit comme ayant une probabilité de production proche des métabolites expérimentalement identifiés.

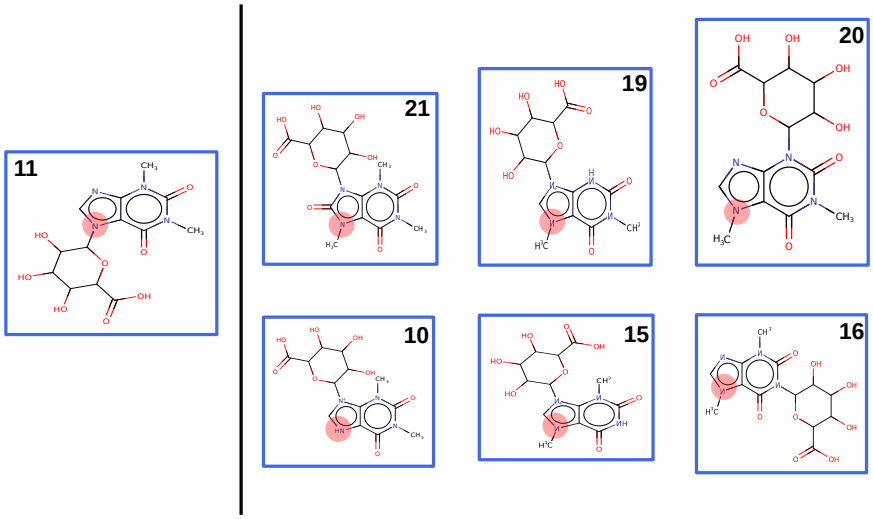


FIGURE 3.5 – **Comparaison de la structure chimique des métabolites inconnus de la caféine issus de glucuronidation.** À gauche le métabolite du noeud 11 qui est conservé après la filtration et à droite les métabolites éliminés par la filtration. Chaque métabolite est identifié par le numéro du noeud auquel il est associé. La zone rouge marque l'azote auquel le groupe glucuronyl est associé dans le métabolite 11.

La carte filtrée obtenue à partir d'un seuil de 0,70 est représentée dans la figure 3.4. Elle contient 14 composés dont la caféine et 18 réactions. Parmi les 13 métabolites de la caféine, on retrouve les 11 métabolites identifiés expérimentalement précédemment retrouvés. Les réactions conservées entre ces métabolites sont les mêmes que celles que l'on pouvait observer dans la première carte, la filtration ne filtrant pas directement les réactions mais n'éliminant que celles qui ne sont plus rattachées au composé d'origine. Seuls deux nouveaux métabolites ont été conservés, le métabolite associé au noeud 9 et le métabolite associé au noeud 11, tous deux proviennent de la transformation du métabolite 13X. L'ensemble des autres métabolites prédits ont été filtrés.

En comparant la structure des nouveaux métabolites éliminés et conservés on remarque que deux groupes se distinguent. Le groupe des métabolites issus d'une réaction de glucuronidation, c'est le cas des métabolites des noeuds 10, 11, 15, 16, 19, 20, et 21, et le groupe de métabolites issus de réactions de phase I, c'est le cas des noeuds 5, 9, 14, 18 et 22.

Tout d'abord il faut noter que nous n'avons pas trouvé dans la littérature de métabolites glucuronoconjugués issus de la caféine. Le faible score de ces métabolites glucuronoconjugués, c'est-à-dire le premier groupe que nous avons précédemment défini, et l'élimination de ces conjugués par la filtration est en accord avec ce constat.
Dans ce groupe, seul le noeud 11 est conservé tandis que les autres sont éliminés. Le noeud 11 est conservé du fait que le score SOM associé à la réaction de 13X vers 11 est de 0.824 ce qui est

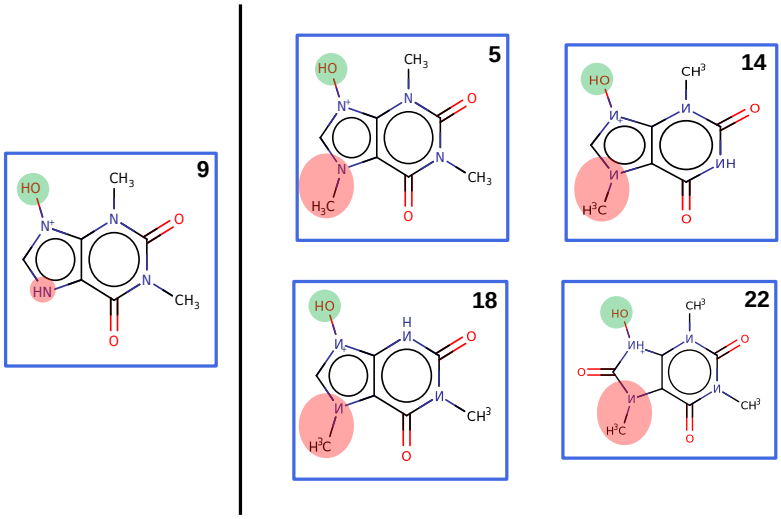


FIGURE 3.6 – **Comparaison de la structure chimique des métabolites inconnus de la caféine produits par des réactions de phase I**. À gauche le métabolite du noeud 9 qui est conservé après la filtration et à droite les métabolites éliminés par la filtration. Chaque métabolite est identifié par le numéro du noeud auquel il est associé. La zone rouge marque l'azote du groupe imidazole de la caféine auquel est associé un groupe méthyle. Il indique que la caféine a dû subir une déméthylation avant de pouvoir produire le métabolite du noeud 9. La zone verte indique le groupe hydroxyl que l'on retrouve dans l'ensemble des métabolites inconnus de la caféine produits par des réactions de phase I.

du même ordre de grandeur que les scores de la caféine vers ses trois métabolites majoritaires 17X (0.926), 13X (0.892) et 37X (0.951). Ce résultat suggère donc qu'il y a une différence entre la structure du noeud 11 et des autres noeuds issus de glucuronidations qui explique cette différence de score. Après une comparaison des structures nous avons constaté que le métabolite du noeud 11 est le seul des métabolites de ce groupe où le groupement glucuronyl est associé à un atome, précédemment associé à un groupe méthyl, de la partie imidazole de la caféine. La comparaison de ces structures chimiques est représentée dans la figure 3.5. Cette particularité pourrait expliquer le score SOM particulièrement haut associé à ce nouveau métabolite, cependant aucune de nos recherches dans la littérature ne vient conforter cette hypothèse. Nous avons noté qu'aucun métabolite glucuronyl-conjugué de la caféine n'a été identifié à ce jour.

Dans le second groupe de métabolites filtrés, seul le noeud 9 est conservé. De la même manière que pour le noeud 11 ce métabolite est conservé du fait du score SOM entre 13X et le noeud 9 de 0,714, qui est du même ordre de grandeur que d'autres réactions connues du métabolisme de la caféine comme la réaction de 13X vers 13U dont le score SOM est de 0,606. En procédant comme pour le noeud 11 nous avons comparé les structures de ce métabolite aux structures des autres métabolites de ce groupe afin d'identifier une particularité qui pourrait expliquer ce score SOM particulièrement fort. Cette comparaison est décrite dans la figure 3.6. Nous avons identifié une particularité que partage les noeuds 5, 14, 18 et 22. Chacun de ces métabolites

présente une structure N=C-N ou N-C-N (dans le cas du noeud 22) sur la partie imidazole de la caféine où le premier azote est méthylé tandis que le second est lié à un groupe hydroxyl. Le noeud 9 a perdu ce groupement méthyl. De la même manière que pour le noeud 11 cette particularité pourrait expliquer le fort score de probabilité de production associé au noeud 9 mais, à notre connaissance, aucune étude n'a établi la particularité de cette structure.

Une autre hypothèse justifiant de la conservation de ces métabolites est que ce n'est pas la structure de ces deux métabolites qui est spécifique mais bien celle du 13X qui est le métabolite qui permet la formation des métabolites du noeud 9 et du noeud 11. Si tel est le cas alors la faible formation de 13X chez l'homme (3,7% du métabolisme de la caféine) et la compétition avec la formation de 13U (50% du métabolisme de 13X), de 1MX et de 3MX pourrait expliquer l'absence d'identification des métabolites 9 et 11 dans la littérature.

3.3 Conclusion

Nous avons appliqué notre pipeline de prédiction du métabolisme et d'enrichissement des cartes du métabolisme à la caféine. La caféine est un xénobiotique dont le métabolisme chez l'homme est proche de celui des AHAs, notamment du fait de l'implication des mêmes enzymes. Nous avons donc utilisé l'application du pipeline à la caféine comme élément de validation de notre approche.

Nous avons pu tout d'abord évaluer la qualité de la prédiction que permet SyGMa. Nous avons constaté d'une part que le pipeline a permis de retrouver la majorité des métabolites identifiés expérimentalement de la caféine. D'autre part nous avons mis en lumière le fait que SyGMa n'avait pas pu prédire les métabolites de la caféine produits par des réactions catalysées par les NATs. Ce résultat nous a permis d'identifier des lacunes dans les prédictions que propose SyGMa. SyGMa est un outil modulable, ce qui signifie que les prédictions qu'il propose peuvent être modifiées. Pour cela nous pouvons par exemple retirer ou ajouter des règles de biotransformations. L'identification des lacunes dans les prédictions de SyGMa permettra d'améliorer notre pipeline de prédiction en ajoutant les règles qui pourront combler ces lacunes.

Ensuite nous avons évalué le *score de probabilité de production* que propose notre pipeline. Nous avons montré que ce score permet de discriminer les métabolites expérimentalement identifiés, associés à un score fort, et les métabolites inconnus, associés à un score faible. Nous avons alors proposé d'utiliser ce score comme un outil de filtration des métabolites. Deux métabolites inconnus étaient associés à un score fort, c'est-à-dire supérieur au seuil du filtre, et ont été conservés. Les structures de ces métabolites ont été étudiées afin de proposer des hypothèses permettant d'expliquer pourquoi ces métabolites sont bien soutenus par les outils de prédictions

de SOMs. Nous avons également recherché ces métabolites dans la littérature, pour le moment ils n'ont pas été identifiés. Une solution que nous proposons est d'utiliser des outils prédicteurs de spectre de masse afin de prédire le spectre de masse de ces métabolites. À partir de ces spectres de masse on pourrait rechercher ces métabolites dans les données expérimentales d'autres études afin de s'assurer que ces métabolites n'ont jamais été caractérisés.

In fine l'application du pipeline à la caféine a démontré l'intérêt d'annoter les cartes du métabolisme et de calculer notre *score de probabilité de production*.

Chapitre 4

Etude du métabolisme des AHAs

Les AHAs sont des contaminants de l'environnement. Elles sont produites lors de réactions de pyrolyse comme la cuisson de viande et poissons mais peuvent également être retrouvées dans les fumées de cigarettes ou dans les gaz d'échappement [73]. À ce jour, 30 AHAs différents on été identifiés, la liste de ces 30 AHAs est disponible en annexe 2. Si ces contaminants, ou xénobiotiques, sont étudiés c'est parce que leur pouvoir cancérogène est encore mal défini chez l'homme. En effet, les AHAs sont catégorisées comme mutagènes chez la bactérie et cancérogènes chez l'animal [69]. Cependant on manque de données épidémiologiques chez l'homme. Elles sont classées par l'International Agency for Research on Cancer (IARC) comme des cancérogènes possibles (2B) et probables (2A). Ainsi, il est nécessaire d'étudier ces AHAs pour déterminer leur capacité à former des adduits à l'ADN. Trois AHAs ont été bien décrites chez l'homme, c'est-à-dire que les métabolites, l'activation métabolique et la formation des adduits à l'ADN ont été expérimentalement identifiés chez les hépatocytes humains primaires, pour ces trois AHAs [41, 40, 52, 6]. Il s'agit des AHAs AαC, PhIP et MeIQx. L'étude de Delannée et al. [19] a permis d'établir la carte du métabolisme de chacune des 30 AHAs. Ces cartes étaient orientées dans la recherche de métabolites capables de former des adduits à l'ADN. Nous avons proposé (cf chapitre 2) un pipeline de prédiction du métabolisme des xénobiotiques. Les cartes obtenues par ce pipeline, à la différence des cartes de Delannée et al., sont annotées par un score qui décrit la probabilité qu'un métabolite soit produit.
En utilisant ce score comme base de comparaison entre les métabolites nous avons pu explorer les conditions physiopathologiques favorisant l'apparition des métabolites réactifs vis-à-vis de l'ADN. Nous avons dans un premier temps appliqué la partie de prédiction du métabolisme de notre pipeline aux 30 AHAs, ensuite 6 AHAs d'intérêt ont été sélectionnées et l'ensemble du pipeline a été appliqué. Nous avons alors obtenu des cartes du métabolisme annotées de ces 6 AHAs. Nous avons ensuite introduit la notion de contexte enzymatique, permettant de décrire un contexte physiopathologique, et la notion de signature enzymatique optimale, c'est-à-dire un ou des contextes enzymatiques permettant d'optimiser un objectif. L'objectif que nous avons fixé était de maximiser la chance de produire des métabolites réactifs à l'ADN.

4.1 Prédiction du métabolisme des 30 AHAs par la première étape du pipeline (SyGMa)

Nous avons commencé par appliquer la prédiction d'une carte du métabolisme pour chacune des 30 AHAs. Cette prédiction utilise l'outil SyGMa et est détaillée dans le chapitre 2. En résumé, SyGMa utilise un ensemble de règles de transformation biochimique à appliquer à une structure chimique. À chaque fois qu'une règle est appliquée, un métabolite est produit. Nous avons choisi d'utiliser les règles du groupe *phase I* et du groupe *phase II* que propose SyGMa. Ensuite nous avons choisi de paramétrer le nombre d'itérations de SyGMa à 2. Cela signifie que SyGMa a prédit un ensemble de métabolites une première fois puis a prédit de nouveaux métabolites à partir des métabolites précédemment obtenus. Enfin SyGMa a besoin d'une structure chimique de départ pour effectuer la prédiction de métabolites. Pour chacune des 30 AHAs la structure a été fournie sous la forme d'une formule SMILES disponible en annexe 2.

Les principales caractéristiques des 30 cartes du métabolisme obtenues sont décrites dans la table 4.1.

4.1.1 Identification des métabolites susceptibles de réagir avec l'ADN

Afin de déterminer la réactivité à l'ADN des métabolites, nous avons utilisé l'outil XenoSite Reactivity v1 [29, 30]. Cet outil permet de prédire les sites de réactivité (SORs) d'une molécule. De la même manière que les prédicteurs de SOMs, XenoSite Reactivity v1 propose de calculer un score pour chaque atome à partir des descripteurs moléculaires de l'atome. Ce score représente la chance que l'atome soit effectivement un SOR, c'est-à-dire l'atome qui réagit lors de la réaction de fixation à l'ADN. Un seul score annote le métabolite mais il existe un score par atome du métabolite. Le score qui annote le métabolite est le score le plus grand parmi l'ensemble des scores des atomes de la molécule. L'avantage d'utiliser XenoSite Reactivity v1 par rapport à d'autres outils de prédictions est que cet outil a déjà été appliqué aux métabolites des AHAs. L'étude de Delannée et al. [19] propose de prédire la réactivité à l'ADN des métabolites d'AHAs via XenoSite Reactivity. Elle a proposé un seuil de 0,85 permettant de discriminer, par le score annotant les métabolites, les métabolites réactifs des autres métabolites. Ce seuil a été défini à l'aide d'un jeu de données d'entraînement contenant des molécules connues pour former des adduits à l'ADN. Pour dénombrer les métabolites réactifs à l'ADN, on regarde pour chaque métabolite si le *score de réactivité à l'ADN* annoté est supérieur ou égal à 0,85. Si tel est le cas le métabolite est considéré comme réactif vis-à-vis de l'ADN.

4.1.1.1 Analyse des caractéristiques des cartes du métabolisme prédites

La table 4.1 décrit pour chaque carte du métabolisme des AHAs le nombre de métabolites, le nombre de réactions, le nombre de métabolites réactifs à l'ADN et le ratio entre ce nombre

AHA	Métabolites	Réactions	Métabolites Réactifs à l'ADN	Ratio entre les métabolites réactifs et les métabolites de la carte
4,7,8-TriMeIQx	194	282	91	46,9
4-CH2OH-8-MeIQx	189	266	80	42,3
7,8-DiMeIQx	174	250	81	46,6
7,9-DiMeIgQx	174	250	81	46,6
4,8-DiMeIQx	174	250	79	45,4
6,7-DiMeIgQx	169	245	76	45,0
AMPNH	157	225	16	10,1
7-MeIgQx	155	220	70	45,2
MeIQx	155	220	70	45,2
GluP1	142	202	64	45,1
IQx	137	192	62	45,3
IgQx	133	188	56	42,1
TrP1	129	179	61	47,3
PhIP	128	177	57	44,5
MeIQ	125	175	64	51,2
4'-OH-PhIP	123	166	49	39,8
3,5,6-TMIP	122	172	57	46,7
APNH	120	165	10	8,3
MeAαC	113	154	48	42,5
TrP2	113	154	49	43,4
GluP2	110	151	46	41,8
IQ	109	150	59	54,1
IQ[4,5-b]	109	150	59	54,1
1,5,6-TMIP	107	153	59	55,1
1,6-DMIP	95	132	53	55,8
IFP	90	123	37	41,1
AαC	85	111	33	38,8
PheP1	76	97	35	46,1
Harman	70	98	9	12,7
NorHarman	50	65	0	0,0
Moyenne	127,6	178,7	53,7	41,0
Médiane	124	173,5	58	45,2

TABLE 4.1 – **Table descriptive des cartes du métabolisme prédite par SyGMa avant calcul du score de confiance de chaque métabolite.**

de métabolites réactifs et le nombre total de métabolites prédits. On considère que le nombre de métabolites au sein de la carte du métabolisme représente la taille de la carte. Les AHAs 4,7,8-TriMeIQx et 4-CH2OH-8-MeIQx sont associées aux cartes les plus grandes avec 194 et 189 métabolites. Les AHAs associées aux cartes les plus petites sont NorHarman et Harman avec 50 et 70 métabolites respectivement. Le nombre de métabolites semble bien corrélé avec le nombre de réactions. Lors de la prédiction des métabolites, l'outil SyGMa recherche dans la structure chimique de la molécule des sous-structures, c'est-à-dire une structure chimique plus petite et complètement intégrée à la première. Lorsqu'une sous-structure correspond à une structure décrite dans les règles de biotransformations, la règle est appliquée et un métabolite est généré. Le nombre de métabolites décrits est donc directement dépendant des sous-structures que SyGMa identifie dans la structure de l'AHA dont on prédit le métabolisme. Nos résultats suggèrent qu'il existe moins de sous-structures pouvant être transformées chez NorHarmanet Harman que chez 4,7,8-TriMeIQx ou 4-CH2OH-8-MeIQx.

Parmi les AHAs associées aux cartes du métabolisme de grande taille nous avons noté 7,8-DIMeIQx et 7,9-DiMeIgQx qui partagent exactement les mêmes caractéristiques en termes de nombre de métabolites, de réactions ou encore de métabolites réactifs à l'ADN. 7,9-DiMeIgQx est l'isomère plan de 7,8-DIMeIQx, nous avons cherché à identifier d'autres couples d'isomères plan. Ainsi nous avons constaté que les couples MeIQx et 7,MeIgQx ou IQ et IQ[4,5-b] partageaient également les mêmes caractéristiques. Une hypothèse expliquant ce résultat est que les structures chimiques des isomères plan sont très proches. Chaque structure d'un couple d'isomères plan formerait alors exactement le même nombre de sous-structures chimiques ce qui impliquerait l'utilisation des mêmes règles de transformation par SyGMa et aboutirait au même nombre de réactions et de métabolites.

Les plus grandes cartes du métabolisme, associées à 4,7,8-TriMeIQx, 4-CH2OH-8-MeIQx, 7,8-DIMeIQx et 7,9-DiMeIgQx contiennent également le plus grand nombre de métabolites réactifs à l'ADN soit 91, 80, 81 et 81 respectivement. Dans le cas des petites cartes du métabolisme, NorHarman ne contient pas de métabolites réactifs à l'ADN et Harman est lui associé à 9 métabolites réactifs à l'ADN. Puisque les cartes sont de tailles variables nous avons introduit le ratio des métabolites réactifs par le nombre total de métabolites prédits par une carte. On constate que NorHarman, Harman, APNH et AMPNH sont associées à un score faible par rapport aux autres AHAs de 0%, 12,7%, 8,3% et 10,1%. Le ratio des autres AHAs oscille entre 38,8% et 55,8%. Ce faible ratio peut s'expliquer d'une part par la petite taille des cartes de NorHarman et Harman qui sont proches structurellement. Dans le cas de APNH et AMPNH les cartes sont de taille moyenne, mais le nombre de métabolites réactifs associés à chaque carte est bas. Ce résultat suggère que les structures de ces deux AHAs présentent une ou des particularités qui rendent peu réactifs à l'ADN les métabolites dérivés de ces AHAs

Nous nous sommes intéressés à la taille et au nombre de métabolites réactifs que proposent les cartes de PhIP, AαC et MeIQx qui sont les trois AHAs dont le métabolisme et les adduits sont les mieux décrits dans les hépatocytes humains primaires. Les cartes de PhIP et de MeIQx sont de tailles de 128 et 155 métabolites respectivement ce qui est proche de la moyenne d'environ 128 métabolites et de la médiane de 124 métabolites. La carte de AαC quant à elle, est plutôt petite avec 85 métabolites, ce qui représente deux tiers des cartes de PhIP ou MeIQx. On note également que le nombre de métabolites réactifs à l'ADN de AαC est deux fois moins grand que ceux de PhIP et MeIQx. Nous avons alors comparé le ratio du nombre de métabolites réactifs divisés par le nombre total de métabolites. Il est de 38,8 pour AαC et de 45,2% et 44,5% pour MeIQx et PhIP. Nous constatons que AαC présente bel et bien un nombre plus faible de métabolites réactifs à l'ADN que PhIP et MeIQx y compris lorsque ce nombre est relatif. C'est un résultat étonnant car il a été établi expérimentalement qu'AαC produit une quantité plus importante d'adduits à l'ADN que PhIP ou MeIQx [51].
Ce résultat étonnant pourrait s'expliquer par la faible taille de la carte du métabolisme de AαC, en effet cela suggère qu'un certain nombre de métabolites ont pu échapper à la prédiction de SyGMa. Cette hypothèse des prédictions incomplètes est en parie supportée par les résultats obtenus lors de l'application de la prédiction à la caféine. Nous avons pu établir que la prédiction des métabolites dérivés de NAT et notamment de NAT2 était lacunaire. Or NAT2 est directement responsable de la production de certains adduits à l'ADN dérivés des AHAs comme AαC []. Cependant, ci cette hypothèse est la bonne alors elle implique que les conséquences de cette prédiction lacunaire n'est pas la même pour tous les AHAs puisque les cartes de PhIP et MeIQx sont bien plus grandes que celle d'AαC. Une autre hypothèse est que les métabolites réactifs à l'ADN n'ont pas les mêmes capacités à effectivement produire des adduits lorsque l'on passe en condition expérimentale.

4.2 Annotation des cartes du métabolisme, calcul et analyse du score de probabilités de production

À partir des caractéristiques des cartes du métabolisme des AHAs nous avons sélectionné six AHAs d'intérêt pour appliquer la suite du pipeline et calculer le score de probabilités de production. En effet, une étape du pipeline est particulièrement chronophage et ne nous à pas permis d'appliquer le pipeline à l'ensemble des AHAs. Il s'agit de l'annotation des réactions par le *numéro d'atome de la réaction*. Cette annotation est pour le moment manuelle et requiert donc un temps d'annotation proportionnel à la taille des cartes que l'on annote.

4.2.1 Identification de six AHAs d'intérêt

Tout d'abord nous avons choisi d'appliquer la suite de notre pipeline aux cartes du métabolisme des trois AHAs les mieux décrites chez les hépatocytes humains primaires, c'est-à-dire AαC, PhIP et MeIQx. Ensuite nous avons complété cette liste par les trois cartes les plus grandes, et qui comportent le plus de métabolites réactifs à l'ADN, 4,7,8-TriMeIQx, 4-CH2OH-8-MeIQx et 7,8-DIMeIQx. 7,8-DIMeIQx représentant la paire d'isomères plan 7,8-DIMeIQx et 7,9-DiMeIgQx. La suite du pipeline de prédiction a été appliquée à ces six cartes.

4.2.2 Annotation des cartes du métabolisme des six AHAs

La première annotation que l'on fait est celle du *numéro d'atome de la réaction*. C'est une annotation manuelle qui est un prérequis pour annoter les réactions par un score SOM. Pour cela nous utilisons l'outil de visualisation MarvinView [11] qui nous permet de comparer les structures des métabolites consommés et produits d'une réaction. Une fois que l'on a repéré l'atome qui réagit dans le métabolite consommé la réaction est annotée par l'index de ce métabolite. L'indexe de l'atome est le numéro qui permet d'identifier l'atome. La numérotation des atomes dans MarvinView suit la norme IUPAC [22] c'est la même numérotation qui est utilisée par les outils prédicteurs de SOMs que sont Way2Drugs SOMP[61] et FAME 3[65].

Lors du parcours des cartes de AαC, PhIP et MeIQx pour annoter le *numéro d'atome* nous avons également repéré les métabolites correspondants aux métabolites expérimentalement identifiés de ces trois AHAs. Nous avons utilisé les tables 1.1, 1.2 et 1.3 qui répertorient les métabolites de chaque AHAs. La table 4.2 reprend la liste de ces métabolites et précise si le métabolite est retrouvé ou non dans les métabolites prédits par SyGMa.

4.2.3 Métabolites connus des trois AHAs de références dans les cartes prédites du métabolisme

Parmi les 11 métabolites identifiés expérimentalement de AαC, 9 sont retrouvés dans la carte du métabolisme. 7 des 9 métabolites identifiés expérimentalement de PhIP sont retrouvés dans sa carte du métabolisme et 6 des 10 métabolites connus de MeIQx sont retrouvés dans sa carte du métabolisme. Au total ce sont 8 métabolites dérivés de ces trois AHAs qui ne sont pas prédits par SyGMa. Parmi eux on retrouve les trois dérivés N-Sulfonyl et les trois dérivés N-Acetoxy de AαC, PhIP et MeIQx. Ce résultat nous a incité à explorer en détail les règles de biotransformations SMIRKS que propose SyGMa. Nous avons constaté qu'il était possible, selon ces règles, de prédire des métabolites sulfonyl ou acetoxy mais seulement depuis un oxygène lié à un carbone. Hors pour ces métabolites l'oxygène qui réagit est lié à un azote. Nous avons ainsi mis en lumière

Métabolite	Retrouvé dans la carte du métabolisme
AαC	OUI
AαC-3-O-Gluc	OUI
AαC-3-OH	OUI
AαC-3-O-SO3H	OUI
AαC-6-O-Gluc	OUI
AαC-6-OH	OUI
AαC-6-O-SO3H	OUI
AαC-HN2-O-Gluc	OUI
AαC-HN2-OH	OUI
AαC-N2-Gluc	OUI
N-Acetoxy-AαC	NON
N-Sulfonyloxy-AαC	NON
MeIQx	OUI
7-oxo-MeIQx	OUI
8-CH2OH-Iqx	OUI
HON-MeIQx	OUI
HON-MeIQx-N2-Gluc	OUI
Iqx-8-COOH	OUI
MeIQx-N2-Gluc	OUI
MeIQx-N2-SO3H	OUI
N-Acetoxy-MeIQx	NON
N-desmethyl-7-oxo-MeIQx	NON
N-Sulfonyloxy-MeIQx	NON
PhIP	OUI
4'HO-PhIP	OUI
4'-Ogluc-PhIP	OUI
4'-OSO3H-PhIP	OUI
HON-PhIP	OUI
HON-PhIP-N2-Gluc	OUI
PhIP-N2-Gluc	OUI
PhIP-N3-Gluc	OUI
N-Acetoxy-PhIP	NON
N-Sulfonyloxy-PhIP	NON

TABLE 4.2 – **Table descriptive des métabolites identifiés de AαC, PhIP et MeIQx.**

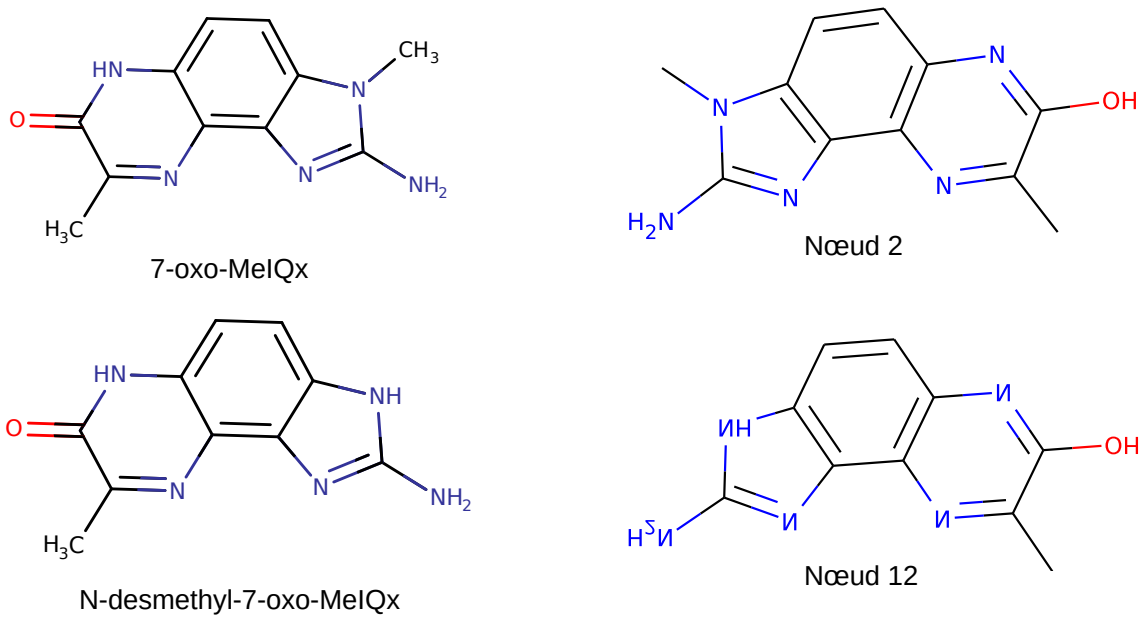
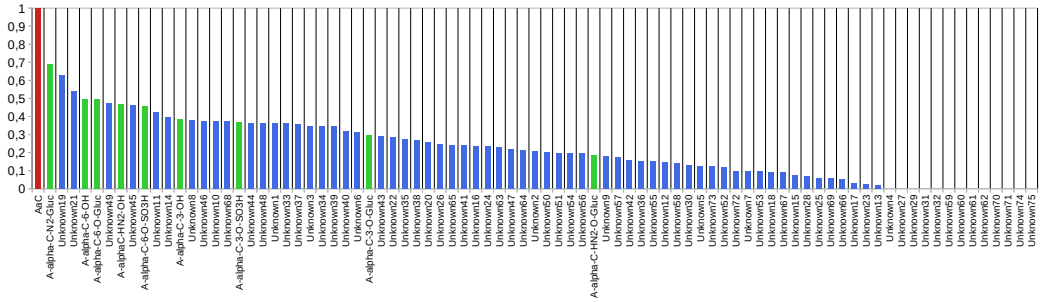


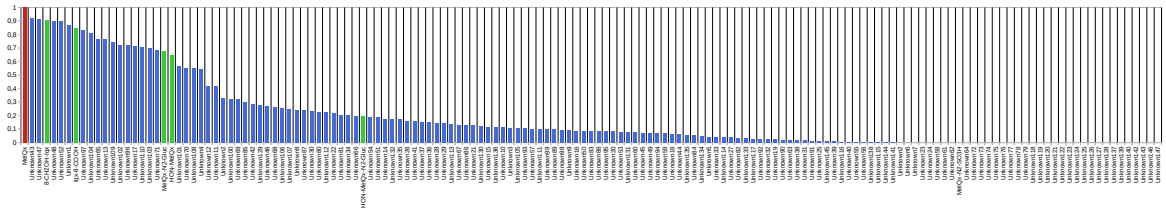
FIGURE 4.1 – **Comparaison des structures chimiques des métabolites de MeIQx.** Les métabolites expérimentalement identifiés de MeIQx sont décrits à gauche de la figure. Les métabolites de droite sont des métabolites provenant de la carte du métabolisme prédite, par SyGMa, de MeIQx.

deux lacunes des règles SMIRKS que propose SyGMa.

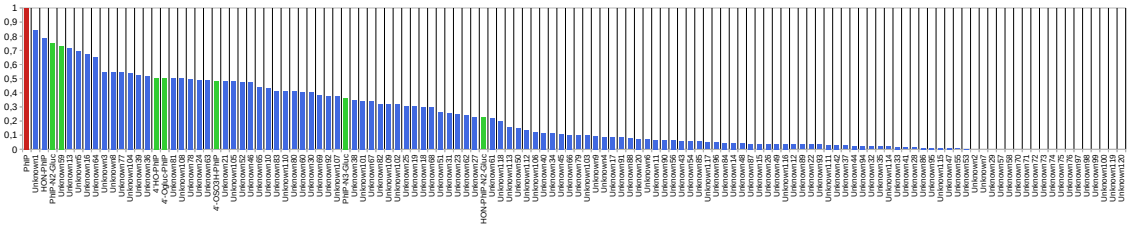
Les deux derniers métabolites non prédits sont le 7-oxo-MeIQx et le N-desmethyl-7-oxo-MeIQx. Le second est un dérivé du premier, il est donc normal qu'il ne soit pas prédit si le 7-oxo-MeIQx n'a lui-même pas été prédit. Lors de l'annotation des réactions de la carte de MeIQx nous avons relevé deux métabolites ayant une structure proche, le métabolite associé au noeud 2 de la carte et le métabolite associé au noeud 12. La figure 4.1 montre la comparaison entre ces métabolites. On remarque que le métabolite du noeud 2 est proche du 7-oxo-MeIQx, l'unique différence entre les deux et que le groupe cétone du 7-oxo-MeIQx est remplacé par un groupe hydroxyl pour le métabolite du noeud 2. La remarque est identique entre le N-desmethyl-7-oxo-MeIQx et le métabolite du noeud 12. Ce résultat comme le résultat précédent suggère une nouvelle lacune de prédiction de SyGMa des groupes cétone, associés à un carbone d'un hétérocycle. L'ensemble de ces résultats peuvent permettre d'améliorer les prédictions de SyGMa en identifiant précisément les règles SMIRKS qu'il faudrait rajouter afin de permettre la prédiction de ce genre de métabolites.



(a) Distribution du score de probabilités de production de chaque métabolite de la carte prédite du métabolisme d'AαC



(b) Distribution du score de probabilités de production de chaque métabolite de la carte prédite du métabolisme de MeIQx



(c) Distribution du score de probabilités de production de chaque métabolites de la carte prédite du métabolisme de PhIP

FIGURE 4.2 – **Distribution des scores de probabilités de production des AHAs de référence** Distribution des scores de probabilités de production de AαC (a), MeIQx (b) et PhIP (c). Les scores ont été rangés du plus grand au plus petit.

4.2.4 Calcul et distribution du score de confiance des AHAs de référence.

Les six cartes du métabolisme ont donc été annotées manuellement, puis le reste du pipeline a été appliqué. Les réactions on donc été annotées par un *nom de réaction*, une *famille d'enzyme*, un *numéro d'atome de la réaction*, un *score SOM* et une *enzyme*. Les annotations *score SOM* et *enzyme* ont permis de calculer un score de probabilités de production pour chaque métabolite. Le détail du calcul de ce score est décrit dans le chapitre 2. En résumé, nous utilisons les propriétés des réseaux Bayésien afin de calculer la probabilité qu'un métabolite soit produit. Pour cela on utilise les scores SOMs annotant les réactions comme des probabilités. Pour une réaction donnée le score SOM traduit la probabilité d'obtenir le métabolite produit à la condition que le métabolite consommé par la réaction soit lui-même produit. Le score de probabilités de production est une synthèse de ces probabilités qui traduit la chance pour un métabolite donné d'être produit. Ce score annote finalement les métabolites et une carte du métabolisme complètement annotée est obtenue en sortie du pipeline.

Nous avons ensuite analysé la distribution des scores de probabilités de production des métabolites des cartes de AαC, MeIQx et PhIP. Ces distributions sont décrites dans les figures 4.2a, 4.2b et 4.2c. Dans ces distributions le score du composé original est repéré par une barre rouge, les scores des métabolites expérimentalement identifiés sont repérés par une barre verte et les scores des métabolites inconnus sont repérés par une barre bleu. Le nom des métabolites est soit le nom usuel que nous avons défini dans les tables 1.1, 1.2 et 1.3, soit par le nom *Unknown* suivi du numéro d'identification du noeud auquel le métabolite est associé.

Dans un premier temps nous avons remarqué que les scores des métabolites expérimentalement identifiés sont compris entre 0,9 et 0,2. Un métabolite a retenu notre attention, il s'agit de MeIQx-N2-SO3H. Ce métabolite est annoté par un score de probabilités de production de 0,0. Un score de 0,0. Comme nous avons pu le décrire pour les métabolites de la caféine avec un score de 0,0 (cf chapitre 3) un score de 0,0 signifie qu'au moins une réaction est annotée par un score de 0,0. Hors ici le métabolite est expérimentalement identifié. Ce résultat suggère alors que les outils de prédictions de sites du métabolisme ont mal prédit le score SOM associé à la réaction concernée. Nous avons analysé la carte de MeIQx et nous avons remarqué que MeIQx-N2-SO3H est produit par une réaction de XX consommant MeIQx et produisant MeIQx-N2-SO3H. Cette réaction est donc de rang 1 et est associée à la famille d'enzymes SULTs selon le dictionnaire qui relie les noms des réactions, décrit en annexe 1. Cependant dans notre pipeline, on annote par un score SOM les réactions associées à cette famille d'enzymes, uniquement si cette réaction est de rang 2. En effet on ne peut annoter par un score SOM ces réactions qu'en utilisant FAME 3 et cet outil annote uniquement les réactions de rang 2. Pour autant la réaction peut être techniquement annotée par FAME 3 elle n'a pas été éliminée par le pipeline. Ce résultat permet d'invalider l'hypothèse d'une mauvaise prédiction de SOM de la part des outils utilisés

AHA	Métabolites dans la carte filtrée	Métabolites Filtrés	Réactions dans la carte filtrée	Réactions filtrées	Métabolites réactifs à l'ADN dans la carte filtrée	Métabolites réactifs à l'ADN filtrés
4,7,8-TriMeIQx	87	107	120	162	41	50
7,8-DiMeIQx	90	84	116	134	43	38
MeIQx	72	83	90	130	27	43
PhIP	63	65	74	103	22	35
4-CH2OH-8-MeIQx	63	126	78	188	31	49
AαC	59	26	72	39	20	13

TABLE 4.3 – **Table décrivant le nombre de métabolites, de réactions et de métabolites réactifs à l'ADN des 6 AHAs sélectionnées.** Cette table décrit le nombre de métabolites, de métabolites réactifs à l'ADN et de réactions dénombrés dans les cartes filtrées des 6 AHAs. Elle décrit, pour chacun de ces paramètre, le nombre qui a été éliminé par la filtration. La filtration consiste en l'élimination des métabolites dont le score de probabilité de production était strictement inférieur au seil, ici le seuil était de 0,10.

mais met en lumière une des caractéristiques inhérente au modèle que l'on utilise.

Ensuite nous avons constaté que MeIQx et PhIP présentent un grand nombre de métabolites associés à un score de probabilités de production inférieur à 0,1 comparativement à AαC. Plus précisément ce sont 83 métabolites avec un score inférieur à 0,1 pour MeIQx (53,5% des métabolites de la carte). Parmi ces métabolites, 36 sont associés à un score de 0,0 (23,2% de la carte). Pour PhIP ce sont 65 métabolites (50,8% de la carte) qui sont annotés par un score inférieur à 0,1 dont 18 métabolites (14,1% de la carte) avec un score nul. Alors que pour AαC ce sont 26 métabolites (30,6% de la carte) avec un score inférieur à 0,1 dont 13 (15,3% de la carte) avec un score nul. Ce résultat indique qu'une grande partie des cartes du métabolisme prédites est peu supporté par notre score de probabilités de production. Nous proposons alors d'utiliser notre score comme un outil de filtration de ces cartes afin de les réduire aux seuls métabolites avec un support suffisant.

4.3 Filtration des cartes du métabolisme

4.3.1 Effet de la filtration sur les métabolites

Nous avons donc utilisé notre score de probabilités de production comme outil de filtration. Pour cela nous avons éliminé tout les métabolites des six cartes dont le score de probabilités de production était inférieur à 0,1. Ensuite l'ensemble des métabolites et réactions isolés, c'est-à-dire que les métabolites et réactions ne sont plus liés au composé original. La table4.3 dénombre les métabolites et métabolites réactifs à l'ADN dans les cartes filtrées ainsi que les réactions. Elle dénombre également la quantité de métabolites, métabolites réactifs et de réactions filtrées. L'annexe 3 détaille les métabolites des 6 cartes du métabolisme des AHAs. Nous avons constaté

que la filtration des cartes avec un seuil sur le score de probabilités de production de 0,1 a eu un fort impact sur les cartes. Par exemple la carte associée au 4-CH2OH-MeIQx a été filtré de 66% des métabolites (126 métabolites) qu'elle contenait, les cartes de PhIP, MeIQx, 4,7,8-TriMeIQx et 7,8-DiMeIQx ont été réduites d'environ la moitié des métabolites qui les composaient soit 51%, 54%, 55% et 48% des métabolites respectivement. AαC n'a pas été autant affecté par la filtration des métabolites. En effet 26 métabolites ont été filtrés ce qui correspond à 31% des métabolites de la carte de AαC. Nos résultats suggèrent d'une part qu'un nombre important de métabolites et de réactions est peu ou pas soutenu par les informations qu'apportent les outils de prédictions de sites du métabolismes. Ensuite qu'il existe chez 4-CH2OH-8-MeIQx mais aussi PhIP, MeIQx, 4,7,8-TriMeIQx et 7,8-DiMeIQx des sous structures chimiques qui produisent un grand nombre de métabolites qui sont ensuite peu crédibles. La structure chimique de AαC semble être suffisamment différente des 5 autres AHAs ce qui a eu pour conséquence une plus faible prédiction de métabolites mais avec une plus grande proportion de métabolites soutenue par nos outils.

4.3.2 Effet de la filtration sur les métabolites réactifs à l'ADN

Environ la moitié des métabolites filtrés sont des métabolites réactifs à l'ADN, à l'exception du cas de 4-CH2OH-8-MeIQx. La filtration a donc un impact modéré sur les ratios du nombre de métabolites réactifs sur le nombre total de métabolites. Cependant les AHAs AαC, PhIP et MeIQx présentent maintenant un nombre proche de métabolites réactifs à l'ADN dans leur carte (entre 20 et 27). De plus on note que pour les 6 AHAs ce sont la moitié des métabolites réactifs de la carte initiale qui ont été filtrés. Ce résultat montre l'intérêt d'utiliser notre score de probabilité comme outil de filtration qui permet de conserver un ensemble de métabolites réactifs à l'ADN soutenus par nos outils. Cela est particulièrement intéressant pour nos analyses suivantes qui se concentrent sur l'étude des conditions permettant la formation de métabolites réactifs à l'ADN chez les AHAs.

Nous avons pu vérifier que le MeIQx-N2-SO3H de la carte du MeIQx a bien été éliminé du fait de son score de probabilités de production nul.

4.4 Conditions favorisant la production de métabolites réactifs chez les AHAs

Nous proposons ici une autre utilisation du score de probabilités de production. Celui-ci peut être utilisé pour comparer deux métabolites et ainsi dire quel métabolite a plus de probabilité d'être produit. On peut également comparer les scores d'un même métabolite dans différentes conditions. Comme nous l'avons plusieurs fois évoqué, notre méthode de calcul du score de probabilités de production nécessite à la fois les annotations *score SOM* et *enzyme* des réactions.

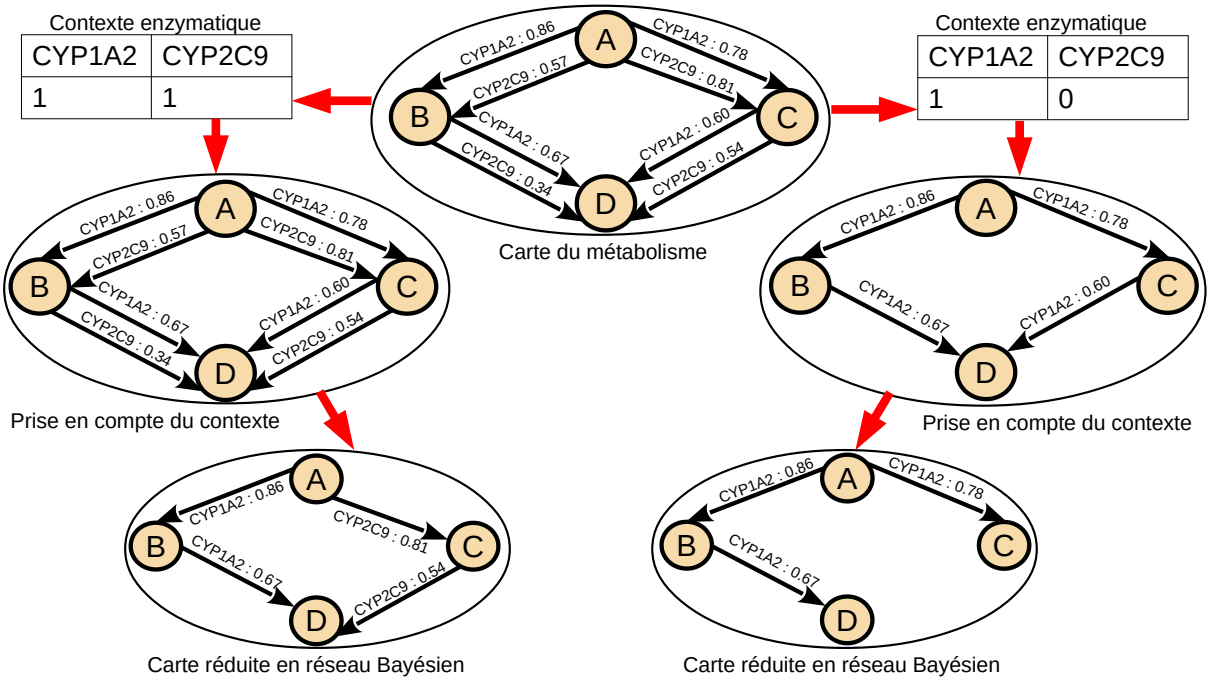


FIGURE 4.3 – **Effet des contextes enzymatiques sur la réduction des cartes du métabolisme en réseau bayésien.** La carte du métabolisme annotée, au centre, peut servir de base à la formation d'un réseau bayésien pour calculer le score de probabilités de production de chaque métabolite. Ici les contextes enzymatiques, décrits par les tables de gauche et droite, influencent ce calcul en réduisant une première fois la carte en éliminant les réactions associées aux enzymes indisponibles. Ici 1 représente une enzyme disponible et 0 une enzyme indisponible.

Ce score est donc dépendant des enzymes. Nous avons défini des contextes enzymatiques qui permettent de calculer des scores de probabilités de production en fonction de contextes physiopathologiques différents. Puis nous avons recherché, à l'aide de comparaisons des scores de probabilités de production, les conditions qui favorisent l'apparition de métabolites réactifs à l'ADN. Pour cela nous avons introduit le concept de *signatures optimales en termes de réactivité*. Cette ou ces signatures représentent des contextes enzymatiques particuliers qui maximisent la chance de produire des métabolites réactifs à l'ADN.

4.4.1 Influence des contextes enzymatiques sur le calcul du score de probabilités de production

Nous introduisons ici le concept de *contextes enzymatiques*. Un contexte enzymatique décrit pour plusieurs enzymes si elles sont disponibles ou non. Lorsqu'une enzyme n'est pas disponible, par exemple parce que si on décrit une condition *in vitro* dans laquelle on ajoute les enzymes CYPs mais pas UGTs, alors il est absurde de prendre en compte des réactions qui sont catalysées par ces enzymes. Notre modèle prend en compte ce paramètre. Lorsqu'une enzyme est "indispo-

nible" les réactions associées ne sont plus prises en compte lors de la réduction de la carte vers un réseau Bayésien. Ceci est décrit dans la figure 4.3. Le pipeline produit deux réseaux bayésiens différents à partir de la même carte du métabolisme en fonction du contexte. À gauche le premier contexte décrit que toutes les enzymes que l'on peut rencontrer dans la carte annotée sont disponibles. Dans le second à droite 2C9 est considéré comme indisponible. Ainsi avant la réduction de la carte en réseau bayésien, les réactions associées aux enzymes indisponibles sont éliminées. La carte de droite n'a plus que des réactions catalysées par le CYP1A2. Ensuite la réduction de la carte en réseau bayésien intervient, l'ensemble des étapes de cette réduction est décrite dans le chapitre 2. Les cartes réduites en réseaux bayésiens sont alors très différentes et les scores de probabilités de production de C et D en seront modifiés.

La *table des contextes enzymatiques* décris l'ensemble des contextes enzymatiques possible pour une carte du métabolisme. A chaque contexte on peut calculer un score de probabilité de réaction pour chaque métabolite. En comparant les score entre eux on peut alors identifier les contextes qui favorisent la production d'un métabolite donné.

Nous avons utilisé les 6 cartes filtrées des AHAs que nous avons précédemment décrit afin de calculer le score de probabilités de production des métabolite et cela dans chacun des 512 contextes enzymatique possible.

4.4.2 Calcul des signature optimale en terme de réactivité

Nous nous sommes alors intéressé aux contextes enzymatiques favorisant la production des métabolites réactifs à l'ADN. Pour cela nous calculons un nouveau score, celui de la *production globale de métabolites réactifs*. Ce score consiste en la somme des scores de probabilités de production des métabolites considérés comme réactifs. On peut calculer ce score pour chaque contexte enzymatique et ainsi isoler les contextes enzymatiques pour lesquels ce score est maximal.

Ensuite, dans cet ensemble de contexte on recherche le ou les contextes les plus petits pour lequel ou lesquels ce score est maximal. Cet ensemble de petits contextes constitue la ou les *signature(s) optimales en termes de réactivité*. On peut traduire ces signatures par la plus petite combinaison d'enzymes *disponibles* permettant de maximiser la chance de produire des métabolites réactifs à l'ADN.

Nous avons donc calculé les *signatures optimales en termes de réactivité* pour chacune des 6 cartes filtrées. La figure 4.4 décrit les signatures associées à chaque AHAs. Nous avons constaté que les enzymes SULTs, UGTs et CYP1A2 sont communes aux signatures des 6 AHAs. Ce résultat est en accord avec la littérature qui a décrit l'implication des SULTs et du CYP1A2 dans la formation des adduits à l'ADN des AHAs[73, 12]. De plus une étude de 2017 a montré une nouvelle voie de formation de certains adduits à l'ADN de AαC impliquant les UGTs[6]. Ensuite les enzymes NATs et GSTs sont communément absentes des signatures. Nous avons recherché

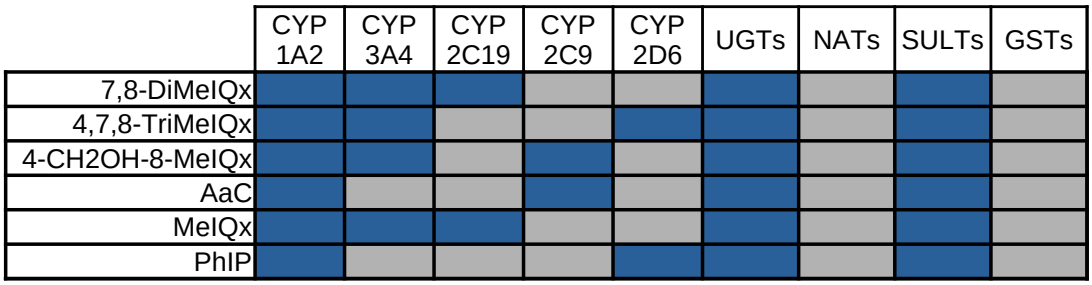


FIGURE 4.4 – **Signatures enzymatiques optimales en termes de réactivité.** Une cellule bleue correspond à une enzyme disponible. Une cellule grise correspond à une enzyme indisponible

dans les 6 cartes si il existait des réactions annotées par GSTs ou NATs. Nous avons montré qu'il n'existait aucune réaction annotée par les GSTs y compris dans les cartes avant filtration. À l'inverse, il n'existe aucune carte filtrée ou non, qui ne contient pas au moins une réaction annotée par NATs, il en va de même pour les UGTs, SULTs, CYP1A2, CYP2C19, CYP2C9, CYP2D6 et CYP3A4. On explique donc l'absence des GSTs dans les signatures optimales par l'absence de cette enzyme dans les cartes. En revanche, l'absence des NATs dans les signatures optimales est étonnant. En effet il a été montré que les NATs catalysent des voies de production des adduits à l'ADN chez les AHAs []. Ce résultat pourrait être expliqué par une hypothèse que nous avons faite lors de l'analyse de la carte prédite de la caféine. Notre hypothèse propose que SyGMa est incomplet concernant certaines réactions et notamment au niveau des réactions catalysées par les NATs. Ainsi cette lacune n'a pas permis de prédire des métabolites réactifs et dérivés de NATs.

Nos résultats proposent également des signatures enzymatiques différentes en fonction des AHAs, en effet la présence des CYPs CYP2C9, CYP2C19, CYP2D6 et CYP3A4 est variable en fonction de l'AHA. Par exemple on constate que pour MeIQx, 7,8-DiMeIQx, 4,7,8-TriMeIQx et 4-CH2OH-MeIQx l'enzyme CYP3A4 est dans la signature. Elle discriminerait les AHAs avec une structure proche de MeIQx. Les différentes isoformes de CYPs ne seraient donc pas impliquées de la même manière dans la formation des adduits en fonction de la structure de l'AHA. Ce résultat reste à confirmer en appliquant le reste du pipeline aux 24 autres AHAs.

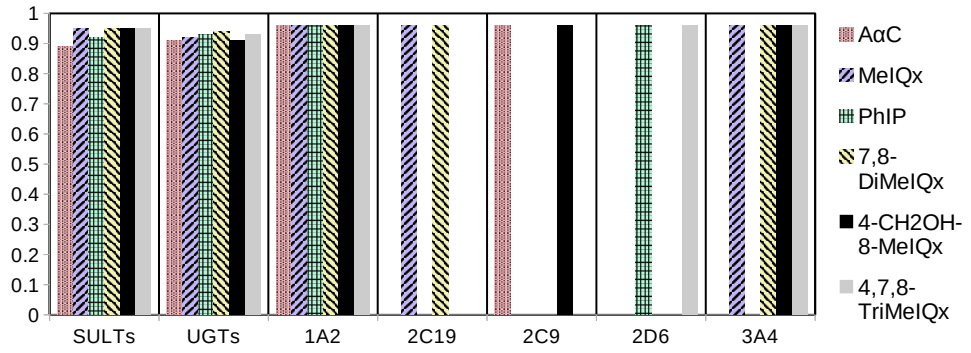


FIGURE 4.5 – **Impact du seuil définissant les métabolites réactifs sur les signatures optimales en termes de réactivité.** Les seuils sur les scores de XenoSite Reactivity sont décrits en ordonnée. Les enzymes sont décrites en abscisse. Pour chaque AHA une barre pleine signifie que l'enzyme est présente dans la signature et l'absence de barre signifie que l'enzyme est absente dans la signature.

4.4.3 Impact du seuil de définition des métabolites réactifs à l'ADN

Un métabolite réactif à l'ADN est un métabolite dont le score de réactivité à l'ADN, provenant de XenoSite Reactivityy v1, est supérieur ou égal à 0,85. Nous avons voulu vérifier la robustesse de nos signatures en recalculant nos signatures pour chaque seuil sur le score de XenoSite Reactivity entre 0,0 et 1,00. Notre démarche est la suivante, à chaque seuil sur le score XenoSite Reactivity un ensemble de métabolites est considéré comme réactif à l'ADN. Ce sont les métabolites annotés par un score de réactivité supérieur ou égal au seuil. Cet ensemble de métabolite est également annoté par un *score de probabilités de production*. Ces scores permettent le calcul du *score production globale de métabolites réactifs* qui les additionnent et ce score global permet de déterminer les signatures. À chaque seuil sur le score de réactivité, l'ensemble des métabolites réactifs à l'ADN peut changer entraînant un changement dans le *score production globale de métabolites réactifs* et *in fine* un changement dans les *signatures optimales en termes de réactivité*.

Au vu des résultats précédents concernant les prédictions sur les réactions associées au NATs, nous avons décidé de retirer de notre analyse les NATs. De même les GSTs ne seront plus représentées car elles n'annotent aucune réaction.

On cherche donc à connaître l'influence de ce seuil sur les signatures optimales. La figure 4.5 décrit pour chaque seuil sur le score XenoSite Reactivity, la signature optimale en termes de réactivité pour chaque AHAs. On remarque que les signatures ne varient que très peu selon

le seuil XenoSite Reactivity. Les signatures se stabilisent lorsque le seuil est de 0,89. Tous les seuils inférieurs produisent les mêmes signatures. On retrouve donc les résultats de la figure 4.4, les enzymes CYP1A2, SULTs et GSTs sont communes aux signatures des 6 AHAs et le CYP3A4 est spécifique des AHAs avec une structure proche de MeIQx. Ce résultat suggère que les enzymes des signatures optimales sont suffisantes pour prendre en charge l'ensemble des voies métaboliques menant aux métabolites réactifs. La signature ne changeant pas non plus lorsque le seuil est à 0,0 cela signifie que la signature optimale décrit l'ensemble des enzymes qui sont les plus à même de métaboliser l'ensemble de la carte du métabolisme de l'AHA correspondant. La figure 4.6 détaille la partie supérieure de la figure 4.5. On peut voir dans cette nouvelle figure les seuils sur le score XenoSite Reactivity à partir duquel les enzymes sont intégrées à la signature optimale en termes de réactivité.

Nous rappelons ici que le score XenoSite Reactivity représente la chance que le métabolite soit réactif à l'ADN. Plus il est haut plus le métabolite peut former un adduit à l'ADN. On remarque dans la figure 4.6 que le seuil le plus bas à partir duquel l'ensemble des signatures optimales deviennent stables est 0,89 avec AαC. Ensuite on remarque que les premières enzymes qui intègrent la signature optimale sont les CYPs, que ce soit le CYP1A2 ou les autres isoformes en fonction de l'AHA. Les CYPs sont intégrées à la signature lorsque le seuil XenoSite Reactivity est à 0,98 pour PhIP et 0,97 pour les 5 autres AHAs. Ce résultat signifie que les métabolites considérés comme les plus réactifs sont tous des dérivés de réactions catalysées uniquement par les cytochromes P450. Nous avons aussi constaté une différence entre les signatures de AαC et PhIP et celles de MeIQx, 4,7,8-TriMeIQx, 7,8-DiMeIQx et 4-CH2OH-8-MeIQx. Pour le groupe d'AαC ce sont les métabolites dérivés des UGTs qui sont les métabolites les plus réactifs après les dérivés des CYPs. Pour le groupe de MeIQx ce sont les métabolites dérivés des SULTs qui sont les plus réactifs à l'ADN après les dérivés de CYPs. Ce résultat suggère donc que la structure des AHAs proche de MeIQx est suffisamment différente de la structure des autres AHAs et forme ainsi des dérivés via les SULTs plus réactifs que les dérivés SULTs des autres AHAs. Cette hypothèse nécessite cependant l'application du pipeline complet à l'ensemble des 30 AHAs afin de tester cette hypothèse à travers les autres AHAs proches de MeIQx.

4.5 Conclusion

Notre pipeline a reconstruit les cartes du métabolisme des 30 AHAs identifiées à ce jour. 6 d'entre elles ont été sélectionnées pour être annotées manuellement, soit en fonction des connaissances dont nous disposions sur leur bioactivation et sur la formation des adduits à l'ADN, soit en fonction des critères des cartes. Nous avons ainsi pu mettre en lumière la qualité des prédictions de métabolites de SyGMa à partir des cartes de AαC, PhIP et MeIQx. Nous avons constaté d'une

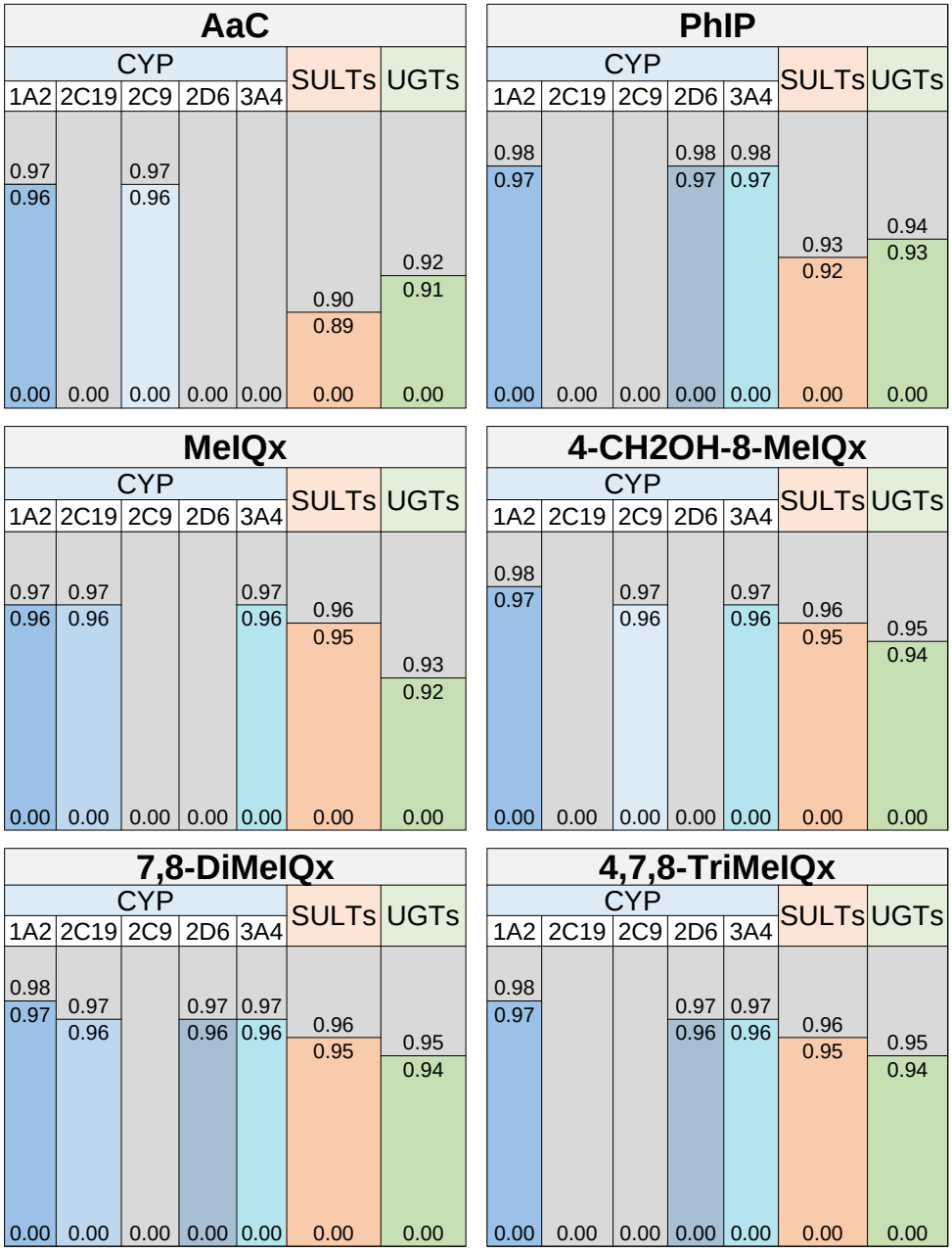


FIGURE 4.6 – **Signatures optimales en termes de réactivité en fonction du seuil XenoSite Reactivity utilisé.** Pour chaque enzyme, la valeur du score XenoSite qui permet l'intégration de l'enzyme dans la signature optimale est indiquée.

part, que la majorité des métabolites identifiés expérimentalement étaient prédits par SyGMa. D'autre part que les métabolites identifiés expérimentalement et non retrouvés sont associés à trois types de réactions spécifiques que SyGMa ne peut pas prédire. Tout comme l'application du pipeline à la caféine, ce résultat met en avant certaines lacunes dans les prédictions que propose SyGMa, l'identification de trois types de réactions spécifiquement non prédites permet de cibler ces lacunes. Cela permet d'identifier précisément les règles de biotransformations manquantes parmi l'ensemble des règles que propose SyGMa. Cela démontre l'avantage d'utiliser SyGMa, cet outil est adaptable et permet l'ajout ou le retrait de règles de biotransformations. Nos résultats sur les AHAs et la caféine nous permettront de définir de nouvelles règles de biotransformations qui pourront être intégrées à SyGMa et améliorer ses qualités de prédictions.

Ensuite nous avons calculé le score de probabilité de réaction des métabolites de 6 cartes sélectionnées. La distribution de ces scores et la filtration de ces cartes sur ce score nous a permis de mettre en évidence qu'il existe une grande quantité de métabolites prédits par SyGMa et peu soutenus par les outils prédicteurs de SOMs. Ces résultats nous ont permis de montrer l'intérêt de filtrer les métabolites à partir de notre score. Ensuite nous avons utilisé notre score et les concepts de *contexte enzymatique* et de *signature optimale en termes de réactivité* pour explorer les liens entre la disponibilité des enzymes et la production de métabolites réactifs à l'ADN. Nous avons montré l'implication des enzymes CYP1A2, UGTs et SULTs dans la production des métabolites réactifs à l'ADN de chacune des AHAs. Dans le même temps, nous avons montré la spécificité de certaines enzymes notamment le CYP3A4 exclusif aux signatures des AHAs proches de MeIQx. Ces résultats préliminaires demandent d'être approfondis par l'application du pipeline aux 24 autres AHAs. Ils constituent une piste de recherche intéressante pour identifier les métabolites réactifs inconnus des 27 AHAs non décrites dans les hépatocytes humains primaires. Le concept de *signature optimale en termes de réactivité* peut permettre d'identifier des conditions physiopathologiques qui peuvent être comparées aux données d'expressions disponibles dans les bases de données comme TCGA [48] ou GTEx [44]. On pourrait ainsi déterminer si une condition physiopathologique aurait tendance à augmenter ou diminuer la production de métabolites réactifs par rapport à une autre condition. Ces résultats montrent également un nouvel intérêt de notre *score de probabilité de production*. En effet celui-ci peut être utilisé pour comparer les métabolites entre eux. De plus il peut être utilisé pour définir des objectifs, comme la maximisation des métabolites réactifs à l'ADN. Cet objectif peut être adapté suivant d'autres critères par exemple les métabolites susceptibles de se lier à un type de récepteur membranaire spécifique.

Notre pipeline de prédiction du métabolisme et d'enrichissement des cartes du métabolisme présente donc des résultats intéressants à plusieurs niveaux et nous voulons maintenant l'étendre à

la prédiction d'autres xénobiotiques d'intérêts comme l'aflatoxine B1 ou encore le glyphosate. Ce pipeline a permis de répondre à notre problématique qui était de déterminer le rôle des enzymes dans la production des métabolites réactifs à l'ADN chez les AHAs.

Chapitre 5

Conclusion & Perspectives

Afin de prédire le métabolisme et la formation des adduits à l'ADN des AHAs, nous avons produit un pipeline. Celui-ci permet de prédire les cartes du métabolisme des xénobiotiques et de les enrichir par des annotations. Ces cartes enrichies nous permettent de retrouver les conditions physiopathologiques les plus favorables à la formation de métabolites réactifs vis-à-vis de l'ADN. Ce pipeline a été appliqué aux AHAs ce qui a permis de répondre à notre problématique qui consistait à déterminer le rôle des enzymes dans la production des métabolites réactifs vis-à-vis de l'ADN.

5.1 Contributions méthodologiques

Nous avons montré que l'état de l'art des méthodes de prédictions du métabolisme et de la réactivité ne permettait pas de répondre à la problématique du rôle des enzymes dans la production des métabolites réactifs vis-à-vis de l'ADN. Nous nous sommes spécifiquement intéressés à deux études très proches de notre problématique, celle de Delannée et al. 2019 et celle de Dang et al . 2018 [19, 17].

Nous avons montré que Delannée et al. proposait un pipeline de prédiction du métabolisme et de la réactivité à l'ADN des AHAs. Celui-ci permettait d'évaluer le potentiel à former des adduits sur la base du nombre de métabolites réactifs présents dans les cartes prédites. Cependant, rien ne permet d'interroger le rôle des enzymes sur la production des métabolites réactifs. De plus, rien ne permet non plus de comparer les métabolites entre eux afin d'avoir une meilleure visualisation de la formation des métabolites capables de former des adduits à l'ADN. Ce pipeline permet toutefois de répondre à la prédiction du métabolisme et de la réactivité à l'ADN qui plus est pour les xénobiotiques qui nous intéressent.

L'étude de Dang et al. 2018 nous a permis de déterminer une utilisation inédite des scores provenant d'outils de prédictions des sites du métabolisme. En effet ces scores annotaient les réactions et représentaient la probabilité que la réaction se produise. Ensuite ces probabilités permettaient de calculer la probabilité d'emprunter des voies métaboliques. Cependant, cette étude ne permettait pas de comparer les métabolites entre eux mais les voies métaboliques, de

plus cette étude ne contextualisait pas les probabilités avec des enzymes ce qui ne permet pas d'obtenir des informations sur le rôle des enzymes.

Nous avons donc développé notre propre pipeline capable de dépasser ses limites en intégrant les enzymes comme un élément des cartes ayant de l'influence sur la réalisation des réactions. Pour cela, notre pipeline prédit, dans un premier temps, des cartes du métabolisme. Pour cela il utilise l'outil SyGMa qui est un outil intégrable et modulable. Nous avons choisi cet outil car il permet à la fois d'être amélioré par l'ajout de règles de biotransformations plus précises mais également car il ne peut plus devenir inaccessible comme cela a déjà été le cas dans l'étude de Delannée et al. Une fois le métabolisme du xénobiotique prédit, celui-ci est retranscrit sous la forme d'une carte du métabolisme. Celle-ci est inspirée de la structure des cartes de Delannée et al. 2019. À la différence de cette étude, nous enrichissons nos cartes via plusieurs annotations. Cet enrichissement a pour but d'utiliser les résultats des prédicteurs de SOMs comme Dang et al. 2018. Ces résultats peuvent annoter les réactions et être utilisés comme des probabilités. Cependant nous ne nous contentons pas d'annoter les réactions avec ce score. Elles sont également annotées par des enzymes, liant ainsi le score à une condition biologique qui est la disponibilité de l'enzyme. C'est ce qui nous permet de retrouver le rôle des enzymes dans la production des métabolites d'intérêts, ici les métabolites réactifs vis-à-vis de l'ADN. Nous avons ajouté à notre méthode la modélisation de nos cartes sous la forme de réseaux Bayesiens. Cette modélisation a permis d'utiliser les probabilités pour dépasser le cadre de Dang et al et calculer la probabilité des métabolites en plus de la probabilité des voies métaboliques. De cette manière nous avons pu déterminer le rôle des enzymes sur la production des métabolites et non plus seulement sur les réactions ou voies métaboliques.

5.2 Contributions biologiques

Notre pipeline a été appliqué à sept molécules, six AHAs et la caféine. La caféine a été utilisée pour valider notre approche. Nous avons choisi la caféine car c'est un xénobiotique et que son métabolisme chez l'homme est proche du métabolisme des AHAs. Dans ce métabolisme certains métabolites peuvent être produits par différentes voies métaboliques et notre méthode peut prendre en compte la multiplicité des voies métaboliques. La caféine était donc parfaitement adaptée pour évaluer notre approche. Une fois notre approche validée nous avons appliqué le pipeline aux AHAs. Cette application à permis de répondre à la question du rôle des enzymes dans la production des adduits à l'ADN.

Lorsque le pipeline a été appliqué à la caféine nous avons pu montrer plusieurs résultats intéressants. Tout d'abord, nous avons montré que la majorité des métabolites de la caféine identifiés expérimentalement sont retrouvés par les prédictions de SyGMa. Deux d'entre eux

n'étaient pas prédits et sont des dérivés issus de réactions catalysées par les N-acetyl transferases. Cela nous à conduit à deux conclusions, tout d'abord que SyGMa était bien un outil adapté pour prédire le métabolisme des xénobiotiques. Ensuite que SyGMa avait malgré tout quelques lacunes dans les prédictions qu'il propose et notamment des réactions catalysées par les NATs. Ce résultat a été renforcé lorsque le pipeline a été appliqué aux AHAs qui présentaient une possible lacune sur la prédiction des réactions catalysées par les NATs.

Les résultats obtenus sur la caféine nous ont permis de déterminer que notre *score de probabilité de production* permet de discriminer les métabolites identifiés expérimentalement des autres métabolites. Seuls deux métabolites inconnus étaient moins discriminés et nous avons alors proposé des hypothèses structurelles pour expliquer leurs forts scores. Nous avons pu montrer que la valeur de notre score n'était pas corrélé à la quantité de métabolites que l'on retrouve expérimentalement dans le métabolisme de la caféine chez l'homme. Du fait de sa capacité à discriminer les métabolites, nous avons proposé une utilisation de ce score comme outil de filtration des cartes du métabolisme.

L'intérêt de notre approche ayant été démontré par l'application à la caféine, nous avons alors appliqué notre pipeline aux AHAs. Six d'entre elles ont été sélectionnées afin de permettre la production de cartes enrichies. Parmi ces six AHAs, trois ont été sélectionnées car leur métabolisme et les adduits à l'ADN qu'elles forment ont été décrits chez les hépatocytes humains primaires. Pour les trois autres, elles ont été sélectionnées car les cartes du métabolisme prédites pour ces trois AHAs étaient les cartes contenant le plus de métabolites.

Nous avons pu montrer que la majorité des métabolites identifiés expérimentalement chez les AHAs est prédite par SyGMa. Les métabolites manquants nous ont permis de déterminer des lacunes dans les prédictions de SyGMa. En effet, la majorité des métabolites manquant étaient issus de réactions spécifiques. Ces réactions ne peuvent pas être prédites par les règles de biotransformatons que propose SyGMa. Ensuite deux métabolites de MeIQx étaient également manquants, ceux-ci étaient également issus d'une réaction spécifique formant un groupe cétone que SyGMa ne peut pas prédire avec les règles de biotransformations qu'il propose. De plus l'analyse ultérieure des signatures favorisant la formation d'adduits à l'ADN nous a permis de conforter les résultats obtenus par l'application du pipeline à la caféine . En effet il semble que SyGMa ait des lacunes a prédire des réactions catalysées par les NATs. Nous avons ensuite analysé les scores de probabilité de production des métabolites de ces six AHAs. Nous avons tout d'abord constaté qu'un métabolite de MeIQx, identifié expérimentalement, était associé à un score SOM de zéro. Nous avons pu montrer que ce score n'était pas dû à un faible support de la part des outils prédicteurs de SOMs. Ce score est directement issu de notre méthode qui annote les réactions en fonction des rangs. Puisque ce métabolite était issu d'une réaction de rang 1 mais catalysé par les SULTs, le score de la réaction a été automatiquement considéré comme

un zéro. Les réactions de rang 1 ne sont annotées que par WayToDrugs SOMP qui ne prédit pas de score SOM pour les SULTs. Pour autant la réaction pouvait être annotée par FAME3 et n'a donc pas été éliminée. En utilisant le score de probabilité de production comme outil de filtration tel que décrit lors de l'application à la caféine, nous avons montré qu'un grand nombre de métabolites prédits par SyGMa étaient peu soutenus par les outils prédicteurs de SOMs. Une fois les cartes filtrées, nous avons de nouveau calculé le score de probabilité de production afin de déterminer les conditions favorisant la production des métabolites réactifs vis-à-vis de l'ADN. Pour cela, nous avons introduit des *contextes enzymatiques* qui influencent le calcul des scores de probabilité de réactions. Ces contextes précisent les enzymes disponibles ou non et si l'enzyme est indisponible alors les scores SOMs associés à ces enzymes ne sont pas pris en compte dans le calcul du score de probabilité de production. Un score de probabilité de production peut être calculé pour chaque métabolite et chaque condition, c'est en comparant les probabilités des métabolites réactifs vis-à-vis de l'ADN selon différentes conditions que l'on peut identifier les conditions favorables à l'apparition d'adduits à l'ADN. Ces conditions constituent les *signatures optimales en termes de réactivité à l'ADN*. Ces signatures permettent alors d'identifier le rôle des enzymes dans la production des métabolites réactifs vis-à-vis de l'ADN. Nous avons ainsi identifié que les enzymes CYP1A2 (une isoforme de CYP), UGTs et SULTs permettent la formation de métabolites réactifs chez les 6 AHAs. L'enzyme CYP3A4 quant à elle serait spécifique de la production de métabolites réactifs chez les AHAs ayant une structure chimique proche de MeIQx. Nous avons également montré, à travers l'utilisation de différents seuils pour déterminer la réactivité à l'ADN des métabolites, que les métabolites les plus réactifs vis-à-vis de l'ADN des AHAs sont des dérivés de réactions catalysées par les CYPs.

Notre pipeline a donc permis de répondre à notre problématique et complète bien la méthodologie des études précédentes en permettant notamment de comparer les métabolites entre eux et d'interroger l'implication des enzymes dans la formation de métabolites clés. L'ensemble de ces travaux ont fait l'objet d'une publication, actuellement en processus de review et disponible en annexe 4, et d'une communication dans le cadre de la conférence internationale *Environmental Mutagenesis and Genomics Society* (EMGS).

5.3 Perspectives

L'application de notre pipeline aux AHAs et à la caféine nous a permis d'identifier plusieurs aspects d'amélioration et d'application de notre approche. Nous présenterons tout d'abord les améliorations que nous allons appliquer à notre pipeline à partir des lacunes que nous avons identifiées. Ensuite nous présenterons les applications biologiques que permet notre pipeline. Nous détaillerons notamment l'exploration des cartes enrichies de l'ensemble des AHAs. Mais aussi les applications à d'autres métabolites et l'exploration des données de protéomiques et de

génomiques au travers de notre méthode.

5.3.1 Amélioration du pipeline

Comme nous l'avons montré précédemment il existe plusieurs points sur lesquels notre pipeline peut être amélioré. Dans un premier temps nous détaillerons les améliorations qui permettront de prédire plus précisément le métabolisme des AHAs. Ensuite nous reviendrons sur l'un de nos choix de modélisation et détaillerons des solutions permettant de dépasser les limites qu'impliquent ces choix. Enfin nous discuterons des améliorations qui permettraient de tester la robustesse des signatures que l'on prédit.

Ajout de règles de biotransformations Tout d'abord nous avons pu identifier plusieurs réactions qui ne sont pas prédites par SyGMa lors de l'application du pipeline aux AHAs et à la caféine. Une possibilité d'amélioration est donc de déterminer quelles sont les réactions non prédites par SyGMa. Ensuite il faudrait synthétiser ces réactions sous la forme de règles de biotransformations au format SMIRKS pour que ces règles soient prises en compte par SyGMa. Cela permettra de prédire plus précisément les métabolites des xénobiotiques.

Changer la notion de rang Nous avons pu montrer avec le MeIQx-N2-SO3H que notre stratégie d'annotation des réactions par rang n'était pas idéale et nous prive de certains métabolites pourtant identifiés expérimentalement. Une solution pour améliorer cette stratégie pourrait être d'annoter les réactions non plus à partir des rangs mais directement à partir de l'annotation *famille d'enzymes*. Il nous faudrait alors choisir quel outil prédicteur de SOM utiliser pour chaque famille puis annoter en fonction de ce choix.

Ajout d'autres outils prédicteurs de SOMs Une dernière amélioration que l'on pourrait apporter à plus long terme serait d'utiliser un plus grand nombre d'outils prédicteurs de SOMs. Cela permettrait d'une part de pouvoir décrire plus d'enzymes, par exemple XenoSite Metabolism [80] propose un modèle pour d'autres isoformes de CYPs et ADMET Predictor [32] propose un score pour 9 isoformes d'UGTs. La grande diversité de ces outils entraîne une diversité de scores qui peut avoir une influence sur notre score de probabilité de prédiction. Si l'on prend l'exemple de deux prédicteurs de SOMs associées aux UGTs XenoSite UGT [16] et Way2Drugs SOMP [61]. On pourrait déterminer deux types de signatures pour la même condition enzymatique, la signature où les réactions de glucuronidations ont été annotées par XenoSite UGT et la signature où ces réactions on été annotées par Way2Drugs SOMP. Cette stratégie aurait un coût non négligeable en temps car cela implique de multiplier les prédictions de SOMs pour les mêmes métabolites. De plus la combinatoire d'enzymes possibles suivant les outils disponibles pour pré dire les SOMs des CYPs, des UGTs, des GSTs, des NATs ou des SULTs peut vite de-

venir importante ce qui entraînerai un grand nombre de signatures. Cependant cela permettrait de tester la robustesse de nos signatures mais aussi d'établir des signatures consensus. Celles-ci pourraient être utilisées pour déterminer encore plus efficacement le rôle des enzymes dans la formation des adduits à l'ADN ou pour d'autres cibles.

5.3.2 Applications Biologiques

Notre pipeline peut permettre de nouveaux apports en l'état actuel. Nous reviendrons sur la prédiction du métabolisme des AHAs et les apports de l'application de notre pipeline aux 30 AHAs. Ensuite, nous montrerons l'intérêt de notre méthode pour explorer des données d'expression et de protéomique afin de comparer des conditions physio-pathologiques par le prisme de la prédiction de la réactivité vis-à-vis de l'ADN. Enfin, nous développerons comment notre méthode peut être démocratisée et étendue à d'autres objets d'études que les adduits à l'ADN dérivés des xénobiotiques.

Exploration des cartes enrichies des 30 AHAs Notre objectif à court terme est d'appliquer notre pipeline aux 26 autres AHAs afin de déterminer les signatures optimales en termes de réactivité vis-à-vis de l'ADN. Pour cela il est nécessaire de réaliser l'annotation manuelle *atome de la réaction*. C'est une étape indispensable du pipeline qui est cependant chronophage car elle nécessite la comparaison manuelle des structures chimiques des métabolites consommés et produits de chaque réaction. Une solution pour réduire ce temps pourrait être d'automatiser cette annotation. Notre proposition est de comparer automatiquement les structures chimiques des molécules consommées et produites sous la forme de graphes. En recherchant les sous-graphes, c'est-à-dire de plus petits graphes intégrés aux graphes de base, communs entre les deux structures, on pourrait déterminer l'atome qui a réagi.

Une fois l'ensemble des cartes enrichies des AHAs produites, nous pourrions confirmer certaines de nos pistes par exemple la spécificité de l'isoforme CYP3A4 dans l'activation métabolique de certaines AHAs, comme nos résultats le suggèrent pour MeIQx. On pourrait également confirmer l'implication du CYP1A2 dans la production des métabolites réactifs de tous les AHAs ou au contraire déterminer quelles sont les AHAs qui n'utilisent pas cette enzyme.

Une fois les cartes des 30 AHAs générées, nous nous intéresserons à la structure chimique de métabolites spécifiques. Nous avons montré, lors de l'application de la méthode aux AHAs, l'effet des seuils sur le score de réactivité sur les signatures optimales en terme de réactivité. On propose alors d'isoler les premiers métabolites considérés comme réactifs par ces seuils. Comparer les structures de ces métabolites permettrait de déterminer les structures des métabolites les plus réactives dérivés des AHAs. Si des structures communes apparaissent mais ne sont pas toutes partagées par les 30 AHAs, cela permettrait de classer les AHAs en fonction des métabolites

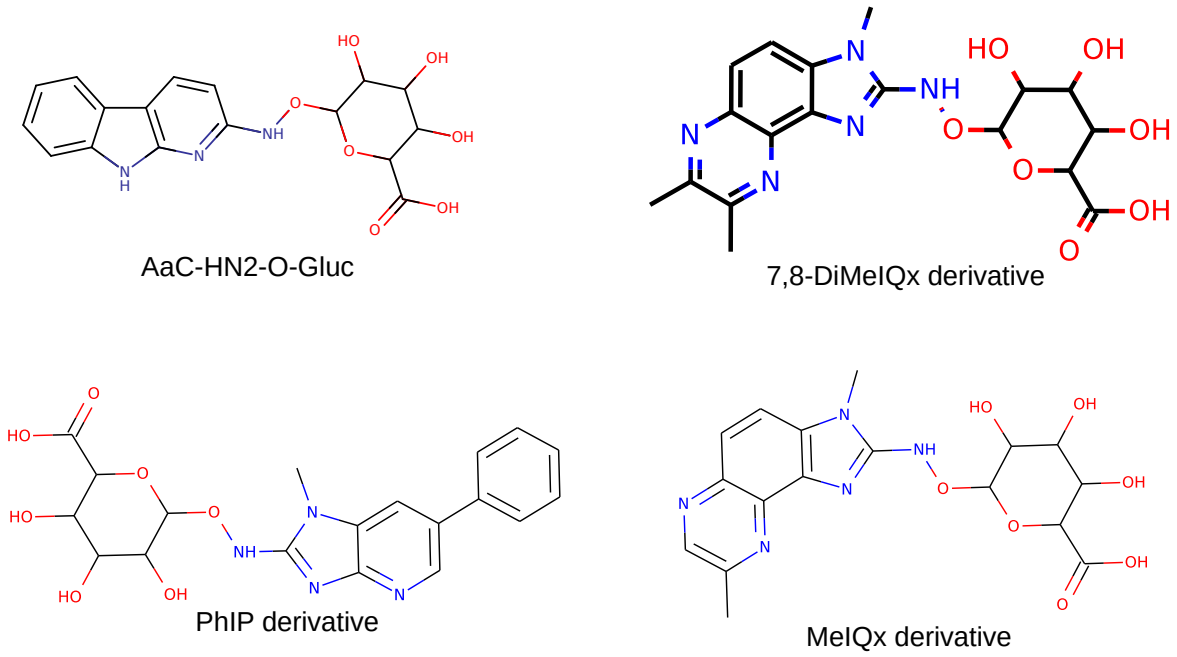


FIGURE 5.1 – **Comparaison des structures chimiques des métabolites les plus réactifs vis-à-vis de l'ADN et dérivés d'une glucuronidation de PhIP, MeIQx et 7,8-DiMeIQx.** Comparaison de la structure chimique des dérivés les plus réactifs de PhIP, MeIQx et 7,8-DiMeIQx avec la structure de AαC-HN2-O-Gluc qui est un métabolite dérivé d' AαC à l'origine de la formation d'adduits à l'ADN. Chaque métabolite provient d'une glucuronidation qui s'est effectuée sur l'oxygène d'un groupe hydroxyl lié à un azote.

réactifs qu'elles formeraient. Nous avons commencé ce travail à partir des six cartes que nous avons générées et nous avons pu noter un résultat préliminaire interressant en comparant la structure chimique de métabolites dérivés de PhIP, MeIQx et 7,8-DiMeIQx avec la structure d'un métabolite identifié de AαC. Les structures de ces métabolites sont décrites dans la figure 5.1. Nous avons noté que la structure chimique des métabolites les plus réactifs vis-à-vis de l'ADN, issus d'une glucuronidation, dérivés de PhIP, MeIQx et 7,8-DiMeIQx, est similaire à la structure de AαC-HN2-O-Gluc. Ce métabolite dérivé de AαC a été décrit comme étant capable de former des adduits à l'ADN [6]. Ce résultat préliminaire nous conforte dans l'idée de comparer les structures des métabolites les plus réactifs des cartes du métabolisme des 30 AHAs.

Exploration de différentes conditions biologiques À moyen terme, nous voulons pouvoir appliquer notre méthode afin de caractériser la capacité à former des adduits à l'ADN de différentes conditions physio-pathologiques. Dans notre approche, ces conditions sont les conditions enzymatiques qui décrivent la disponibilité des enzymes. L'idée serait donc de discrétiser des données de protéomique décrivant l'expression des enzymes du métabolisme des xénobiotiques ou, dans une moindre mesure, l'expression de gènes codant ces mêmes enzymes. Ces données peuvent être spécifiques de certains tissus ou encore être spécifiques de conditions physio-pathologiques. On pourrait alors comparer ces conditions et déterminer, pour un xénobiotique donné, si celui-ci a plus de chances de former des métabolites réactifs vis-à-vis de l'ADN dans un tissu spécifique ou une condition physio-pathologique particulière. Nous avons tenté de comparer les conditions *foie sain* et *foie atteint par un carcinome hépatocellulaire* (CHC) à partir de la discrétisation des données d'expression de gènes. Ces données sont extraites de la base de données TCGA pour les foies atteints par un CHC et GTEx pour les foies sains. Une étude a normalisée les données d'expression des deux bases de données et propose une discrétisation des données d'expression pour chaque gène et conditions [23]. Des résultats préliminaires ne nous ont pas permis de différencier les conditions car la discrétisation des enzymes, que l'on retrouve dans notre modèle, était la même pour les deux conditions. Cependant, on sait que les CHC sont très hétérogènes et cette particularité a pu fortement impacter la discrétisation de l'expression des gènes du métabolisme des xénobiotiques. Nous prévoyons de rechercher des sous-populations au sein des données d'expression de CHC disponibles dans TCGA, pour isoler les données d'expression de ces populations, les discrétiser et ainsi les comparer afin de déterminer si certaines populations seraient plus favorables à la formation d'adduits à l'ADN.

Construction de cartes pour d'autres contaminants de l'environnement À plus long terme, nous voudrions appliquer cette méthode à d'autres xénobiotiques dont le métabolisme dépend des mêmes enzymes. Nous pensons notamment aux mycotoxines comme l'aflatoxine B1 et aux pesticides et herbicides de la famille du glyphosate, du parathion/malathion et des pentachlorophénol (PCP) et 2,4,6trichlorophénol (TCP), de la famille de l'aldrine, le dieldrine,

et le 3,3',4,4'tétrachloroazobenzène (TCAB) tous aujourd'hui classés cancérogènes possibles ou probables (2A,2B) par l'IARC. Notre méthode permettrait d'identifier des métabolites d'intérêts dérivés de ces molécules avec à la fois une forte probabilité d'altérer l'ADN mais aussi une forte probabilité d'être produit à travers notre score.

Prise en compte d'autres critères de toxicité Il est également possible de déterminer de nouvelles cibles avec lesquelles nous pourrions utiliser notre score de probabilité de production. L'exemple que nous prenons est celui des métabolites capables de lier les récepteurs des oestrogènes : ERα et ERβ. En effet, les modèles QSAR sont développés afin de montrer la probabilité du lien entre une structure chimique et une activité comme la génotoxicité par exemple. En caractérisant l'activité de liaison à ces récepteurs, on peut alors prédire les conditions qui favoriseraient la production de métabolites pouvant lier ces récepteurs en lien avec les perturbateurs endocriniens.

Notre objectif à plus long terme est donc de démocratiser l'utilisation de notre méthode qui peut dépasser le contexte de la réactivité vis-à-vis de l'ADN et des maladies chroniques hépatiques. Cela pourrait permettre de déterminer des conditions favorables à l'apparition de critères d'intérêts ou encore déterminer si une condition donnée est plus favorable à l'apparition d'un critère d'intérêt qu'une autre condition.

Chapitre 6

Annexe

Annexe 1

Cette annexe détaille les 30 AHAs leur structure 2D et formules SMILES associées. Il s'agit du fichier supplémentaire numéro 2 déposé avec l'article soumis à *BMC Bioinformatics*.

Additional file 2. **Description of thirty HAAs and Caffeine structure used as pipeline input**. This pdf file describes the SMILES formula and 2D-structure of each of the thirty HAAs and the Caffeine studied in the paper.

Xenobiotic Name	SMILES formula	2D Structure
1,5,6-TMIP	N=1C=2C=C(C(=CC2N(C1N)C)C)C	
1,6-DMIP	N=1C=2C=CC(=CC2N(C1N)C)C	
3,5,6-TMIP	N=1C=2C=C(C(=NC2N(C1N)C)C)C	
4-CH2OH-8-MeIQx	OCC1=CC=2N=CC(=NC2C=3N=C(N)N(C31)C)C	
4,7,8-TriMeIQx	N=1C=2C=C(C3=C(N=C(N)N3C)C2N=C(C1C)C)C	
4,8-DiMeIQx	N1=CC(=NC2=C1C=C(C3=C2N=C(N)N3C)C)C	
4'-OH-PhIP	OC=1C=CC(=CC1)C=2C=NC=3N=C(N)N(C3C2)C	
6,7-DiMeIgQx	N=1C=2C=C3N=C(C(=NC3=CC2N(C1N)C)C)C	
7-MeIgQx	N1=CC(=NC2=CC3=C(N=C(N)N3C)C=C12)C	
7,8-DiMeIQx	N=1C=2C=CC3=C(N=C(N)N3C)C2N=C(C1C)C	
7,9-DiMeIgQx	N1=CC(=NC=2C1=CC=3N=C(N)N(C3C2C)C)C	
AαC	N=1C(N)=CC=C2C1NC=3C=CC=CC32	
AMPNH	N=1C=CC=2C=3C=CC=CC3N(C4=CC=C(N)C(=C4)C)C2C1	
APNH	N=1C=CC=2C=3C=CC=CC3N(C4=CC=C(N)C=C4)C2C1	
GluP1	N1=C(N)C=CC=2N=C3C(=CC=CN3C12)C	
GluP2	N1=C(N)C=CC=2N=C3C=CC=CN3C12	
Harman	N=1C=CC=2C=3C=CC=CC3NC2C1C	
IFP	N1=C2N=C(N)N(C2=CC=3OC(=CC13)C)C	
IgQx	N1=CC=NC2=CC3=C(N=C(N)N3C)C=C12	
IQ	N1=CC=CC2=C1C=CC3=C2N=C(N)N3C	
IQ[4,5-b]	N1=C2N=C(N)N(C2=CC3=CC=CC=C13)C	
IQx	N1=CC=NC2=C1C=CC3=C2N=C(N)N3C	
MeAαC	N=1C(N)=C(C=C2C1NC=3C=CC=CC32)C	
MeIQ	N1=CC=CC2=C1C=C(C3=C2N=C(N)N3C)C	
MeIQx	N1=CC(=NC2=C1C=CC3=C2N=C(N)N3C)C	
NorHarman	N=1C=CC2=C(C1)NC=3C=CC=CC32	
PheP1	N1=CC(=CC=C1N)C2=CC=CC=C2	
PhIP	N1=CC(=CC2=C1N=C(N)N2C)C=3C=CC=CC3	
TrP1	N=1C(N)=C(C=2NC=3C=CC=CC3C2C1C)C	
TrP2	N1=C(N)C=C2NC=3C=CC=CC3C2=C1C	
Caffeine	CN1C=NC2=C1C(=O)N(C(=O)N2C)C	

Annexe 2

Dictionnaire des règles SMIRKS de SyGMA. Ce dictionnaire a été construit manuellement et relie le nom des règles SMIRKS qu'utilise SyGMa. A chaque règles est associé une famille d'enzymes. Lorsque les réactions sont annotées par un socre SOM nous utilisons un outil en fonction de la famille d'enzymes et du rang de la réaction. Cette table répertorie également l'outil utilisé en fonction du rang et la ou les valeurs que peut prendre l'annotation *enzymes*.

Nom de la règle SMIRKS	Famille d'enzymes annotée	Outil prédicteur de SOMs utilisé	Enzyme annotée
N-deacetylation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
N-dealkylation_(R-NHCH2-alkyl)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
N-dealkylation_(c-NHCH2-alkyl)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

N-dealkylation-_(morpholine)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
N-dealkylation-_(nCH2)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
N-dealkylation-_(piperazine)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
N-dealkylation-_(quarternary_N)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
N-dealkylation-_(tertiaryN-CH2-alkyl)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

N-deformylation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
N-deglycosidation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
N-demethylation_(R-N(CH3)2)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
N-demethylation_(R-N(CR)CH3)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
N-demethylation_(R-NHCH3)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

N-demethylation_(c-N(CH3)2)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
N-demethylation_(c-NHCH3)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
N-demethylation-_(nCH3)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
N-depropylation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
N-oxidation_(-N=)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

N-oxidation- _(RN(CH3)2)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
N-oxidation_(aniline)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
N-oxidation_(tertiary- _N)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
N-oxidation_(tertiary- _NCH3)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
O-deacetylation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1

O-dealkylation-_(aliphatic)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
O-dealkylation-_(aromatic)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
O-dealkylation-_(methylenedioxyphenyl)a	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
O-dealkylation-_(methylenedioxyphenyl)b	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
O-deglycosidation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

O-demethylation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
S-dealkylation_c-SCH2-R	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
acetyl_shift	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
aldehyde_oxidation-_(aliphatic)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
aldehyde_oxidation-_(aromatic)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

aldehyde_reduction- _(aliphatic)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
aldehyde_reduction- _(aromatic)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
aliphatic- _dehalogenation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
aliphatic- _hydroxylation- _(primary_carbon- _next_to_SP2_or- _SP1)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
aliphatic- _hydroxylation- _(primary_carbon- _next_to_quart- _carbon)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1

aliphatic- _hydroxylation- _(primary_carbon- _next_to_sec- _carbon)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
aliphatic- _hydroxylation- _(primary_carbon- _next_to_tert- _carbon)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
aliphatic- _hydroxylation_(sec- _carbon,next_to- _CH3)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
aliphatic- _hydroxylation_(sec- _carbon_both_sides- _next_to_SP2,in_a- _ring)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
aliphatic- _hydroxylation_(sec- _carbon_in_a_ringA)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1

aliphatic-_hydroxylation_(sec-_carbon_in_a_ringB)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
aliphatic-_hydroxylation_(sec-_carbon_next_to-_SP2,in_a_ring)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
aliphatic-_hydroxylation_(sec-_carbon_next_to-_SP2,not_in_a_ring)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
aliphatic-_hydroxylation_(tert-_carbon_linked_to-_two_CH3_groups)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
aliphatic-_hydroxylation_(tert-_carbon_next_to-_SP2)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

all_aliph_hydr	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
all_dehydro	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
aniline_to_nitro	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
aromatic-_dechlorination	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
aromatic-_dehydroxylation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

aromatic-_hydroxylation-_(meta_to_carbon)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
aromatic-_hydroxylation-_(ortho_to_2-_substituents)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
aromatic-_hydroxylation-_(ortho_to_nitrogen)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
aromatic-_hydroxylation-_(ortho_to_oxygen)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
aromatic-_hydroxylation_(para-_to_carbon)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

aromatic- _hydroxylation_(para- _to_nitrogen)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
aromatic- _hydroxylation_(para- _to_oxygen)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
aromatic- _hydroxylation- _(sulfur_containing- _5ring)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
aromatic_oxidation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
aromatic_oxidation- _(nitrogen_containing- _5ring)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1

azide_cleavage	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
benzylic-_hydroxylation_(c-CH1-CH3)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
benzylic-_hydroxylation_(c-CH1-CR)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
benzylic-_hydroxylation_(c-CH2-CH3)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
benzylic-_hydroxylation_(c-CH2-CR)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

benzylic-_hydroxylation_(c-CH2-N)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
benzylic-_hydroxylation_(c-CH3)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
beta-oxidation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
carbonyl_reduction-_(aliphatic)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
carbonyl_reduction-_(both_sides_next-_to_aromatic_carbon)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

carbonyl_reduction-_(next_to_SP2-_carbon)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
carbonyl_reduction-_(next_to_aromatic-_carbon)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
carboxylation-_(benzylic_CH3)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
carboxylation-_(primary_carbon-_next_to_SP2)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
carboxylation-_(primary_carbon-_next_to_quart-_carbon)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

carboxylation-_(primary_carbon-_next_to_sec-_carbon)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
carboxylation-_(primary_carbon-_next_to_tert-_carbon)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
cyclic_hemiacetal-_ring_opening	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
decarboxylation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
dehydration_next_to-_SP2_a	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

dehydration_next_to-_SP2_b	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
dehydration_next_to-_SP2_both_sides	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
dehydrogenation-_(CH1-CH3->C=CH2)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
dehydrogenation-_(CH2-CH3->C=CH2)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
dehydrogenation-_(alpha,beta_to_SP2)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

dehydrogenation-_(alpha,beta_to_SP2-_both_sides)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
dehydrogenation-_(amine)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
dehydrogenation-_(aromatization_of-_1,4-dihydropyridine)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
deiodonidation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
diazene_cleavage	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

double_bond- _reduction	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
double_bond- _reduction_(aromatic)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
double_bond- _reduction_(benzylic)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
epoxidation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
epoxide_hydrolysis	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1

haloacid_hydrolysis	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
het-O-demethylation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
hydrazone_hydrolysis	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
hydrolysis-_(CNC(OH)R)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
hydrolysis_(N-substituted-pyridine)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

hydrolysis_(X=X-X-_exclude_phosphate)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
hydrolysis_(ester)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
hydrolysis-_(heteroatom_bonded-_amide)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
hydrolysis-_(methoxyester)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
hydrolysis_(primary-_amide)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

hydrolysis_(secondary-_amide)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
hydrolysis_(tertiary-_amide)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
hydrolysis_(urea_or-_carbonate)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
hydroxyl-amide_5ring-_closure	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
hydroxyl-amide_5ring-_rearr	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

hydroxyl-amide_6ring-_rearr	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
imine_hydrolysis	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
isopropenyl_oxidation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
n-deglycosidation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
nitrile_to_amide	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

nitro_to_aniline	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
nitro_to_nitroso	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
oxidation_(C=N)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
oxidation_(amine_in_a_ring)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
oxidation_to_quinone	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

oxidative- _deamination- _(amidine)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
oxidative- _deamination- _(aromatic)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
oxidative- _deamination_(on- _primary_carbon)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
oxidative- _deamination_(on- _secondary_carbon)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
oxidative- _decarboxylation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1

oxidative- _dehalogenation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
phosphine_sulphide- _hydrolysis	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
primary_alcohol- _oxidation_(aliphatic)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
primary_alcohol- _oxidation_(benzylic)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
ring_closure_(NH1- 5bonds-carboxyl)2	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1

ring_closure_(NH1-6bonds-carboxyl)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
ring_closure-_(hydroxyl-5bonds-carboxyl)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
ring_closure-_(hydroxyl-6bonds-carboxyl)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
secondary_N-depropylation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
secondary_alcohol-_oxidation_(aliphatic)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

secondary_alcohol- _oxidation_(benzylic)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
secondary_aliphatic- _carbon- _hydroxylation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
steroid_17hydroxy_to- _keto	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
steroid_d5d4	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1
sulfide_oxidation_(C- S-C)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes- CYPs de rang 1

sulfide_oxidation_(c-S-C)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
sulfide_oxidation_(c-S-c)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
sulfoxide_oxidation-_(C-S-C)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
sulfoxide_oxidation-_(c-S-C)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
sulfoxide_oxidation-_(c-S-c)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

sulfoxide_reduction	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
tautomerisation_(keto->enol)	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
tertiary_N-depropylation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
thiophene_oxidation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
try	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1

vinyl_oxidation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
xanthine_oxidation	CYPs	Réaction rang 1 : **Way2Drug SOMP**	CYP1A2 ; CYP2C19 ; CYP2C9 ; CYP2D6 ; CYP3A4
		Réaction rang 2 : **FAME 3**	Toutes isoformes-CYPs de rang 1
N-acetylation_(NH1)	NATs	Réaction rang 1 : **No Tool**	Pas d'annotations
		Réaction rang 2 : **FAME 3**	NATs
N-acetylation_(NH1-CH3)	NATs	Réaction rang 1 : **No Tool**	Pas d'annotations
		Réaction rang 2 : **FAME 3**	NATs
N-acetylation_(aniline)	NATs	Réaction rang 1 : **No Tool**	Pas d'annotations
		Réaction rang 2 : **FAME 3**	NATs
N-acetylation-_(aromatic_-nH-)	NATs	Réaction rang 1 : **No Tool**	Pas d'annotations
		Réaction rang 2 : **FAME 3**	NATs
N-acetylation-_(heteroatom_bonded-NH2)	NATs	Réaction rang 1 : **No Tool**	Pas d'annotations
		Réaction rang 2 : **FAME 3**	NATs
N-glucuronidation-_(N(CH3)2)	UGTs	Réaction rang 1 : **Way2Drug SOMP**	UGTs
		Réaction rang 2 : **FAME 3**	UGTs
N-glucuronidation-_(NCH3_in_a_ring)	UGTs	Réaction rang 1 : **Way2Drug SOMP**	UGTs
		Réaction rang 2 : **FAME 3**	UGTs
N-glucuronidation-_(NH_in_a_ring)	UGTs	Réaction rang 1 : **Way2Drug SOMP**	UGTs
		Réaction rang 2 : **FAME 3**	UGTs
N-glucuronidation-_(aliphatic_NH2)	UGTs	Réaction rang 1 : **Way2Drug SOMP**	UGTs
		Réaction rang 2 : **FAME 3**	UGTs
N-glucuronidation-_(aniline)	UGTs	Réaction rang 1 : **Way2Drug SOMP**	UGTs
		Réaction rang 2 : **FAME 3**	UGTs
N-glucuronidation-_(aniline_NH1-R)	UGTs	Réaction rang 1 : **Way2Drug SOMP**	UGTs
		Réaction rang 2 : **FAME 3**	UGTs

N-glucuronidation-_(aromatic_-nH-)	UGTs	Réaction rang 1 : **Way2Drug SOMP**	UGTs
		Réaction rang 2 : **FAME 3**	UGTs
N-glucuronidation-_(aromatic_=n-)	UGTs	Réaction rang 1 : **Way2Drug SOMP**	UGTs
		Réaction rang 2 : **FAME 3**	UGTs
O-glucuronidation_(N-hydroxyl)	UGTs	Réaction rang 1 : **Way2Drug SOMP**	UGTs
		Réaction rang 2 : **FAME 3**	UGTs
O-glucuronidation-_(aliphatic_carboxyl)	UGTs	Réaction rang 1 : **Way2Drug SOMP**	UGTs
		Réaction rang 2 : **FAME 3**	UGTs
O-glucuronidation-_(aliphatic_hydroxyl)	UGTs	Réaction rang 1 : **Way2Drug SOMP**	UGTs
		Réaction rang 2 : **FAME 3**	UGTs
O-glucuronidation-_(aromatic_carboxyl)	UGTs	Réaction rang 1 : **Way2Drug SOMP**	UGTs
		Réaction rang 2 : **FAME 3**	UGTs
O-glucuronidation-_(aromatic_hydroxyl)	UGTs	Réaction rang 1 : **Way2Drug SOMP**	UGTs
		Réaction rang 2 : **FAME 3**	UGTs
aliphatic_N_sulfation1	SULTs	Réaction rang 1 : **No Tool**	Pas d'annotations
		Réaction rang 2 : **FAME 3**	SULTs
aliphatic_N_sulfation2	SULTs	Réaction rang 1 : **No Tool**	Pas d'annotations
		Réaction rang 2 : **FAME 3**	SULTs
aromatic_N_sulfation2	SULTs	Réaction rang 1 : **No Tool**	Pas d'annotations
		Réaction rang 2 : **FAME 3**	SULTs
sulfation_(aliphatic-_hydroxyl)	SULTs	Réaction rang 1 : **No Tool**	Pas d'annotations
		Réaction rang 2 : **FAME 3**	SULTs
sulfation_(aniline)	SULTs	Réaction rang 1 : **No Tool**	Pas d'annotations
		Réaction rang 2 : **FAME 3**	SULTs
sulfation_(aromatic-_hydroxyl)	SULTs	Réaction rang 1 : **No Tool**	Pas d'annotations
		Réaction rang 2 : **FAME 3**	SULTs
Glutathionation(+SX)	GSTs	Réaction rang 1 : **No Tool**	Pas d'annotations
		Réaction rang 2 : **FAME 3**	GSTs

Annexe 3

Cette annexe détail les métabolites que l'on retrouve dans les cartes filtrées des 6 AHAs. Il s'agit du fichier supplémentaire numéro 3 déposé avec l'article soumis à *BMC Bioinformatics*.

Constructing xenobiotic maps of metabolism to predict the role of enzymes in DNA adduct formation
Conan M. , Théret N., Langouet S. and Siegel, A

Additional file3. **Description of six HAA maps of metabolism**
The file provides a detailed description of metabolites for the six HAAs filtered maps of metabolism built in the paper. For each metabolite of each map, the file provides the identifier of the metabolite, its SMILES formula, its production probability score, its reactivity to DNA and the score of XenoSite Reactivity.

4,7,8-TriMeIQx

Metabolites_ID	SMILES_Formula	Production probability score	Reactive_to_DNA (>=0.85)	XenoSite Reactivity score
0	Cc1nc2cc(C)c3c(nc(N)n3C)c2nc1C	1.0000	True	0.9519193669740386
1	Cc1nc2cc(C)c3[nH]c(N)nc3c2nc1C	0.8170	True	0.952519771988244
3	Cc1nc2cc(C(=O)O)c3c(nc(N)n3C)c2nc1C	0.3929	True	0.9364105737792748
4	Cc1nc2cc(C)c3c(nc(N)n3C)c2nc1C(=O)O	0.3578	True	0.936508784540186
5	Cc1nc2c(cc(C)c3c2nc(N)n3C)nc1C(=O)O	0.3727	True	0.936508784540186
6	Cc1nc2cc(CO)c3c(nc(N)n3C)c2nc1C	0.6010	True	0.9449386351114564
7	Cc1nc2cc(C)c3c(nc(N)n3C)c2nc1CO	0.5520	True	0.9450602682223948
8	Cc1nc2c(cc(C)c3c2nc(N)n3C)nc1CO	0.5520	True	0.9450602682223948
9	Cc1nc2cc(C)c3c([nH]c(=O)n3C)c2nc1C	0.5771	False	0.3856827655849179
13	Cc1nc2cc(C)c3c(nc(NO)n3C)c2nc1C	0.6900	True	0.9688695242313216
15	Cc1nc2cc(C(=O)O)c3[nH]c(N)nc3c2nc1C	0.1255	True	0.9370681306254384
17	Cc1nc2c(cc(C)c3[nH]c(N)nc32)nc1C(=O)O	0.3557	True	0.9371990578502958
18	Cc1nc2cc(CO)c3[nH]c(N)nc3c2nc1C	0.1425	True	0.945475586892378
20	Cc1nc2c(cc(C)c3[nH]c(N)nc32)nc1CO	0.3500	True	0.9456217987811018
21	Cc1nc2cc(C)c3[nH]c(=O)[nH]c3c2nc1C	0.2919	False	0.4254405686966013
25	Cc1nc2cc(C)c3[nH]c(NO)nc3c2nc1C	0.2148	True	0.966448534166162
37	Cc1nc2c(cc(C(=O)O)c3c2nc(N)n3C)nc1C(=O)O	0.2944	True	0.9138368184981132
38	Cc1nc2cc(C(=O)O)c3c(nc(N)n3C)c2nc1C(=O)O	0.2987	True	0.9022678720097804
39	Cc1nc2c(cc(C(=O)O)c3c2nc(N)n3C)nc1CO	0.2868	True	0.926430934305395
40	Cc1nc2cc(C(=O)O)c3c(nc(N)n3C)c2nc1CO	0.2900	True	0.9166227608850012
42	Cc1nc2cc(C(=O)O)c3c([nH]c(=O)n3C)c2nc1C	0.5734	False	0.2955217825470901
46	Cc1nc2cc(C(=O)O)c3c(nc(NO)n3C)c2nc1C	0.5589	True	0.9622959616671906
47	Cc1cc2nc(C(=O)O)c(C(=O)O)nc2c2nc(N)n(C)c12	0.3176	True	0.9138368143328124

48	Cc1cc2nc(CO)c(C(=O)O)nc2c2nc(N)n(C)c12	0.2942	True	0.9264309308723871
49	Cc1nc2cc(CO)c3c(nc(N)n3C)c2nc1C(=O)O	0.2920	True	0.916622757113866
51	Cc1nc2cc(C)c3c([nH]c(=O)n3C)c2nc1C(=O)O	0.5059	False	0.3472195794616843
55	Cc1nc2cc(C)c3c(nc(NO)n3C)c2nc1C(=O)O	0.5997	True	0.9622959606347984
56	Cc1cc2nc(C(=O)O)c(CO)nc2c2nc(N)n(C)c12	0.3050	True	0.9264309308723871
57	Cc1nc2c(cc(CO)c3c2nc(N)n3C)nc1C(=O)O	0.2819	True	0.9264309309932304
58	Cc1cnc2cc(C)c3c(nc(N)n3C)c2n1	0.2136	True	0.9527193370927972
59	Cc1nc2c(cc(C)c3c2[nH]c(=O)n3C)nc1C(=O)O	0.6618	False	0.2793007631212143
62	Cc1nc2c(cc(C)c3c2[n+]([O-])c(N)n3C)nc1C(=O)O	0.1133	True	0.9332082812204421
63	Cc1nc2c(cc(C)c3c2nc(NO)n3C)nc1C(=O)O	0.6399	True	0.962971720228064
64	Cc1nc2c(cc(CO)c3c2nc(N)n3C)nc1CO	0.2742	True	0.9366974574009704
65	Cc1nc2cc(CO)c3c(nc(N)n3C)c2nc1CO	0.2833	True	0.9283426870883372
66	Cc1nc2cc(CO)c3c([nH]c(=O)n3C)c2nc1C	0.5930	False	0.7455365925176584
70	Cc1nc2cc(CO)c3c(nc(NO)n3C)c2nc1C	0.5792	True	0.9657085043560121
71	Cc1cc2nc(CO)c(CO)nc2c2nc(N)n(C)c12	0.2812	True	0.936697457301072
72	Cc1nc2cc(C)c3c([nH]c(=O)n3C)c2nc1CO	0.4615	False	0.8193547082416229
76	Cc1nc2cc(C)c3c(nc(NO)n3C)c2nc1CO	0.5638	True	0.9657085043252516
77	Cc1nc2c(cc(C)c3c2[nH]c(=O)n3C)nc1CO	0.6215	False	0.8185217294546052
80	Cc1nc2c(cc(C)c3c2[n+]([O-])c(N)n3C)nc1CO	0.1009	True	0.941109202007716
81	Cc1nc2c(cc(C)c3c2nc(NO)n3C)nc1CO	0.5969	True	0.9663466347214794
82	Cc1nc2cc(C)c3c([nH]c(=O)n3C)c2[n+]([O-])c1C	0.1329	False	0.5036994480756635
83	Cc1nc2c3[nH]c(=O)n(C)c3c(C)cc2[n+]([O-])c1C	0.1200	False	0.538418109175139
90	Cc1nc2cc(C)c3c(nc(NC4OC(C(=O)O)C(O)C(O)C4O)n3C)c2nc1C	0.7340	False	0.13837005958543425
96	Cc1nc2cc(C)c3[nH]c(NC4OC(C(=O)O)C(O)C(O)C4O)nc3c2nc1C	0.1732	False	0.1387371709508488
99	Cc1nc2c3nc(N)[nH]c3c(C)cc2[n+](C2OC(C(=O)O)C(O)C(O)C2O)c1C	0.1961	True	0.8510328768395194
111	Cc1nc2cc(C(=O)OC3OC(C(=O)O)C(O)C(O)C3O)c3c(nc(N)n3C)c2nc1C	0.6175	False	0.7457106445531515
112	Cc1nc2cc(C(=O)O)c3c(nc(NC4OC(C(=O)O)C(O)C(O)C4O)n3C)c2nc1C	0.5593	False	0.09681831803413032
114	Cc1nc2cc(C(=O)O)c3c(c2nc1C)[n+](C1OC(C(=O)O)C(O)C(O)C1O)c(N)n3C	0.1139	False	0.7888191176469449
116	Cc1nc2cc(C(=O)O)c3c(nc(NS(=O)(=O)O)n3C)c2nc1C	0.3593	False	0.17689654946988007
117	CC(=O)Nc1nc2c3nc(C)c(C)nc3cc(C(=O)O)c2n1C	0.4884	False	0.28902633542034223
118	Cc1nc2cc(C)c3c(nc(N)n3C)c2nc1C(=O)OC1OC(C(=O)O)C(O)C(O)C1O	0.4647	False	0.7851814314058655
119	Cc1nc2cc(C)c3c(nc(NC4OC(C(=O)O)C(O)C(O)C4O)n3C)c2nc1C(=O)O	0.5010	False	0.0968183140947095

122	Cc1nc2cc(C)c3c(c2nc1C(=O)O)[n+](C1OC(C(=O)O)C(O)C(O)C1O)c(N)n3C	0.1452	False	0.7888191098064813
123	Cc1nc2cc(C)c3c(nc(NS(=O)(=O)O)n3C)c2nc1C(=O)O	0.3461	False	0.2249407154685828
124	CC(=O)Nc1nc2c3nc(C(=O)O)c(C)nc3cc(C)c2n1C	0.4671	False	0.28976940727666495
125	Cc1nc2c(cc(C)c3c2nc(N)n3C)nc1C(=O)OC1OC(C(=O)O)C(O)C(O)C1O	0.4861	False	0.7852704398378394
126	Cc1nc2c(cc(C)c3c2nc(NC2OC(C(=O)O)C(O)C(O)C2O)n3C)nc1C(=O)O	0.5208	False	0.10589644229212053
129	Cc1nc2c(cc(C)c3c2[n+](C2OC(C(=O)O)C(O)C(O)C2O)c(N)n3C)nc1C(=O)O	0.1562	False	0.8017170182277169
130	Cc1nc2c(cc(C)c3c2nc(NS(=O)(=O)O)n3C)nc1C(=O)O	0.3646	False	0.18364682721014144
131	CC(=O)Nc1nc2c3nc(C)c(C(=O)O)nc3cc(C)c2n1C	0.4787	False	0.2890263250501654
132	Cc1nc2cc(COC3OC(C(=O)O)C(O)C(O)C3O)c3c(nc(N)n3C)c2nc1C	0.5313	False	0.7909614324400446
133	Cc1nc2cc(CO)c3c(nc(NC4OC(C(=O)O)C(O)C(O)C4O)n3C)c2nc1C	0.5193	False	0.4532113173412775
134	Cc1nc2c3nc(N)n(C)c3c(CO)cc2[n+](C2OC(C(=O)O)C(O)C(O)C2O)c1C	0.1010	False	0.822721280524058
135	Cc1nc2cc(CO)c3c(c2nc1C)[n+](C1OC(C(=O)O)C(O)C(O)C1O)c(N)n3C	0.1058	False	0.8185963707547755
137	Cc1nc2cc(CO)c3c(nc(NS(=O)(=O)O)n3C)c2nc1C	0.3462	False	0.6353760737667271
138	Cc1nc2cc(COS(=O)(=O)O)c3c(nc(N)n3C)c2nc1C	0.5577	True	0.9362816567048096
139	CC(=O)Nc1nc2c3nc(C)c(C)nc3cc(CO)c2n1C	0.4640	False	0.6882510334487474
140	Cc1nc2cc(C)c3c(nc(N)n3C)c2nc1COC1OC(C(=O)O)C(O)C(O)C1O	0.4107	False	0.8242326161539762
141	Cc1nc2cc(C)c3c(nc(NC4OC(C(=O)O)C(O)C(O)C4O)n3C)c2nc1CO	0.4571	False	0.6104779437776809
143	Cc1cc2c(nc(CO)c(C)[n+]2C2OC(C(=O)O)C(O)C(O)C2O)c2nc(N)n(C)c12	0.1148	False	0.8227212802089878
144	Cc1nc2cc(C)c3c(c2nc1CO)[n+](C1OC(C(=O)O)C(O)C(O)C1O)c(N)n3C	0.1281	False	0.8185963705156074
145	Cc1nc2cc(C)c3c(nc(NS(=O)(=O)O)n3C)c2nc1CO	0.3157	False	0.7492928735688927
146	Cc1nc2cc(C)c3c(nc(N)n3C)c2nc1COS(=O)(=O)O	0.4681	True	0.9579709478002996
147	CC(=O)Nc1nc2c3nc(CO)c(C)nc3cc(C)c2n1C	0.4261	False	0.7879146520038296
148	Cc1nc2c(cc(C)c3c2nc(N)n3C)nc1COC1OC(C(=O)O)C(O)C(O)C1O	0.4217	False	0.8243413148736591
149	Cc1nc2c(cc(C)c3c2nc(NC2OC(C(=O)O)C(O)C(O)C2O)n3C)nc1CO	0.4637	False	0.6089069790986317
152	Cc1nc2c(cc(C)c3c2[n+](C2OC(C(=O)O)C(O)C(O)C2O)c(N)n3C)nc1CO	0.1413	False	0.8299753219232127
153	Cc1nc2c(cc(C)c3c2nc(NS(=O)(=O)O)n3C)nc1CO	0.3246	False	0.7482482571695571
154	Cc1nc2c(cc(C)c3c2nc(N)n3C)nc1COS(=O)(=O)O	0.4725	True	0.957809790047806
155	CC(=O)Nc1nc2c3nc(C)c(CO)nc3cc(C)c2n1C	0.4261	False	0.7869881850369813
157	Cc1nc2c3[nH]c(=O)n(C)c3c(C)cc2[n+](C2OC(C(=O)O)C(O)C(O)C2O)c1C	0.1677	False	0.12733594423837818
174	Cc1nc2cc(C)c3c(nc(NOC4OC(C(=O)O)C(O)C(O)C4O)n3C)c2nc1C	0.2981	True	0.9394393402141152
175	Cc1nc2cc(C)c3c(nc(N(O)C4OC(C(=O)O)C(O)C(O)C4O)n3C)c2nc1C	0.2042	False	0.4351992592542831
178	Cc1nc2c3nc(NO)n(C)c3c(C)cc2[n+](C2OC(C(=O)O)C(O)C(O)C2O)c1C	0.1380	True	0.9298838804031396

4-CH2OH-8-MeIQx

Metabolites_ID	SMILES_Formula	Production probability score	Reactive_to_DNA (>=0.85)	XenoSite Reactivity score
0	Cc1cnc2cc(CO)c3c(nc(N)n3C)c2n1	1.0000	True	0.939246545782168
1	Cc1cnc2cc(CO)c3[nH]c(N)nc3c2n1	0.6940	True	0.9399486976355284
4	Cn1c(N)nc2c3nc(C(=O)O)cnc3cc(CO)c21	0.7416	True	0.9183886343549283
5	Cn1c(N)nc2c3nc(CO)cnc3cc(CO)c21	0.8330	True	0.9297950666666231
6	Cc1cnc2cc(C(=O)O)c3c(nc(N)n3C)c2n1	0.7970	True	0.9297634019142198
7	Cc1cnc2cc(CO)c3c([nH]c(=O)n3C)c2n1	0.2681	False	0.7525994388021336
11	Cc1cnc2cc(CO)c3c(nc(NO)n3C)c2n1	0.4870	True	0.9661762365971523
12	Cc1nc2c(cc(CO)c3[nH]c(N)nc32)nc1O	0.1915	True	0.9396626815064018
14	Nc1nc2c([nH]1)c(CO)cc1ncc(C(=O)O)nc12	0.3061	True	0.9192069137181936
15	Nc1nc2c([nH]1)c(CO)cc1ncc(CO)nc12	0.3022	True	0.9304529094585507
16	Cc1cnc2cc(C(=O)O)c3[nH]c(N)nc3c2n1	0.1206	True	0.9306062036714235
17	Cc1cnc2cc(CO)c3[nH]c(=O)[nH]c3c2n1	0.1314	False	0.7344730497861981
21	Cc1cnc2cc(CO)c3[nH]c(NO)nc3c2n1	0.1632	True	0.9633861582056056
24	Cn1c(N)nc2c3nc(CO)c(O)nc3cc(CO)c21	0.1466	True	0.9329028728863292
25	Cc1nc2c(cc(C(=O)O)c3c2nc(N)n3C)nc1O	0.1782	True	0.9295012404074886
26	Cc1nc2c(cc(CO)c3c2[nH]c(=O)n3C)nc1O	0.1388	False	0.7367474363970398
30	Cc1nc2c(cc(CO)c3c2nc(NO)n3C)nc1O	0.1500	True	0.9631115497659792
39	Cn1c(N)nc2c3nccnc3cc(CO)c21	0.3137	True	0.9470328096060072
40	Cn1c(N)nc2c3nc(C(=O)O)cnc3cc(C(=O)O)c21	0.3607	True	0.9044177826890109
41	Cn1c(=O)[nH]c2c3nc(C(=O)O)cnc3cc(CO)c21	0.5381	False	0.6968627824778062
44	Cn1c(N)[n+]([O-])c2c3nc(C(=O)O)cnc3cc(CO)c21	0.1470	True	0.9150345060270454
45	Cn1c(NO)nc2c3nc(C(=O)O)cnc3cc(CO)c21	0.5346	True	0.9585368267450812
46	Cn1c(N)nc2c3nc(CO)cnc3cc(C(=O)O)c21	0.3556	True	0.9183886380336036
47	Cn1c(=O)[nH]c2c3nc(CO)cnc3cc(CO)c21	0.5181	False	0.813710764792595
50	Cn1c(N)[n+]([O-])c2c3nc(CO)cnc3cc(CO)c21	0.1224	True	0.9252476842839548
51	Cn1c(NO)nc2c3nc(CO)cnc3cc(CO)c21	0.5145	True	0.9624737595749632
53	Cc1cnc2cc(C(=O)O)c3c([nH]c(=O)n3C)c2n1	0.4601	False	0.3076820818757737
54	Cc1c[n+]([O-])c2cc(C(=O)O)c3c(nc(N)n3C)c2n1	0.1782	True	0.9214313394054104
57	Cc1cnc2cc(C(=O)O)c3c(nc(NO)n3C)c2n1	0.1829	True	0.9628554874324928

66	Cc1cnc2cc(COC3OC(C(=O)O)C(O)C(O)C3O)c3c(nc(N)n3C)c2n1	0.1430	False	0.7959041892790614
67	Cc1cnc2cc(CO)c3c(nc(NC4OC(C(=O)O)C(O)C(O)C4O)n3C)c2n1	0.4470	False	0.4781837291519712
74	Cc1cnc2cc(COC3OC(C(=O)O)C(O)C(O)C3O)c3[nH]c(N)nc3c2n1	0.6190	False	0.7974615695700105
78	Cc1c[n+](C2OC(C(=O)O)C(O)C(O)C2O)c2cc(CO)c3[nH]c(N)nc3c2n1	0.1277	False	0.8281445139036444
81	Cc1cnc2cc(COS(=O)(=O)O)c3[nH]c(N)nc3c2n1	0.5996	True	0.9281737677695644
103	Cn1c(N)nc2c3nc(C(=O)O)cnc3cc(COC3OC(C(=O)O)C(O)C(O)C3O)c21	0.7856	False	0.7034278531573016
104	Cn1c(N)nc2c3nc(C(=O)OC4OC(C(=O)O)C(O)C(O)C4O)cnc3cc(CO)c21	0.6940	False	0.7361105200113731
105	Cn1c(NC2OC(C(=O)O)C(O)C(O)C2O)nc2c3nc(C(=O)O)cnc3cc(CO)c21	0.6482	False	0.3648961922925914
108	Cn1c(N)[n+](C2OC(C(=O)O)C(O)C(O)C2O)c2c3nc(C(=O)O)cnc3cc(CO)c21	0.1409	False	0.7562250686735003
109	Cn1c(NS(=O)(=O)O)nc2c3nc(C(=O)O)cnc3cc(CO)c21	0.4509	False	0.5801416341311411
110	Cn1c(N)nc2c3nc(C(=O)O)cnc3cc(COS(=O)(=O)O)c21	0.7187	True	0.9200551274244104
111	CC(=O)Nc1nc2c3nc(C(=O)O)cnc3cc(CO)c2n1C	0.6341	False	0.6402471927586828
112	Cn1c(N)nc2c3nc(COC4OC(C(=O)O)C(O)C(O)C4O)cnc3cc(CO)c21	0.6497	False	0.7826073222968977
113	Cn1c(N)nc2c3nc(CO)cnc3cc(COC3OC(C(=O)O)C(O)C(O)C3O)c21	0.7430	False	0.7474417380900512
114	Cn1c(NC2OC(C(=O)O)C(O)C(O)C2O)nc2c3nc(CO)cnc3cc(CO)c21	0.6131	False	0.5840922307976594
116	Cn1c(N)nc2c3nc(CO)c[n+](C4OC(C(=O)O)C(O)C(O)C4O)c3cc(CO)c21	0.1433	False	0.7957213472607796
117	Cn1c(N)[n+](C2OC(C(=O)O)C(O)C(O)C2O)c2c3nc(CO)cnc3cc(CO)c21	0.1266	False	0.7899565344739715
118	Cn1c(NS(=O)(=O)O)nc2c3nc(CO)cnc3cc(CO)c21	0.4265	False	0.7389240175157654
119	Cn1c(N)nc2c3nc(COS(=O)(=O)O)cnc3cc(CO)c21	0.7031	True	0.9513218613965329
120	Cn1c(N)nc2c3nc(CO)cnc3cc(COS(=O)(=O)O)c21	0.6797	True	0.9301629974208848
121	CC(=O)Nc1nc2c3nc(CO)cnc3cc(CO)c2n1C	0.5998	False	0.780440363453988
122	Cc1cnc2cc(C(=O)OC3OC(C(=O)O)C(O)C(O)C3O)c3c(nc(N)n3C)c2n1	0.7715	False	0.7516321470738067
123	Cc1cnc2cc(C(=O)O)c3c(nc(NC4OC(C(=O)O)C(O)C(O)C4O)n3C)c2n1	0.6025	False	0.09934650446541436
125	Cc1cnc2cc(C(=O)O)c3c(c2n1)[n+](C1OC(C(=O)O)C(O)C(O)C1O)c(N)n3C	0.1307	False	0.7929786868238702
127	Cc1cnc2cc(C(=O)O)c3c(nc(NS(=O)(=O)O)n3C)c2n1	0.3762	False	0.187152862044287
128	CC(=O)Nc1nc2c3nc(C)cnc3cc(C(=O)O)c2n1C	0.5675	False	0.2787233326697437
129	Cc1cnc2cc(COC3OC(C(=O)O)C(O)C(O)C3O)c3c([nH]c(=O)n3C)c2n1	0.4165	False	0.6179684916244721
133	Cc1cnc2cc(COS(=O)(=O)O)c3c([nH]c(=O)n3C)c2n1	0.4683	True	0.9331380220338998
155	Cc1cnc2cc(COC3OC(C(=O)O)C(O)C(O)C3O)c3c(nc(NO)n3C)c2n1	0.4325	True	0.9186227958903936
156	Cc1cnc2cc(CO)c3c(nc(NOC4OC(C(=O)O)C(O)C(O)C4O)n3C)c2n1	0.2104	True	0.9187100465128704
157	Cc1cnc2cc(CO)c3c(nc(N(O)C4OC(C(=O)O)C(O)C(O)C4O)n3C)c2n1	0.1442	False	0.4397655038488933
161	Cc1cnc2cc(COS(=O)(=O)O)c3c(nc(NO)n3C)c2n1	0.4539	True	0.9528983217648508

162	Cc1cnc2cc(COC3OC(C(=O)O)C(O)C(O)C3O)c3c(nc(NC4OC(C(=O)O)C(O)C(O)C4O)n3C)c2n1	0.4305	False	0.2674484837322748
172	Cc1cnc2cc(COS(=O)(=O)O)c3c(nc(NC4OC(C(=O)O)C(O)C(O)C4O)n3C)c2n1	0.4327	False	0.8236623799216444

7,8-DiMeIQx

Metabolites_ID	SMILES_Formula	Production probability score	Reactive_to_DNA (>=0.85)	XenoSite Reactivity score
0	Cc1nc2ccc3c(nc(N)n3C)c2nc1C	1.0000	True	0.9526303225571204
1	Cc1nc2ccc3[nH]c(N)nc3c2nc1C	0.9130	True	0.9532412905725546
3	Cc1nc2cc(O)c3c(nc(N)n3C)c2nc1C	0.3300	True	0.947257793228356
4	Cc1nc2ccc3c(nc(N)n3C)c2nc1C(=O)O	0.3974	True	0.937552266604346
5	Cc1nc2c(ccc3c2nc(N)n3C)nc1C(=O)O	0.4111	True	0.937552266604346
6	Cc1nc2ccc3c(nc(N)n3C)c2nc1CO	0.5860	True	0.945917356926128
7	Cc1nc2c(ccc3c2nc(N)n3C)nc1CO	0.5860	True	0.945917356926128
8	Cc1nc2ccc3c([nH]c(=O)n3C)c2nc1C	0.6472	False	0.3931118524235226
12	Cc1nc2ccc3c(nc(NO)n3C)c2nc1C	0.7390	True	0.9692538972925264
13	Cc1nc2c(O)cc3[nH]c(N)nc3c2nc1C	0.1132	True	0.947622997351468
14	Cc1nc2cc(O)c3[nH]c(N)nc3c2nc1C	0.1413	True	0.947813590358718
15	Cc1nc2ccc3[nH]c(N)nc3c2nc1C(=O)O	0.3195	True	0.9382566407991466
16	Cc1nc2c(ccc3[nH]c(N)nc32)nc1C(=O)O	0.4326	True	0.9382566407991466
18	Cc1nc2c(ccc3[nH]c(N)nc32)nc1CO	0.4233	True	0.946491592095027
19	Cc1nc2ccc3[nH]c(=O)[nH]c3c2nc1C	0.7965	False	0.433071014730353
23	Cc1nc2ccc3[nH]c(NO)nc3c2nc1C	0.2631	True	0.9668948123415207
26	Cc1nc2c(O)cc3c(nc(N)n3C)c2nc1C(=O)O	0.0922	True	0.9293640973340352
29	Cc1nc2c(O)cc3c([nH]c(=O)n3C)c2nc1C	0.0370	False	0.3642730234787643
33	Cc1nc2c(O)cc3c(nc(NO)n3C)c2nc1C	0.0562	True	0.966706651341194
34	Cc1nc2cc(O)c3c(nc(N)n3C)c2nc1C(=O)O	0.1149	True	0.9296412848269409
35	Cc1nc2c(cc(O)c3c2nc(N)n3C)nc1C(=O)O	0.1676	True	0.9296412848269409
36	Cc1nc2cc(O)c3c(nc(N)n3C)c2nc1CO	0.1147	True	0.9393408281727236
37	Cc1nc2c(cc(O)c3c2nc(N)n3C)nc1CO	0.1549	True	0.9393408281727236
38	Cc1nc2cc(O)c3c([nH]c(=O)n3C)c2nc1C	0.1882	False	0.36207412502176894
42	Cc1nc2cc(O)c3c(nc(NO)n3C)c2nc1C	0.1891	True	0.9667196195161194
43	Cn1c(N)nc2c3nc(C(=O)O)c(C(=O)O)nc3ccc21	0.3400	True	0.9153814781071824
44	Cn1c(N)nc2c3nc(C(=O)O)c(CO)nc3ccc21	0.3142	True	0.9276954450903158

45	Cc1cnc2c(ccc3c2nc(N)n3C)n1	0.2077	True	0.9534974792514564
46	Cc1nc2ccc3c([nH]c(=O)n3C)c2nc1C(=O)O	0.6144	False	0.3537918781352531
49	Cc1nc2ccc3c(c2nc1C(=O)O)[n+]([O-])c(N)n3C	0.1818	True	0.9340630817091758
50	Cc1nc2ccc3c(nc(NO)n3C)c2nc1C(=O)O	0.6963	True	0.9628312566982632
51	Cn1c(N)nc2c3nc(CO)c(C(=O)O)nc3ccc21	0.3291	True	0.927695445090316
52	Cc1cnc2ccc3c(nc(N)n3C)c2n1	0.2147	True	0.9020011272283596
53	Cc1nc2c(ccc3c2[nH]c(=O)n3C)nc1C(=O)O	0.3320	False	0.28502616028699546
56	Cc1nc2c(ccc3c2[n+]([O-])c(N)n3C)nc1C(=O)O	0.1504	True	0.9340630817091758
57	Cc1nc2c(ccc3c2nc(NO)n3C)nc1C(=O)O	0.7151	True	0.9635006209839014
58	Cn1c(N)nc2c3nc(CO)c(CO)nc3ccc21	0.3029	True	0.9377346761986114
59	Cc1nc2ccc3c([nH]c(=O)n3C)c2nc1CO	0.5626	False	0.8229736559749736
62	Cc1nc2ccc3c(c2nc1CO)[n+]([O-])c(N)n3C	0.1231	True	0.9418315309031982
63	Cc1nc2ccc3c(nc(NO)n3C)c2nc1CO	0.6554	True	0.9661630273872972
64	Cc1nc2c(ccc3c2[nH]c(=O)n3C)nc1CO	0.6911	False	0.8221506428125687
67	Cc1nc2c(ccc3c2[n+]([O-])c(N)n3C)nc1CO	0.1349	True	0.9418315309031982
68	Cc1nc2c(ccc3c2nc(NO)n3C)nc1CO	0.6686	True	0.9667957493222371
69	Cc1nc2ccc3c([nH]c(=O)n3C)c2[n+]([O-])c1C	0.1411	False	0.5106579759858405
70	Cc1nc2c3[nH]c(=O)n(C)c3ccc2[n+]([O-])c1C	0.1579	False	0.545304641006959
77	Cc1nc2ccc3c(nc(NC4OC(C(=O)O)C(O)C(O)C4O)n3C)c2nc1C	0.7900	False	0.14128924533621762
83	Cc1nc2ccc3[nH]c(NC4OC(C(=O)O)C(O)C(O)C4O)nc3c2nc1C	0.1972	False	0.14163223897622915
84	Cc1nc2ccc3[nH]c(N)[n+](C4OC(C(=O)O)C(O)C(O)C4O)c3c2nc1C	0.1972	False	0.8496759833252281
85	Cc1nc2ccc3[nH]c(N)nc3c2[n+](C2OC(C(=O)O)C(O)C(O)C2O)c1C	0.1570	True	0.8541000543492518
86	Cc1nc2c3nc(N)[nH]c3ccc2[n+](C2OC(C(=O)O)C(O)C(O)C2O)c1C	0.2337	True	0.8540999219182439
87	Cc1nc2ccc3c(nc(N)n3C3OC(C(=O)O)C(O)C(O)C3O)c2nc1C	0.3177	False	0.8335984380668678
88	Cc1nc2ccc3[nH]c(NS(=O)(=O)O)nc3c2nc1C	0.5405	False	0.29014961211242624
89	CC(=O)Nc1nc2c(ccc3nc(C)c(C)nc32)[nH]1	0.7231	False	0.39624710208252295
98	Cc1nc2cc(OC3OC(C(=O)O)C(O)C(O)C3O)c3c(nc(N)n3C)c2nc1C	0.2086	False	0.7975230911352285
99	Cc1nc2cc(O)c3c(nc(NC4OC(C(=O)O)C(O)C(O)C4O)n3C)c2nc1C	0.1964	False	0.11980117236897055
103	Cc1nc2cc(OS(=O)(=O)O)c3c(nc(N)n3C)c2nc1C	0.2732	True	0.906323873862639
104	Cc1nc2cc(O)c3c(nc(NS(=O)(=O)O)n3C)c2nc1C	0.1214	False	0.23067120891923004
105	CC(=O)Nc1nc2c3nc(C)c(C)nc3cc(O)c2n1C	0.2442	False	0.3488387270156466
106	Cc1nc2ccc3c(nc(N)n3C)c2nc1C(=O)OC1OC(C(=O)O)C(O)C(O)C1O	0.4915	False	0.7895415567149885

107	Cc1nc2ccc3c(nc(NC4OC(C(=O)O)C(O)C(O)C4O)n3C)c2nc1C(=O)O	0.5735	False	0.09902060770467656
110	Cc1nc2ccc3c(c2nc1C(=O)O)[n+](C1OC(C(=O)O)C(O)C(O)C1O)c(N)n3C	0.2278	False	0.7922689419572471
111	Cc1nc2ccc3c(nc(NS(=O)(=O)O)n3C)c2nc1C(=O)O	0.3866	False	0.2303419735567399
112	CC(=O)Nc1nc2c3nc(C(=O)O)c(C)nc3ccc2n1C	0.4967	False	0.2959348465199415
113	Cc1nc2c(ccc3c2nc(N)n3C)nc1C(=O)OC1OC(C(=O)O)C(O)C(O)C1O	0.5125	False	0.7896259199915192
114	Cc1nc2c(ccc3c2nc(NC2OC(C(=O)O)C(O)C(O)C2O)n3C)nc1C(=O)O	0.5857	False	0.10831323730539713
117	Cc1nc2c(ccc3c2[n+](C2OC(C(=O)O)C(O)C(O)C2O)c(N)n3C)nc1C(=O)O	0.2013	False	0.804986177950973
118	Cc1nc2c(ccc3c2nc(NS(=O)(=O)O)n3C)nc1C(=O)O	0.4027	False	0.18819526740987594
119	CC(=O)Nc1nc2c3nc(C)c(C(=O)O)nc3ccc2n1C	0.5073	False	0.2951971305132569
120	Cc1nc2ccc3c(nc(N)n3C)c2nc1COC1OC(C(=O)O)C(O)C(O)C1O	0.4360	False	0.8278196414052801
121	Cc1nc2ccc3c(nc(NC4OC(C(=O)O)C(O)C(O)C4O)n3C)c2nc1CO	0.5251	False	0.6177843899021765
123	Cc1c(CO)nc2c3nc(N)n(C)c3ccc2[n+]1C1OC(C(=O)O)C(O)C(O)C1O	0.1406	False	0.8263333282328998
124	Cc1nc2ccc3c(c2nc1CO)[n+](C1OC(C(=O)O)C(O)C(O)C1O)c(N)n3C	0.1641	False	0.8215309603633622
125	Cc1nc2ccc3c(nc(NS(=O)(=O)O)n3C)c2nc1CO	0.3539	False	0.7544096199052549
126	Cc1nc2ccc3c(nc(N)n3C)c2nc1COS(=O)(=O)O	0.4969	True	0.958565947543656
127	CC(=O)Nc1nc2c3nc(CO)c(C)nc3ccc2n1C	0.4547	False	0.7922359747966528
128	Cc1nc2c(ccc3c2nc(N)n3C)nc1COC1OC(C(=O)O)C(O)C(O)C1O	0.4477	False	0.8279253001213381
129	Cc1nc2c(ccc3c2nc(NC2OC(C(=O)O)C(O)C(O)C2O)n3C)nc1CO	0.5251	False	0.6162198140862697
130	Cc1nc2c3nc(N)n(C)c3ccc2[n+](C2OC(C(=O)O)C(O)C(O)C2O)c1CO	0.1078	False	0.8263333282328998
132	Cc1nc2c(ccc3c2[n+](C2OC(C(=O)O)C(O)C(O)C2O)c(N)n3C)nc1CO	0.1805	False	0.8327448843846529
133	Cc1nc2c(ccc3c2nc(NS(=O)(=O)O)n3C)nc1CO	0.3586	False	0.7533741708342333
134	Cc1nc2c(ccc3c2nc(N)n3C)nc1COS(=O)(=O)O	0.5016	True	0.9584058263601922
135	CC(=O)Nc1nc2c3nc(C)c(CO)nc3ccc2n1C	0.4547	False	0.7913194225854521
136	Cc1nc2ccc3c([nH]c(=O)n3C)c2[n+](C2OC(C(=O)O)C(O)C(O)C2O)c1C	0.1378	False	0.13163256054954975
137	Cc1nc2c3[nH]c(=O)n(C)c3ccc2[n+](C2OC(C(=O)O)C(O)C(O)C2O)c1C	0.2016	False	0.1306499178923529
138	Cc1nc2ccc3c(c2nc1C)n(C1OC(C(=O)O)C(O)C(O)C1O)c(=O)n3C	0.1546	False	0.09426602764544464
154	Cc1nc2ccc3c(nc(NOC4OC(C(=O)O)C(O)C(O)C4O)n3C)c2nc1C	0.3163	True	0.940142849022794
155	Cc1nc2ccc3c(nc(N(O)C4OC(C(=O)O)C(O)C(O)C4O)n3C)c2nc1C	0.2187	False	0.4418016730104229
156	Cc1nc2ccc3c(c2nc1C)[n+](C1OC(C(=O)O)C(O)C(O)C1O)c(NO)n3C	0.1301	True	0.9479245326875604
158	Cc1nc2c3nc(NO)n(C)c3ccc2[n+](C2OC(C(=O)O)C(O)C(O)C2O)c1C	0.1803	True	0.9312877406288887
162	Cc1nc2c3nc(NC4OC(C(=O)O)C(O)C(O)C4O)n(C)c3ccc2[n+](C2OC(C(=O)O)C(O)C(O)C2O)c1C	0.1264	False	0.04152612811541051

AalphaC

Metabolites_ID	SMILES_Formula	Production probability score	Reactive_to_DNA (>=0.85)	XenoSite Reactivity score
0	Nc1ccc2c(n1)[nH]c1ccccc12	1.0000	True	0.8594642870288121
1	Nc1ccc2c(n1)[nH]c1cc(O)ccc12	0.3620	True	0.9178122922308036
2	Nc1nc2[nH]c3ccccc3c2cc1O	0.3850	True	0.9104374669419212
3	Nc1ccc2c(n1)[nH]c1ccc(O)cc12	0.4980	True	0.9178122922308036
4	Nc1ccc2c(n1)[nH]c1c(O)cccc12	0.2050	True	0.9173389807704077
5	O=c1ccc2c([nH]1)[nH]c1ccccc12	0.3483	False	0.07521168888614659
7	ONc1ccc2c(n1)[nH]c1ccccc12	0.4700	True	0.9320796713748204
8	Nc1ccc2c(n1)[nH]c1cc(O)c(O)cc12	0.1246	True	0.9054297852222036
9	Nc1nc2[nH]c3cc(O)ccc3c2cc1O	0.3144	True	0.8971425081962023
11	O=c1ccc2c([nH]1)[nH]c1cc(O)ccc12	0.3786	False	0.04921904388986637
12	Nc1ccc2c3ccc(O)cc3[nH]c2n1O	0.1796	False	0.21362896295664915
13	ONc1ccc2c(n1)[nH]c1cc(O)ccc12	0.3743	True	0.9557483537545116
14	Nc1nc2[nH]c3ccc(O)cc3c2cc1O	0.4256	True	0.8977844487769387
15	Nc1nc2[nH]c3c(O)cccc3c2cc1O	0.1480	True	0.8971425081962023
17	O=c1[nH]c2[nH]c3ccccc3c2cc1O	0.3958	False	0.061405818683500576
19	ONc1nc2[nH]c3ccccc3c2cc1O	0.2380	True	0.9657290904460372
22	O=c1ccc2c([nH]1)[nH]c1ccc(O)cc12	0.6273	False	0.1004533015347112
23	Nc1ccc2c3cc(O)ccc3[nH]c2n1O	0.2550	False	0.21362896295664915
24	ONc1ccc2c(n1)[nH]c1ccc(O)cc12	0.5408	True	0.9557483537545116
25	Nc1ccc2c(n1)N=C1C=CC(=O)C=C12	0.2868	False	0.8409736893124711
27	O=c1ccc2c([nH]1)[nH]c1c(O)cccc12	0.2344	False	0.08511848795086216
29	ONc1ccc2c(n1)[nH]c1c(O)cccc12	0.2490	True	0.9557483537545116
32	O=C(O)C1OC(Nc2ccc3c(n2)[nH]c2ccccc23)C(O)C(O)C1O	0.6880	False	0.06958169896151717
34	Nc1ccc2c3ccccc3n(C3OC(C(=O)O)C(O)C(O)C3O)c2n1	0.1280	False	0.7506542647056539
37	Nc1ccc2c(n1)[nH]c1cc(OC3OC(C(=O)O)C(O)C(O)C3O)ccc12	0.3606	False	0.7259178798306849
38	O=C(O)C1OC(Nc2ccc3c(n2)[nH]c2cc(O)ccc23)C(O)C(O)C1O	0.3446	False	0.05942302710234624
39	Nc1ccc2c3ccc(O)cc3[nH]c2n1C1OC(C(=O)O)C(O)C(O)C1O	0.2766	False	0.017567897784690138
40	Nc1ccc2c3ccc(O)cc3n(C3OC(C(=O)O)C(O)C(O)C3O)c2n1	0.1520	False	0.7110168267507706
41	Nc1ccc2c(n1)[nH]c1cc(OS(=O)(=O)O)ccc12	0.3548	True	0.8725659998255538
42	O=S(=O)(O)Nc1ccc2c(n1)[nH]c1cc(O)ccc12	0.2664	False	0.08317776219627293

43	CC(=O)Nc1ccc2c(n1)[nH]c1cc(O)ccc12	0.3446	False	0.1868702590928962
44	Nc1nc2[nH]c3ccccc3c2cc1OC1OC(C(=O)O)C(O)C(O)C1O	0.2988	False	0.6744863352644157
45	O=C(O)C1OC(Nc2nc3[nH]c4ccccc4c3cc2O)C(O)C(O)C1O	0.3203	False	0.06661743405262767
46	Nc1c(O)cc2c3ccccc3[nH]c2n1C1OC(C(=O)O)C(O)C(O)C1O	0.2402	False	0.050903698198718726
47	Nc1nc2c(cc1O)c1ccccc1n2C1OC(C(=O)O)C(O)C(O)C1O	0.1586	False	0.6904836807414154
48	Nc1nc2[nH]c3ccccc3c2cc1OS(=O)(=O)O	0.3696	True	0.8976396777784863
49	O=S(=O)(O)Nc1nc2[nH]c3ccccc3c2cc1O	0.2880	False	0.07463840074684512
50	CC(=O)Nc1nc2[nH]c3ccccc3c2cc1O	0.3634	False	0.1850721970410205
51	Nc1ccc2c(n1)[nH]c1ccc(OC3OC(C(=O)O)C(O)C(O)C3O)cc12	0.4940	False	0.7259178798306849
52	O=C(O)C1OC(Nc2ccc3c(n2)[nH]c2ccc(O)cc23)C(O)C(O)C1O	0.4602	False	0.05942302710234624
53	Nc1ccc2c3cc(O)ccc3[nH]c2n1C1OC(C(=O)O)C(O)C(O)C1O	0.3745	False	0.017567897784690138
54	Nc1ccc2c3cc(O)ccc3n(C3OC(C(=O)O)C(O)C(O)C3O)c2n1	0.2171	False	0.6807681397450976
55	Nc1ccc2c(n1)[nH]c1ccc(OS(=O)(=O)O)cc12	0.4582	True	0.8725659998255538
56	O=S(=O)(O)Nc1ccc2c(n1)[nH]c1ccc(O)cc12	0.3625	False	0.07219347334092874
57	CC(=O)Nc1ccc2c(n1)[nH]c1ccc(O)cc12	0.4741	False	0.1868702590928962
58	Nc1ccc2c(n1)[nH]c1c(OC3OC(C(=O)O)C(O)C(O)C3O)cccc12	0.2009	False	0.7259178798306849
59	O=C(O)C1OC(Nc2ccc3c(n2)[nH]c2c(O)cccc23)C(O)C(O)C1O	0.1960	False	0.05942302710234624
60	Nc1ccc2c3cccc(O)c3[nH]c2n1C1OC(C(=O)O)C(O)C(O)C1O	0.1205	False	0.017567897784690138
62	Nc1ccc2c(n1)[nH]c1c(OS(=O)(=O)O)cccc12	0.1960	True	0.8721892283792341
63	O=S(=O)(O)Nc1ccc2c(n1)[nH]c1c(O)cccc12	0.1501	False	0.07528770637472898
64	CC(=O)Nc1ccc2c(n1)[nH]c1c(O)cccc12	0.1952	False	0.17505455742338466
65	O=C(O)C1OC(n2c(=O)ccc3c4ccccc4[nH]c32)C(O)C(O)C1O	0.1768	False	0.05114923693441379
66	O=C(O)C1OC(n2c3ccccc3c3ccc(=O)[nH]c32)C(O)C(O)C1O	0.1434	False	0.036301666725559366
71	O=C(O)C1OC(ONc2ccc3c(n2)[nH]c2ccccc23)C(O)C(O)C1O	0.1861	True	0.9004352701152516
72	O=C(O)C1OC(N(O)c2ccc3c(n2)[nH]c2ccccc23)C(O)C(O)C1O	0.2275	False	0.27448964713364665
73	O=C(O)C1OC(n2c(NO)ccc3c4ccccc4[nH]c32)C(O)C(O)C1O	0.2106	False	0.34194329742413737
74	O=C(O)C1OC(n2c3ccccc3c3ccc(NO)nc32)C(O)C(O)C1O	0.2425	True	0.9100635799042868
77	O=C(O)C1OC(Nc2ccc3c4ccccc4n(C4OC(C(=O)O)C(O)C(O)C4O)c3n2)C(O)C(O)C1O	0.3735	False	0.02091430158696055
82	CC(=O)Nc1ccc2c3ccccc3n(C3OC(C(=O)O)C(O)C(O)C3O)c2n1	0.1224	False	0.04508733890936721

MeIQx

Metabolites_ID	SMILES_Formula	Production probability score	Reactive_to_DNA (>=0.85)	XenoSite Reactivity score

0	Cc1cnc2ccc3c(nc(N)n3C)c2n1	1.0000	True	0.9020011272283596
1	Cc1cnc2ccc3[nH]c(N)nc3c2n1	0.8660	True	0.9540024480766478
3	Cc1cnc2cc(O)c3c(nc(N)n3C)c2n1	0.3290	True	0.9481434830784878
4	Cc1cnc2c(O)cc3c(nc(N)n3C)c2n1	0.1080	True	0.9479487042622924
5	Cn1c(N)nc2c3nc(C(=O)O)cnc3ccc21	0.8479	True	0.9388094727833032
6	Cn1c(N)nc2c3nc(CO)cnc3ccc21	0.9070	True	0.9469520865325473
7	Cc1cnc2ccc3c([nH]c(=O)n3C)c2n1	0.5417	False	0.418435141090156
11	Cc1cnc2ccc3c(nc(NO)n3C)c2n1	0.6500	True	0.9498784008521852
12	Cc1nc2c(ccc3[nH]c(N)nc32)nc1O	0.2425	True	0.9486334206121112
14	Cc1cnc2cc(O)c3[nH]c(N)nc3c2n1	0.1131	True	0.9486803877383232
15	Nc1nc2c(ccc3ncc(C(=O)O)nc32)[nH]1	0.4157	True	0.9395171969398016
16	Nc1nc2c(ccc3ncc(CO)nc32)[nH]1	0.4176	True	0.9475300341434734
17	Cc1cnc2ccc3[nH]c(=O)[nH]c3c2n1	0.7647	False	0.4592935267484921
18	Cc1cnc2ccc3[nH]c(N)n(O)c3c2n1	0.1760	False	0.5843451083119229
19	Cc1cnc2ccc3[nH]c(N)nc3c2n1O	0.1178	False	0.40200726318413016
21	Cc1cnc2ccc3[nH]c(NO)nc3c2n1	0.7134	True	0.9673234151486991
25	Cn1c(N)nc2c3nc(CO)c(O)nc3ccc21	0.1270	True	0.9403177904594091
26	Cc1nc2c(ccc3c2[nH]c(=O)n3C)nc1O	0.2154	False	0.5632902743933411
30	Cc1nc2c(ccc3c2nc(NO)n3C)nc1O	0.1607	True	0.9671489112527334
32	Cn1c(N)nc2c3nc(C(=O)O)cnc3cc(O)c21	0.1462	True	0.931103786031207
33	Cn1c(N)nc2c3nc(CO)cnc3cc(O)c21	0.1453	True	0.9405427932760242
34	Cc1cnc2cc(O)c3c([nH]c(=O)n3C)c2n1	0.2293	False	0.3854993100237397
38	Cc1cnc2cc(O)c3c(nc(NO)n3C)c2n1	0.1997	True	0.967161448656272
39	Cn1c(N)nc2c3nc(C(=O)O)cnc3c(O)cc21	0.1728	True	0.9308316163365276
40	Cn1c(N)nc2c3nc(CO)cnc3c(O)cc21	0.1482	True	0.940317790580855
41	Cc1cnc2c(O)cc3c([nH]c(=O)n3C)c2n1	0.1525	False	0.3854993128551255
45	Cc1cnc2c(O)cc3c(nc(NO)n3C)c2n1	0.1588	True	0.9671489112903668
46	Cn1c(N)nc2c3nccnc3ccc21	0.2840	True	0.9542529809532858
47	Cn1c(=O)[nH]c2c3nc(C(=O)O)cnc3ccc21	0.9197	False	0.1992342790924088
50	Cn1c(N)n(O)c2c3nc(C(=O)O)cnc3ccc21	0.2685	False	0.4093439372469093
51	Cn1c(NO)nc2c3nc(C(=O)O)cnc3ccc21	0.9144	True	0.9633731331898584
52	Cn1c(=O)[nH]c2c3nc(CO)cnc3ccc21	0.9019	False	0.8300962026492659

55	Cn1c(N)n(O)c2c3nc(CO)cnc3ccc21	0.1906	False	0.4799896955080949
56	Cn1c(NO)nc2c3nc(CO)cnc3ccc21	0.8954	True	0.9666158494872537
58	Cc1cn(O)c2ccc3c([nH]c(=O)n3C)c2n1	0.1566	False	0.0765082336780704
61	Cc1cn(O)c2ccc3c(nc(NO)n3C)c2n1	0.0650	False	0.7917549020978537
65	Cc1cnc2ccc3c(nc(NC4OC(C(=O)O)C(O)C(O)C4O)n3C)c2n1	0.6790	False	0.1446962821904339
71	Cc1cnc2ccc3[nH]c(NC4OC(C(=O)O)C(O)C(O)C4O)nc3c2n1	0.7690	False	0.14497195402967128
72	Cc1cnc2ccc3[nH]c(N)n(C4OC(C(=O)O)C(O)C(O)C4O)c3c2n1	0.1974	False	0.09406177705658296
73	Cc1cnc2ccc3[nH]c(N)nc3c2n1C1OC(C(=O)O)C(O)C(O)C1O	0.1316	False	0.09715657659495487
74	Cc1cn(C2OC(C(=O)O)C(O)C(O)C2O)c2ccc3[nH]c(N)nc3c2n1	0.2598	False	0.09715657680399216
75	Cc1cnc2ccc3c(nc(N)n3C3OC(C(=O)O)C(O)C(O)C3O)c2n1	0.3187	False	0.8204868913526789
76	Cc1cnc2ccc3[nH]c(NS(=O)(=O)O)nc3c2n1	0.5542	False	0.3086755846244513
77	CC(=O)Nc1nc2c(ccc3ncc(C)nc32)[nH]1	0.6859	False	0.3890355639166726
86	Cc1cnc2cc(OC3OC(C(=O)O)C(O)C(O)C3O)c3c(nc(N)n3C)c2n1	0.2250	False	0.8023074573739809
87	Cc1cnc2cc(O)c3c(nc(NC4OC(C(=O)O)C(O)C(O)C4O)n3C)c2n1	0.2023	False	0.12277015897840465
91	Cc1cnc2cc(OS(=O)(=O)O)c3c(nc(N)n3C)c2n1	0.3014	True	0.8951855339987089
92	Cc1cnc2cc(O)c3c(nc(NS(=O)(=O)O)n3C)c2n1	0.1290	False	0.2451431906838536
93	CC(=O)Nc1nc2c3nc(C)cnc3cc(O)c2n1C	0.2421	False	0.3375214971698557
99	Cc1cnc2c(OS(=O)(=O)O)cc3c(nc(N)n3C)c2n1	0.1037	True	0.9186650779421685
102	Cn1c(N)nc2c3nc(C(=O)OC4OC(C(=O)O)C(O)C(O)C4O)cnc3ccc21	0.7197	False	0.7946026634538524
103	Cn1c(NC2OC(C(=O)O)C(O)C(O)C2O)nc2c3nc(C(=O)O)cnc3ccc21	0.8353	False	0.10158491150115644
105	Cn1c(N)nc2c3nc(C(=O)O)cn(C4OC(C(=O)O)C(O)C(O)C4O)c3ccc21	0.1007	False	0.0554217411453125
106	Cn1c(N)n(C2OC(C(=O)O)C(O)C(O)C2O)c2c3nc(C(=O)O)cnc3ccc21	0.3244	False	0.0552739976181928
107	Cn1c(NS(=O)(=O)O)nc2c3nc(C(=O)O)cnc3ccc21	0.5631	False	0.14500233148252945
108	CC(=O)Nc1nc2c3nc(C(=O)O)cnc3ccc2n1C	0.7234	False	0.3020083644432914
109	Cn1c(N)nc2c3nc(COC4OC(C(=O)O)C(O)C(O)C4O)cnc3ccc21	0.6966	False	0.8319718029748999
110	Cn1c(NC2OC(C(=O)O)C(O)C(O)C2O)nc2c3nc(CO)cnc3ccc21	0.8127	False	0.6232642624468664
111	Cn1c(N)nc2c1ccc1ncc(CO)n(C3OC(C(=O)O)C(O)C(O)C3O)c12	0.1052	False	0.2974558001724153
112	Cn1c(N)nc2c3nc(CO)cn(C4OC(C(=O)O)C(O)C(O)C4O)c3ccc21	0.2576	False	0.2421130103601805
113	Cn1c(N)n(C2OC(C(=O)O)C(O)C(O)C2O)c2c3nc(CO)cnc3ccc21	0.2467	False	0.19825894633489416
114	Cn1c(NS(=O)(=O)O)nc2c3nc(CO)cnc3ccc21	0.5478	False	0.7633624995670347
115	Cn1c(N)nc2c3nc(COS(=O)(=O)O)cnc3ccc21	0.7437	True	0.9566845806231064
116	CC(=O)Nc1nc2c3nc(CO)cnc3ccc2n1C	0.7038	False	0.8001797422224037

117	Cc1cnc2ccc3c([nH]c(=O)n3C)c2n1C1OC(C(=O)O)C(O)C(O)C1O	0.1015	False	0.02553909311788967
118	Cc1cn(C2OC(C(=O)O)C(O)C(O)C2O)c2ccc3c([nH]c(=O)n3C)c2n1	0.2246	False	0.02158353560208915
119	Cc1cnc2ccc3c(c2n1)n(C1OC(C(=O)O)C(O)C(O)C1O)c(=O)n3C	0.1385	False	0.09763457340287646
135	Cc1cnc2ccc3c(nc(NOC4OC(C(=O)O)C(O)C(O)C4O)n3C)c2n1	0.2782	True	0.9299795372322708
136	Cc1cnc2ccc3c(nc(N(O)C4OC(C(=O)O)C(O)C(O)C4O)n3C)c2n1	0.1950	False	0.4486358508635255
137	Cc1cnc2ccc3c(c2n1)n(C1OC(C(=O)O)C(O)C(O)C1O)c(NO)n3C	0.1222	False	0.6092029312434818
139	Cc1cn(C2OC(C(=O)O)C(O)C(O)C2O)c2ccc3c(nc(NO)n3C)c2n1	0.1742	False	0.5345948478603479
143	Cc1cn(C2OC(C(=O)O)C(O)C(O)C2O)c2ccc3c(nc(NC4OC(C(=O)O)C(O)C(O)C4O)n3C)c2n1	0.1141	False	0.00757452767551932

PhIP

Metabolites_ID	SMILES_Formula	Production probability score	Reactive_to_DNA (>=0.85)	XenoSite Reactivity score
0	Cn1c(N)nc2ncc(-c3ccccc3)cc21	1.0000	True	0.9057459333311574
1	Nc1nc2ncc(-c3ccccc3)cc2[nH]1	0.8410	True	0.959387637670526
2	Cn1c(N)nc2ncc(-c3ccc(O)cc3)cc21	0.5050	True	0.9537529526142764
4	Cn1c(N)nc2ncc(-c3cccc(O)c3)cc21	0.5480	True	0.9537195442905044
6	Cn1c(=O)[nH]c2ncc(-c3ccccc3)cc21	0.6931	False	0.0642376242299052
9	Cn1c(NO)nc2ncc(-c3ccccc3)cc21	0.7880	True	0.9669464022397948
10	COc1cc(-c2cnc3nc(N)n(C)c3c2)ccc1O	0.5480	True	0.9480287919697864
12	Nc1nc2nc(O)c(-c3ccccc3)cc2[nH]1	0.4340	True	0.9544928470545476
15	O=c1[nH]c2cc(-c3ccccc3)cnc2[nH]1	0.7144	False	0.06579856824366599
18	ONc1nc2ncc(-c3ccccc3)cc2[nH]1	0.6741	True	0.971709627243492
20	Cn1c(N)nc2nc(O)c(-c3ccc(O)cc3)cc21	0.3010	True	0.9479677590433104
21	Cn1c(N)nc2ncc(-c3ccc(O)c(O)c3)cc21	0.3027	True	0.9479677589401496
23	Cn1c(=O)[nH]c2ncc(-c3ccc(O)cc3)cc21	0.4801	False	0.05810069832844205
25	Cn1c(N)n(O)c2ncc(-c3ccc(O)cc3)cc21	0.2464	False	0.5520313639664687
26	Cn1c(NO)nc2ncc(-c3ccc(O)cc3)cc21	0.4893	True	0.9713002167285292
27	Cn1c(N)nc2nc(O)c(-c3cccc(O)c3)cc21	0.3039	True	0.9479075231870252
29	Cn1c(=O)[nH]c2nc(O)c(-c3ccccc3)cc21	0.2264	False	0.07095711372210785
32	Cn1c(NO)nc2nc(O)c(-c3ccccc3)cc21	0.4035	True	0.9713002167592756
33	COc1cc(-c2cc3c(nc2O)nc(N)n3C)ccc1O	0.2537	True	0.9412249892655148
36	Cn1c(N)nc2ncc(-c3cccc(O)c3O)cc21	0.1137	True	0.9479075230839712
38	Cn1c(=O)[nH]c2ncc(-c3cccc(O)c3)cc21	0.5191	False	0.05752792224046956

40	Cn1c(N)n(O)c2ncc(-c3cccc(O)c3)cc21	0.3507	False	0.5530915186655545
41	Cn1c(NO)nc2ncc(-c3cccc(O)c3)cc21	0.5219	True	0.9713002167285292
42	Cn1c(=O)[nH]c2ncc(-c3ccccc3)c(O)c21	0.0826	False	0.07484085325393733
47	Cn1c(=O)[nH]c2c1cc(-c1ccccc1)cn2O	0.0886	False	0.1845209476878811
48	COc1cc(-c2cnc3[nH]c(=O)n(C)c3c2)ccc1O	0.4753	False	0.10366246742473892
52	Cn1c(NO)n(O)c2ncc(-c3ccccc3)cc21	0.1478	True	0.9262429526232714
53	COc1cc(-c2cnc3c(c2)n(C)c(N)n3O)ccc1O	0.2652	False	0.5007644859889226
54	COc1cc(-c2cnc3nc(NO)n(C)c3c2)ccc1O	0.4783	True	0.96897099789443
58	Cn1c(NC2OC(C(=O)O)C(O)C(O)C2O)nc2ncc(-c3ccccc3)cc21	0.7530	False	0.12145374210618748
60	Cn1c(N)n(C2OC(C(=O)O)C(O)C(O)C2O)c2ncc(-c3ccccc3)cc21	0.3640	False	0.09464373107451456
63	O=C(O)C1OC(Nc2nc3ncc(-c4ccccc4)cc3[nH]2)C(O)C(O)C1O	0.7300	False	0.12252888605716475
64	Nc1[nH]c2cc(-c3ccccc3)cnc2n1C1OC(C(=O)O)C(O)C(O)C1O	0.4037	False	0.0972211591231239
65	Nc1nc2c(cc(-c3ccccc3)cn2C2OC(C(=O)O)C(O)C(O)C2O)[nH]1	0.2220	False	0.09169036538536668
66	Nc1nc2ncc(-c3ccccc3)cc2n1C1OC(C(=O)O)C(O)C(O)C1O	0.2388	False	0.8412783598134883
67	O=S(=O)(O)Nc1nc2ncc(-c3ccccc3)cc2[nH]1	0.4878	False	0.08766060423839672
68	CC(=O)Nc1nc2ncc(-c3ccccc3)cc2[nH]1	0.6493	False	0.358261565747931
69	Cn1c(N)nc2ncc(-c3ccc(OC4OC(C(=O)O)C(O)C(O)C4O)cc3)cc21	0.5050	False	0.8492078425506253
70	Cn1c(NC2OC(C(=O)O)C(O)C(O)C2O)nc2ncc(-c3ccc(O)cc3)cc21	0.4423	False	0.10225086050154618
72	Cn1c(N)n(C2OC(C(=O)O)C(O)C(O)C2O)c2ncc(-c3ccc(O)cc3)cc21	0.3434	False	0.08401986383673721
73	Cn1c(N)nc2ncc(-c3ccc(OS(=O)(=O)O)cc3)cc21	0.4808	True	0.9283077736841796
74	Cn1c(NS(=O)(=O)O)nc2ncc(-c3ccc(O)cc3)cc21	0.2969	False	0.08032209556006872
75	CC(=O)Nc1nc2ncc(-c3ccc(O)cc3)cc2n1C	0.3818	False	0.3209659080057035
83	Cn1c(N)nc2ncc(-c3cccc(OC4OC(C(=O)O)C(O)C(O)C4O)c3)cc21	0.5436	False	0.8492078425506252
84	Cn1c(NC2OC(C(=O)O)C(O)C(O)C2O)nc2ncc(-c3cccc(O)c3)cc21	0.4954	False	0.10225086050154618
86	Cn1c(N)n(C2OC(C(=O)O)C(O)C(O)C2O)c2ncc(-c3cccc(O)c3)cc21	0.4099	False	0.07908372490666958
87	Cn1c(N)nc2ncc(-c3cccc(OS(=O)(=O)O)c3)cc21	0.5042	True	0.928371131062364
88	Cn1c(NS(=O)(=O)O)nc2ncc(-c3cccc(O)c3)cc21	0.3222	False	0.07681060739882067
89	CC(=O)Nc1nc2ncc(-c3cccc(O)c3)cc2n1C	0.4143	False	0.3089632602528161
98	Cn1c(=O)n(C2OC(C(=O)O)C(O)C(O)C2O)c2ncc(-c3ccccc3)cc21	0.3785	False	0.041588007492785815
107	Cn1c(NOC2OC(C(=O)O)C(O)C(O)C2O)nc2ncc(-c3ccccc3)cc21	0.3436	True	0.9398086493004248
108	Cn1c(N(O)C2OC(C(=O)O)C(O)C(O)C2O)nc2ncc(-c3ccccc3)cc21	0.2238	False	0.4115043471698325
109	Cn1c(NO)n(C2OC(C(=O)O)C(O)C(O)C2O)c2ncc(-c3ccccc3)cc21	0.3215	False	0.6370610518756213

111	COc1cc(-c2cnc3nc(N)n(C)c3c2)ccc1OC1OC(C(=O)O)C(O)C(O)C1O	0.5414	False	0.824970717275429
112	COc1cc(-c2cnc3nc(NC4OC(C(=O)O)C(O)C(O)C4O)n(C)c3c2)ccc1O	0.4799	False	0.08754793690023116
113	COc1cc(-c2cc3c(nc(N)n3C)n(C3OC(C(=O)O)C(O)C(O)C3O)c2)ccc1O	0.1184	False	0.062141228656356774
114	COc1cc(-c2cnc3c(c2)n(C)c(N)n3C2OC(C(=O)O)C(O)C(O)C2O)ccc1O	0.3726	False	0.06837664241127442
115	COc1cc(-c2cnc3nc(N)n(C)c3c2)ccc1OS(=O)(=O)O	0.5042	True	0.9179597233568728
116	COc1cc(-c2cnc3nc(NS(=O)(=O)O)n(C)c3c2)ccc1O	0.3222	False	0.08307232175487678
117	COc1cc(-c2cnc3nc(NC(C)=O)n(C)c3c2)ccc1O	0.4143	False	0.27830314455341865
119	Cn1c(NC2OC(C(=O)O)C(O)C(O)C2O)n(C2OC(C(=O)O)C(O)C(O)C2O)c2ncc(-c3ccccc3)cc21	0.1321	False	0.010254377675017844
120	Cn1c(NC2OC(C(=O)O)C(O)C(O)C2O)nc2c1cc(-c1ccccc1)cn2C1OC(C(=O)O)C(O)C(O)C1O	0.1265	False	0.007643960286787444
125	CC(=O)Nc1n(C)c2cc(-c3ccccc3)cnc2n1C1OC(C(=O)O)C(O)C(O)C1O	0.1966	False	0.05268674209226962

Annexe 4

Le développement d'un pipeline de prédiction du métabolisme, son application à la caféine et son application aux amines hétérocycliques aromatiques font l'objet d'une valorisation sous la forme d'un article scientifique soumis dans la revue *BMC Bioinformatics*. L'article est transposé ci-après :

RESEARCH

Constructing xenobiotic maps of metabolism to predict the role of enzymes in DNA adduct formation

Mael Conan[1,2], Nathalie Théret[1,2], Sophie Langouet[1*†] and Anne Siegel[2*†]

[*]Correspondence:
sophie.langouet@univ-rennes1.fr;
anne.siegel@irisa.fr
[1]Irset, UMR S1085, Univ Rennes, Inserm, EHESP, Rennes, FR
[2]Irisa, UMR 6074, Univ Rennes, Inria, CNRS, Rennes, FR
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

Background: The liver plays a major role in the metabolic activation of xenobiotics (drugs, chemicals such as pollutants, pesticides, food additives...). Among environmental contaminants of concern, heterocyclic aromatic amines (HAA) are xenobiotics classified as possible or probable carcinogens (2A or 2B) by IARC for which low information exist in humans. While HAA is a family of more than thirty identified chemicals, the metabolism activation and DNA adduct formation have been fully characterized in human liver for few of them (MeIQx, PhIP, AαC).

Results: We developed a modeling approach in order to predict all the possible metabolite derivatives of a xenobiotic. Our approach relies on the construction of an enriched and annotated map of derivative metabolites from an input metabolite. The pipeline assembles reaction prediction tools (SyGMa), sites of metabolism prediction tools (Way2Drug, SOMP and Fame 3), a tool to estimate the ability of a xenobotics to form DNA adducts (XenoSite Reactivity V1), and a filtering procedure based on Bayesian framework. This prediction pipeline was evaluated using caffeine and then applied to HAAs. The method was applied to determine enzyme profiles associated with the maximization of DNA adducts formation derived from each HAA. These profiles could be very different depending on the chemicals allowing to classify HAAs which have been grouped by their associated profiles.

Conclusions: Overall, such a predictive toxicological model based on a *in silico* systems biology approach open perspectives to estimate genotoxicity of various chemical classes of environmental contaminants. Moreover, our approach based on enzymes profile determination open the perspective to predict various xenobiotics derived metabolites susceptible to bind DNA adducts in both normal and physiopathological situations.

Keywords: Metabolism; Heterocyclic aromatic amines; xenobiotics; DNA binding ability; Site of metabolism

Background

Heterocyclic Aromatic Amines (HAA) and their metabolic derivatives The liver plays a major role in the metabolic activation of xenobiotics (drugs, pollutants, pesticides, food additives...). HAA are environmental contaminants formed during the cooking of meat or fish, in cigarette smoke or exhaust gas [1, 2, 3]. HAA are contaminants of concern because previous studies have shown that they are mutagenic in bacteria, carcinogen in animals and due to a lack of epidemiological studies there

are classified as possible and probable carcinogens by the International Agency for Research on Cancer [4].

Among 30 HAAs have been identified so far. They are divided in two classes depending on their formation reaction. First there is the pyrolytic ones which are formed by a pyrolysis reaction of amino acids at temperature as high as 250°C. As an example amine of this group there is AαC (2-Amino-9H-pyrido[2,3-*b*]indole). Amino Imidazo Arene correspond to the second class of HAA which are produced by Maillard reaction between hexose and amino acids at a temperature greater than 150°C. MeIQx (2-amino-3,8-dimethylimidazo[4,5-textitf]quinoxaline), PhIP (2-amino-1-methyl-6-phenylimidazo[4,5-*b*]pyridine) and IQ (2-amino-3-methylimidazo[4,5-*f*]quinoline) belong to this class with quinoxaline, pyridine and quinoline core, respectively.

In human, HAA metabolism includes two transformation steps as illustrared in Figure 1. The first one is catalyzed by phase I metabolism enzymes which consist of an oxidation mainly catalyzed by cytochromes P450 (CYPs). The oxidative metabolite is then conjugated with phase II xenobiotic metabolism enzymes such as UDP glucuronyl transferase (UGTs) but also by glutathione S transferase (GSTs), N-acetyltransferase (NATs) and sulfotransferase (SULTs). Conjugate metabolites can be either excreted or in some case they can form aryl nitrenium ion throught heterolytic cleavage; this ion is very nucleophile, can bind DNA and form DNA adducts [4, 2].

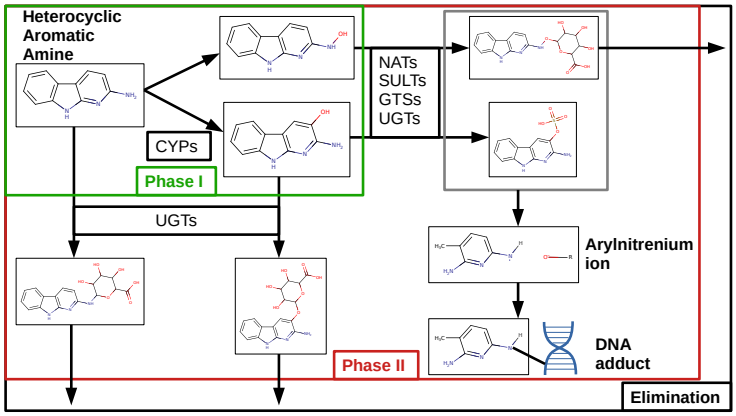


Figure 1: **Representation of Heterocyclic Aromatic Amine Metabolism.** This metabolism is divided in two steps: Phase I (in green) is known as "oxidation reaction" and catalyzed by cytochromes P450 (CYPs) and Phase II (in red) is known as a "conjugation reaction", catalyzed by UDP-glucuronyltransferases (UGTs), N-actetyltransferases (NATs), Sulfotransferases (SULTs) or Glutathione S-transferase (GSTs)

.

In order to predict the genotoxicity of a metabolite, different tools infer the possibility for a compound to bind DNA (potential DNA adduct). A first strategy is to search for specific chemical structure assumed to bind DNA because a known compound, with the similar structures, has been shown to form DNA-adducts [5, 6, 7].

Another strategy to determine if a compound can form DNA adducts is based on a Quantitative Structure-Toxicity Relationship (QSAR) score that models toxicity according to molecular descriptors of compounds [5, 6, 8]. More recent tools use deep learning to infer from the descriptors of each atom if it can bind DNA [9, 10]. The tool predict site of reactivity (SOR) associated with a SOR score representing the probability to bind DNA.

The main cornerstone to use toxicity prediction tools on HAA is that compounds which may bind to DNA result from one or several metabolic transformations, which are unstable and cannot be experimentally characterized. As a consequence, the bioactivation metabolism and DNA adduct formation are fully characterized for only three HAAs i.e., AαC, MeIQx and PhIP in human liver [11, 12, 13, 14]. This advocates for the use of in silico methods to predict HAA derivatives and potential DNA adducts derived from HAAs bioactivation in order to drive research about toxicity of HAAs family.

Prediction of metabolites To overcome the lack of information about metabolic bioactivation of HAAs and potential formation of DNA adducts, tools for metabolism prediction have been developed allowing identification of potential biomarkers of exposure in humans. Methods for the prediction of metabolites and reactions use biochemical transformation rules describing chemical reactions and linking an input chemical structure to an output chemical structure. For a given compound, the prediction tool searches for chemical structures matching with such input structures and when they are found, the rule is applied and the resulting derivatives are predicted as metabolites. Several tools implement such methods of prediction including MetaSite, METEOR, META, PROXIMAL, TIMES, UM-PPS BioTransformer or SyGMa [15, 16, 17, 18, 19, 20, 21, 22]. The predictions are often represented as a *metabolism map* containing the predicted metabolites and the reactions that link them. The main drawback of these approaches is that the use of a high number of transformation rules can lead to a great number of predictions with a high number of unknown metabolites [23].

Prediction of sites of metabolism (SOM) Another method for predicting metabolite structures uses prediction of site of metabolism (SOM) that can reduce the high number of unknown predicted metabolites in metabolism map. SOM-based tools predict the reaction of an atom by using a set of specific reactions. This set is generated by associating reactions catalysed by the same enzyme. It results in models that predict the probability for an atom to interact with specific enzymes or isoforms. These methods use molecular descriptors which describe different parameters of each atom of a compound. Some tools such as QMBO, CypScore, SMARTCyp or MetaSite [24, 25, 26, 27, 15, 28] rely on the hydrogen abstraction reaction which is the energy necessary to remove an hydrogen linked to the atom. Other tools such as Way2Drug SOMP, FAst MEtabolizer (FAME) or XenoSite Metabolism 1.0 [29, 30, 31, 9, 10] use structure parameters such as the atom nature and the nearest neighbour atoms. In these tools, machine learning methods are used to determine a score based on atom molecular descriptors, which represents the probability of an atom to be a SOM. Others SOM predictors such as IDSite, IMPACTS or MLite

[32, 33, 34] use docking methods and similarities between ligand structure and structure of the compound of interest.

The literature highlights two strategies for using SOM prediction to predict metabolic maps. A first strategy classifies and evaluates the confidence of different predicted pathways by interpreting SOM as the probability of a reaction to occur in the map. Ranking pathways with these probabilities permits to analyse the predicted metabolites and reactions. To the state of our knowledge, this recent method was only applied to the metabolism of Terbinafine (TBF), permitting to detect a new pathway that can explain the formation of TBF-A from TBF [35]. Another strategy, detailed in [36] uses SOM predictions to filter metabolic maps by removing predicted reactions which are not supported by an accurate SOM prediction. In this study, the SOM-filter threshold is determined by using a training set of analog chemicals to the chemicals of interest. The method was applied to predict HAA metabolism using SOM predictors of CYPs and UGTs enzymes. Metabolites predictions and DNA reactivity prediction were then used to predict potential DNA adducts derived from each HAA. The potential of each HAA to form DNA adducts was finally characterized by the ratio between the number of metabolites predicted to bind DNA and the number of total predicted metabolites. The main limitation of this approach is that the filtration of the metabolic maps relied on SOM scores associated with the reaction producing the putative DNA binding metabolites, getting rid of both the predecessor reactions which are required to produce intermediary metabolites and the possible multiple pathways that produce the same metabolite, as evidenced in [35].

Contribution To get further in the prediction of formation of DNA adducts by HAA, we introduce a new method which combines the concept of filtered metabolic map introduced in [36] and the concept of ranked pathways introduced in [35]. Instead of filtering metabolic maps according to individual reaction SOM scores, we introduce a *production probability score* which describes the probability for a metabolite to be produced according to one or several chains of reactions weighted by SOM scores.

Our method consists in a three steps pipeline: first step is the prediction of metabolite derivatives of the compound of interest, second step is the annotation of the resulted *metabolic map* using SOM scores and the third step is the computation of the production probability score for each metabolite with Bayesian networks in order to rank and filter metabolite maps.

We used caffeine as a matter of validation of our modeling approach based on SOM predictions of phase I and phase II xenobiotic metabolism enzymes. Indeed, caffeine metabolism is well described and shares enzymes with HAAs metabolism such as phase I enzyme especially CYP1A2, the main enzyme of caffeine metabolism, but also CYP3A4, CYP2E1, CYP2D6 and phase II enzymes including NATs. In addition, some caffeine metabolites can be produced through distinct pathways similarly to HAAs predicted metabolites. After validation of the method using caffeine, the method was applied to HAAs to predict DNA adducts formation and to identify associated enzymes signature.

Results

Definition and construction of enriched metabolic maps

Map of metabolism We define the concept of *enriched maps of metabolism* to be oriented graphs where nodes represent chemicals compounds and edges represent reactions that model the transformation of the input compound into the output compound. In these maps, different information is added as labels of reactions and nodes in order to enable the exploration of predicted metabolism results. As detailed below, in enriched metabolic maps, nodes are labeled by *smile formula DNA reactivity label* and *production probability score* and edges are labeled by *rule name*, *atom number*, *rank label*, *enzyme name* and *enzyme family*. Consequently, two edges with the same enzyme family but different enzymes are associated with different edges. An example is shown in Fig. 2.

More precisely, nodes of enriched maps of metabolism represent chemical compounds. They are associated with a SMILE formula, which is interpreted in a 2D structure allowing to label atoms with numbers according to the International Union of Pure and Applied Chemistry (IUPAC) standard conventions [37]. An other label of the node is its *DNA reactivity label*, an information provided by site of reactivity (SOR) predictors.

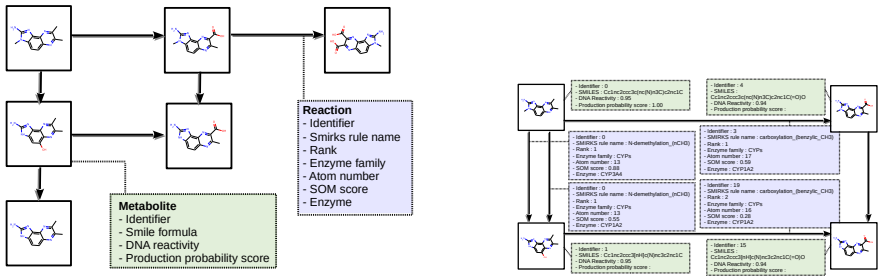
Edges are first labeled by a unique identifier and by a *rule name* that refers to a SMIRKS rule which encodes the transformation [38]. Each reaction is also associated with an *atom number of reaction*, the label of the atom in 2D structure of the input compound which is transformed by the reaction to form the output compound. Finally, each edge is labeled by an *enzyme name*, which catalyzes the reaction. The *production probability score* of the edge is determined by using site of metabolism (SOM) predictors (see methods for details) and the atom number of reaction.

A specific node is identified in the graph, named the *original compound*. It is defined to be the source of the map of metabolism, and a compound is described by its SMILES canonical formula available in PubChem database [39]. We introduce a *rank* label for each edge, which describes the position of the reaction in the graph with respect to the original compound: *first rank reactions* correspond to edges having the original compound as input, *second rank reactions* corresponds to edges whose input is the output of a first rank reaction, etc...

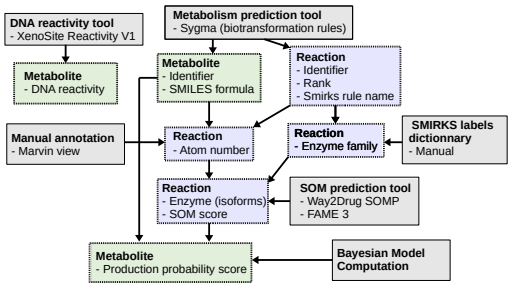
Pipeline for building a map of metabolism Maps of metabolism were built by combining several tools, which are precisely described in the Methods section.

The pipeline starts with the selection of an original compound described by its SMILES canonical formula available in PubChem database. In this paper, the method was applied to 31 original compounds: caffeine (for the sake of validation of the method), and 30 HAAs, see Results below.

For each original compound, the SyGMa python package [22] is applied to compute nodes (e.g, metabolic derivatives of the original compound) and edges (e.g., transformations between metabolites) of the associated map of metabolism. In our studies, SyGMa was iterated twice, in order to predict first-rank and second-rank reactions with respect to the original compound, e.g, all possible derivatives of the original compound with at most two transformations.



(a) Labels in an enriched map of metabolism



(b) Pipeline for building an enriched map of metabolism

Figure 2: **Definition and construction of enriched maps of metabolism** (a) **Labels of a map of metabolism**. Metabolites are depicted by their 2D structure in black squares. Arrows between metabolites represent reactions, which consume a metabolite to produce another one. Left side: labels are shown in dashed squares (green for metabolites and purple for reactions). Right side: examples of values of labels for a part of the left side network. (b) **Pipeline to obtain labels for metabolites and reactions**. Labels are depicted in green (metabolites) and purple (reactions) squares. They are computed thanks to different tools and methods (grey squares) which require as input the information provided by other labels. SMIRKS labels dictionary is detailed in Additional file 1.

The *SMILES label* (and its associated 2D structure) of each node is generated by the RDKit package implemented in SyGMa. To define the *DNA reactivity label* of each node, we consider that a node is reactive with DNA if at least one of the atom of the metabolite has a score of reactivity computed by XenoSite Reactivity [9] greater than 0.85.

For each edge, a manual curation procedure is undergone to determine the *atom number of reaction*. This label represents the number of the atom, according to IUPAC numbering [37], of the input metabolite of the edge, on which the reaction occurs to produce the output metabolite of the edge. These atom numbers of reactions are obtained by manually comparing the structure of the input and output metabolites of each reaction provided by the MarvinView tool [40], in order to identify the IUPAC numbering of the transformed atom.

The next step of the pipeline consists in annotating edges with rank, rule name and enzyme labels. For each edge, the rank label is defined to be the number of iterations of SyGMa from the original compound allowing to predict the reaction. The rule name label is also provided by SyGMa, according to a catalogue of 176 SMIRKS rules (149 for phase I xenobiotic metabolism reactions and 27 for phase II xenobiotic metabolism reactions).

In order to label each edge with an enzyme, we created a dictionary mapping every SMIRKS rule label to an *enzyme family label* (see Additional file 1). The 149 SMIRKS rules corresponding to phase I reactions are mapped to the *CYPs* enzyme family. Among the 27 SMIRKS rule labels corresponding to phase II reactions, 25 label rules are associated with the *UGTs* (13 SMIRKS labels), *NATs* (5 SMIRKS labels), *SULTs* (6 SMIRKS labels) and *GSTs* enzyme family (1 SMIRKS labels). The two remaining SMIRKS rule labels are not related to CYPs, UGTs, NATs, GSTs or SULTs and are out of the scope of the method. The corresponding edges are removed from the metabolic map. In addition, nodes appearing to be isolated in the map after this curation are also removed from the map.

The pipeline continues with a procedure used to annotate each edge with a site of metabolism (SOM) prediction score. This procedure depends on its *rank*, *enzyme label* and *atom number of reaction*. (a) Based on knowledge about xenobiotics metabolism, we assume that reactions of first-rank can be considered mostly as phase I reactions, catalyzed by different isoforms of CYPs. Therefore, the tool Way2Drug SOMP [29] is used to compute SOM scores for edges of first-rank, because it provides refine annotation of CYP isoforms (CYP1A2, CYP3A4, CYP2D6, CYP2C9 and CYP2C19), involved in phase I metabolism. As the tool also provides annotations for reactions catalyzed by UGTs, the predicted scores for such reactions are also conserved. (b) Assuming that reactions of second-rank can be considered mostly as reactions of phase II (catalyzed by SULT, UGT, NAT, GST), the tool FAME3 [31] is used to annotate reactions of second rank, because it is associated with the largest family of phase II enzymes. Note that reactions of second-rank catalyzed by UGT are annotated with a different score than reaction of first-rank catalyzed by UGTs, in order to have homogeneous and comparable scores for reactions with the same rank. (c) Notably, when an edge can be annotated with SOMs associated with different enzymes (especially for isoform predictions), the edge is duplicated for each enzyme to avoid confusion. All reactions which could not be annotated with a SOM score are removed from the metabolic map, as well as the resulting isolated nodes.

As a final step, all the metabolites of the map are associated with a SOM-based pathway production probability score (production probability score). This score depicts the probability for each node to be formed for each metabolites according to all the annotations of the reactions of the metabolic map. The approach relies on the formalism of Bayesian networks [41], a relevant framework to ensure that all possible production pathways are contributing factor to a probability of metabolite production (See methods for details).

Validation of the method: construction and analysis of a map of metabolism for caffeine

As a matter of validation of our modeling approach based on SOM predictions of phase I and phase II xenobiotics metabolsim enzymes, we applied the pipeline to caffeine. Indeed, caffeine metabolism is well described and share enzymes with HAA metabolism, linked to xenobiotics metabolism enzymes, such as CYP1A2, the main enzyme of caffeine metabolism, but also CYP3A4, CYP2E1 and CYP2D6[42]. Caffeine metabolism is also known to involve metabolites produced by NATs, the phase II enzymes of xenobiotic metabolism.

Fig. 3 shows maps of metabolism obtained as several steps of the pipeline applied the molecule of caffeine, modelled by its SMILES formula extracted from pubchem[39]. The first step of the pipeline consisted in prediction of caffeine metabolites according to two transformation steps using SyGMa[22]. SyGMa can make chemical structures predictions using two parameters that define a scenario: (i) The first parameter is a group of SMIRKS reactions to use in the predictions. We choose to use reactions related to xenobiotics metabolism available in SyGMa as two set of reactions named "phase I" and "phase II"[22]. (ii) The second parameter is a maximum number of transformations that can occur between the original chemical and a predicted metabolite. We choose to set this parameter at 2 due to the fact that the first two reactions of xenobiotics metabolism[2] (phase I and phase II) are the main biotransformation steps which are included in SyGMa SMIRKS reaction sets. The resulting map contained 23 metabolites and 31 reactions shown in Fig. 3.

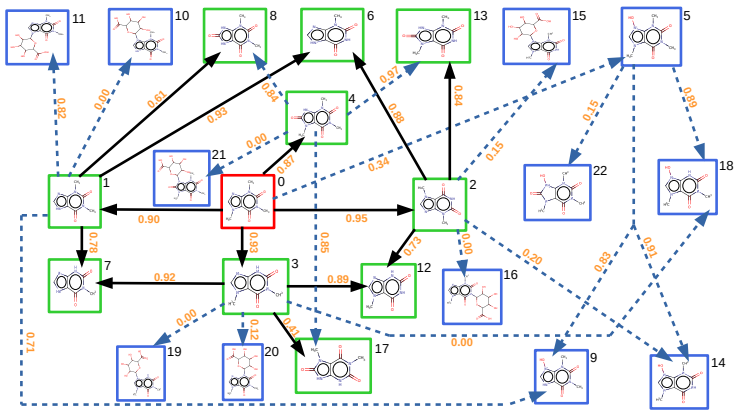


Figure 3: **Map of metabolism of caffeine predicted by the tools SyGMA** Among the 23 metabolites, the caffeine node is shown in red and metabolites which have been experimentally observed in previous studies are shown in green. Among the 31 reactions depicted by arrows, the 13 black arrows depict transformations with a literature-based evidence. Site of metabolism (SOM) score annotating reactions are shown in orange bolt on arrows. They are provided by SOM prediction tools (Way2Drug SOMP, FAME 3).

Predicted metabolites Our results are consistent with the literature knowledge [43, 44, 42, 45, 46] since 11 of the 16 known derivatives of caffeine, including caf-

feine itself, are effectively recovered according to two steps of reactions associated with phase I and phase II reactions. This metabolites are shown as green in Fig 3, except for caffeine which is red in the figure. Only two metabolite 5-acetylamino-6-formylamino-3-methyluracil, i.e., AFMU, and 6-amino-5-(N-formylmethylamino)-1,3-dimethyluracil, i.e., 137-TAU are not identified and are both associated to a NAT catalysed reaction. The other known metabolites of caffeine, such as 3-methyluric acid (i.e., 3MU), 7-methyluric acid (i.e., 7MU), and 1-methyluric acid (i.e., 1MU), are out of the scope of the method because they correspond to other conditions including for instance three steps of reactions.

The method predicted 16 reactions between known metabolites (green nodes), including the 13 reactions supported by the literature (black arrows) [45, 46, 42]. Therefore, our method predicted three new possible transformations from 137U (1,3,7-trimethyluric acid, node 4 in the figure) to 13U (node 8), 17U (node 17) and 37U (node 13), through a demethylation. The latter is similar to other reactions of the model. This suggests that the metabolite 137U has the capability to be transformed into several metabolites although this hypothesis has not been tested because of the low quantities produced in human metabolism and its elimination [42].

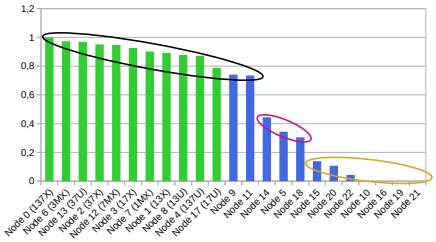
In addition to 11 known metabolites of caffeine, the method predicted that 12 other compounds are potential caffeine derivatives, which are called *caffeine metabolites* (blue nodes). Indeed the metabolite node 5 (Cn1c(=O)c2c(n(C)c1=O)[n+](O)cn2C) is predicted to be derived from caffeine after one step reaction. Ten other compounds are predicted to be derived from caffeine after two step reactions. Among these ten metabolites, three (nodes 9, 14, 18) are also predicted to be derivatives of node 5. A single metabolite (node 22) is predicted to be derived only from the newly metabolite node 5, which has the specificity to lose the aromatic structure of the imidazole, a chemical structure which is improbable according to our knowledge.

Production probability scores of predicted metabolites In order to estimate the confidence of the predicted metabolites, we computed production probability scores for each of them. As detailed in the Methods section, the production probability score first takes into account scores associated with reactions, which are computed from several site of metabolism (SOM) scores predicted by dedicated tools (Way2Drug SOMP, FAME3), e.g. the chance that a transformation occurs on a given atom. In addition, the production probability score for the metabolites also take into account the different pathways, e.g. chains of reaction, from caffeine to the considered metabolite, combined according to a Bayesian framework.

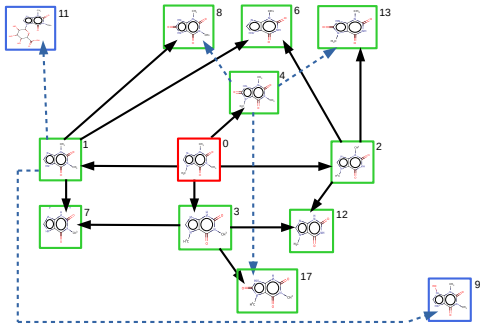
The SOM score associated to reactions are indicated as orange label on arrows in Fig. 3. We notice that five reactions have a null score. The production probability score for metabolites are shown in Fig. 4a, where metabolites are ordered from top-score metabolites to lowest scores. The 11 known metabolites (green nodes) have the largest scores. The 12 unknown predicted metabolites are represented with a blue bar.

We observe that these predicted metabolites can be gathered into three groups, each of them is shown by an ellipse with a specific color. The first group (black ellipse) describes metabolites with a score greater than 0.70: it contains all known

metabolites predicted by our method and two unknown metabolites, node 9 (Cn1c(=O)c2[nH]c[n+](O)c2n(C)c1=O) and node 11 (Cn1c(=O)c2c(ncn2C2OC-(C(=O)O)C(O)C(O)C2O)n(C)c1=O). The second group (violet ellipse) contains three derivatives with a medium score (between 0.20 and 0.70) corresponding to the unknown metabolites node 14, node 5 and node 18. The last group of seven unknown metabolites (yellow ellipse) with low score (¡ 0.20) corresponding to the nodes 15, 20, 22, 10, 16, 19 and 21. The remaining nodes 10, 16, 19 and 21 have a score of 0.00, which is explained by the fact that all the pathways which produce them contain at least a reaction with a null score.



(a) Selection of metabolites according to Bayesian model



(b) Filtered predicted caffeine map of metabolism

Figure 4: **Filtered map of metabolism for caffeine**. (a) **Production probability score of predicted metabolites** The production probability score of each metabolite is computing according to a Bayesian model based on the reaction production probability scores. Red, green and blue bars are associated to caffeine, known and unknown metabolites, respectively. Metabolites are grouped into three groups according to their range of scores (black, red and yellow ellipses). (b) **Filtered map of metabolism of caffeine** This map is the filtration of the previous predicted map in order to keep metabolites with a high production probability score (black ellipse, score >0.70). The map contains 13 metabolites and 17 reactions. Two unknown metabolites (nodes 9 and 11) are predicted to be equivalently producible than the known metabolites according to the production probability scores.

Filtered caffeine metabolic map The production probability scores were used to filter the caffeine map of metabolism as follows. The metabolites of the filtered map are all metabolites of the first group (black ellipse) in Fig. 4a, and the reactions

in the map are those of Fig. 3 transforming nodes in the black ellipse group. The filtered map is shown in Fig. 4b. We observe that all the known 11 metabolites and 13 reactions belong to this map and therefore were conserved by the filtration procedure.

The filtering procedure removed metabolites associated with nodes 5, 14, 18 and 22, which have a specific configuration regarding nitrogen atoms on the imidazole part. These four metabolites are the only ones that are oxidized derivatives of caffeine with a pattern N=C-N or N-C-N such that one of the nitrogens is methylated and the other nitrogen is linked to an oxygen atom.

The other metabolites that were eliminated by the filtration procedures are nodes 10, 15, 16, 19, 20, and 21. They are all glucuronyl-conjugates of caffeine and caffeine metabolites whose glucoronyl group is not associated with the methylated nitrogen atom of the caffeine imidazole part. The fact that most of the glucuronyl conjugates are filtered is consistent with literature because there is no glucuronyl-conjugate metabolite of caffeine described in human yet. The only glucuronyl-conjugate appearing in the final map is node 11 (with SMILES formula : Cn1c(=O)c2c(ncn2C2OC(C(=O)O)C(O)C(O)C2O)n(C)c1=O, or IUPAC name: 6-(1,3-dimethyl-2,6-dioxo-2,3,6,7-tetrahydro-1H-purin-7-yl)-3,4,5-trihydroxyoxane-2-carboxylic acid), whose glucoronyl group is linked to the imidazole part.

This suggests that node 1 (theophylline or 13x), which is already known to be metabolized into nodes 6 (3-methylxanthine or 3MX), 7 (1-methylxanthine or 1MX) and 8 (1,3-dimethyluric acid or 13U) could have also the potential to be biotransformed into another new metabolite. However the 13U and 1MX which are the most known produced metabolites from 13X [42], could compete with this new metabolite and make it undetectable.

Apart to node 11, the only unknown metabolite of the initial map conserved after the filtration is node 9 (Cn1c(=O)c2[nH]c[n+](O)c2n(C)c1=O or 9-hydroxy-1,3-dimethyl-2,6-dioxo-2,3,6,7-tetrahydro-1H-purin-9-ium). According to our method, this metabolite is a derivative of node 1, e.g. 1,3-dimethylxanthine (theophylline or 13x), with a SOM score predicted by Way2Drug of 0.714 for enzyme CYP3A4. This reaction from node 1 is an oxydation on an nitrogen atom of the imidazole part. Contrary to the oxidation of the carbon between nitrogen atom in the imidazole part which produced node 8 (13U) from node 1 with a predicted SOM score of 0.606 for enzyme CYP2D6, with the same order of magnitude, this reaction occurs on an un-methylated nitrogen atom of the imidazole part. This suggests than oxidation on nitrogen atoms of caffeine could theoretically occurs although it has never been experimentally observed.

Application to Heterocyclic Aromatc Amines (HAA) and DNA reactivity predictions

The caffeine example suggests that there is an added-value to build map of metabolism by combining several approaches such as assembling reaction prediction tools, predicting sites of metabolism and filtering the map according to a production probability score. Based on this validation, we further investigated how this method may facilitate the prediction of DNA adducts formation derived from xenobiotics. To that matter, we first constructed the predicted maps of metabolism of

the 30 human HAAs (see Additional file 2 for details). Then we annotated maps of metabolism for six HAAs of interest in order to study DNA adducts formation prediction.

Unfiltered maps of metabolism of HAAs The pipeline was applied to predict maps of metabolism of the 30 identified HAAs. The characteristics of the maps predicted by the tool SygMa are described in Table 1. HAAs are ordered according to the number of metabolites in the maps predicted by the SyGMa tool. HAAs associated with the largest map are 4-CH2OH-8-MeIQx and 4,7,8-TriMeIQx (194 predicted metabolites). The smallest maps correspond to the HAA Harman (70 metabolites) and NorHarman (50 metabolites). The size of these maps could be explained by the chemical structure of both HAAs that have few sub-structures on which SyGMa transformation rules can be applied. The maps predicted for two of the three well characterized HAAs in primary human hepatocytes [11, 12, 14], MeIQx - 155 metabolites - and PhIP - 128 metabolites, have a medium size. On the contrary, AαC, the third one, is associated with one of the smallest metabolic map.

We also noticed that the maps associated with pairs of HAAs such as 7,8-DIMeIQx and 7,9-DiMeIgQx, MeIQx and 7,MeIgQx or IQ and IQ[4,5-b] are associated with maps with similar characteristics: they have the same number of metabolites, reactions and metabolites reactive to DNA. The closeness of the chemical structure of isomers could explain such a similarity between maps: as plane isomers are composed of the same chemical substructures, the transformation rules contained in the SyGMa database have a high probability to occur equivalently on the derivatives of both HAAs.

For each HAAs, we estimated the number of metabolites reactive to DNA by assuming that each metabolite with a *XenoSite Reactivity* score greater than 0.85 is reactive to DNA, following the criteria introduced in [36]. The four HAAs, 4-CH2OH-8-MeIQx, 7,8-DiMeIQx, 7,9-DiMeIgQx and 4,7,8-TriMeIQx have both the largest map and the greatest amount of metabolites reactive to DNA (91, 80, 81 and 81). NorHarman HAA characterized by the smallest map contains no metabolite reactive to DNA. Harman, AMPNH and APNH are also associated with a very small ratio of metabolites reactive to DNA (12,7% 10,1% and 8,1%). The ratio for the other HAAs ranges from 38,8% to 55,8%. MeIQx and PhIP have a relatively high ratio of metabolites reactive to DNA (45,2% and 44,5%) while AαC has a lower ratio (38,8%) in spite of its known higher reactivity towards DNA compared with MeIQx or PhIP [14]. This observation might be related to its smallest map of metabolism (average and median of numbers of metabolites and reactions in the map). Several hypotheses can be made about the variability of the sizes of the maps of metabolism: (a) the metabolites reactive to DNA do not have the same importance *in vitro*, (b) the reactions predictions performed by SyGMa may be incomplete, as we observed it for NAT2 reactions in the case of caffeine, (c) the reactions performed by SyGMa may not be homogeneous. As detailed below, the analysis of the production probability scores of the maps suggests that the two last hypotheses are highly probable.

Manual annotation and filtering of six maps of metabolism As the pipeline for the study of maps of metabolism encompasses a part of manual annotations for atom

HAA	Metabolites	Reactions	Metabolites which are reactive to DNA	Ratio of metabolites reactive to DNA
4,7,8-TriMeIQx	194	282	91	46,9
4-CH2OH-8-MeIQx	189	266	80	42,3
7,8-DiMeIQx	174	250	81	46,6
7,9-DiMeIgQx	174	250	81	46,6
4,8-DiMeIQx	174	250	79	45,4
6,7-DiMeIgQx	169	245	76	45,0
AMPNH	157	225	16	10,1
7-MeIgQx	155	220	70	45,2
MeIQx	155	220	70	45,2
GluP1	142	202	64	45,1
IQx	137	192	62	45,3
IgQx	133	188	56	42,1
TrP1	129	179	61	47,3
PhIP	128	177	57	44,5
MeIQ	125	175	64	51,2
4'-OH-PhIP	123	166	49	39,8
3,5,6-TMIP	122	172	57	46,7
APNH	120	165	10	8,3
MeAαC	113	154	48	42,5
TrP2	113	154	49	43,4
GluP2	110	151	46	41,8
IQ	109	150	59	54,1
IQ[4,5-b]	109	150	59	54,1
1,5,6-TMIP	107	153	59	55,1
1,6-DMIP	95	132	53	55,8
IFP	90	123	37	41,1
AαC	85	111	33	38,8
PheP1	76	97	35	46,1
Harman	70	98	9	12,7
NorHarman	50	65	0	0,0
Average	127,6	178,7	53,7	41,0
Median	124	173,5	58	45,2

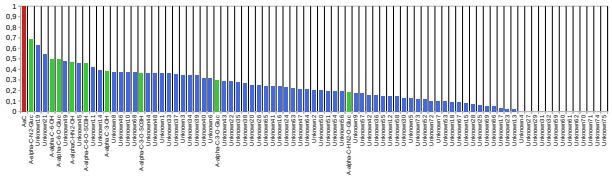
Table 1: Characteristics of the maps of metabolism predicted according to biotransformations rules by the SyGMa tool for 30 HAA.

number of reactions, we applied the pipeline to a selection of six HAAs among the 30 HAAs. We first selected the three well desccribed HAAs in human hepatocytes AαC, PhIP and MeIQx [11, 12, 14, 13]. We complemented this list with the two HAAs with the largest map (4-CH2OH-8-MeIQx and 4,7,8-TriMeIQx). We finally selected 7,8-DiMeIQx which represents the pair of isomers (7,8-DIMeIQx and 7,9-DiMeIgQx) having large maps. Note that the three selected HAAs (4-CH2OH-8-MeIQx, 4,7,8-TriMeIQx and 7,8-DIMeIQx) have also the largest amount of metabolites reactive with DNA. We applied the prediction pipeline and the annotated maps of metabolism obtained were explored to identify enzyme families and isoforms associated with reactions. We confirmed that most enzymes families (SULTs, NATs, CYPs, UGTs) and CYPs isoforms (CYP1A2, CYP2C19, CYP2C9, CYP2D6 and CYP3A4) annotate at least one reaction in each map with the exception of GSTs that are not found in annotated map.

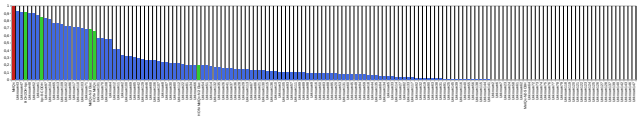
The annotated maps of metabolism obtained from AαC, PhIP and MeIQx were explored in order to identify the metabolites corresponding to the metabolites described in humans of these three AHAs. Upon the 11, 10 and 9 derivatives experimentally shown for AαC, MeIQx and PhIP 9, 6 and 7 are found in the annotated maps of metabolism. Among the 8 known derivatives not present in annotated maps of metabolism, three are N-sulfonyl derivatives of each HAA and three are N-acetoxy derivatives of each HAA. This suggests that N-sulfonyl and N-acetoxy derivatives may not be predicted using SyGMa's SMIRKS rules, supporting the hypothesis (b) above. The two last known derivatives of MeIQx, 7-oxo-MeIQx and N-desmethyl-7-oxo-MeIQx are not present in annotated maps of metabolism but we identified two metabolites, with SMILES formula Cc1nc2c(ccc3c2nc(N)n3C)nc1O and Cc1nc2c(ccc3[nH]c(N)nc32)nc1O, which are close to these known missing derivatives. The main structural difference between those metabolites and 7-oxo-MeIQx and N-desmethyl-7-oxo-MeIQx is that the ketone group of the 7-oxo-MeIQx part is replaced by an hydroxyl group. This result suggests also that SyGMa's SMIRKS rules are not able to predict ketone group linked to a carbon of an heterocycle, still supporting the hypothesis (b) above.

Figs. 5a, 5b and 5c, show that the distribution of the production probability scores of the known metabolites of each AHA is rather scattered with scores ranging from 0.9 to 0.2. Surprisingly, the metabolite MeIQx-N2-SO3H, in the map of metabolism of MeIQx, is the only experimentally identified metabolite associated with a production probability score of 0.0. A null score suggests that the metabolic pathways leading to the metabolite contains at least one reaction that has not been annotated either by Way2Drug for a reaction of rank 1, or by FAME 3 for a reaction of rank 2.

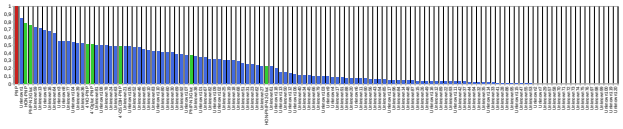
More generally, we notice that PhIP and MeIQx have a high amount of metabolites associated with a low production probability score, compared to AαC. 83 metabolites (53,5% of the map) of the MeIQx map have a score lower than 0.1, including 36 metabolites with a nul score (23,2% map). Similarily, 65 metabolites (50,8% of the map) of the PhIP map have a score lower than 0.1 including 18 metabolites with a nul score (14,1% map) and no known metabolite. On the contrary, only 26 metabolites (30,6% of the map) of the AαC map have a score lower than 0.1, including 13 metabolites with a nul score (15,3% map). Based on this remark, and



(a) Distribution of the prediction probability score of each metabolite of the annotated map of metabolism of AαC



(b) Distribution of the prediction probability score on each metabolite of the annotated map of metabolism of MeIQx



(c) Distribution of the prediction probability score on each metabolite of the annotated map of metabolism of PhIP

Figure 5: **Distribution of prediction probability scores in annotated maps of metabolism** of AαC (**a**), MeIQx (**b**) and PhIP (**c**). The origin compound of each map is shown with a red bar. Metabolites known as metabolites experimentally described are associated to a green bar and other metabolite are associated to a blue bar. The names of metabolites shown in X-axis are either the name found for known metabolites or a two parts name : *Unknown* + *X* where X describes the identifier of the node in the map associated with this metabolite.

after comparing the effect of different thresholds, we used the value 0.10 to filter metabolites with little support according to our model.

Table 2 describes the characteristics of the six selected filtered maps of metabolism. Details about metabolites contained in these maps are provided in Additional file 3. We observe that the filtration procedure has a strong impact on sizes of the maps of metabolism. 126 metabolites derived from 4-CH2OH-8-MeIQx map of metabolism are filtered representing 66% of the map and more than half of metabolites are filtered in the maps of PhIP, MeIQx, 7,8-DiMeIQx and 4,7,8-TriMeIQx. By contrast, the filtration procedure had a lower impact on AαC since only 26 metabolites are filtered. This suggests that many reactions which were predicted according to existing transformation rules were currently not supported by sites of metabolism.

We observe that the filtering procedure based of prediction probability scores tends to homogeneise the size of the final maps of metabolism while keeping a similar ratio of metabolites predicted to be reactive to DNA (between 40% and 50%). This

HAA	Metabolites after filtration	Filtered metabolites	Reactions after filtration	Filtered Reactions	DNA Reactive Metabolites after filtration	Filtered DNA Reactive Metabolites
4,7,8-TriMeIQx	87	107	120	162	41	50
7,8-DiMeIQx	90	84	116	134	43	38
MeIQx	72	83	90	130	27	43
PhIP	63	65	74	103	22	35
4-CH2OH-8-MeIQx	63	126	78	188	31	49
AaC	59	26	72	39	20	13

Table 2: Characteristics of six HAA maps of metabolism filtrated according the production probability scores computed after the annotation of all metabolites of each map.

suggests that there is a strong interest in using SOM scores (as included in the production probability scores) to homogeneise maps of metabolism and eliminate unsupported DNA reactive metabolites.

Optimal enzymatic signature in terms of DNA reactivity As described previously, the pipeline relies on the computation for SOM scores on reactions involving manually annotated metabolites to reduce the maps of metabolism according to a Bayesian prediction probability. As SOM scores are directly related to enzymes, the production probability score is influenced by enzymes availability. We define *enzymatic contexts* to be tables describing all the possible combinations of enzymes that may be considered as available. In our study, there are 512 such different enzymatic context. Each enzymatic context is associated with a specific distribution of production probability scores. Indeed, when an enzyme is described as *unavailable* in an enzymatic context, the reactions annotated with this enzyme cannot be taken in account for the calculation of the production probability scores.

Based on this assumption, for each of the six filtered maps of metabolism described above, and for each of the 512 enzymatic context, we computed the production probability scores of all metabolites of the map. This allowed us to determine the *global reactivity score of a HAA in a given context* that we defined as the sum of the production probability score (in the considered context) of all metabolites reactive to DNA in the considered map of metabolism.

In this framework, it becomes possible to define *optimal enzymatic signatures in terms of reactivity*, which are all the enzymatic contexts where the *global reactivity score* is maximized while they contain the smallest number of activated enzymes. Intuitively, optimal enzymatic signatures therefore correspond to enzymatic contexts where the chance to obtain at least one metabolite reactive to DNA is maximal according to our models.

Impact of enzymatic context to the production probability score and application on DNA adduct formation of HAAs Fig. 6 shows all optimal enzymatic signatures for the six HAA.

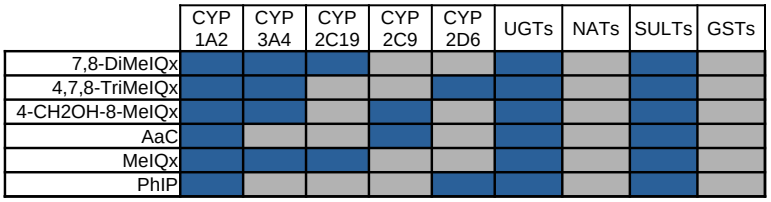


Figure 6: **Optimal enzymatic signatures in terms of reactivity.** A blue cell corresponds to an available enzyme. A grey cell corresponds to an unavailable enzyme. The threshold to determine if a metabolite is reactive to DNA was 0.85.

We observe that the enzymes UGTs, SULTs and CYP1A2 are present in all optimal signatures of the six HAA. This is consistent with the literature which describes the implication of SULTs and CYP1A2 in the formation of HAA DNA adducts [2]. In addition, it has been recently shown that UGTs an also be involved in a pathway leading to DNA adduct formation for AαC[14].

Conversely, the enzymes NATs and GSTs are absent in all optimal signatures. While the absence of GSTs is explained by the fact that annotated maps of metabolism do not contain any GSTs, the absence of NATs suggest that the resulting metabolites are not reactive to DNA since NATs are present in all maps of metabolism. However, NAT2 has been previously involved in formation of DNA adducts derived from HAAs [2]. We hypothesise that NATs reaction implicates other isoforms in the maps of metabolism. In accordance with this hypothesis, NAT2 enzymes did not appear in the map of metabolism predicted for caffeine since it is involved in the only known missing metabolite (AFMU).

Fig. 6 suggests that HAA is characterized by a specific enzymatic profile. The profiles according to the availability of the CYP isoforms other than CYP1A2: CYP3A4 is available in four of the six optimal signatures and correspond to all HAAs with a MeIQx chemical structure. This suggests a different involvement of CYP isoforms, other than CYP1A2, in the formation of metabolites that are highly regarded as reactive and depending on the structure of the HAA.

Impact of the XenoSite Reactivity threshold In order to test the impact of the reactivity threshold chosen to characterize all the compounds of a DNA reactive map, we decided to explore all the thresholds from 0 to 1.0. For the 101 threshold values considered (step of 0.01), we recalculated all the metabolites considered as "reactive" (i.e. associated with a XenoSite Reactivity score greater than or equal to the threshold), and then calculated the optimal signatures for reactivity to DNA. According with our previous results we did not consider NATs and GSTs. As shown in Fig. 7, the enzymes CYP1A2, UGTs and SULTs are present in the optimal signature whatever the threshold thereby suggesting that the result described in

Fig. 6 for the reactivity threshold value 0.85 are robust. The Fig. 7 also confirms the specificity of CYP3A4 in optimal signatures of HAAs with an MeIQx chemical structure.

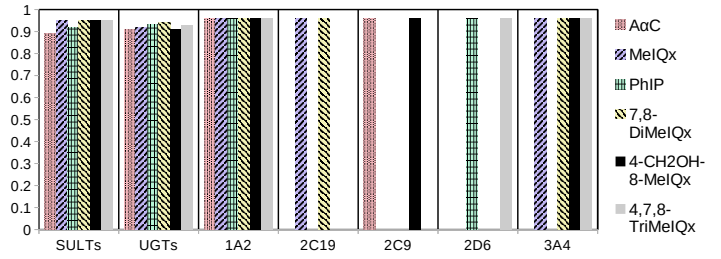


Figure 7: **Optimal enzymatic signatures in term of reactivity for each reactivity score thresholds**. Each line describe an optimal enzymatic signature for each HAA for a reactivity score threshold.

We further observed a low variability between HAAs since the largest optimal signature is reached at a reactivity threshold of 0.89 for AαC, 0.91 for 4-CH2OH-8-MeIQx, 0.92 for MeIQx and PhIP, 0.93 for 4,7,8-TriMeIQx and 0.95 for 7,8-DiMeIQx. In addition, the use of any reactivity threshold lower than 0.89 returns the same optimal signature for each HAA. This suggests that enzymes involved in metabolic pathways leading to the most reactive metabolites (given by XenoSite Reactivity scores), are sufficient to activate all the pathways leading to the less reactive metabolites.

The Fig. 8 is a counter-part of Fig. 7 to compare the values of XenoSite reactivity associated with the apparition of each enzyme in an optimal signature. When considering high XenoSite Reactivity thresholds, we observed that cytochromes P450 isoforms are the only enzymes present in all optimal signatures (thresholds from 0.96 to 0.97). This suggests that CYPs are responsible of the production of most of DNA-reactive metabolites, especially CYP1A2 found in all HAAs. Therefore, reactive metabolites derived from CYPs-annotated reactions may form DNA-adduct more easily than reactive metabolites derived from phase II enzymes-annotated reaction. In addition, we observed that the enzymes UGTs and SULTs were present in the optimal signatures for reactivity thresholds between 0.95 and 0.89. For MeIQx, 4-CH2OH-8-MeIQx, 7,8-DiMeIQx and 4,7,8-TriMeIQx, we noted that the reactivity thresholds for which SULTs is present in the optimal signature is greater than the one for UGTs. This suggests that SULTs-associated metabolites are more likely to form DNA adducts than UGTs-associated metabolites when the chemical structure of the HAA is close to the MeIQx structure.

Discussion and conclusion

In this study, we introduced a pipeline for predicting the metabolism of xenobiotics in humans. The pipeline was applied to six HAA of interest and caffeine. Our predic-

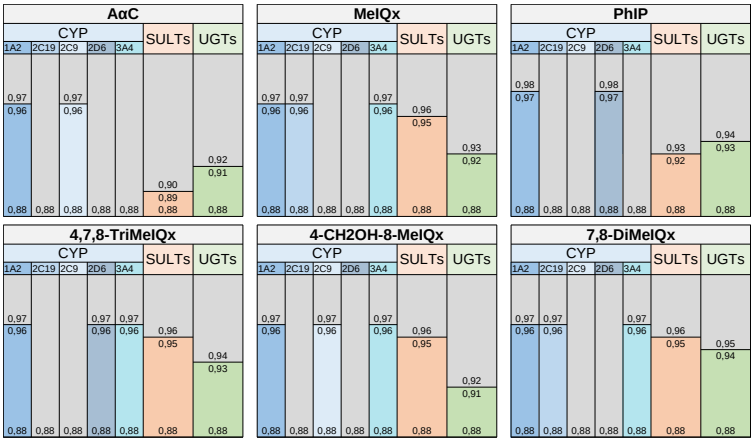


Figure 8: **Optimal signatures in terms of reactivity for the highest reactivity thresholds**. A grey cell means that the enzyme is not a part of the optimal signature and a colored cell mean the enzyme is available in the optimal signature. Each lines represent the optimal signature at a specific reactivity score threshold. Thresholds marking the integration of new enzyme in the optimal signature are annotated on each bar.

tion pipeline is based on the construction of so-called *enriched maps of metabolism*. The main specificity of the pipeline is to use *production probability score* to sort metabolites according both to prediction of site of metabolisms and the topology of the maps predicted by bio-transformation rules. This score allows comparing metabolite production in different physiopathological conditions that permit to explore the role of enzymes in the production of specific metabolites. In this study we focus on metabolites and their DNA adduct formation capacity.

The pipeline was used to reconstruct the maps of caffeine metabolism and of six HAAs. Among the four xenobiotics for which the metabolism was known i.e;, MeIQx, PhIP, AαC and caffeine, the majority of their metabolites described in humans were found. Among these known metabolites no one were removed from the maps after the filtration on the production probability scores, with the exception of MeIQx-N2-SO3H.

In the maps of metabolism predicted by our pipeline, most of experimentally identified caffeine and HAA metabolites are associated with a high prediction probability score. On the contrary a large number of predicted metabolites, which can be considered as over-predicted metabolites, are not supported by site of metabolism predictions.

This allowed us to use prediction probability scores as filters that importantly reduced the size of the maps. In the case of HAAs, half of the predicted metabolites were filtered out. Thus map of metabolism produced by other tools could be filtered by this production probability score.

The prediction of the map of caffeine and HAA metabolism highlighted areas for improvement in our pipeline. First, the study of the caffeine map of metabolism evidenced that a metabolite (AFMU), resulting from a NAT2-directed reaction, is not predicted by the SyGMa tool, suggesting that the tool is incomplete in predicting reactions catalyzed by N-acetyl transferases. This assumption is also supported by the analysis of the optimal HAAs signatures as detailed in figure6. The prediction of HAA maps of metabolism also suggested lacks of prediction of N-Sulfonyl and N-Acetoxy derivatives of AαC, PhIP and MeIQx, which were not predicted because there are no SMIRKS rules in SyGMa adapted to the prediction of these metabolites. These different unpredicted metabolites advocate for relying on the capability of the SyGMa tool to add new SMIRKS biotransformation rules to its prediction rules in order to complete the maps of metabolism with relevant enzymes.

A characteristics of the map of metabolism of MeIQx is that MeIQx-N2-SO3H, an experimentally identified metabolite, is associated with a nul prediction probability score. This is explained by the fact that the reaction producing MeIQx-N2-SO3H is labeled as a rank 1 reaction, catalyzed by the SULTs. Our pipeline differentiate rank 1 and rank 2 reactions and could not annotate reactions of rank 1 with SO score only available for tools annotating reactions of rank2. To overcome this issue, we plan to differentiate the SOMs prediction tools according to the enzymes annotating the reactions instead of the rank of the reactions. This however requires to homogeneize the level of information about isoforms.

The production probability score that we defined allowed us to analyze the influence of enzymes on the production of DNA reactive metabolites and to propose a specificity of the CYP3A4 enzyme in the production of DNA adducts derived from AHAs close to MeIQx, which will be the subject of further experiments.

In conclusion, our study describes a new method for the construction and analysis of maps of metabolism by combining prediction of biotransformation rules, predictions of site of metabolisms, and prediction of reactivity to DNA. The method was validated and applied to six xenobiotics. The further study will consist in applying the pipeline to the 24 other human HAAs, which requires to automatize the annotation of metabolites predicted by transformation rules. Moreover, our approach based on enzymes profile determination open the perspective to predict various xenobiotics derived metabolites susceptible to bind DNA adducts in both normal and physiopathological situations. The enzymatic contexts extracted from data repositories such as TCGA [47] and GTEX [48] makes this goal achievable.

Method

Tool for the prediction of edges and nodes of maps of metabolism: SyGMa

The SyGMa python package (Systematic Generation of potential Metabolites) [22] is a rule-based method to predict metabolite (e.g. derivative compounds) from an input chemical compound. In the paper, the input compound was either caffeine or HAA. The method relies on an internal set of metabolic reactions (biotransformation rules) in SMIRKS format which can be applied to the input compound. If the input chemical structure of a SMIRKS reaction is detected in the compound, the reaction is applied and the resulting structure obtained is a predicted metabolite.

The first parameter required by SyGMa to make prediction is the set of SMIRKS reactions to use. In this paper, we used the two sets of SMIRKS reactions, named

"phase I" and "phase II", obtained by data mining the Metabolite Database[22] and corresponding to reactions involved in phase I and II metabolism of xenobiotics.

The second parameter required by SyGMa is the number of reactions (which we call *rank*) that separates the original compound and a metabolite. If this maximal rank number is greater than 1, the metabolites obtained after a first iteration are used as source for new reactions to obtain second rank metabolites, until the maximal rank is reached. In the paper, the maximal rank number was equal to 2. This allows reproducing the main observed xenobiotic metabolism with a first reaction associated to phase I enzymes of xenobiotic metabolism and a second reaction that conjugates the oxidized metabolite by enzymes of phase II metabolism of xenobiotics.

Tools for the prediction of sites of metabolisms (SOM)

We use tools for site of metabolism (SOM) prediction tools to annotate reactions that could be catalysed by phase I or phase II enzymes of xenobiotics metabolism. All methods we used for predicting SOMs consists in inferring models which can be applied to an atom configuration according to a set of valid reactions which is specific to each model, extracted from a database. Using these models, each method can evaluate for each atom of a compound if it can be involved in a reaction similar to those described in the reaction database. This analysis results in a score for each atom of a compound that describes the probability to be transformed by each reaction. Therefore, the methods differ according to the parameters describing the atom configuration and the enzymes involved in the reactions (different enzyme family and/or isoforms).

In order to take advantage of the panel of existing methods, our procedure include different tools to annotate each edge (reaction) with a site of metabolism (SOM) prediction score depending on its *rank*, *enzyme label* and *atom number of reaction* (see results section).

Way2Drug SOMP [29] is used to predict SOMs associated with reactions applied to the original compound as source (first-rank reactions). This SOM predictor has different CYP models and can provide a SOM score for each atom of a compound for five the main cytochrome isoforms: 1A2, 3A4, 2D6, 2C9 and 2C19. It also predicts SOM scores for the UGT enzyme family but does not specify any isoforms.

When input of a reaction is not the original compound, we used FAME 3 [31] tool, which provides SOM score for phase II enzymes of xenobiotics metabolism. We use specific models restrained to specific reactions catalyzed by the five enzyme families: N-acetyltransferase (NATs), Sulfotransferases (SULTs) and Glutathione S-transferases (GTSs), UDP-glucuronosyltransferase (UGTs) and cytochromes P450 (CYPs).

Prediction of SOR

XenoSite Reactivity [9, 10] is used to annotate the DNA reactivity of each metabolite. Based on the SOM scores, this tool computes a score of reactivity (SOR) for each atom of the metabolite meaning ability of the atom to bind DNA. It also relies on the atom configuration, described by molecular descriptors, and uses deep-learning to infer a model that predicts the probability for an atom be involved in a

specific set of reactions that are DNA-binding reactions. To determine if a metabolite is considered as reactive to DNA, we apply a threshold on the SOR scores if at least one SOR score of an atom of the metabolite is retained the metabolite is considered as reactive to DNA. We use the same threshold as in [36] where XenoSite Reactivity where a threshold of 0.85 was learned according to metabolites known to be reactive to DNA.

Scoring metabolites with a SOM-based pathway production probability score A specific method was designed to compute a probability for each node to be formed for each metabolites according to all the annotations of the reactions of the metabolic map of metabolism. The approach relied on the formalism of Bayesian networks [41] which are probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG). Bayesian networks were used to predict the the production probability score of each metabolite of the metabolic map, assuming that all possible production pathways are contributing factor to this score.

The graph (DAG) on which probabilities were computed was derived from the map of metabolism based on the principles that (1) when several isoforms of the same enzyme can transform a compound to another, the total affinity of the compound with the enzyme family can be approximated by the maximal isoform enzyme affinity (2) when several members of several enzyme families are competing to produce the same target metabolite from different input metabolites, the recruitment of enzymes in reactions follows an exclusivity principle such that each enzyme family can catalyze the production of the targeted metabolite with at most one reaction.

Consequently, a variable of the Bayesian model was build for each node (e.g. metabolites) of the map of metabolism. It was therefore associated with the event *production of the metabolite*. The DAG was built according to the following rules, which are illustrated in Fig.9. (a) For reactions which has the same input and output and therefore varied only by the isoform of their enzyme family (in our case, CYPs), we selected the edge with the maximal SOM score and removed the other edges of the graph. (b) For each metabolite which is the output of several edges with different input, we selected the enzyme family (UGTs, NATs, SULTs, CYPs, GSTs) with the largest SOM score and removed all the other edges producing the targeted metabolite with the same enzyme family. Then we selected the remaining enzyme family producing the targeted metabolite with the second largest score and again removed from the graph all the edges leading to the targeted metabolite with an enzyme of the same enzyme family. This was repeated until each enzyme family could produce the targeted metabolite from at most an input compound.

For each edge, the SOM score was interpreted as a conditional probability, which is the probability to get the output compound of the edge assuming the presence of input compound. The structure of the graph, which is both acyclic (because the pipeline for building metabolic maps cannot create cycles) and such that enzymes do not compete for the production of a same metabolite, yields that conditional probabilities are independent. This allowed creating a complete probability table associated with each metabolite production event.

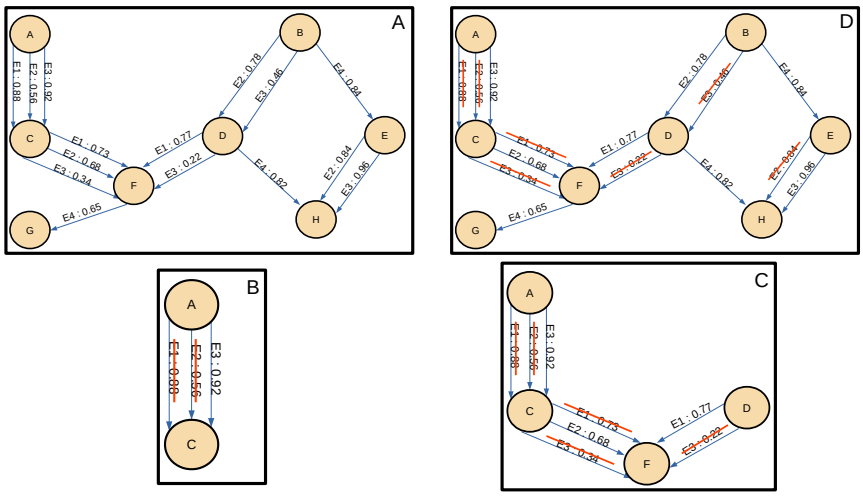


Figure 9: **Creating a Bayesian network from an annotated map of metabolism** Considering A) An annotated map of metabolism with enzymes named E1, E2, E3 and E4 and a SoM score associated to each enzyme for each reactions. B) Illustrate the application of the rule: "only one enzyme can catalyze a reaction". There is a reaction from A to C, divided in three reactions one for each enzyme annotating the reaction. Those reactions need to be reduced in one with a SoM score used as probability for the Bayesian Network. SoM score annotating reactions are indicators of enzymatic affinity so the reaction with maximal SoM score is selected. Here it is the reaction annotated by E3. C) Illustrate a different case where there is reactions from D and from C leading to production of F. This is the application of "if an enzyme is recruited by a reaction, another enzyme has to be recruited for a second reaction with the same product". Here the maximal SoM score annotating reactions is 0.77 for the reaction from D to F annotated by E1. Because this reaction is the retain reaction from D to F, E1 is no more allowed to annotate the reduced reaction from C to F. The reaction retain from C to F, is the second reaction with the maximal SoM score which is the reaction annotated by E2. D) Illustrate the final reduced graph used as Bayesian Network to calculate production probability score of metabolites.

The *production probability score* of each metabolite was defined to be the probability of a metabolite production. These scores were computed by using the conditional probability tables to expand joint probability function (Bayes formula) [41].

Acknowledgements
Not applicable

Funding
The work has supported by the Institut National de la Santé et de la Recherche Médicale (Inserm), University of Rennes 1, Ligue contre le Cancer du Grand Ouest, PNREST Anses cancer TMOI AVIESAN 2013/1/166.

Abbreviations
HAA: heterocyclic aromatic amines

Availability of data and materials
The description of HAAs by SMILE formula were provided in [36]. The SMILES formula of caffeine was extracted from PubChem: https://pubchem.ncbi.nlm.nih.gov/. They are all available in the Supp. data 2.

Ethics approval and consent to participate
Not applicable

Competing interests
The authors declare that they have no competing interests.

Consent for publication
The authors declare that they consent for publication.

Authors' contributions
M.C developed and implemented the pipeline and produced all the results. N.T. and S.L. analyzed the biological consistency of the results. A.S. contributed to the design of the pipeline, the choice of the different parameters and the data analysis. All authors contributed to the manuscript. A.S. and S.L. contributed equally to the direction of the work.

Author details
[1]Irset, UMR S1085, Univ Rennes, Inserm, EHESP, Rennes, FR. [2]Irisa, UMR 6074, Univ Rennes, Inria, CNRS, Rennes, FR.

References
1. Ni, W., McNaughton, L., LeMaster, D.M., Sinha, R., Turesky, R.J.: Quantitation of 13 heterocyclic aromatic amines in cooked beef, pork, and chicken by liquid chromatography-electrospray ionization/tandem mass spectrometry. J Agric Food Chem **56**(1), 68–78 (2008)
2. Turesky, R.J., Le Marchand, L.: Metabolism and biomarkers of heterocyclic aromatic amines in molecular epidemiology studies: lessons learned from aromatic amines. Chem Res Toxicol **24**(8), 1169–1214 (2011)
3. OZ, F., Kaya, M.: Heterocyclic Aromatic Amines in Meat. Journal of Food Processing and Preservation **35**(6), 739–753 (2011)
4. Gibis, M.: Heterocyclic Aromatic Amines in Cooked Meat Products: Causes, Formation, Occurrence, and Risk Assessment. Compr Rev Food Sci Food Saf **15**(2), 269–302 (2016)
5. Marchant, C.A., Briggs, K.A., Long, A.: In silico tools for sharing data and knowledge on toxicity and metabolism: derek for windows, meteor, and vitic. Toxicol Mech Methods **18**(2-3), 177–187 (2008)
6. Jeliazkova, N., Jeliazkov, V.: AMBIT RESTful web services: an implementation of the OpenTox application programming interface. J Cheminform **3**, 18 (2011)
7. Patlewicz, G., Jeliazkova, N., Safford, R.J., Worth, A.P., Aleksiev, B.: An evaluation of the implementation of the Cramer classification scheme in the Toxtree software. SAR QSAR Environ Res **19**(5-6), 495–524 (2008)
8. Rudik, A.V., Bezhentsev, V.M., Dmitriev, A.V., Druzhilovskiy, D.S., Lagunin, A.A., Filimonov, D.A., Poroikov, V.V.: MetaTox: Web Application for Predicting Structure and Toxicity of Xenobiotics' Metabolites. J Chem Inf Model **57**(4), 638–642 (2017)
9. Hughes, T.B., Miller, G.P., Swamidass, S.J.: Site of reactivity models predict molecular reactivity of diverse chemicals with glutathione. Chem Res Toxicol **28**(4), 797–809 (2015)
10. Hughes, T.B., Dang, N.L., Miller, G.P., Swamidass, S.J.: Modeling Reactivity to Biological Macromolecules with a Deep Multitask Network. ACS Cent Sci **2**(8), 529–537 (2016)
11. Langouet, S., Welti, D.H., Kerriguy, N., Fay, L.B., Huynh-Ba, T., Markovic, J., Guengerich, F.P., Guillouzo, A., Turesky, R.J.: Metabolism of 2-amino-3,8-dimethylimidazo[4,5-f]quinoxaline in human hepatocytes: 2-amino-3-methylimidazo[4,5-f]quinoxaline-8-carboxylic acid is a major detoxification pathway catalyzed by cytochrome P450 1A2. Chem Res Toxicol **14**(2), 211–221 (2001)
12. Langouet, S., Paehler, A., Welti, D.H., Kerriguy, N., Guillouzo, A., Turesky, R.J.: Differential metabolism of 2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine in rat and human hepatocytes. Carcinogenesis **23**(1), 115–122 (2002)
13. Nauwelaers, G., Bellamri, M., Fessard, V., Turesky, R.J., Langou?t, S.: DNA adducts of the tobacco carcinogens 2-amino-9H-pyrido[2,3-b]indole and 4-aminobiphenyl are formed at environmental exposure levels and persist in human hepatocytes. Chem Res Toxicol **26**(9), 1367–1377 (2013)
14. Bellamri, M., Le Hegarat, L., Turesky, R.J., Langou?t, S.: Metabolism of the Tobacco Carcinogen 2-Amino-9H-pyrido[2,3-b]indole (AÎ±C) in Primary Human Hepatocytes. Chem Res Toxicol **30**(2), 657–668 (2017)
15. Cruciani, G., Carosati, E., De Boeck, B., Ethirajulu, K., Mackie, C., Howe, T., Vianello, R.: MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist. J Med Chem **48**(22), 6970–6979 (2005)
16. Marchant, C.A., Briggs, K.A., Long, A.: In silico tools for sharing data and knowledge on toxicity and metabolism: derek for windows, meteor, and vitic. Toxicol Mech Methods **18**(2-3), 177–187 (2008)
17. Klopman, G., Dimayuga, M., Talafous, J.: META. 1. A program for the evaluation of metabolic transformation of chemicals. J Chem Inf Comput Sci **34**(6), 1320–1325 (1994)
18. Yousofshahi, M., Manteiga, S., Wu, C., Lee, K., Hassoun, S.: PROXIMAL: a method for Prediction of Xenobiotic Metabolism. BMC Syst Biol **9**, 94 (2015)
19. Mekenyan, O.G., Dimitrov, S.D., Pavlov, T.S., Veith, G.D.: A systematic approach to simulating metabolism in computational toxicology. I. The TIMES heuristic modelling framework. Curr Pharm Des **10**(11), 1273–1293 (2004)
20. Gao, J., Ellis, L.B., Wackett, L.P.: The University of Minnesota Biocatalysis/Biodegradation Database: improving public access. Nucleic Acids Res **38**(Database issue), 488–491 (2010)

21. Djoumbou-Feunang, Y., Fiamoncini, J., Gil-de-la-Fuente, A., Greiner, R., Manach, C., Wishart, D.S.: BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. J Cheminform **11**(1), 2 (2019)
22. Ridder, L., Wagener, M.: SyGMa: combining expert knowledge and empirical scoring in the prediction of metabolites. ChemMedChem **3**(5), 821–832 (2008)
23. Bugrim, A., Nikolskaya, T., Nikolsky, Y.: Early prediction of drug metabolism and toxicity: systems biology approach and modeling. Drug Discovery Today **9**(3), 127–135 (2004). doi:10.1016/S1359-6446(03)02971-4
24. Hennemann, M., Friedl, A., Lobell, M., Keldenich, J., Hillisch, A., Clark, T., G?ller, A.H.: CypScore: Quantitative prediction of reactivity toward cytochromes P450 based on semiempirical molecular orbital theory. ChemMedChem **4**(4), 657–669 (2009)
25. Afzelius, L., Arnby, C.H., Broo, A., Carlsson, L., Isaksson, C., Jurva, U., Kjellander, B., Kolmodin, K., Nilsson, K., Raubacher, F., Weidolf, L.: State-of-the-art tools for computational site of metabolism predictions: comparative analysis, mechanistical insights, and future applications. Drug Metab Rev **39**(1), 61–86 (2007)
26. Rydberg, P., Gloriam, D.E., Olsen, L.: The SMARTCyp cytochrome P450 metabolism prediction server. Bioinformatics **26**(23), 2988–2989 (2010)
27. Rydberg, P., Gloriam, D.E., Zaretzki, J., Breneman, C., Olsen, L.: SMARTCyp: A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism. ACS Med Chem Lett **1**(3), 96–100 (2010)
28. Cruciani, G., Milani, N., Benedetti, P., Lepri, S., Cesarini, L., Baroni, M., Spyrakis, F., Tortorella, S., Mosconi, E., Goracci, L.: From Experiments to a Fast Easy-to-Use Computational Methodology to Predict Human Aldehyde Oxidase Selectivity and Metabolic Reactions. J Med Chem **61**(1), 360–371 (2018)
29. Rudik, A., Dmitriev, A., Lagunin, A., Filimonov, D., Poroikov, V.: SOMP: web server for in silico prediction of sites of metabolism for drug-like compounds. Bioinformatics **31**(12), 2046–2048 (2015)
30. Kirchmair, J., Williamson, M.J., Afzal, A.M., Tyzack, J.D., Choy, A.P., Howlett, A., Rydberg, P., Glen, R.C.: FAst MEtabolizer (FAME): A rapid and accurate predictor of sites of metabolism in multiple species by endogenous enzymes. J Chem Inf Model **53**(11), 2896–2907 (2013)
31. ??cho, M., Stork, C., Mazzolari, A., de Bruyn Kops, C., Pedretti, A., Testa, B., Vistoli, G., Svozil, D., Kirchmair, J.: FAME 3: Predicting the Sites of Metabolism in Synthetic Compounds and Natural Products for Phase 1 and Phase 2 Metabolic Enzymes. J Chem Inf Model **59**(8), 3400–3412 (2019)
32. Li, J., Schneebeli, S.T., Bylund, J., Farid, R., Friesner, R.A.: IDSite: An accurate approach to predict P450-mediated drug metabolism. J Chem Theory Comput **7**(11), 3829–3845 (2011)
33. Campagna-Slater, V., Pottel, J., Therrien, E., Cantin, L.D., Moitessier, N.: Development of a computational tool to rival experts in the prediction of sites of metabolism of xenobiotics by p450s. J Chem Inf Model **52**(9), 2471–2483 (2012)
34. Oh, W.S., Kim, D.N., Jung, J., Cho, K.H., No, K.T.: New combined model for the prediction of regioselectivity in cytochrome P450/3A4 mediated metabolism. J Chem Inf Model **48**(3), 591–601 (2008)
35. Dang, N.L., Hughes, T.B., Miller, G.P., Swamidass, S.J.: Computationally Assessing the Bioactivation of Drugs by N-Dealkylation. Chem Res Toxicol **31**(2), 68–80 (2018)
36. Delannee, V., Langouet, S., Siegel, A., Theret, N.: In silico prediction of Heterocyclic Aromatic Amines metabolism susceptible to form DNA adducts in humans. Toxicol Lett **300**, 18–30 (2019)
37. Favre, H.A., Powell, W.H.: Nomenclature of Organic Chemistry, pp. 001–1568. The Royal Society of Chemistry, ??? (2014). doi:10.1039/9781849733069. http://dx.doi.org/10.1039/9781849733069
38. Inc., D.: Daylight Theory: SMIRKS - A Reaction Transform Language. https://www.daylight.com/dayhtml/doc/theory/theory.smirks.html Accessed (accessed: 24.01.2021)
39. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., Zaslavsky, L., Zhang, J., Bolton, E.E.: PubChem 2019 update: improved access to chemical data. Nucleic Acids Res **47**(D1), 1102–1109 (2019)
40. ChemAxon.com: Marvin — ChemAxon. https://chemaxon.com/products/marvin Accessed (accessed: 24.01.2021)
41. Jensen, F.V., Nielsen, T.D.: Bayesian Networks and Decision Graphs, 2nd edn. Springer, ??? (2007)
42. Smith, P.F., Smith, A., Miners, J., McNeil, J., Proudfoot, A.: Safety Aspects of Dietary Caffeine – Report from the Expert Working Group. Australia New Zealand Food Authority, 20–3 (2000)
43. Thorn, C.F., Aklillu, E., Klein, T.E., Altman, R.B.: PharmGKB summary: very important pharmacogene information for CYP1A2. Pharmacogenet Genomics **22**(1), 73–77 (2012)
44. Perera, V., Gross, A.S., McLachlan, A.J.: Measurement of CYP1A2 activity: a focus on caffeine as a probe. Curr Drug Metab **13**(5), 667–678 (2012)
45. Cornelis, M.C., Kacprowski, T., Menni, C., Gustafsson, S., Pivin, E., Adamski, J., Artati, A., Eap, C.B., Ehret, G., Friedrich, N., Ganna, A., Guessous, I., Homuth, G., Lind, L., Magnusson, P.K., Mangino, M., Pedersen, N.L., Pietzner, M., Suhre, K., V?lzke, H., Bochud, M., Spector, T.D., Grabe, H.J., Ingelsson, E.: Genome-wide association study of caffeine metabolites provides new insights to caffeine metabolism and dietary caffeine-consumption behavior. Hum Mol Genet **25**(24), 5472–5482 (2016)
46. Ngueta, G.: Caffeine and caffeine metabolites in relation to hypertension in U.S. adults. Eur J Clin Nutr **74**(1), 77–86 (2020)
47. McLendon, R., Friedman, A., Bigner, D., Van Meir, E.G., Brat, D.J., Mastrogianakis, G.M., Olson, J.J., Mikkelsen, T., Lehman, N., Aldape, K., Yung, W.K., Bogler, O., Weinstein, J.N., VandenBerg, S., Berger, M., Prados, M., Muzny, D., Morgan, M., Scherer, S., Sabo, A., Nazareth, L., Lewis, L., Hall, O., Zhu, Y., Ren, Y., Alvi, O., Yao, J., Hawes, A., Jhangiani, S., Fowler, G., San Lucas, A., Kovar, C., Cree, A., Dinh, H., Santibanez, J., Joshi, V., Gonzalez-Garay, M.L., Miller, C.A., Milosavljevic, A., Donehower, L., Wheeler, D.A., Gibbs, R.A., Cibulskis, K., Sougnez, C., Fennell, T., Mahan, S., Wilkinson, J., Ziaugra, L., Onofrio, R., Bloom, T., Nicol, R., Ardlie, K., Baldwin, J., Gabriel, S., Lander, E.S., Ding, L., Fulton, R.S., McLellan, M.D., Wallis, J., Larson, D.E., Shi, X., Abbott, R., Fulton, L., Chen, K., Koboldt, D.C., Wendl, M.C., Meyer, R., Tang, Y., Lin, L., Osborne, J.R., Dunford-Shore, B.H., Miner, T.L., Delehaunty, K., Markovic, C., Swift, G., Courtney, W., Pohl, C., Abbott, S., Hawkins, A., Leong, S., Haipek, C., Schmidt, H., Wiechert, M., Vickery, T., Scott, S.,

Dooling, D.J., Chinwalla, A., Weinstock, G.M., Mardis, E.R., Wilson, R.K., Getz, G., Winckler, W., Verhaak, R.G., Lawrence, M.S., O'Kelly, M., Robinson, J., Alexe, G., Beroukhim, R., Carter, S., Chiang, D., Gould, J., Gupta, S., Korn, J., Mermel, C., Mesirov, J., Monti, S., Nguyen, H., Parkin, M., Reich, M., Stransky, N., Weir, B.A., Garraway, L., Golub, T., Meyerson, M., Chin, L., Protopopov, A., Zhang, J., Perna, I., Aronson, S., Sathiamoorthy, N., Ren, G., Yao, J., Wiedemeyer, W.R., Kim, H., Kong, S.W., Xiao, Y., Kohane, I.S., Seidman, J., Park, P.J., Kucherlapati, R., Laird, P.W., Cope, L., Herman, J.G., Weisenberger, D.J., Pan, F., Van den Berg, D., Van Neste, L., Yi, J.M., Schuebel, K.E., Baylin, S.B., Absher, D.M., Li, J.Z., Southwick, A., Brady, S., Aggarwal, A., Chung, T., Sherlock, G., Brooks, J.D., Myers, R.M., Spellman, P.T., Purdom, E., Jakkula, L.R., Lapuk, A.V., Marr, H., Dorton, S., Choi, Y.G., Han, J., Ray, A., Wang, V., Durinck, S., Robinson, M., Wang, N.J., Vranizan, K., Peng, V., Van Name, E., Fontenay, G.V., Ngai, J., Conboy, J.G., Parvin, B., Feiler, H.S., Speed, T.P., Gray, J.W., Brennan, C., Socci, N.D., Olshen, A., Taylor, B.S., Lash, A., Schultz, N., Reva, B., Antipin, Y., Stukalov, A., Gross, B., Cerami, E., Wang, W.Q., Qin, L.X., Seshan, V.E., Villafania, L., Cavatore, M., Borsu, L., Viale, A., Gerald, W., Sander, C., Ladanyi, M., Perou, C.M., Hayes, D.N., Topal, M.D., Hoadley, K.A., Qi, Y., Balu, S., Shi, Y., Wu, J., Penny, R., Bittner, M., Shelton, T., Lenkiewicz, E., Morris, S., Beasley, D., Sanders, S., Kahn, A., Sfeir, R., Chen, J., Nassau, D., Feng, L., Hickey, E., Barker, A., Gerhard, D.S., Vockley, J., Compton, C., Vaught, J., Fielding, P., Ferguson, M.L., Schaefer, C., Zhang, J., Madhavan, S., Buetow, K.H., Collins, F., Good, P., Guyer, M., Ozenberger, B., Peterson, J., Thomson, E.: Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature **455**(7216), 1061–1068 (2008)

48. authors listed, N.: The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science **369**(6509), 1318–1330 (2020)

Additional Files

Additional file 1 - Dictionary of SMIRKS rules
The file provides a manually constructed catalogue of SMIRK rules required to map SMIRKS rules to enzyme family labels by taking into account the rank of the reaction in the pipeline for building enriched maps of metabolism.

Additional file 2 - Description of thirty HAAs and Caffeine structure used as pipeline input
This pdf file describes the SMILES formula and 2D-structure of each of the thirty HAAs and the Caffeine studied in the paper.

Additional file 3 - Description of six HAA maps of metabolism
The file provides a detailed description of metabolites for the six HAAs filtered maps of metabolism built in the paper. For each metabolite of each map, the file provides the identifier of the metabolite, its SMILES formula, its production probability score, its reactivity to DNA and the score of XenoSite Reactivity.

BIBLIOGRAPHIE

[1] H. U. Aeschbacher et R. J. Turesky. « Mammalian cell mutagenicity and metabolism of heterocyclic aromatic amines ». In : *Mutat Res* 259.3-4 (1991), p. 235-250.

[2] L. Afzelius et al. « State-of-the-art tools for computational site of metabolism predictions : comparative analysis, mechanistical insights, and future applications ». In : *Drug Metab Rev* 39.1 (2007), p. 61-86.

[3] L. E. Armstrong et G. L. Guo. « Understanding Environmental Contaminants' Direct Effects on Non-alcoholic Fatty Liver Disease Progression ». In : *Curr Environ Health Rep* 6.3 (sept. 2019), p. 95-104.

[4] R. Bataller et D. A. Brenner. « Liver fibrosis ». In : *J Clin Invest* 115.2 (fév. 2005), p. 209-218.

[5] BayesServer.com. *Asia Bayesian network | Live demo*. (accessed : 28.01.2021). url : https://www.bayesserver.com/examples/networks/asia.

[6] M. Bellamri et al. « Metabolism of the Tobacco Carcinogen 2-Amino-9H-pyrido[2,3-b]indole (AÎ±C) in Primary Human Hepatocytes ». In : *Chem Res Toxicol* 30.2 (fév. 2017), p. 657-668.

[7] Andrej Bugrim, Tatiana Nikolskaya et Yuri Nikolsky. « Early prediction of drug metabolism and toxicity : systems biology approach and modeling ». In : *Drug Discovery Today* 9.3 (2004), p. 127-135. issn : 1359-6446. doi : https://doi.org/10.1016/S1359-6446(03)02971-4. url : http://www.sciencedirect.com/science/article/pii/S1359644603029714.

[8] T. Cai, L. Yao et R. J. Turesky. « Bioactivation of Heterocyclic Aromatic Amines by UDP Glucuronosyltransferases ». In : *Chem Res Toxicol* 29.5 (mai 2016), p. 879-891.

[9] V. Campagna-Slater et al. « Development of a computational tool to rival experts in the prediction of sites of metabolism of xenobiotics by p450s ». In : *J Chem Inf Model* 52.9 (sept. 2012), p. 2471-2483.

[10] L. Carlsson et al. « Use of historic metabolic biotransformation data as a means of anticipating metabolic sites using MetaPrint2D and Bioclipse ». In : *BMC Bioinformatics* 11 (juill. 2010), p. 362.

[11] ChemAxon.com. *Marvin | ChemAxon*. (accessed : 24.01.2021). url : https://chemaxon.com/products/marvin.

[12] M. Chevereau et al. « Role of human sulfotransferase 1A1 and N-acetyltransferase 2 in the metabolic activation of 16 heterocyclic amines and related heterocyclics to genotoxicants in recombinant V79 cells ». In : *Arch Toxicol* 91.9 (sept. 2017), p. 3175-3184.

[13] M. C. Cornelis et al. « Genome-wide association study of caffeine metabolites provides new insights to caffeine metabolism and dietary caffeine-consumption behavior ». In : *Hum Mol Genet* 25.24 (déc. 2016), p. 5472-5482.

[14] G. Cruciani et al. « From Experiments to a Fast Easy-to-Use Computational Methodology to Predict Human Aldehyde Oxidase Selectivity and Metabolic Reactions ». In : *J Med Chem* 61.1 (jan. 2018), p. 360-371.

[15] G. Cruciani et al. « MetaSite : understanding metabolism in human cytochromes from the perspective of the chemist ». In : *J Med Chem* 48.22 (nov. 2005), p. 6970-6979.

[16] N. L. Dang et al. « A simple model predicts UGT-mediated metabolism ». In : *Bioinformatics* 32.20 (oct. 2016), p. 3183-3189.

[17] N. L. Dang et al. « Computationally Assessing the Bioactivation of Drugs by N-Dealkylation ». In : *Chem Res Toxicol* 31.2 (fév. 2018), p. 68-80.

[18] V. Delannee. *EPPIGRAPH (ExPlore PredIcted GRAPH)*. (accessed : 27.01.2021). url : `http://eppigraph.genouest.org/`.

[19] V. Delannee et al. « In silico prediction of Heterocyclic Aromatic Amines metabolism susceptible to form DNA adducts in humans ». In : *Toxicol Lett* 300 (jan. 2019), p. 18-30.

[20] Victorien Delannée. « Intégrer les échelles moléculaires et cellulaires dans l'inférence de réseaux métaboliques : application aux xénobiotiques ». Theses. Université Rennes 1, nov. 2017. url : `https://tel.archives-ouvertes.fr/tel-01659375`.

[21] Y. Djoumbou-Feunang et al. « BioTransformer : a comprehensive computational tool for small molecule metabolism prediction and metabolite identification ». In : *J Cheminform* 11.1 (jan. 2019), p. 2.

[22] Henri A Favre et Warren H Powell. *Nomenclature of Organic Chemistry. IUPAC Recommendations and Preferred Names 2013*. The Royal Society of Chemistry, 2014, P001-1568. isbn : 978-0-85404-182-4. doi : `10.1039/9781849733069`. url : `http://dx.doi.org/10.1039/9781849733069`.

[23] F. Foerster et al. « The immune contexture of hepatocellular carcinoma predicts clinical outcome ». In : *Sci Rep* 8.1 (mars 2018), p. 5351.

[24] B. B. Fredholm et al. « Actions of caffeine in the brain with special reference to factors that contribute to its widespread use ». In : *Pharmacol Rev* 51.1 (mars 1999), p. 83-133.

[25] J. Gao, L. B. Ellis et L. P. Wackett. « The University of Minnesota Biocatalysis/Biodegradation Database : improving public access ». In : *Nucleic Acids Res* 38.Database issue (jan. 2010), p. D488-491.

[26] M. Gibis. « Heterocyclic Aromatic Amines in Cooked Meat Products : Causes, Formation, Occurrence, and Risk Assessment ». In : *Compr Rev Food Sci Food Saf* 15.2 (mars 2016), p. 269-302.

[27] K. L. Hayward et R. A. Weersink. « Improving Medication-Related Outcomes in Chronic Liver Disease ». In : *Hepatol Commun* 4.11 (nov. 2020), p. 1562-1577.

[28] M. Hennemann et al. « CypScore : Quantitative prediction of reactivity toward cytochromes P450 based on semiempirical molecular orbital theory ». In : *ChemMedChem* 4.4 (avr. 2009), p. 657-669.

[29] T. B. Hughes, G. P. Miller et S. J. Swamidass. « Site of reactivity models predict molecular reactivity of diverse chemicals with glutathione ». In : *Chem Res Toxicol* 28.4 (avr. 2015), p. 797-809.

[30] T. B. Hughes et al. « Modeling Reactivity to Biological Macromolecules with a Deep Multitask Network ». In : *ACS Cent Sci* 2.8 (août 2016), p. 529-537.

[31] IARC. *Red Meat and Processed Meat. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans Volume 114*. 2018. isbn : 978-92-832-0152-6. url : https://publications.iarc.fr/Book-And-Report-Series/Iarc-Monographs-On-The-Identification-Of-Carcinogenic-Hazards-To-Humans/Red-Meat-And-Processed-Meat-2018.

[32] ADMET Predictor Simulations Plus Inc. *CYP metabolite prediction | CYP kinetic parameters | CYP inhibition | UGT*. (accessed : 04.02.2021). url : https://www.simulations-plus.com/software/admetpredictor/metabolism/.

[33] N. Jeliazkova et V. Jeliazkov. « AMBIT RESTful web services : an implementation of the OpenTox application programming interface ». In : *J Cheminform* 3 (mai 2011), p. 18.

[34] Finn V. Jensen et Thomas D. Nielsen. *Bayesian Networks and Decision Graphs*. 2nd. Springer Publishing Company, Incorporated, 2007. isbn : 9780387682815.

[35] S. Kim et al. « PubChem 2019 update : improved access to chemical data ». In : *Nucleic Acids Res* 47.D1 (jan. 2019), p. D1102-D1109.

[36] J. Kirchmair et al. « Computational prediction of metabolism : sites, products, SAR, P450 enzyme dynamics, and mechanisms ». In : *J Chem Inf Model* 52.3 (mars 2012), p. 617-648.

[37] J. Kirchmair et al. « FAst MEtabolizer (FAME) : A rapid and accurate predictor of sites of metabolism in multiple species by endogenous enzymes ». In : *J Chem Inf Model* 53.11 (nov. 2013), p. 2896-2907.

[38] J. Kirchmair et al. « Predicting drug metabolism : experiment and/or computation ? » In : *Nat Rev Drug Discov* 14.6 (juin 2015), p. 387-404.

[39] G. Klopman, M. Dimayuga et J. Talafous. « META. 1. A program for the evaluation of metabolic transformation of chemicals ». In : *J Chem Inf Comput Sci* 34.6 (1994), p. 1320-1325.

[40] S. Langouet et al. « Differential metabolism of 2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine in rat and human hepatocytes ». In : *Carcinogenesis* 23.1 (jan. 2002), p. 115-122.

[41] S. Langouet et al. « Metabolism of 2-amino-3,8-dimethylimidazo[4,5-f]quinoxaline in human hepatocytes : 2-amino-3-methylimidazo[4,5-f]quinoxaline-8-carboxylic acid is a major detoxification pathway catalyzed by cytochrome P450 1A2 ». In : *Chem Res Toxicol* 14.2 (fév. 2001), p. 211-221.

[42] S. L. Lauritzen et D. J. Spiegelhalter. « Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems ». In : *Journal of the Royal Statistical Society. Series B (Methodological)* 50.2 (1988), p. 157-224. issn : 00359246. url : http://www.jstor.org/stable/2345762.

[43] J. Li et al. « IDSite : An accurate approach to predict P450-mediated drug metabolism ». In : *J Chem Theory Comput* 7.11 (nov. 2011), p. 3829-3845.

[44] No authors listed. « The GTEx Consortium atlas of genetic regulatory effects across human tissues ». In : *Science* 369.6509 (sept. 2020), p. 1318-1330.

[45] C. A. Marchant, K. A. Briggs et A. Long. « In silico tools for sharing data and knowledge on toxicity and metabolism : derek for windows, meteor, and vitic ». In : *Toxicol Mech Methods* 18.2-3 (2008), p. 177-187.

[46] C. A. Marchant, K. A. Briggs et A. Long. « In silico tools for sharing data and knowledge on toxicity and metabolism : derek for windows, meteor, and vitic ». In : *Toxicol Mech Methods* 18.2-3 (2008), p. 177-187. doi : https://doi.org/10.1002/minf.200900006.

[47] M. K. Matlock, T. B. Hughes et S. J. Swamidass. « XenoSite server : a web-available site of metabolism prediction tool ». In : *Bioinformatics* 31.7 (avr. 2015), p. 1136-1137.

[48] R. McLendon et al. « Comprehensive genomic characterization defines human glioblastoma genes and core pathways ». In : *Nature* 455.7216 (oct. 2008), p. 1061-1068.

[49] O. G. Mekenyan et al. « A systematic approach to simulating metabolism in computational toxicology. I. The TIMES heuristic modelling framework ». In : *Curr Pharm Des* 10.11 (2004), p. 1273-1293.

[50] A. M. Moon, A. G. Singal et E. B. Tapper. « Contemporary Epidemiology of Chronic Liver Disease and Cirrhosis ». In : *Clin Gastroenterol Hepatol* 18.12 (nov. 2020), p. 2650-2666.

[51] G. Nauwelaers et al. « DNA adduct formation of 4-aminobiphenyl and heterocyclic aromatic amines in human hepatocytes ». In : *Chem Res Toxicol* 24.6 (juin 2011), p. 913-925.

[52] G. Nauwelaers et al. « DNA adducts of the tobacco carcinogens 2-amino-9H-pyrido[2,3-b]indole and 4-aminobiphenyl are formed at environmental exposure levels and persist in human hepatocytes ». In : *Chem Res Toxicol* 26.9 (sept. 2013), p. 1367-1377.

[53] G. Ngueta. « Caffeine and caffeine metabolites in relation to hypertension in U.S. adults ». In : *Eur J Clin Nutr* 74.1 (jan. 2020), p. 77-86.

[54] W. S. Oh et al. « New combined model for the prediction of regioselectivity in cytochrome P450/3A4 mediated metabolism ». In : *J Chem Inf Model* 48.3 (mars 2008), p. 591-601.

[55] G. Patlewicz et al. « An evaluation of the implementation of the Cramer classification scheme in the Toxtree software ». In : *SAR QSAR Environ Res* 19.5-6 (2008), p. 495-524.

[56] V. Perera, A. S. Gross et A. J. McLachlan. « Measurement of CYP1A2 activity : a focus on caffeine as a probe ». In : *Curr Drug Metab* 13.5 (juin 2012), p. 667-678.

[57] S. Preissner et al. « SuperCYP : a comprehensive database on Cytochrome P450 enzymes including a tool for analysis of CYP-drug interactions ». In : *Nucleic Acids Res* 38.Database issue (jan. 2010), p. D237-243.

[58] RDKit. *RDKit*. (accessed : 28.01.2021). url : `https://www.rdkit.org/`.

[59] International Agency for Research on Cancer (IARC). *IARC Monographs on the Evaluation of Carcinogenic Risk to Humans, Vol. 56, Some Naturally Occurring Substances : Food Items and Constituents, Heterocyclic Aromatic Amines and Mycotoxins*. T. 294. 3. 1994, p. 341-.

[60] L. Ridder et M. Wagener. « SyGMa : combining expert knowledge and empirical scoring in the prediction of metabolites ». In : *ChemMedChem* 3.5 (mai 2008), p. 821-832.

[61] A. Rudik et al. « SOMP : web server for in silico prediction of sites of metabolism for drug-like compounds ». In : *Bioinformatics* 31.12 (juin 2015), p. 2046-2048.

[62] A. V. Rudik et al. « MetaTox : Web Application for Predicting Structure and Toxicity of Xenobiotics' Metabolites ». In : *J Chem Inf Model* 57.4 (avr. 2017), p. 638-642.

[63] P. Rydberg, D. E. Gloriam et L. Olsen. « The SMARTCyp cytochrome P450 metabolism prediction server ». In : *Bioinformatics* 26.23 (déc. 2010), p. 2988-2989.

[64] P. Rydberg et al. « SMARTCyp : A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism ». In : *ACS Med Chem Lett* 1.3 (juin 2010), p. 96-100.

[65] M. Sicho et al. « FAME 3 : Predicting the Sites of Metabolism in Synthetic Compounds and Natural Products for Phase 1 and Phase 2 Metabolic Enzymes ». In : *J Chem Inf Model* 59.8 (août 2019), p. 3400-3412.

[66] S. B. Singh et al. « A model for predicting likely sites of CYP3A4-mediated metabolism on drug-like molecules ». In : *J Med Chem* 46.8 (avr. 2003), p. 1330-1336.

[67] P. F. Smith et al. « Safety Aspects of Dietary Caffeine – Report from the Expert Working Group ». In : *Australia New Zealand Food Authority* (jan. 2000), p. 20-3.

[68] O. Spjuth et al. « XMetDB : an open access database for xenobiotic metabolism ». In : *J Cheminform* 8 (2016), p. 47.

[69] T. Sugimura et al. « Heterocyclic amines : Mutagens/carcinogens produced during cooking of meat and fish ». In : *Cancer Sci* 95.4 (avr. 2004), p. 290-299.

[70] C. F. Thorn et al. « PharmGKB summary : very important pharmacogene information for CYP1A2 ». In : *Pharmacogenet Genomics* 22.1 (jan. 2012), p. 73-77.

[71] S. P. Tighe et al. « Chronic Liver Disease and Silymarin : A Biochemical and Clinical Review ». In : *J Clin Transl Hepatol* 8.4 (déc. 2020), p. 454-458.

[72] R. J. Turesky. « Interspecies metabolism of heterocyclic aromatic amines and the uncertainties in extrapolation of animal toxicity data for human risk assessment ». In : *Mol Nutr Food Res* 49.2 (fév. 2005), p. 101-117.

[73] R. J. Turesky et L. Le Marchand. « Metabolism and biomarkers of heterocyclic aromatic amines in molecular epidemiology studies : lessons learned from aromatic amines ». In : *Chem Res Toxicol* 24.8 (août 2011), p. 1169-1214.

[74] J. Verma, V. M. Khedkar et E. C. Coutinho. « 3D-QSAR in drug design–a review ». In : *Curr Top Med Chem* 10.1 (2010), p. 95-115.

[75] D. S. Wishart et al. « DrugBank 5.0 : a major update to the DrugBank database for 2018 ». In : *Nucleic Acids Res* 46.D1 (jan. 2018), p. D1074-D1082.

[76] D. S. Wishart et al. « DrugBank : a comprehensive resource for in silico drug discovery and exploration ». In : *Nucleic Acids Res* 34.Database issue (jan. 2006), p. D668-672.

[77] Q. Xu et al. « ADMETNet : The knowledge base of pharmacokinetics and toxicology network ». In : *J Genet Genomics* 44.5 (mai 2017), p. 273-276.

[78] M. Yousofshahi et al. « PROXIMAL : a method for Prediction of Xenobiotic Metabolism ». In : *BMC Syst Biol* 9 (déc. 2015), p. 94.

[79] J. Zaretzki, M. Matlock et S. J. Swamidass. « XenoSite : accurately predicting CYP-mediated sites of metabolism with neural networks ». In : *J Chem Inf Model* 53.12 (déc. 2013), p. 3373-3383.

[80] J. Zaretzki, M. Matlock et S. J. Swamidass. « XenoSite : accurately predicting CYP-mediated sites of metabolism with neural networks ». In : *J Chem Inf Model* 53.12 (déc. 2013), p. 3373-3383.

UNIVERSITÉ DE RENNES 1

Titre : Construction de cartes du métabolisme des xénobiotiques enrichies pour prédire le rôle des enzymes dans la formation des adduits à l'ADN

Mot clés : Prédiction du métabolisme ; AHAs ; Réactivité à l'ADN ; Réseaux Bayésiens

Résumé : Le foie joue un rôle majeur dans l'activation métabolique des contaminants de l'environnement (médicaments, produits chimiques comme les polluants, pesticides, additifs alimentaires...). Parmi les xénobiotiques préoccupants, les amines hétérocycliques aromatiques(AHA) sont classés, comme cancérigènes possibles ou probables (2A ou 2B) par l'IARC, pour lesquels il existe peu d'informations chez l'homme. 30 AHAs ont été identifiés à ce jour mais la bioactivation du métabolisme et la formation d'adduits à l'ADN n'ont été entièrement caractérisées dans le foie humain que pour trois d'entre elles (MeIQx, PhIP, AαC). Nous avons développé une approche de modélisation afin de prédire à la fois le métabolisme (métabolites et réactions), la réactivité de l'ADN et la probabilité de production des métabolites. Notre approche repose sur la construction de cartes du métabolisme enrichies. Nous rassemblons des outils de prédiction des réactions et des métabolites (SyGMa), de prédiction des sites de métabolisme (Way2Drug SOMP, Fame 3), de prédiction de la réactivité de l'ADN (XenoSite Reactivity V1) et le calcul d'un score de probabilité de production basé sur les propriétés des réseaux bayésiens. Ce pipeline de prédiction a été évalué et validé à l'aide de la caféine, puis appliqué à six AHAs. Les principaux résultats montrent que notre approche permet de prédire le métabolisme des xénobiotiques et que le score de probabilité de production a différentes propriétés qui peuvent conduire à la filtration de la carte du métabolisme ou à la détermination des profils enzymatiques associés à la maximisation de la formation des adduits à l'ADN. Cette approche de toxicologie prédictive ouvre des perspectives pour estimer la génotoxicité de divers xénobiotiques dans des conditions normales ou physiopathologiques.

Title: Constructing xenobiotic enriched maps of metabolism to predict the role of enzymes in DNA adduct formation

Keywords: Metabolism prediction ; HAA ; DNA reactivity ; Bayesian Networks

Abstract: The liver plays a major role in the metabolic activation of xenobiotics (drugs, chemicals such as pollutants, pesticides, food additives, etc.). Among environmental contaminants of concern, heterocyclic aromatic amines (HAAs) are xenobiotics classified as possible or probable carcinogens (2A or 2B) by IARC, for which low information exists in humans. 30 AHAs have been identified to date, but the bioactivation pathways, metabolites and DNA adducts have been fully characterised in the human liver for only three of them (MeIQx, PhIP, AαC). We have developed a modelling approach to predict both metabolism (metabolites and reactions), DNA reactivity and the production probability of metabolite. Our approach is based on the construction of enriched metabolism maps. We bring together tools for predicting reactions and metabolites (SyGMa), predicting metabolism sites (Way2Drug SOMP, Fame 3), predicting DNA reactivity (XenoSite Reactivity V1) and calculating a production probability score based on the properties of Bayesian networks. This prediction pipeline was evaluated and validated using caffeine and then applied to six AHAs. Main results show that our approach allows us to predict the metabolism of xenobiotics and that the production probability score has different properties that can lead to the filtration of the metabolism map or to the determination of the enzymatic profiles associated with maximising the formation of DNA adducts. This predictive toxicology approach opens up prospects for estimating the genotoxicity of various environmental contaminants in normal or pathophysiological situations.