université PARIS-SACLAY

Simulation-Based Inference for the study of Higgs Self-Coupling in the $HH \to b\bar{b}\tau^+\tau^-$ decay channel with the ATLAS Experiment and Machine Learning Benchmarks for Uncertainty Quantification

Inférence basée sur la simulation pour l'étude de l'auto-couplage du boson de Higgs dans le canal de désintégration ($HH \to b\bar{b}\tau^+\tau^-$) avec l'expérience ATLAS et benchmarks d'apprentissage automatique pour la quantification des incertitudes.

École doctorale n°576 : Particules Hadrons Énergie et Noyau : Instrumentation, Imagerie, Cosmos et Simulation (PHENIICS)
Spécialité de doctorat: Physique des particules
Graduate School : Physique
Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **IJCLab** (Université Paris-Saclay, CNRS), sous la direction de **David ROUSSEAU**, Directeur de recherche

Thèse soutenue à Paris-Saclay, le 16 octobre 2026, par

Ragansu CHAKKAPPAI

Composition du jury

Membres du jury avec voix délibérative

Sandra Kortner Principal Investigator, Max Planck Institute for Physics	Présidente & Rapportrice
Robert Schöfbeck Privatdozent, Marietta Blau Institute for Particle Physics	Rapporteur & Examinateur
Julien Donini Professeur, Université Clermont Auvergne & LPC Clermont Auvergne	Examinateur
Lukas Alexander Heinrich Professeur, Technical University Munich	Examinateur
Anja Butter Chargé de recherche, Sorbonne Université & LPNHE Paris	Examinatrice
Thomas Strebler Chargé de recherche, Aix-Marseille Université & CPPM Marseille	Examinateur

THESE DE DOCTORAT

NNT : 2026UPASP114

université PARIS-SACLAY

ÉCOLE DOCTORALE
Particules, hadrons, énergie et noyau : instrumentation, imagerie, cosmos et simulation (PHENIICS)

Titre: Inférence basée sur la simulation pour l'étude de l'auto-couplage du boson de Higgs dans le canal de désintégration ($HH \to b\bar{b}\tau^+\tau^-$) avec l'expérience ATLAS et des benchmarks d'apprentissage automatique pour la quantification des incertitudes.

Mots clés: Physique des Hautes Énergies (PHE), Grand collisionneur de hadrons (LHC), Intelligence Artificielle (IA), Boson de Higgs, Inférence Basée sur Simulation, Quantification des Incertitudes (QI).

Résumé: La découverte du boson de Higgs en 2012 et la mesure de sa masse qui s'en est suivie constituent des étapes importantes pour le mécanisme BEH, qui confère une masse aux particules fondamentales. Cependant, la masse du boson de Higgs n'est que l'un des deux paramètres libres du potentiel de Higgs ; l'autre est la constante d'auto-couplage de Higgs, lambda. Une étude du processus Di-Higgs pourrait permettre une mesure directe du paramètre lambda, et ainsi fournir une mesure plus précise du potentiel de Higgs. La mesure du paramètre lambda peut être réalisée en étudiant la production de Di-Higgs, car la section efficace de cette production dépend du couplage linéaire du boson de Higgs. L'objet de cette thèse est d'étudier la production de Di-Higgs dans le canal $HH \to b\bar{b}\tau^+\tau^-$. La production de Di-Higgs implique une interférence quantique entre un diagramme en boîte, indépendant du paramètre lambda, et un diagramme en arbre, dépendant du paramètre lambda. Cette interférence destructive entraîne une perte d'information lorsqu'elle est analysée à l'aide de méthodes traditionnelles fondées sur des histogrammes.

L'inférence basée sur la simulation (SBI) s'est imposée comme un cadre solide d'apprentissage automatique pour l'estimation de paramètres à haute dimensionnalité, permettant de contourner la perte d'information associée aux méthodes traditionnelles fondées sur les histogrammes. Suite à la récente application réussie de la SBI dans l'analyse ATLAS du boson de Higgs hors resonance[1, 2], nous présentons une étude de validation de principe qui applique ces techniques au canal $HH \to b\bar{b}\tau^+\tau^-$. Ce travail vise à quantifier les gains de sensibilité dans l'estimation de κ_λ par rapport à l'analyse traditionnelle basée sur les histogrammes[3], qui utilisait le score du classificateur GNN comme discriminant. Les résultats préliminaires et la validation de divers estimateurs de rapport de densité mettent en évidence le potentiel de SBI pour l'analyse des paires de Higgs.

L'estimation des paramètres en haute dimensionnalité est également nécessaire pour réduire les incertitudes systématiques. Diverses méthodes ont été proposées au cours de la dernière décennie pour estimer avec précision l'incertitude systématique. Malheureusement, il est difficile de comparer ces méthodes, car chacune utilise des ensembles de données, des objectifs et des a priori différents. Un benchmark équitable est nécessaire pour comparer correctement ces méthodes dans le cadre de la quantification des incertitudes. Pour y remédier, en s'appuyant sur les fondements du défi Kaggle HiggsML, le projet Fair Universe a organisé le Fair Universe HiggsML Uncertainty Challenge[4], qui s'est déroulé de septembre 2024 au 14 mars 2025 et a été accepté comme compétition officielle de NeurIPS 2024. Dans cette thèse, nous aborderons en détail l'ensemble de données constitué, les métriques utilisées, les solutions gagnantes et les stratégies d'analyse employées pour finaliser les résultats. L'apprentissage automatique fait désormais partie intégrante de l'analyse en physique des hautes énergies. Cette thèse vise à étudier comment évaluer et comparer efficacement ces modèles afin de progresser vers de nouvelles solutions, telles que l'inférence par simulation.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Particules, hadrons, énergie et noyau : instrumentation, imagerie, cosmos et simulation (PHENIICS)

Title: Simulation-based inference for the study of the Higgs self-coupling in the ($HH \to b\bar{b}\tau^+\tau^-$) decay channel with the ATLAS experiment and machine learning benchmarks for uncertainty quantification.

Keywords: High Energy Physics (HEP), Large Hadron Collider (LHC), Artificial Intelligence (AI), Higgs Boson, Simulation-based Inference(SBI), Uncertainty Quantification (UQ).

Abstract:

The 2012 Higgs discovery and the subsequent mass measurement are important milestones for the BEH mechanism, which endows fundamental particles with mass. However, the Higgs mass is only one of the two free parameters in the Higgs potential; the other is the Higgs self-coupling constant, lambda. A study of the Di-Higgs process could provide a direct measurement of the lambda parameter, thereby yielding a more precise measurement of the Higgs potential. The measurement of the lambda parameter can be achieved by studying the production of Di-Higgs, as the cross-section for Di-Higgs production involves the Higgs linear coupling. The focus of this thesis is to study Di-Higgs production in the $HH \to b\bar{b}\tau^+\tau^-$ channel. The Di-Higgs production involves quantum interference between a box diagram, which is independent of the lambda parameter, and a tree diagram, which is dependent on the lambda parameter. This destructive interference results in information loss when analysed using traditional histogram-based methods.

Simulation-Based Inference (SBI) has emerged as a robust machine learning framework for high-dimensional parameter estimation, circumventing the information loss associated with traditional histogram-based methods. Following the recent successful application of SBI in the ATLAS off-shell Higgs analysis[1, 2], we present a proof of concept study applying these techniques to the $HH \to b\bar{b}\tau^+\tau^-$ channel. This work aims to quantify the gains in sensitivity in κ_λ estimation relative to traditional histogram-based analysis[3], which utilised the GNN classifier score as discriminant. Preliminary results and validation of various density ratio estimators, highlighting the potential of SBI for Di-Higgs Analysis.

High-dimensional parameter estimation is also necessary to reduce systematic uncertainties. Various methods have been proposed over the last decade to accurately estimate systematic uncertainty. Unfortunately, comparing these methods is difficult because each has different datasets, objectives, and priors. A fair benchmark is required to properly compare such methods in the context of uncertainty quantification. To address this, building on the foundation of the Kaggle HiggsML challenge, the Fair Universe project organised the Fair Universe HiggsML Uncertainty Challenge[4], which ran from Sep 2024 to 14th March 2025 and was accepted as an official NeurIPS2024 competition. In this thesis, we will discuss in detail the built dataset, the metrics used, the winning solutions, and the analysis strategies employed to finalise the results. Machine learning has become an integral part of HEP analysis. This thesis aims to study how we can effectively benchmark and compare such models to progress toward newer solutions like Simulation-Based Inference.

Acknowledgement

For a young boy from a small town in southern India to reach here, it is only thanks to the help and support of many, many very kind people.

First and foremost, I would like to thank the jury, Julien Donini, Sandra Kortner, Robert Schöfbeck, Lukas Heinrich, Anja Butter and Thomas Strebler, most of whom had to travel from faraway cities, if not countries, to be present here. I had the fortune of meeting each of them at various conferences and workshops and learning from them well before they were part of the jury. I thank them for the many discussions we had, both before and after the thesis submission.

I have waited for this moment for the last 3 years to properly express my gratitude to my thesis director, David Rousseau. About 5 years ago, David decided to give a chance to an Indian kid who barely knew Python and thought jupyter was a planet. Nurtured him, protected him, and guided him to what he is today. David had a unique way of guiding me to the right answer and pushing me to greater challenges.

The 3 years have given me the opportunity to work with very impressive people. Starting from the FAIR Universe, I would like to thank Isabelle Guyon, who was both a mentor and a guiding light for me and for the challenge during its early days, when we were going from ideas on a whiteboard to actual code. Without her constant guidance in each step, I don't think we could have done what we did.

On that note, I like to thank the "cool captain of the FAIR Universe Project", Wahid Bhimji. He had the unique ability to take the pressure away from me when I was stuck in a difficult situation. I am very fortunate to have worked under his leadership. Alongside the FAIR Universe team as a whole, there are a few special people whom I want to thank for always having my back: Ihsan, Sasha, Elham, Chris, Po-wen, Yulei and Steven.

The ACTS QP project was quite scary at the start because of the sheer amount of C++ code it contained. But thankfully, I was constantly supported by the amazing members of the ACTS development team. I especially thank Carlo Varni, Noemi Calace, Hadrien Grasland, Andreas Stefl, Andreas Salzburger, and of course, Corentin Allaire. Corentin is yet another person for whom words are not enough to express my gratitude. Our relationship started with me constantly bugging him about every error and bug in the code, and slowly grew into something of a brotherhood. The daily coffee breaks and our discussions of everything from the Roman Empire to AGI made me feel at home. I will never forget the support and motivation you gave when I was at my lowest.

The SBI project, though, was a 2-man team with David and me. I would like to take this opportunity to thank the nSBI4ATLAS group and specifically the 2 people

who helped me understand SBI, Jay and Una. The debates and discussions we had helped shape my understanding of the subject.

Next, I would also like to specially thank Marine Lemorvan, whose input was paramount to our understanding of the Tabular foundation models. On that note, I would like to take this opportunity to thank my interns, Remy, Alexandre, Guilhem, Abdullah, and Ahmed, from whom I have learned a lot.

I often joke that this lab was my second home, and it's true. It's true, thanks to the amazing people here. Starting from Arnaud, my academic big brother, on whom I relied for support and motivation. Matthew, my office buddy for a year, Marco and Salah, my comrades at arms, we fought shoulder to shoulder in the battle of PhD. Oliver, Ines, Omar, Laurine and Justine; the best juniors one could ask for and my seniors, Vera, Giacomo, Mouran, Zuchen, Zhoran, and Tuan. Thank you all for making my lab life extra fun.

I would like to thank the ATLAS permanent staff. Some of them have taught me, many of them have guided me, and each and every one of them has inspired me to be a better physicist.

Life is not just about work (ironic, coming from me), and my friends made sure of that: Stavros, Chaithaiya, Sruthy, Laury, Mehdi, Prasanna, Srijitha, Trisha, Purbhitha, Joshojith, Allenchris, Navin and Rhea. (I am proud, that's a long list)

The three people who are absolutely paramount to my success: My Father, My mother and my dearest brother. 20 years ago, my dad started pursuing his PhD, and since then, he has instilled in me the inspiration that has led to what transpired today. He taught me how to think critically and to ask questions. Thank you, Appa, for building me into what I am today. My mother is the centre of my stability. The hour-long calls we make every day are often the reason I am able to push forward with my PhD. Without constant support and prayers, I would not have lasted through this process.

Six years ago, I almost lost hope in physics. The person who brought me back to reality pushed me forward and reminded me never to stop believing in myself and never to give up, My dearest brother "*Kunjetten*". He is the reason I chose Paris-Saclay for my master's, and he is the reason I have always kept going.

Résumé en français

La découverte du **boson de Higgs** en 2012 au **Grand collisionneur de hadrons** (LHC) par les collaborations **ATLAS** et **CMS** a permis d'établir le mécanisme BEH, qui confère une masse aux particules fondamentales. Depuis lors, de nombreuses mesures réalisées par ATLAS et CMS ont permis de déterminer les différentes constantes de couplage du boson de Higgs, consolidant ainsi la validité du Modèle standard de la physique des particules. La mesure de couplages plus faibles ou de processus plus rares a nécessité le recours à l'apprentissage automatique (ML) afin de définir de meilleures régions d'intérêt (ROI) et de meilleurs discriminants. Afin de réunir des spécialistes de physique et d'apprentissage automatique du monde entier et de favoriser le développement d'algorithmes d'apprentissage automatique plus sophistiqués, le **Kaggle HiggsML Challenge** 2014 a été lancé. Ce concours a été le premier du genre à évaluer les performances d'algorithmes d'apprentissage automatique sur un problème de classification utilisant les données ouvertes d'ATLAS dans le canal $pp \to \tau^+\tau^-$. Les participants ont reçu un ensemble de données sous forme de tableau contenant les quadri-impulsions des particules. Au cours de la dernière décennie, diverses méthodes d'intelligence artificielle tenant compte des incertitudes ont été publiées, et un nouveau benchmark est nécessaire pour traiter correctement cette question. Le chapter 4 répond à ce besoin avec un nouveau défi d'apprentissage automatique axé sur la quantification et/ou la réduction des incertitudes systématiques. Dans le cadre de ce nouveau concours, les participants ont été évalués à la fois sur leur capacité à déterminer avec précision la section efficace correcte et sur leur capacité à fournir des intervalles d'incertitude corrects et bien calibrés. Le concours s'est déroulé de septembre 2024 au 14 mars 2025 et a été reconnu comme un concours officiel de **NeurIPS 2024**.

Cependant, la masse du boson de Higgs n'est que l'un des deux paramètres libres du potentiel de Higgs ; l'autre est la constante d'auto-couplage du Higgs, λ. Le paramètre λ peut être mesuré directement à partir de la production de paires de bosons de Higgs. En juin 2026, la collaboration ATLAS a présenté les résultats de la deuxième campagne de prise de données (Run 2) et d'une partie de la troisième campagne (Run 3) concernant la production de paires de bosons de Higgs dans le canal $\tau^+\tau^-$. Elle a utilisé une méthode d'analyse par bins qui s'appuyait sur plusieurs classificateurs fondés sur des réseaux de neurones graphiques pour définir les bins dans la région du signal. Cela a permis d'établir les limites les plus précises concernant l'intensité du signal μ et la constante d'auto-couplage κ_λ dans le canal $b\bar{b}\tau^+\tau^-$. Cependant, la production de paires de bosons de Higgs dans le canal de fusion de gluons (ggF) implique une interférence quantique entre le diagramme triangulaire

(avec la constante d'auto-couplage) et le diagramme en boîte (sans constante d'auto-couplage). Cette interférence destructive réduit la puissance statistique lorsque l'on utilise une méthode d'inférence par bins. C'est là qu'intervient l'inférence fondée sur la simulation (SBI). L'utilisation de la SBI en 2024 dans l'analyse ATLAS du Higgs hors résonance a démontré avec succès l'intérêt de recourir à la SBI pour les problèmes impliquant une interférence destructive importante. La production de paires de bosons de Higgs présente également des défis similaires. Dans le chapter 7, nous présentons l'étude des techniques de SBI pour le canal $HH \to b\bar{b}\tau^+\tau^-$. Ce travail vise à comparer divers modèles d'apprentissage automatique et à quantifier les gains potentiels en termes de sensibilité dans l'estimation de κ_λ par rapport à l'analyse traditionnelle par histogrammes. Les résultats préliminaires et la validation des estimateurs de rapports de densité seront abordés en détail dans ce chapitre. Enfin, nous discuterons des prochaines étapes nécessaires et de l'avenir de l'inférence fondée sur la simulation dans l'analyse $HH \to b\bar{b}\tau^+\tau^-$.

1 Contexte théorique

Le Modèle standard est la meilleure la théorie la mieux vérifiée expérimentallement dont nous disposons pour modéliser les interactions subatomiques. Ce modèle classe les particules en fonction de leur spin intrinsèque : les **fermions** sont des particules de matière dotées d'un spin demi-entier, tandis que les **bosons** ont un spin entier. Les fermions eux-mêmes sont regroupés en **quarks**, qui interagissent avec la force forte — par exemple, les quarks up et down présents dans les protons et les neutrons —, et en **leptons**, les fermions de couleur neutre. Les meilleurs exemples en sont les **électrons** et les **neutrinos**. Les électrons ont une charge de -1, tandis que les neutrinos sont neutres. Les fermions sont classés en **générations**. Chaque génération comprend un quark de type « up », un quark de type « down », un lepton et un neutrino. Il existe trois générations, qui présentent toutes des propriétés similaires à l'exception de la masse des particules. Le Modèle standard n'explique pas la hiérarchie des masses.

Les bosons sont des vecteurs de force et contribuent à la médiation des interactions entre les particules. Il existe quatre forces fondamentales dans l'univers. Le Modèle standard en explique trois :

1. **La force électromagnétique**, propagée par le **photon** γ, explique l'interaction des particules chargées électriquement (tous les fermions à l'exception des neutrinos).

2. **La force nucléaire forte** (chromodynamique quantique), véhiculée par les **gluons**, explique la liaison entre les quarks.

3. **La force nucléaire faible**, véhiculée par les bosons Z et $W^\pm$, explique le changement de saveur chez des fermions.

Dans les années 1960, les physiciens ont constaté qu'ils ne pouvaient pas inclure de terme de masse dans le lagrangien des bosons, car cela aurait rompu l'invariance

de jauge, qui est fondamentale pour la théorie. Cela ne posait pas de problème pour les photons et les gluons, qui étaient sans masse, mais pour les bosons Z et $W^{\pm}$, dont on savait qu'ils étaient massifs en raison de la courte longueur d'interaction de la force faible. Pour remédier à cela, le mécanisme de Brout-Englert-Higgs (BEH) a été proposé par Robert Brout et François Englert, Peter Higgs et Gerald Guralnik, Carl R. Hagen et Tom Kibble afin d'introduire des termes de masse pour les bosons massifs sans rompre les invariances de Lorentz et de jauge. L'idée centrale de cette théorie était de proposer un boson de Goldstone, un champ scalaire complexe sans masse qui romprait spontanément la symétrie.

Après la rupture de symétrie, il reste trois bosons massifs du secteur faible (Z, W^+ et W^-) et un seul boson scalaire massif appelé boson de Higgs, qui propage le champ de Higgs. Le boson de Higgs est également responsable du mécanisme de production de masse chez les fermions par le biais du couplage de Yukawa. Ainsi, plus le couplage au boson de Higgs est important, plus le couplage est fort. Voici les principaux modes de production des bosons de Higgs.

- **Fusion de gluons (ggF) :** La ggF présente la section efficace la plus élevée parmi tous les modes de production du boson de Higgs. Le diagramme de Feynman correspondant est présenté dans Figure 2 (a).
- **Fusion de bosons vecteurs (VBF) :** Dans ce processus, les quarks des protons émettent des bosons vecteurs ($W^{\pm}$, Z^0) qui fusionnent pour former un boson de Higgs. La signature de ce canal est constituée de deux jets vers

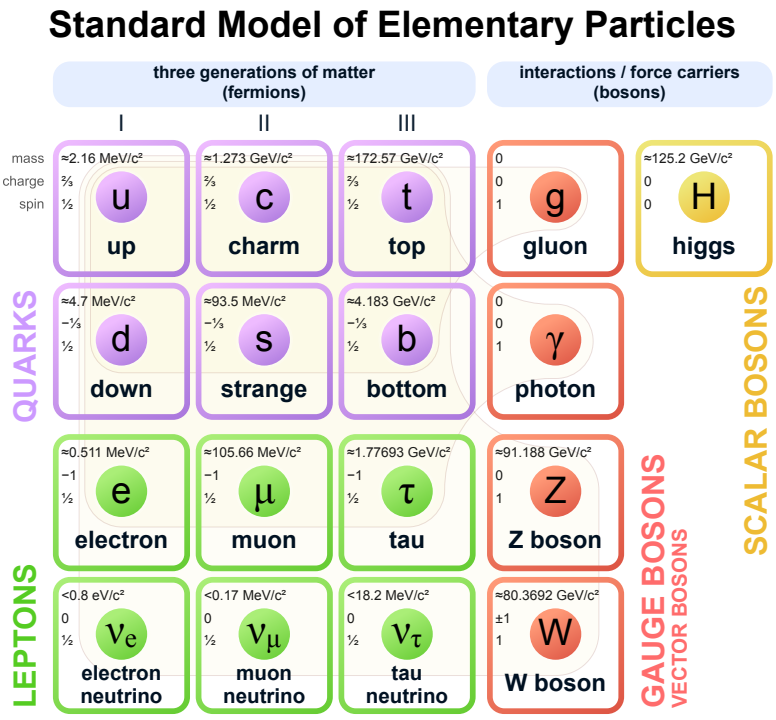


Figure 1: Résumé du Modèle standard des particules élémentaires [5]

l'avant. Le diagramme de Feynman correspondant est présenté dans Figure 2 (b).

- **Production associée de bosons vecteurs (VH) (Higgsstrahlung) :** Ici, une paire de quarks crée d'abord un boson vecteur, qui émet ensuite un boson de Higgs. Le diagramme de Feynman correspondant est présenté dans Figure 2 (c).
- **Production associée de quarks top (bottom) (ttH + tH, bbH + bH) :** À l'instar du processus ggF, ce processus tire également parti du couplage fort et du couplage important du boson de Higgs aux quarks top, mais contrairement au ggF, il implique la création de quarks top (bottom) ; il est donc supprimé par l'espace de phase. Le diagramme de Feynman correspondant est présenté dans Figure 2 (d, e).

Les principaux canaux de désintégration sont $H \to W/ZW/Z$ Figure 2(f), $H \to \gamma\gamma$ Figure 2(f), $H \to b\bar{b}$ Figure 2(h), et $H \to \tau^+\tau^-$ Figure 2(i).

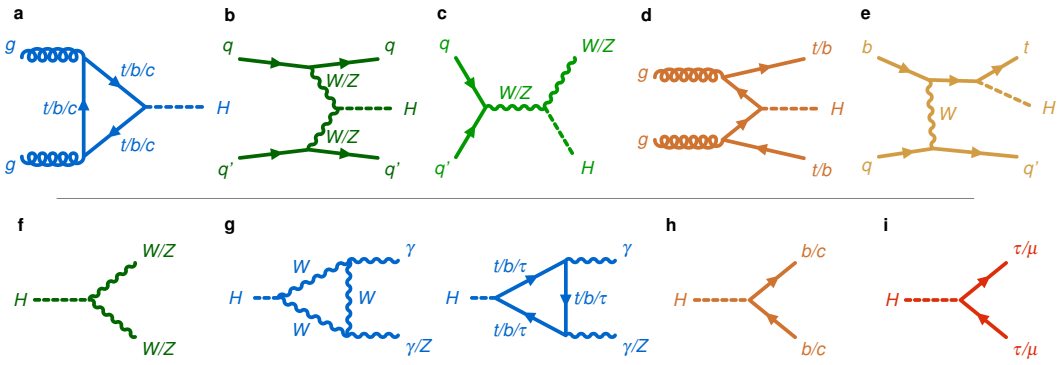


Figure 2: Diagrammes de Feynman illustrant les différents modes de production et canaux de désintégration du boson de Higgs. Production : (a) fusion de gluons (ggF), (b) fusion de bosons vecteurs (VBF), (c) production associée de bosons vecteurs (VH), (d) production associée de quarks top (ttH), (e) production associée de quarks bottom (bbH). Désintégration : (f) $H \to W/ZW/Z$ (g) $H \to \gamma\gamma$ (h) $H \to b\bar{b}$, et (i) $H \to \tau^+\tau^-$. [6]

Comme dans le cas du Higgs unique, différents modes de production nous donnent accès à différents couplages du Modèle standard. Un schéma récurrent pour la production de deux Higgs est la formation d'un Higgs unique, suivie d'un couplage trilinéaire. Nous nous concentrerons principalement sur la **fusion de gluons (ggF) :** dans ce mode de production, on observe le Higgs unique issu de la ggF mentionné précédemment se désintégrer en deux autres bosons de Higgs (Figure 4a) et le diagramme en boîte (Figure 4b). Mais ce qui rend le Di-Higgs intéressant, c'est que ces deux processus partagent les mêmes états initiaux et finaux ; par conséquent, l'état Di-Higgs observé est une superposition des deux diagrammes. L'objectif principal de cette thèse est de comprendre les conséquences de cette superposition et de développer des méthodes pour surmonter ce défi.

2 Dispositif expérimental

Des théories aussi ambitieuses ne peuvent être testées qu'à l'aide d'un dispositif expérimental tout aussi sophistiqué. C'est ce qu'offre le Grand collisionneur de hadrons (LHC) du CERN. Le LHC est un collisionneur circulaire de 27 km, situé à Genève, à la frontière franco-suisse. Il provoque des collisions de protons à une énergie dans le centre de masse de 13,6 TeV. Le LHC comporte quatre points d'interaction principaux, chacun abritant l'une des quatre expériences principales du LHC.

- **ATLAS** (**A** **T**oroidal **L**HC **A**pparatu**S**): ATLAS est un détecteur polyvalent qui poursuit un large programme de physique, tels que la mesure des propriétés du boson de Higgs, l'étude de la matière noire, la symétrie de super-symétrie (SUSY) et d'autres recherches en physique au-delà du Modèle standard.
- **ALICE** (**A** **L**arge **I**on **C**ollider **E**xperiment) : L'expérience ALICE[8] est consacrée à l'étude du plasma quark-gluon (QGP) par le biais de collisions d'ions lourds.
- **CMS** (**C**ompact **M**uon **S**olenoid) : CMS[9], tout comme ATLAS, est un détecteur polyvalent, mais utilise des technologies différentes dans ses sous-détecteurs. CMS mène également des études similaires à celles d'ATLAS.
- **LHCb** (**L**arge **H**adron **C**ollider **b**eauty) : LHCb[10] est un détecteur exclusivement orienté vers la région avant. LHCb se concentre sur la physique des saveurs et dispose donc d'un détecteur de vertex dédié, VELO, et d'un détecteur d'identification des particules, le RICH.

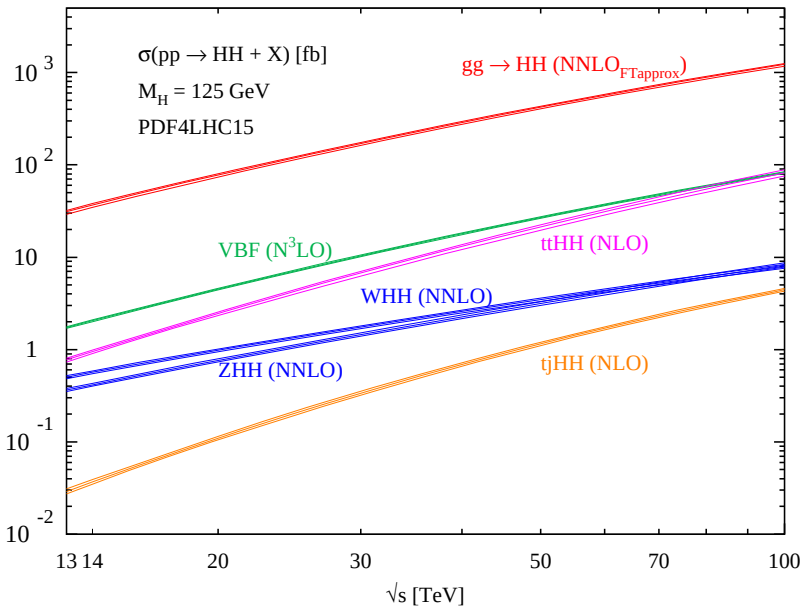


Figure 3: $pp \to HH + X$, section efficace de production dans divers canaux en fonction de l'énergie dans le centre de masse ($\sqrt{s}$)[7].

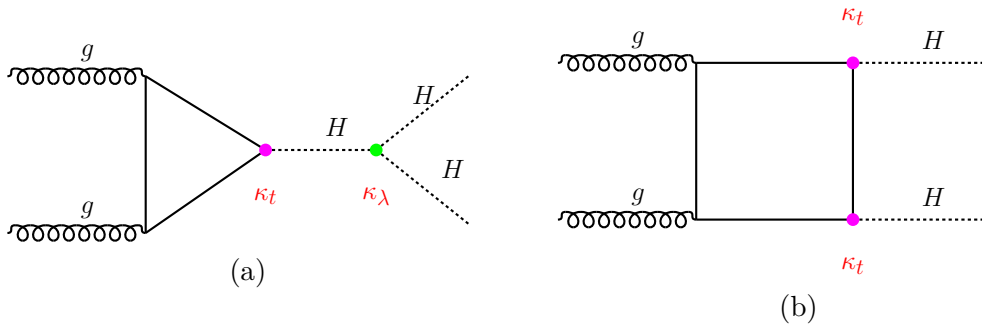


Figure 4: Production de paires de bosons de Higgs par fusion de gluons (ggF) avec un auto-couplage trilinéaire du boson de Higgs (κ_λ (triangle) (a) et impliquant uniquement des contributions de boucle du quark top (encadré) (b).

L'expérience ATLAS

Le détecteur ATLAS est composé de plusieurs sous-détecteurs, chacun conçu pour identifier et mesurer différents types de particules issues des événements de collision. Le **détecteur interne**, immergé dans un champ magnétique solénoïdal de 2 T, reconstitue les trajectoires et les impulsions des particules chargées proches du point d'interaction. Les calorimètres à **argon liquide** et à **tuiles** mesurent les énergies des particules via leurs interactions avec le matériau du détecteur, ce qui permet de reconstruire les électrons, les photons et les jets hadroniques. Entourant ces systèmes, le **spectromètre à muons** utilise un vaste champ magnétique toroïdal et des chambres de trajectographe de précision pour identifier et mesurer les muons. Ensemble, ces composants du détecteur fournissent des informations complètes sur les collisions, permettant la reconstruction d'états finaux complexes ainsi que des mesures de précision et la recherche de nouveaux phénomènes en physique des particules à haute énergie. Les données collectées par le détecteur ATLAS sont ensuite traitées par des logiciels de reconstruction et d'identification avant d'être utilisées pour l'analyse.

Exploitation du LHC et phase de haute luminosité

La période d'exploitation est divisée en cycles d'exploitation. Un cycle dure environ 2 à 3 ans, période durant laquelle la plupart des technologies utilisées restent inchangées, ce qui garantit la cohérence des résultats physiques. Chaque cycle est suivi d'un long arrêt destiné à la mise en œuvre de mises à niveau majeures et à la maintenance. La luminosité est la mesure de l'intensité du collisionneur de particules et, par extension, des données collectées. La Run 1 a été la toute première campagne du LHC ; le faisceau était réglé à une énergie au centre de masse de 7 TeV. La run 1 présentait une luminosité intégrée de 30fb^{-1}. Elle a été marquée par la découverte du boson de Higgs lors des expériences ATLAS et CMS, par des mesures de section efficace pour les bosons W et Z, ainsi que par la production de quarks top. Les Run 2 et 3 ont marqué l'ère des mesures de précision pour le LHC. L'énergie au centre de masse a été fixée respectivement à 13 et 13,6 TeV. La Run 2 a permis de collecter environ 140fb^{-1} et, en 2025, la Run 3 avait collecté environ 56fb^{-1}. La

prochaine phase du LHC est la phase à haute luminosité, au cours de laquelle la luminosité instantanée devrait être 10 fois supérieure aux valeurs enregistrées lors des Run 2 et 3. Cela signifierait qu'il y aurait davantage de collisions proton-proton par croisement de paquets, comme le quantifie le pile-up (μ). Les cycles 2 et 3 présentaient un $\mu \approx 50$, tandis que lors du Run 4 (HL-LHC), le pile-up atteindrait environ $\mu \approx 200$. Pour y remédier, diverses améliorations ont été mises en œuvre dans l'expérience ATLAS. L'une d'entre elles, qui présente un intérêt particulier pour cette thèse, est le nouveau trajectographe interne (ITk) entièrement en silicium du détecteur ATLAS, qui remplacera complètement l'actuel détecteur interne (ID). Il dispose d'une plage de pseudorapidité plus large, $|\eta| < 4$, et d'une granularité bien supérieure à celle de l'ID.

3 Le résolveur d'ambiguïtés ACTS à base de scores et son intégration dans Athena

La reconstruction des trajectoires comporte quatre étapes principales. La première, **la reconstruction et le regroupement des points de mesure** (*space points*), consiste à récupérer les données du détecteur, à les regrouper et à les convertir en coordonnées tridimensionnelles. La deuxième étape, ***seeding***, consiste à regrouper au moins trois points de mesure susceptibles de former une trajectoire. La troisième étape, **la recherche et l'ajustement des trajectoires**, consiste à propager les trajectoires initialisées à partir de graines (*seeds*), puis à en ajuster les paramètres. Enfin, la quatrième étape, **la résolution des ambiguïtés**, consiste à éliminer les fausses trajectoires résultant d'associations incorrectes de mesures, ainsi que les trajectoires reconstruites en double, lorsque plusieurs graines conduisent à la reconstruction d'une même trajectoire, mais avec des qualités différentes.

Le nouvel Inner Tracker (ITk) nécessitait également le développement d'un algorithme amélioré de reconstruction des trajectoires de particules chargées. L'algorithme de reconstruction existant est intégré au cadre logiciel Athena, spécifiquement développé pour le détecteur ATLAS. Le nouveau logiciel de suivi repose quant à lui sur ACTS (*A Common Tracking Software*), une pile logiciel de reconstruction de trajectoires indépendant de l'expérience et disponible en open source. Pour mener à bien la migration vers ce nouveau logiciel, il était donc nécessaire de rendre disponibles, dans ce nouveau cadre, les principales fonctionnalités du logiciel existant.

Dans le cadre de cette thèse, j'ai développé une version entièrement configurable du résolveur d'ambiguïtés d'Athena dans ACTS, puis je l'ai intégrée à Athena afin d'en valider les performances pour le détecteur ITk.

Résolveur d'ambiguïtés à base de scores

Le résolveur à base de scores repose sur quatre étapes principales :

- **Sélections fondées sur η :** attribuent un score nul aux trajectoires ne satisfaisant pas les critères définis en fonction de η.
- **Calcul des scores :** attribue un score à chaque trajectoire en fonction de différents paramètres et de leurs pondérations.

- **Nettoyage :** élimine les trajectoires en fonction des mesures qu'elles partagent et de leurs scores respectifs.

- **Finalisation :** applique les dernières coupures globales et renvoie le vecteur contenant les indices des trajectoires sélectionnées.

Ce processus permet de classer les trajectoires en fonction des critères de sélection et des scores qui leur sont attribués. L'étape de nettoyage permet ensuite d'éliminer les trajectoires partageant des mesures avec des trajectoires de meilleure qualité.

Par défaut, l'algorithme propose deux méthodes de calcul des scores : le score simple (*Simple Scoring*), fondé sur une combinaison additive, et le score ambi (*Ambi Scoring*), fondé sur une combinaison multiplicative. En ce qui concerne les critères de sélection, l'algorithme dispose d'un ensemble de coupures de base dépendant de η, qui prennent notamment en compte le nombre de trous, de mesures (*hits*) et de mesures aberrantes.

Afin de rendre ce processus entièrement configurable, deux mécanismes de personnalisation ont été mis en place :

- **Configuration de la résolution des ambiguïtés :** un fichier JSON permet de définir les classes de η et les seuils associés aux coupures dépendant de η, ainsi que les poids et les facteurs d'échelle utilisés par les algorithmes de calcul des scores pour chaque sous-détecteur du trajectographe.

- **Options de personnalisation :** des options permettent également aux utilisateurs de définir leurs propres fonctions, prenant en entrée un objet représentant une trajectoire, afin d'en modifier le score ou d'y appliquer des critères de sélection supplémentaires.

Intégration dans Athena pour l'ITk

Un fichier de configuration dédié a tout d'abord été créé afin d'utiliser l'ensemble des critères de sélection et des modificateurs de score disponibles dans la version autonome d'ACTS. Les fonctionnalités optionnelles ont ensuite été utilisées pour implémenter les critères de sélection et les modificateurs de score qui n'étaient pas encore disponibles.

Les modificateurs de score optionnels suivants ont ainsi été implémentés dans Athena :

1. Score associé aux doubles trous dans les bandes de silicium.

2. Score associé au nombre de couches de pixels contributives.

3. Score associé aux impacts dans la couche de pixels la plus interne.

4. Score associé aux impacts dans le silicium.

Cinq critères de sélection supplémentaires dépendant de η ont également été ajoutés dans l'implémentation Athena :

1. Valeur absolue maximale de η pour une trajectoire.

2. Valeur maximale de l'impact longitudinal (z_0) pour une trajectoire.

3. Valeur maximale du paramètre d'impact transverse (d_0) pour une trajectoire.

4. Nombre minimal de mesures dans le silicium.

5. Nombre maximal de doubles trous.

L'ensemble de ces critères permet de reproduire le comportement du résolveur d'ambiguïtés d'Athena sans recourir à des composants codés en dur. La Figure 5 présente une comparaison entre les scores obtenus avec le résolveur ACTS à base de scores et ceux obtenus avec l'implémentation historique d'Athena. Les deux résolveurs présentent une forme générale et une échelle comparables.

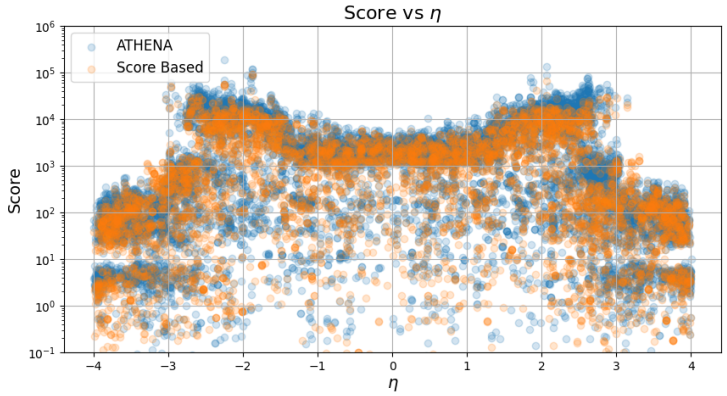


Figure 5: Comparaison des scores obtenus avec le résolveur Athena et le résolveur ACTS à base de scores, en fonction de η.

L'efficacité correspond à la proportion de trajectoires correctement reconstruites parmi les trajectoires sélectionnées. Idéalement, cette valeur devrait être égale à un. La Figure 6a compare l'efficacité du résolveur ACTS à base de scores à celles des résolveurs Athena et ACTS glouton (*Greedy Solver*). Dans l'ensemble, l'efficacité obtenue est satisfaisante, même si des améliorations restent possibles.

Le taux de fausses trajectoires correspond à la proportion de trajectoires incorrectement reconstruites parmi les trajectoires sélectionnées. Plus ce taux est faible, meilleures sont les performances du résolveur. Idéalement, il devrait être nul. La Figure 6b compare le taux de fausses trajectoires obtenu avec le résolveur ACTS à base de scores à ceux des résolveurs Athena et ACTS glouton. Dans l'ensemble, le taux de fausses trajectoires du résolveur ACTS à base de scores est en bon accord avec ceux obtenus avec les deux autres résolveurs.

4 Défi « FAIR Universe HiggsML Uncertainty Challenge »

Le défi « FAIR Universe HiggsML Uncertainty Challenge » (2024-2025) avait pour objectif de développer des méthodes permettant d'estimer l'intensité du signal

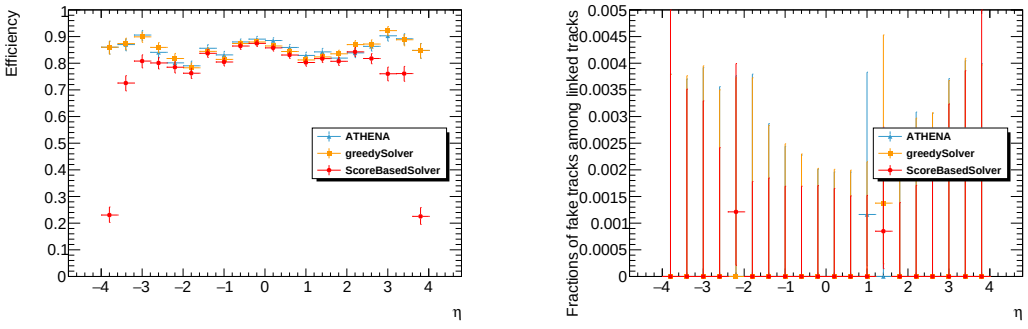
du boson de Higgs tout en tenant compte des incertitudes statistiques et systématiques . Les participants devaient prédire des intervalles de confiance présentant une couverture statistique fiable, plutôt que de se contenter d'optimiser la précision des prédictions. Ce défi a introduit de grands ensembles de données simulées, des distorsions systématiques et un indicateur d'évaluation basé sur la couverture afin d'évaluer la robustesse des méthodes proposées. Le principal paramètre à estimer est la force du signal Higgs, définie comme $\mu = \frac{N_{\text{Higgs}}^{\text{obs}}}{N_{\text{Higgs}}^{\text{SM}}}$, où le dénominateur représente le rendement attendu en événements Higgs selon le Modèle standard. La valeur de référence est $\mu = 1$. Pour chaque pseudo-expérience, les participants doivent fournir un intervalle de confiance à 68,27% pour l'intensité estimée du signal. Ces intervalles doivent tenir compte des fluctuations statistiques, des effets systématiques et des limites de la modélisation. L'objectif est d'obtenir des intervalles aussi étroits que possible tout en conservant une couverture statistique fiable.

Ensemble de données et processus physique

L'ensemble de données est constitué de collisions proton-proton simulées au Grand collisionneur de hadrons (LHC), générées à l'aide de Pythia 8.2 pour la génération d'événements et de Delphes 3.5.0 pour la simulation des détecteurs. Environ 280 millions d'événements ont été produits, chacun étant décrit par 28 caractéristiques contenant des grandeurs cinématiques reconstruites. Le processus de signal correspond à la production d'un boson de Higgs suivie de la désintégration $H \to \tau^+\tau^-$, l'un des leptons tau se désintégrant par voie leptonique et l'autre par voie hadronique. Les principaux processus de fond comprennent $Z \to \tau\tau$, la production top-antitop ($t\bar{t}$) et la production de dibosons. L'objectif est de distinguer le signal Higgs de ces fonds et d'estimer sa contribution aux données observées.

Incertitudes systématiques

Un aspect important de ce défi réside dans l'introduction de six incertitudes systématiques, qui modifient les distributions simulées nominales et peuvent entraîner des différences entre les données d'apprentissage et les données de test. Ces incer-



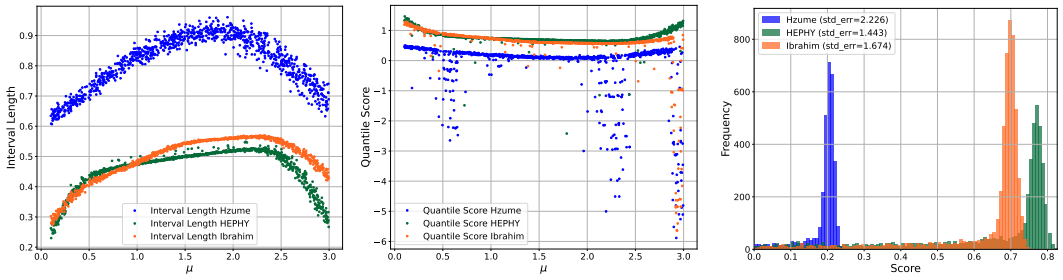
(a) (b)

Figure 6: Comparaison de l'efficacité et du taux de fausses trajectoires obtenus avec les résolveurs Athena, ACTS glouton et ACTS à base de scores : a) efficacité en fonction de η ; b) taux de fausses trajectoires en fonction de η.

titudes concernent l'échelle d'énergie des tau, l'échelle d'énergie des jets, le terme de basse énergie de l'energie transversale manquante, la normalisation globale du fond, la normalisation des dibosons et la normalisation des top-antitop. De telles distorsions peuvent introduire des biais dans l'estimation de l'intensité du signal Higgs. Par conséquent, les participants doivent développer des méthodes qui restent fiables lorsque les distributions d'événements sous-jacentes sont modifiées par ces effets systématiques.

Mesure d'évaluation

Le concours s'est déroulé de septembre 2024 à mars 2025 en deux phases. Lors de la phase publique, les participants ont soumis leurs contributions directement via la plateforme Codabench ; celles-ci ont été testées sur 10 séries de 100 pseudo-expériences, chaque série présentant une valeur fixe de μ. Les graines aléatoires ont été délibérément laissées aléatoires afin d'éviter toute tentative de manipulation du système. Les trois meilleurs modèles issus de la phase publique ont ensuite été sélectionnés pour la phase privée et testés sur 1000 séries de 100 pseudo-expériences. Nous avons utilisé la même graine pour tous les modèles et un nouvel ensemble de données de validation pour cette phase. Afin de pallier l'incertitude liée à l'évaluation du classement du concours, nous avons échantillonné les résultats de 100 séries parmi les 1 000 avec remplacement, reproduisant ainsi une mini-compétition ou un essai. Nous avons répété ces essais 10 000 fois afin de déterminer l'incertitude du classement.



(a) (b) (c)

Figure 7: Étude comparative des trois finalistes (en bleu pour **HZUME**, en vert pour **HEPHY** et en orange pour **IBRAHIME**) avec 1 000 essais de 100 pseudo-expériences. 7a montre la largeur moyenne de l'$IC_{68\%}$, 7b montre le score quantile, et 7c donne la distribution du score final issu des essais.

À l'issue de l'évaluation par la méthode du bootstrap, HEPHY et IBRAHIME ont été déclarés ex æquo, tandis que HZUME a terminé troisième. Voici les résultats qu'ils ont obtenus :

Équipe	Couverture moyenne	Largeur moyenne	Score moyen
HEPHY	0,6683	0,4599	-0,5823
IBRAHIME	0,6698	0,4974	-0,5761
HZUME	0,6659	0,8134	-2,1650

Conclusion

Le défi « FAIR Universe HiggsML Uncertainty Challenge » porte sur l'estimation fiable de l'intensité du signal de Higgs en présence d'incertitudes statistiques et systématiques. Son objectif principal est de garantir que les intervalles de confiance prédits soient à la fois suffisamment précis et statistiquement bien calibrés. Ce concours fournit un référentiel pour le développement de méthodes robustes de quantification des incertitudes et l'évaluation de leurs performances en présence de distorsions systématiques dans des données de collisions simulées.

5 Étude de la production de paires de Higgs dans l'état final $b\bar{b}\ \tau^+\tau^-$

Avant d'aborder l'analyse SBI « Di-Higgs », il convient de passer en revue les procédures et méthodes utilisées dans l'analyse actuelle basée sur les histogrammes, car le prétraitement des données, les déclencheurs et les procédures d'identification sont communs à l'analyse SBI. La dernière analyse réalisée par la collaboration ATLAS utilise une luminosité intégrée totale de $196\,\text{fb}^{-1}$ de données de collisions proton-proton collectées avec le détecteur ATLAS au Grand collisionneur de hadrons du CERN, ce qui correspond à 140 $mathrmfb^{-1}$ provenant de la Run 2 à une énergie dans le centre de masse de 13 TeV et de $56\,\text{fb}^{-1}$ provenant de l'ensemble de données partiel de la Run 3 à $13,6$ TeV. L'analyse porte sur la production non résonante de paires de bosons de Higgs dans l'état final $b\bar{b}\tau^+\tau^-$. Le taux mesuré de production de paires de bosons de Higgs, par rapport à la prédiction du Modèle standard, est de $\mu_{HH} = 2,6^{+1,4}_{-1,0}$, ce qui est conforme aux prévisions du Modèle standard. La significativité observée du signal est de $2,6$ écarts-types au-dessus de l'hypothèse d'un fond seul.

Modèle physique pour κ_λ

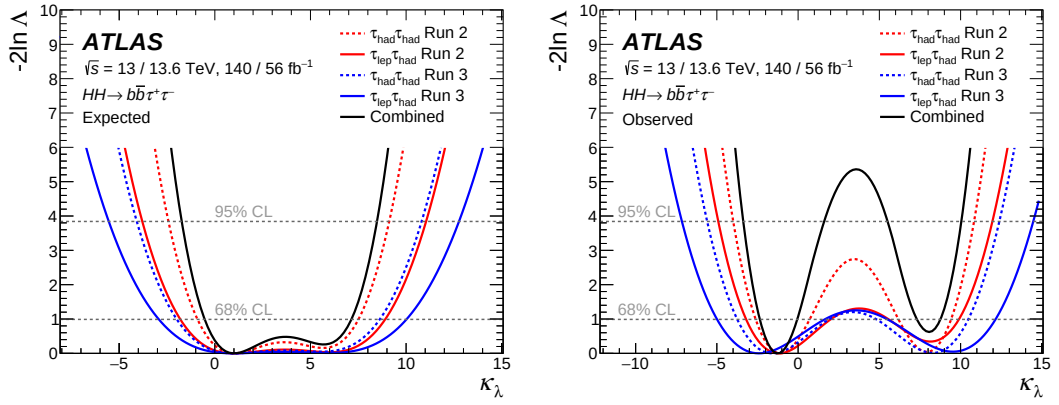
De plus, des contraintes sont obtenues sur le facteur de modification is better d'auto-couplage du boson de Higgs, κ_λ, fournissant des informations supplémentaires sur l'auto-interaction du boson de Higgs. La section efficace totale pour la production de Di-Higgs (σ_{κ_λ}) est donc donnée par

$$\sigma_{\kappa_\lambda} = \kappa_\lambda^2 \cdot \sigma_{HH_\Delta} + \kappa_\lambda \cdot \sigma_{HH_{\text{interférence}}} + \sigma_{HH_\square} \tag{1}$$

Mais comme nous n'avons pas accès aux termes $\sigma_{HH_{\text{interference}}}$ ni σ_{HH_Δ}, nous pouvons simuler des échantillons avec différentes valeurs de κ_λ et les utiliser pour exprimer σ_{κ_λ} sous la forme suivante :

$$\sigma_{\kappa_\lambda} = \sigma_{\kappa_{\lambda 0}} \left(\frac{\kappa_\lambda^2 - 6\kappa_\lambda + 5}{5} \right) + \sigma_{\kappa_{\lambda 1}} \left(\frac{-\kappa_\lambda^2 + 5\kappa_\lambda}{4} \right) + \sigma_{\kappa_{\lambda 5}} \left(\frac{\kappa_\lambda^2 - \kappa_\lambda}{20} \right) \tag{2}$$

Nous utiliserons cette formulation pour construire un modèle statistique tant pour les cas avec regroupement en classes que pour ceux sans regroupement (SBI). La figure Figure 8 présente les valeurs attendues et observées de κ_λ pour différents canaux dans cette analyse.



(a) (b)

Figure 8: Courbe de log-vraisemblance négative pour le scan κ_λ sur différents canaux, pour les valeurs attendues (a) et observées (b). Les lignes pointillées correspondent à $\tau_{\text{had}}\tau_{\text{had}}$ et les lignes pleines à $\tau_{\text{lep}}\tau_{\text{had}}$; la série Run 2 est en bleu et la série Run 3 en rouge. Enfin, les résultats combinés sont représentés en noir.[11]

Ces résultats constituent une référence actualisée pour l'analyse $b\bar{b}\tau^+\tau^-$ et servent de base à l'étude SBI présentée dans ce travail.

Introduction à l'inférence basée sur la simulation

Compte tenu de la nature poissonienne de la distribution des données, on peut exprimer la vraisemblance ($\mathcal{L}$) d'un paramètre d'intérêt, pour un ensemble de données donné, sous la forme

$$\mathcal{L}(\mu|\mathcal{D}) = \frac{\nu(\mu)^n \cdot e^{-\nu(\mu)}}{n!} \prod_i^n p(x_i|\mu) \tag{3}$$

D'après le lemme de Neyman-Pearson, on sait que le meilleur test statistique pour un paramètre d'intérêt est obtenue par le rapport (λ_μ) des vraisemblances sous différentes hypothèses.

$$\lambda_\mu = \frac{\mathcal{L}(\mu|\mathcal{D})}{\mathcal{L}(1|\mathcal{D})} = \frac{\nu(\mu)^n \cdot e^{-\nu(\mu)}}{\nu(1)^n \cdot e^{-\nu(1)}} \prod_i^n \frac{p(x_i|\mu)}{p(x_i|1)} \tag{4}$$

Dans l'inférence basée sur la simulation, nous utilisons des algorithmes d'apprentissage de machine pour estimer le rapport de probabilité $\frac{p(x_i|\mu)}{p(x_i|1)}$. Plutôt que de construire un seul estimateur d'apprentissage de machine pour des problèmes plus complexes, nous pouvons décomposer ce rapport en une somme de rapports de probabilité ; cette approche est appelée « inférence semi-paramétrique basée sur la simulation ».

$$\frac{p(x_i|\mu)}{p_{ref}(x_i)} = \frac{1}{\nu(\mu)} \sum_X^{samples} f_X(\mu) \cdot \nu_X \cdot \frac{p_X(x_i)}{p_{ref}(x_i)} \tag{5}$$

Ici, X représente différents échantillons et μ est le paramètre d'intérêt. Enfin, on peut représenter $\frac{p(x_i|\mu)}{p(x_i|1)}$ sous la forme $\frac{p(x_i|\mu)}{p_{ref}(x_i)} \cdot \frac{p_{ref}(x_i)}{p(x_i|1)}$, ce qui permet d'utiliser

le rapport de densité à la place de la probabilité pour les calculs. Chaque rapport de densité étant indépendant du paramètre d'intérêt, nous pouvons entraîner les classificateurs d'apprentissage automatique pour qu'ils l'approximent. L'efficacité de la méthode SBI semi-paramétrique a été démontrée avec succès dans l'article de 2024 d'ATLAS consacré à la production de Higgs hors résonance en 4 leptons [1, 2].

Inférence par simulation pour le Di-Higgs dans l'analyse $b\bar{b}\tau^+\tau^-$

Nous pouvons reprendre la formulation κ_λ dont nous avons parlé dans Equation 2 et obtenir une expression complète du rapport de probabilité $\frac{p(x_i|\kappa_\lambda)}{p_{\text{ref}}(x_i)}$ sous la forme suivante :

$$\begin{aligned}\frac{p(x_i|\kappa_\lambda)}{p_{\text{ref}}(x_i)} = \frac{1}{\nu(\kappa_\lambda)}\Bigg[&\left(\frac{\kappa_\lambda^2 - 6\kappa_\lambda + 5}{5}\right)\cdot \quad \nu_{\kappa_{\lambda 0}} \cdot \frac{p_{\kappa_{\lambda 0}}(x_i)}{p_{\text{ref}}(x_i)} \\ &+ \left(\frac{-\kappa_\lambda^2 + 5\kappa_\lambda}{4}\right)\cdot \quad \nu_{\kappa_{\lambda 1}} \cdot \frac{p_{\kappa_{\lambda 1}}(x_i)}{p_{\text{ref}}(x_i)} \\ &+ \left(\frac{\kappa_\lambda^2 - \kappa_\lambda}{20}\right)\cdot \quad \nu_{\kappa_{\lambda 5}} \cdot \frac{p_{\kappa_{\lambda 5}}(x_i)}{p_{\text{ref}}(x_i)} \\ &+ \nu_H \cdot \frac{p_H(x_i)}{p_{\text{ref}}(x_i)} + \quad \nu_{VV} \cdot \frac{p_{VV}(x_i)}{p_{\text{ref}}(x_i)} \quad + \nu_{\text{Rest}} \cdot \frac{p_{\text{Rest}}(x_i)}{p_{\text{ref}}(x_i)}\Bigg]\end{aligned} \tag{6}$$

Présélection et distribution de référence

Pour obtenir le rapport de densité le plus précis possible, il convient de sélectionner à la fois une région appropriée pour l'application de la méthode SBI et une distribution de référence adéquate. Pour la présélection, j'ai utilisé une combinaison de seuils simples sur les variables de masse et ΔR, ainsi qu'un discriminant basé sur les scores d'un classificateur multiclasse à réseau neuronal. La principale exigence pour le choix de la distribution de référence est la couverture de l'espace des phases, assurée en utilisant un estimateur de densité par noyau gaussien (KDE) pour déterminer la région de couverture dans un espace à n dimension. Les résultats de nos expériences avec le KDE nous ont amenés à conclure que l'échantillon **Higgs unique** était le meilleur choix de référence.

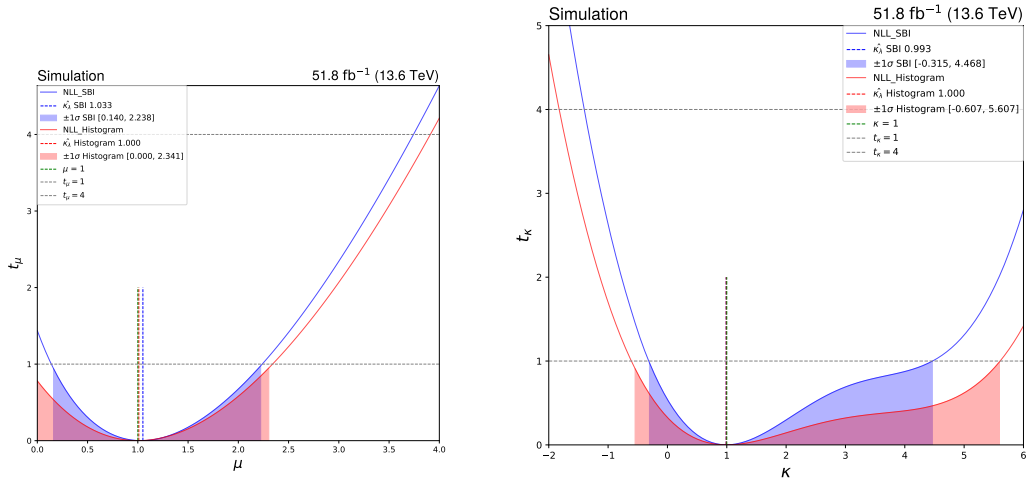
Classificateurs pour l'estimation du rapport de densité

Trois classificateurs différents basés sur l'apprentissage en machine (ML) ont été testés pour l'estimation du rapport de densité. Des arbres de décision boostés (BDT) issus de Scikit-learn, des réseaux neuronaux denses avec TensorFlow et un Graph Attention Transformer ont également été construits avec TensorFlow. À l'aide de courbes d'étalonnage et de repondération, nous avons conclu que les BDT fournissaient la meilleure estimation du rapport de densité pour les échantillons « Di-Higgs », « di-bosons » et « Higgs unique ». Pour les échantillons Z + jets et top, les réseaux neuronaux denses (DNN) ont donné de meilleurs résultats, mais la

qualité était encore trop faible pour être utilisée dans la courbe de vraisemblance finale. Nous les avons donc écartés de nos études et nous nous sommes concentrés uniquement sur les échantillons qui donnaient de bons résultats.

Résultats

Nous avons effectué un balayage de la log-vraisemblance négative et déduit l'intensité du signal μ et la constante d'auto-couplage κ_λ sur un ensemble de test simulé. Afin de démontrer la puissance statistique de la méthode SBI par rapport à la méthode de l'histogramme, nous avons également effectué une analyse par classes du discriminant GNN développé pour l'analyse standard par la collaboration ATLAS. Dans la section précédente, nous avons abordé le problème du mauvais étalonnage de l'échantillon « Rest » malgré tous nos efforts. Enfin, la figure Figure 9 présente la comparaison des courbes de vraisemblance pour les résultats « stat-only », où seuls les échantillons Di-Higgs, Higgs unique et dibosons sont pris en compte entre la méthode SBI et l'histogramme GNN, avec son $\mathrm{CL}_{68\%}$ représenté par une zone ombrée. Nous démontrons ainsi que la méthode SBI présente une puissance statistique supérieure à celle de la méthode de l'histogramme pour l'ensemble des fonds actuellement pris en compte.



(a) (b)

Figure 9: Comparaison des courbes de vraisemblance pour les résultats basés uniquement sur les statistiques, où **seuls** les échantillons Di-Higgs, Higgs unique et diboson sont pris en compte. Courbe de vraisemblance de la méthode SBI en bleu et courbe de vraisemblance de la méthode de l'histogramme en rouge. La ligne verticale pointillée verte indique la valeur réelle du paramètre, tandis que les lignes verticales pointillées bleues et rouges indiquent respectivement les valeurs prédites par les méthodes SBI et histogramme. a) le balayage de la vraisemblance en fonction de l'intensité du signal μ et b) le balayage de la vraisemblance en fonction de la constante d'auto-couplage κ_λ

6 Conclusion

La question à laquelle cette thèse cherche à répondre est de savoir comment évaluer correctement ces avancées pour la recherche en physique des particules, afin de pouvoir à la fois comparer les modèles de manière équitable et mieux comprendre les nouveaux développements en apprentissage automatique pour la physique des hautes énergies. La première partie de la thèse, le **FAIR Universe HiggsML Uncertainty Challenge**[4], aborde directement cette question. Elle met en place un benchmark suffisamment complexe pour représenter la physique que ces modèles sont censés traiter, tout en étant modulable, facile à maintenir et contrôlable, afin qu'une multitude d'études puissent être menées à l'aide de cet outil. Trois points clés méritent d'être soulignés. Le premier est le caractère **FAIR**. Les données et le code permettant leur génération sont accessibles à tous pour être utilisés, modifiés et reproduits. Cela a permis d'utiliser l'ensemble de données FAIR Universe HiggsML [12] pour des études autres que le défi lui-même, telles que l'évaluation comparative de nouveaux modèles comme les modèles de base tabulaires ou à des fins pédagogiques. Le deuxième point concerne l'**open source** et l'**ouverture** : le code du défi est également accessible au public, ce qui permet d'étendre le benchmark pour une analyse plus systématique, de modifier les tests existants ou de changer la corrélation entre les éléments systématiques. Enfin, il s'agit d'un **benchmark à long terme** raisonnable. Les deux propositions gagnantes constituent d'excellentes solutions pouvant être appliquées à des analyses réalistes en physique des hautes énergies. Néanmoins, à mesure que la physique et l'IA progressent, nous aurons besoin d'un benchmark à long terme pour quantifier et classer les modèles, et ce défi remplit parfaitement ce rôle.

La deuxième partie de la thèse consistait également en un exercice de benchmarking, mais d'un autre type. Dans le cas de FAIR Universe, nous cherchions à créer une version FAIR et accessible d'un problème réaliste en physique des hautes énergies. Dans le cadre du SBI pour le Di-Higgs, nous faisions l'inverse. Comment construire un banc d'essai exhaustif pour analyser, comparer et établir des procédures permettant de mettre en place une inférence basée sur la simulation (SBI) pour l'analyse du Di-Higgs dans le canal $b\bar{b}\ \tau^+\tau^-$, dans toute la complexité d'une analyse ATLAS ? Nous avons montré que la méthode SBI actuelle permet d'obtenir un gain d'environ 20% lorsque le top et le $Z+$jets sont ignorés. Cela nous ouvre des perspectives ou nous fournit une feuille de route pour l'avenir. De nombreuses étapes font encore défaut dans cette analyse. L'une des principales contraintes auxquelles nous sommes actuellement confrontés est le nombre insuffisant d'échantillons de Monte Carlo, en particulier pour les échantillons top et $Z+$ jets. Une fois l'analyse purement statistique terminée, nous pourrons commencer à construire des estimateurs de densité pour tenir compte des décalages systématiques et passer d'une analyse purement statistique à une analyse complète combinant les statistiques et l'analyse systématique. La dernière étape consisterait à effectuer la construction de Neyman afin d'obtenir les limites à 1 et 2 σ pour le paramètre κ_λ. Je pense que cette thèse a posé les premiers jalons pour la mise en place d'une inférence basée sur la simulation dans le cadre de l'analyse Di-Higgs dans le canal $b\bar{b}\ \tau^+\tau^-$.

Contents

Acknowledgement **iii**

Résumé en français **v**
1 Contexte théorique . . . vi
2 Dispositif expérimental . . . ix
3 Le résolveur d'ambiguïtés ACTS à base de scores et son intégration dans Athena . . . xi
4 Défi « FAIR Universe HiggsML Uncertainty Challenge » . . . xiii
5 Étude de la production de paires de Higgs dans l'état final $b\bar{b}\,\tau^+\tau^-$ . . . xvi
6 Conclusion . . . xx

Introduction **1**

1 Theoretical Context **3**
1.1 Quantum Field Theory for Particle Physics . . . 3
1.1.1 Particles in Quantum Mechanics . . . 4
1.1.2 Klein-Gordon equation . . . 4
1.1.3 Dirac equation . . . 6
1.1.4 Introduction to Gauge Symmetry . . . 7
1.1.5 Non-Abelian Gauge Theory . . . 8
1.1.6 From Lagrangians to cross-sections . . . 9
1.2 Building the Standard Model of Particle Physics . . . 10
1.2.1 The Electromagnetic Force . . . 12
1.2.2 The Strong Force . . . 13
1.2.3 Weak Force and Electroweak Unification . . . 13
1.2.4 Brout-Englert-Higgs mechanism . . . 16
1.3 Higgs and Di-Higgs in LHC . . . 20
1.3.1 Higgs Production and Decay . . . 20
1.3.2 From Higgs to Di-Higgs . . . 23
1.3.3 Decay Channels for Di-Higgs Analysis . . . 24

2 Experimental Apparatus **29**
2.1 Large Hadron Collider . . . 29
2.1.1 Acceleration Process . . . 29
2.1.2 Luminosity . . . 30
2.1.3 Experiments within the LHC . . . 31
2.1.4 LHC Operations . . . 31
2.2 ATLAS Detector . . . 33

2.2.1 Coordinates in Cylindrical detectors . . . 33
2.2.2 Inner Detector (ID) . . . 34
2.2.3 Calorimeters . . . 36
2.2.4 Muon Spectrometer (MS) . . . 38
2.2.5 Data Acquisition in the ATLAS Experiment . . . 39
2.3 High-Luminosity LHC and ATLAS Upgrades . . . 41
2.3.1 Inner Tracker . . . 41
2.3.2 High Granularity Timing Detector . . . 42

3 ACTS Score-Based Ambiguity Solver and its Implementation in Athena 45
3.1 Tracking in ATLAS and ACTS . . . 45
3.1.1 Space Point Formation and Clustering . . . 46
3.1.2 Track Seeding . . . 47
3.1.3 Track Finding and Track Fitting . . . 47
3.1.4 Ambiguity Resolution . . . 48
3.2 ACTS Score-Based Ambiguity Solver . . . 48
3.2.1 Implementation of Eta Based Cuts . . . 49
3.2.2 Track Scoring for Ambiguity Solving . . . 49
3.2.3 Tracking Cleaning . . . 51
3.2.4 Customisation of the Score-Based Solver . . . 51
3.3 Integrating the ACTS Score-Based Solver into Athena . . . 52
3.3.1 Ambiguity Configuration for ITk Detector . . . 54
3.3.2 Optional cuts and scores in Athena . . . 54
3.4 Results . . . 57
3.4.1 Results from scoring . . . 57
3.4.2 Results from primary comparison . . . 59
3.5 Discussion and Conclusion . . . 61

4 FAIR Universe Higgs Uncertainty Challenge 63
4.1 Towards HiggsML Uncertainty Challenge . . . 64
4.1.1 Systematics Uncertainty in Particle Physics . . . 64
4.1.2 Motivation for an Uncertainty Challenge . . . 64
4.2 Challenge Dataset . . . 65
4.2.1 Events Features . . . 65
4.2.2 Event Selection and Data Processing . . . 67
4.3 Implementation of Systematic Biases . . . 68
4.3.1 Feature Distortion Systematic biases . . . 69
4.3.2 Normalisation Systematics . . . 73
4.4 Metric and Evaluations . . . 73
4.5 Benchmark models and Starting Kit . . . 75
4.5.1 Summary Histogram with Machine Learning Discriminants . . . 77
4.5.2 Quantifying Systematic Uncertainity . . . 77
4.5.3 Profile Likelihood for Uncertainty Estimation . . . 79
4.6 Competition Overview . . . 80
4.6.1 The Final Phase . . . 80

4.7 Conclusions and Outlook . . . 82

5 Simulation-Based Inference with FAIR Universe Dataset 83
5.1 Semi-Parametric Simulation-Based Inference . . . 84
5.1.1 Approximating probability ratios with Classifiers . . . 84
5.1.2 Validation of Density ratio Estimators . . . 85
5.2 Winning Submissions to FAIR Universe challenge . . . 86
5.2.1 `HEPHY`: Unbinned inclusive cross-section measurements with machine-learned systematic uncertainties . . . 87
5.2.2 `IBRAHIM`: Contrastive Normalizing Flows for Uncertainty-Aware Parameter Estimation . . . 91
5.3 ATLAS SBI Workflow with FAIR Universe HiggsML Challenge . . . 94
5.3.1 Modelling systematics in Semi-Parametric SBI . . . 96
5.3.2 Discussion and conclusion . . . 97

6 Study of Higgs pair production in $b\bar{b}\ \tau^+\tau^-$ final state in ATLAS 99
6.1 Theoretical context . . . 99
6.2 Data and Monte Carlo samples . . . 101
6.3 Object reconstruction and Tagging . . . 101
6.3.1 Jet reconstruction . . . 102
6.3.2 b-Jet Tagging . . . 102
6.3.3 τ_{had} Jet Tagging . . . 102
6.4 Event Selection Analysis Region . . . 103
6.5 Multivariate Analysis . . . 103
6.5.1 Network Architecture . . . 103
6.5.2 Training Features and Strategy . . . 104
6.5.3 Event Categorisation . . . 105
6.6 Systematic uncertainties . . . 107
6.6.1 Experimental Uncertainities . . . 107
6.6.2 Theoretical modelling uncertainties . . . 107
6.7 Results and discussion . . . 109

7 Simulation-based Inference for Di-Higgs Analysis 113
7.1 Semi-Parametric SBI for Di-Higgs . . . 113
7.1.1 Preselection . . . 115
7.2 Choice of Reference sample in SBI . . . 116
7.2.1 Gaussian Kernel Density Estimation . . . 118
7.2.2 Uniform Reference Distribution . . . 120
7.3 Density Ratio Estimation Classifiers . . . 121
7.3.1 Dense Neural Network . . . 122
7.3.2 Boosted Decision Trees . . . 122
7.3.3 Graph Attention Transformer . . . 123
7.3.4 Tabular Foundation models for HEP . . . 125
7.3.5 Validation of classifiers . . . 126
7.4 Results . . . 129

7.5 Discussion and Future Steps 131

A FAIR Universe Dataset **151**
A.1 Crossection tables for FAIR Universe Dataset 151
A.2 List of features for FAIR Universe Dataset 152

B Study of Tabular Foundational Models for HEP Analysis **161**
B.1 Tabular Foundational Models 161
B.2 Analysis Methodology 162
B.3 Physics Analysis with Tabular Fundamental Models 164
B.3.1 Approximate Median Significance 165
B.3.2 Summary Histogram with Machine Learning Discriminants 166
B.3.3 Simulation-Based Inference with Tabular Foundational Models 168
B.4 Discussion of Tabular Foundational Models for HEP 168

C Summary Statistics for Di-Higgs Analysis samples **173**

D Analysis plots Di-Higgs SBI **177**
D.1 Analysis plots of Classifier for SBI 177
D.2 SBI pseudo-univertse Anlysis 183

Introduction

The 2012 **Higgs Boson** discovery at the **Large Hadron Collider** (LHC) by the **ATLAS** and **CMS** collaborations has established the BEH mechanism, which endows fundamental particles with mass. Following the discovery of the Higgs boson, the focus of the high-energy physics community shifted from particle discovery to precision measurements. Measuring the properties of the Higgs boson is not merely the study of a single, elusive particle. Subsequent measurements from ATLAS and CMS have parameterised the various coupling constants of the Higgs Boson, thereby cementing the validity of the Standard Model of particle physics.

The second AI revolution in the 2010s is often attributed to the creation of the ImageNet[13] database, which provided an enormous dataset for AI development and testing. Years of LHC operation have given an enormous amount of both simulated and real data, enabling us to build machine learning models for Particle physics. This brought forth the first ML challenge in our field. The 2014 **Kaggle HiggsML Challenge**[14, 15]. This was the first of its kind to benchmark the performance of machine learning algorithms on a classification problem using ATLAS open data in the $pp \to \tau^+\tau^-$ channel.

Despite having excellent simulations to train our AI models, we still can't assure that they will match the real data we collect. This could be a difference in detector response, uncertainty arising from theoretical factors or calibration issues. Collectively, we called this systematic uncertainty. Over the last decade or so, various methods for uncertainty-aware artificial intelligence have been proposed, but comparing them is often difficult because they use different benchmarks, datasets, and problem statements. This motivated us to develop the **FAIR Universe HiggsML Uncertainty Challenge**[4], a machine learning challenge focused on quantifying and/or reducing systematic uncertainty. For the new competition, participants were evaluated on both their ability to precisely determine the correct cross-section and their ability to report correct, well-calibrated uncertainty intervals. The competition ran from Sep 2024 to 14th March 2025 and was accepted as an official **NeurIPS 2024** competition. In this thesis, we will address the dataset, metric, and infrastructure built for this competition. Finally, we will also discuss the winning solution and how the final winners are selected.

The Higgs Boson provides a unique opportunity to probe the Higgs field itself, a fundamental component of the vacuum that permeates the Universe and has existed since the earliest moments after the Big Bang. The Higgs potential has two free parameters: the Higgs Boson mass and the Higgs self-coupling constant, lambda.

The Higgs boson mass has been measured with very high precision over the last decade; unfortunately, the same cannot be said for the self-coupling constant λ. The λ parameter can be directly measured from the Di-Higgs production.

The ATLAS collaboration recently published the latest methods and procedures in analysing the $HH \to b\bar{b}\tau^+\tau^-$ channel[3] at the LHC. We will briefly discuss that work here, as the data preprocessing, triggers, and tagging used are shared with the SBI analysis. The production of Di-Higgs involves large destructive quantum interference. These often make analysis difficult; traditional histogram analysis may not capture high-dimensional information. This is where **Simulation-Based Inference (SBI)** shines. SBI uses collections ML algorithms to directly infer parameters of interest from the high-dimensional data rather than a summary histogram. The 2024 SBI in the ATLAS off-shell Higgs analysis[1, 2] successfully demonstrated the advantage of utilising SBI for problems involving substantial destructive interference. Di-Higgs production also has similar challenges. The core part of this thesis is the application of SBI techniques for the $HH \to b\bar{b}\tau^+\tau^-$ channel. This work aims to compare various machine learning models and quantify the potential gains in sensitivity in κ_λ estimation relative to traditional histogram analysis. Preliminary results and the validation of the density-ratio estimators will be discussed in detail. Finally, we will discuss the next steps required and the future of Simulation-Based Inference in $HH \to b\bar{b}\tau^+\tau^-$ analysis.

Such ambitious theories can only be tested with an equally sophisticated experimental setup. In this thesis, we will discuss the Large Hadron Collider (LHC), the accompanying accelerator facility, and the ATLAS detector, which provides the measurements. As of July 2026, the LHC and the ATLAS detector will undergo a major upgrade to enter the Hi-lumi era, in which we expect the instantaneous luminosity to reach 3-4 times the current value. I played a very small part in this journey by developing and implementing the ACTS score-based ambiguity solver in Athena, the tracking and reconstruction software for ATLAS. Together with developments into new benchmarks and machine learning models, this thesis could pave the way for exciting new physics in the Hi-lumi era of the LHC.

Chapter 1
Theoretical Context

By symmetry, we mean the existence of different viewpoints from which the system appears the same. It is only slightly overstating the case to say that physics is the study of symmetry.
— Philip Warren Anderson

The Standard Model of Particle physics is the best experimentally verified model to describe the physics on the smallest scale in the universe. In this model, we describe all the interactions as interactions of various kinds of elementary particles. The Standard Model is built on principles of Quantum Field Theory (QFT), along with constraints imposed by various symmetry groups and conservation laws, which have been confirmed by decades of experimental data. In this chapter, we will briefly discuss Quantum field theory and symmetry groups that make up the Standard Model of Particle physics.

1.1 Quantum Field Theory for Particle Physics

The Quantum theory of particles is not one elegant theory that completely revolutionised physics, but rather a slow evolution of quantum mechanics over decades of experimental and theoretical breakthroughs. The first full formulation of Quantum Mechanics (QM) was given by Werner Heisenberg, Max Born, and Pascual Jordan as a follow-up to Werner Heisenberg *Umdeutung paper*[16]. The fundamental structure of this interpretation was matrices; hence, this formulation is also called the matrix formulation of quantum mechanics. The core idea behind this goes as follows. Every Observable is represented with a matrix A, and the Hamiltonian is represented with H, then time evolution of A is given by:

$$\frac{dA}{dt} = \frac{i}{\hbar}[H, A] + \frac{\partial A}{\partial t} \tag{1.1}$$

Around the same time, Erwin Schrödinger introduced the equation of motion for a particle, given by the Schrödinger wave equation [17].

$$\hat{E}\psi = \hat{H}\psi \tag{1.2}$$

Here ψ is called the wave function and $\hat{E}$ and $\hat{H}$ are Energy and Hamiltonian operators that act on this wave function. We can expand these operators as differential functions, shown below.

$$\hat{E} = i\hbar\frac{\partial}{\partial t} \quad , \quad H = \frac{p^2}{2m} + V(x) \quad , \quad \hat{p_x} = -i\hbar\frac{\partial}{\partial x} \implies \hat{p} = -i\hbar\nabla \tag{1.3}$$

Substituting $\hat{p}$ to build $\hat{H}$ and replacing the terms in Equation 1.2 gives the full form for the Schrödinger wave equation.

$$i\hbar\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\psi + V(x)\psi \tag{1.4}$$

Later, Dirac unified Schrödinger and Heisenberg's formulation with the help of the bra-ket notation. Where the states is represented by $|\psi\rangle$ and which are column vectors, and a wave function $\psi(x)$ is given by

$$\psi(x) = \langle x|\psi\rangle \quad , \quad A\psi(x) = \langle x|A|\psi\rangle \tag{1.5}$$

1.1.1 Particles in Quantum Mechanics

The Schrödinger wave equations, unfortunately, only apply to non-relativistic scalar particles. The Schrödinger equation Equation 1.4 explicitly violates Lorentz covariance. Discoveries in particle physics in the early 20th century had proven the relativistic nature of particles, thus motivating the creation of a relativistic wave equation. Another interesting property in quantum mechanics that the Schrödinger equation doesn't address is intrinsic spin. Originally proposed to explain the fine spectrum splitting in atomic spectroscopy, where a new *""two-valued quantum degree of freedom"* was required to explain it. Later, Pauli proposed an algebraic model to explain the transformation of such fields[18]. In nature, we can categorise fields by how they transform under rotation. The most well-known ones are scalars, which don't change, and vectors, which obey the SO(3) rotation group. Similarly, there are fields called spinor fields that transform under a different rotation group, specifically SU(2). Fundamentally, we can classify all particles into 2 categories: Bosons and Fermions. Bosons are particles with integer-valued spin (0,1,2 ...). As a result, they follow **Bose-Einstein** Statistics[19]. We will discuss more on the dynamics of spin-0 particles (Scalars) in the next section and spin 1 particles (Vectors) in subsection 1.1.4. Fermions are particles with half-integer spin ($\frac{1}{2}$,$\frac{3}{2}$,$\frac{5}{2}$ ...), these follow **Fermi-Dirac**[20, 21] Statistics. We will discuss them further in subsection 1.1.3.

1.1.2 Klein-Gordon equation

Klein and Gordon started from the relativistic energy–momentum relation to construct a Lorentz invariant term.

$$E^2 = p^2c^2 + m^2c^4 \tag{1.6}$$

Replacing the terms with the respective operators, as from Equation 1.3, we get

$$-\hbar^2\frac{\partial^2\phi}{\partial t^2} = -\hbar^2c^2\nabla^2\phi + m^2c^4\phi \implies \hbar^2\frac{\partial^2\phi}{\partial t^2} - \hbar^2\nabla^2\phi + m^2c^4\phi = 0 \tag{1.7}$$

For convenience, we can rewrite this with the Einstein summation convention. The Einstein convention helps us write a summation over coordinates. Where we use Greek indices (μ, ν, etc.) to mark space-time coordinates, 0 for time and 1-3 for the 3 spatial coordinates. In situations where we explicitly only need spatial coordinates, we can use Latin indices (i, j, k, etc.). If the same letter appears twice, once upstairs and once downstairs, the summation is implied. Further, we take the natural units where $c = \hbar = 1$, thus giving :

$$-\partial_t\partial^t + \partial_i\partial^i + m^2\phi = 0 \tag{1.8}$$

where i goes from 1 to 3 for the 3 spatial coordinates, and μ goes from 0 to 3 for the space-time coordinates, $\eta^{\mu\nu}$ is the Minkowski metric, which governs how dot products are computed in flat space or Minkowski space.

$$\eta^{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{1.9}$$

Finally using 4D indices and using natural units ($c = \hbar = 1$), we get:

$$\hbar^2\partial_\mu\partial^\mu + m^2c^4\phi = 0 \tag{1.10}$$

Finally, we used the box operator as the 4D equivalent of the nabla (∇) operator, thereby yielding the Klein-Gordon equation[22, 23] for a relativistic scalar particle.

$$(\Box + m^2)\phi = 0 \tag{1.11}$$

When dealing with a complex system, it is often easier to work with the Lagrangians rather than the wave equations themselves. One of the tools in the physicist's arsenal is the Euler-Lagrange equation. It is derived from the principle of Least Action. The Euler-Lagrange equation for 4D coordinates is in the following form.

$$\frac{\partial\mathcal{L}}{\partial\phi} - \partial_\mu\left(\frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\right) = 0 \tag{1.12}$$

We can hence construct a Lagrangian ($\mathcal{L}_{\text{free}}$) for the Klein-Gordon equation such that, when we apply the Euler-Lagrange equation, we recover the Klein-Gordon equation.

$$\mathcal{L}_{\text{free}} = \frac{1}{2}(\partial_\mu\phi)(\partial^\mu\phi) - \frac{1}{2}m^2\phi^2 \tag{1.13}$$

One of the most powerful tools in the physicist's arsenal is the Euler-Lagrange equation. It is derived from the Principle of Least Action. The Euler-Lagrange

equation for 4D coordinates is in the following form.

$$\frac{\partial \mathcal{L}}{\partial \phi} - \partial_\mu \left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \right) = 0$$

Applying the Euler-Lagrange equation to Equation 1.13, we get back the Klein-Gordon equation. Equation 1.13 was only for the free particle; we can add a self-interaction term which depends on 4 fields (2 incoming and 2 outgoing) to complete the Lagrangian.

$$\mathcal{L} = \mathcal{L}_{\text{free}} + \mathcal{L}_{\text{int}} = \frac{1}{2}(\partial_\mu \phi)(\partial^\mu \phi) - \frac{1}{2}m^2\phi^2 - \frac{1}{4!}\phi^4 \tag{1.14}$$

1.1.3 Dirac equation

The fundamental issue with the Klein–Gordon equation is that its conserved current cannot be interpreted as a positive-definite single-particle probability density. In particular, the time component of the Klein–Gordon current can take negative values, making a consistent probabilistic interpretation problematic. Paul Dirac wanted a first-order equation for relativistic particles. In order to make the Hamiltonian operator H linear and relativistically invariant, he introduced a term γ which was a 4x4 matrix. This forced the wave function to be a 4-component Spinor, which we now refer to as a Dirac spinor. Equation 1.15 gives the Dirac equation[24] for a relativistic spin-$\frac{1}{2}$ particle.

$$(i\gamma^\mu \partial_\mu - m)\,\psi = 0 \tag{1.15}$$

and its corresponding Lagrangian as

$$\mathcal{L} = \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi \tag{1.16}$$

The 4x4 gamma matrices satisfy the Clifford algebra Equation 1.17.

$$\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu} \tag{1.17}$$

The natural consequences of the 4-component Spinor are that it could be represented as a pair of 2-Spinors (Weyl's spinors), one for left-handed particles ψ_L and the other for right-handed particles ψ_R.

$$\psi = \begin{pmatrix} \psi_L \\ \psi_R \end{pmatrix}$$

The projection operators P_L and P_R transform the Dirac spinor into left- and right-chiral components ($\psi_L = P_L\psi, \qquad \psi_R = P_R\psi$). These can be described with the help of the chiral operator γ^5.

$$P_L = \frac{1}{2}(1 - \gamma^5), \qquad P_R = \frac{1}{2}(1 + \gamma^5) \tag{1.18}$$

and finaly the chiral operator γ^5 is given by Equation 1.19

$$\gamma^5 = i\,\gamma^0\gamma^1\gamma^2\gamma^3 \tag{1.19}$$

1.1.4 Introduction to Gauge Symmetry

In particle physics, we deal with 2 types of symmetries. One is the global symmetry of the Lorentz-Poincaré group, which underlies special relativity, and the other is local symmetry, provided by gauge groups and governing particle interactions. If we perform a local gauge transformation of the field $\phi(x)$, with a gauge group G

$$\phi(x) \to \phi'(x) = G\phi(x)$$

and then its derivative $\partial_\mu \phi'$ is given by :

$$\partial_\mu \phi' = \partial_\mu(G\phi(x)) = (\partial_\mu G\phi(x)) + G(\partial_\mu \phi(x)) \tag{1.20}$$

But the issue is that the term $\partial_\mu \phi(x)$ breaks covariance under local transformations. Instead, we define a covariant derivative:

$$D_\mu = \partial_\mu - igA_\mu \tag{1.21}$$

We introduce the term that depends on A_μ to compensate for the $\partial_\mu G\phi(x)$. A_μ is the vector gauge field associated with this gauge group, and g is the coupling constant. The coupling constant determines the strength of interaction between the vector fields and other fields.

$$\begin{aligned} D'_\mu \phi' = D'_\mu(G\phi) &= (\partial_\mu - igA'_\mu)(G\phi) \\ &= (\partial_\mu G)\phi + G(\partial_\mu \phi) - igA'_\mu(G\phi) \end{aligned} \tag{1.22}$$

A_μ transforms under G as

$$A_\mu \to A'_\mu = GA_\mu G^{-1} - \frac{i}{g}(\partial_\mu G)G^{-1} \tag{1.23}$$

Now combing Equation 1.22 and Equation 1.23 we get

$$\begin{aligned} D'_\mu \phi' &= (\partial_\mu G)\phi + G(\partial_\mu \phi) - igGA_\mu G^{-1}(G\phi) + ig\frac{i}{g}(\partial_\mu G)G^{-1}(G\phi) \\ &= (\partial_\mu G)\phi + G(\partial_\mu \phi) - igGA_\mu \phi - (\partial_\mu G\phi) \end{aligned} \tag{1.24}$$

Due to the introduction the A_μ term we see that $\partial_\mu G\phi$ terms cancel off to give :

$$D'_\mu \phi' = G(\partial_\mu \phi) - igGA_\mu \phi = G(D_\mu \phi) \tag{1.25}$$

Finally, to understand the interaction of gauge bosons with fermions, we can swap the partial derivative in the Dirac equation (Equation 1.16) with the covariant derivative we defined (Equation 1.21).

$$\begin{aligned} \mathcal{L} &= \bar{\psi}(i\gamma^\mu D_\mu - m)\psi \\ &= \bar{\psi} i\gamma^\mu(\partial_\mu - igA_\mu)\psi - m\bar{\psi}\psi \\ &= \bar{\psi} i(\gamma^\mu \partial_\mu - m)\psi + g\bar{\psi}\gamma^\mu \psi A_\mu \end{aligned} \tag{1.26}$$

The first term in Equation 1.26 is the Dirac Lagrangian for a free fermion, and the second gives the Noether current (j_μ). In 1915, German physicist and

mathematician Emmy Noether proved that every continuous symmetry of a system's dynamics corresponds to a conserved quantity[25].

$$j^{\mu} = g\bar{\psi}\gamma^{\mu}\psi \tag{1.27}$$

1.1.5 Non-Abelian Gauge Theory

An Abelian group is any group such that $\forall a, b \in G : a \cdot b = b \cdot a$. An example of this is the $U(1)$ group. On the contrary, the $SU(N)$ group is non-Abelian. Generators form the algebra for the group. Any element of a group can be generated from the basis with the right coefficients. In our case, A_μ can be expressed in terms of the generators of the Group G as:

$$A_\mu(x) = A^a_\mu(x)T^a \tag{1.28}$$

where T^a are generators of the group (G) and A^a_μ is the component in each direction. The generators T^a form the Lie algebra. Any group element connected to the identity can be written as

$$G = e^{ig\theta_a T^a} \qquad \psi \to \psi' = e^{ig\theta_a T^a}\psi \tag{1.29}$$

In the standard model, we deal with unitary groups. For $SU(N)$ groups, the number of generators is given by N^2-1. In Lie algebra, we can define a commutator operator $[T^a, T^b]$ as :

$$[T^a, T^b] = T^aT^b - T^bT^a = if^{abc}T^c \tag{1.30}$$

Now, in differential geometry, the curvature of a connection is measured by the commutator of covariant derivatives. This, in turn, for gauge groups becomes the field strength. We can define the field strength tensor $F_{\mu\nu}$ such that

$$[D_\mu, D_\nu] \propto F_{\mu\nu} \tag{1.31}$$

Now to if we substitute covariant derivative with Equation 1.21, we get

$$[D_\mu, D_\nu] = -ig(\partial_\mu A_\nu - \partial_\nu A_\mu + ig[A_\mu, A_\nu]) \tag{1.32}$$

With that, we can define the field strength tensor as :

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + ig[A_\mu, A_\nu] \tag{1.33}$$

In the case of abelian groups, $[A_\mu, A_\nu] = 0$, giving rise to the Proca Equation[26] for massive vector bosons.

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}m^2A_\mu A^\mu \qquad , \qquad F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{1.34}$$

The mass term breaks gauge invariance; hence, no gauge group can have a mass term as implemented by Proca[1].

[1]Proca Equation is from the 1930s and hence predates gauge theory

Recall we can represent A_μ in terms of its generators (see Equation 1.28) and accounting for the communation relation in Equation 1.30, we get :

$$[A_\mu, A_\nu] = A^a_\mu A^b_\nu [T^a, T^b] = i A^a_\mu A^b_\nu f^{abc} T^c \tag{1.35}$$

Thus giving the expression of field strength as

$$F_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - g A^a_\mu A^b_\nu f^{abc} T^c \tag{1.36}$$

The consequence of the last term is that non-Abelian gauge fields allow for self-interaction. This is paramount to understanding the Weak and Strong Nuclear forces. With this, we can construct the simplest Lagrangian for gauge fields, known as the Yang–Mills Lagrangian[27].

$$\mathcal{L} = -\frac{1}{4} F^a_{\mu\nu} F^{\mu\nu}_a \qquad , \qquad (D_\mu F^{\mu\nu})^a = J^{\nu a}. \tag{1.37}$$

Here $J^{\nu a}$ is the current in the Noether current; in the absence of a source, $J^{\nu a} = 0$.

1.1.6 From Lagrangians to cross-sections

Now that we have some formulation of Lagrangians, we can try to construct a pathway from the Lagrangians to measurable quantities such as cross-sections. The full explanation is beyond the scope of this chapter; hence, I will use a minimal example to convey the key points needed to understand the rest of the chapter.
The scattering cross section between initial state i and final state f for an interaction is

$$S_{fi} = \langle f | S | i \rangle \tag{1.38}$$

Here, S is the S-matrix operator, which can be written using the interaction Lagrangian $\mathcal{L}_{\text{int}}$ as

$$S = T \exp\left(i \int d^4x \, \mathcal{L}_{\text{int}}(x) \right) \tag{1.39}$$

Here, T denotes the time-ordering operator. At high energies, the interactions we consider typically lie in the perturbative regime, allowing the exponential to be expanded as a series and truncated after the first few terms.

$$S = 1 + i \int d^4x \, \mathcal{L}_{\text{int}} + \frac{i^2}{2!} \int d^4x \, d^4y \, T[\mathcal{L}_{\text{int}}(x) \mathcal{L}_{\text{int}}(y)] + \dots \tag{1.40}$$

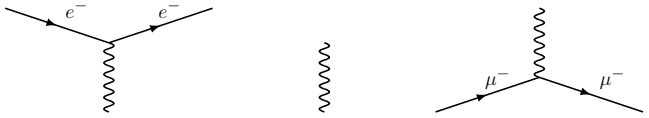
This expansion is called the Dyson expansion. Using Wick's theorem, we can replace the time-ordered terms with the corresponding Green's functions $G_F(x-y)$.

$$G_F(x-y) = \langle 0 | T[\Phi(x) \Phi^\dagger(y)] | 0 \rangle \tag{1.41}$$

Now we will discuss a simple example to illustrate how Feynman propagators are used in physics. Figure 1.1 gives a Feynman diagram for the process of an electron repelling a muon by exchanging a photon.

The matrix element $i\mathcal{M}_i$ for this diagram can be constructed as shown in Equation 1.42. Each part of the matrix element can be constructed from the corresponding part of the diagram using the appropriate propagators.

$$i\mathcal{M}_i = \quad \bar{u}(p_3)(-ie\gamma^\mu)u(p_1) \qquad \frac{-ig_{\mu\nu}}{q^2} \qquad \bar{u}(p_4)(-ie\gamma^\nu)u(p_2) \tag{1.42}$$



vertex propagator vertex

A process can have multiple Feynman diagrams at various orders. The net matrix element is given by its sum.

$$\mathcal{M} = \sum_i \mathcal{M}_i \tag{1.43}$$

We will discuss the different interaction lagrangians and propagators in the following sections. Now we can get the scattering cross-section as

$$S_{fi} = (2\pi)^4 \delta^{(4)}(p_f - p_i)\, i\mathcal{M} \tag{1.44}$$

The total probability for a process is proportional to $|\mathcal{M}|^2$. With that, we can define the differential cross-section as

$$d\sigma = \frac{|\mathcal{M}|^2}{F} d(PS) \tag{1.45}$$

Where F is the flux and PS is for phase space, there 2 quantities that depend on the number in the initial and final states. The Volume integral of the differential cross-section gives the total cross-section σ

1.2 Building the Standard Model of Particle Physics

The period from the 1930s to the 1960s is often referred to as the "particle zoo"; well over 100 particles were discovered during this time, and a theoretical

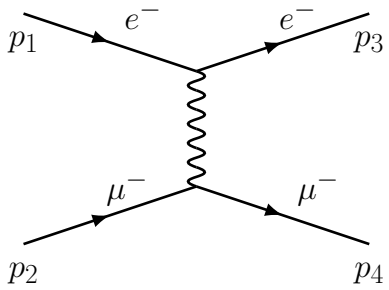


Figure 1.1: Feynman diagram of an electron repelling a muon by exchanging a photon

explanation was required to bring order to this. Gell-Mann and Zweig proposed that these newly discovered particles are not fundamental but are composed of more fundamental particles called **quarks**[28, 29]. For example, the protons found in the nuclei of atoms are made of 2 up (u) quarks and 1 down (d) quark. Apart from quarks, colour-neutral fermions are called **leptons**. The best examples of this are **electrons** and **neutrinos**. Electrons have a charge of -1, and neutrinos are neutral. Fermions are arranged into **generations**. Every generation has an up-type quark, a down-type quark, a lepton and a neutrino. There are 3 generations, all of which have similar properties, apart from the particles' masses. The Standard Model doesn't explain the hierarchy of the masses. In particle physics, we often refer to the type of a fermion as its flavour.

There are four fundamental forces in the universe. The Standard Model explains three of them:

1. **Electromagnetic Force** propagated by the photon γ explains the interaction of electrically charged particles (all fermions except neutrinos)

2. **Strong Nuclear Force** (Quantum Chromodynamics), propagated by gluons, explains the binding of quarks and gluons.

3. **Weak Nuclear Force** propagated by Z and $W^{\pm}$ Bosons, explains flavour-change in fermions

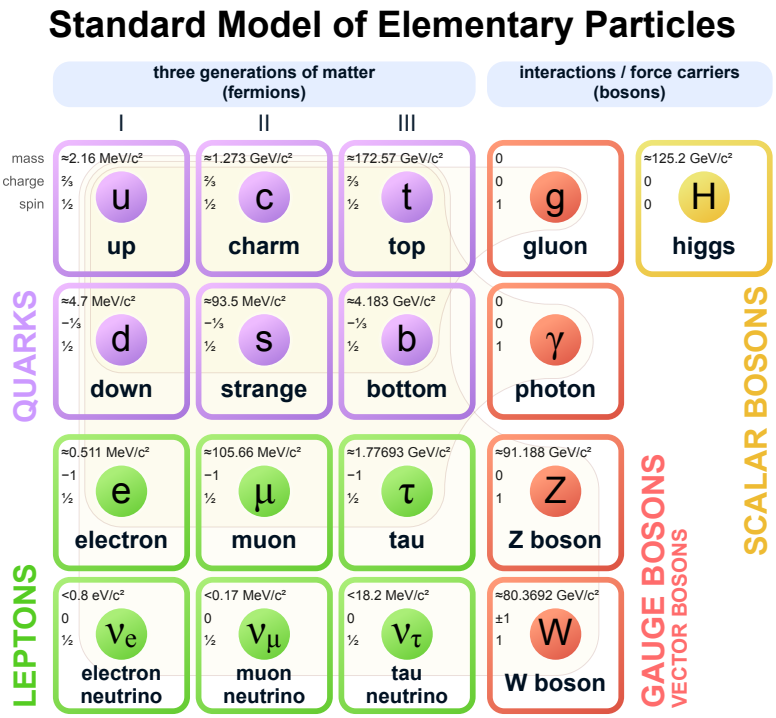


Figure 1.2: Summary of the Standard Model of Elementary particles [5]

The one force not yet in the Standard Model is gravity. Lastly, the Standard Model has a single scalar boson called the Higgs boson, which propagates the Higgs field; the field that gives mass to all fundamental particles. Table 1.1 gives a brief summary of all the fundamental particles in the Standard Model and their Feynman propagators.

Particle	**Representaion**	**Propagator** $G_F(p)$
Fermion		$\frac{i(\gamma^\mu p_\mu + m)}{p^2 - m^2 + i\epsilon}$
Photon (γ)		$\frac{-i\eta_{\mu\nu}}{p^2 + i\epsilon}$
Gluon (g)		$\frac{-i\eta_{\mu\nu}\delta^{ab}}{p^2 + i\epsilon}$
$Z, W^\pm$ boson		$\frac{-i\left(\eta_{\mu\nu} - \frac{p_\mu p_\nu}{M^2}\right)}{p^2 - M^2 + i\epsilon}$
Higgs (H)		$\frac{i}{p^2 - M_H^2 + i\epsilon}$

Table 1.1: Standard Model particles: Feynman diagram lines and propagators.

1.2.1 The Electromagnetic Force

Quantum Electro-Dynamics (QED) is governed by the $U(1)$ gauge group. As the $U(1)$ group is abelian, we can take the abelian formulation of $F_{\mu\nu}$ as seen in Equation 1.34. The $U(1)$ group has only one generator ($T^a = 1$); ergo, the electromagnetic force is mediated by one boson, the photon γ. The coupling for QED is the elementary electric charge e, and the phase is represented as $\alpha(x)$ finally replacing T^a in Equation 1.29, we can rewrite as Equation 1.23 as :

$$A_\mu \to A'_\mu = A_\mu + \frac{1}{\mathrm{e}}\partial_\mu \alpha(x) \tag{1.46}$$

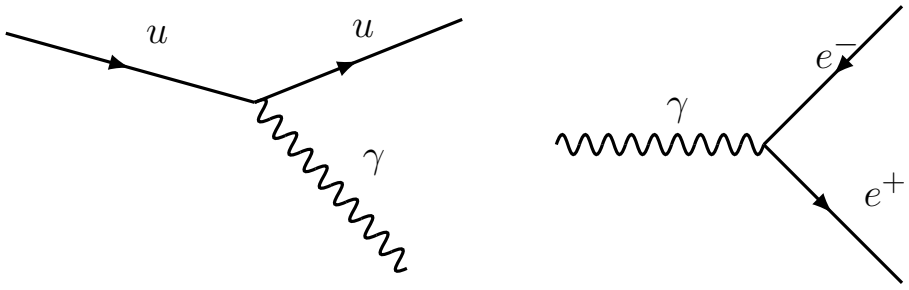
Finally, the complete Lagrangian for the QED is :

$$\boxed{\mathcal{L}_{\mathrm{QED}} = i\bar{\psi}\gamma^\mu\partial_\mu\psi + \mathrm{e}\bar{\psi}\gamma^\mu\psi A_\mu - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}} \tag{1.47}$$

Figure 1.3 gives the Feynman diagrams for Quantum Electro-Dynamics involving photon and electrons.

The Noether charge, which is conserved due to this QED gauge, is the electric charge, and the Noether current is $j^\mu = e\bar{\psi}\gamma^\mu\psi$. As a result, we obtain the classical Continuity equation.

$$\partial_\mu j^\mu = 0 \implies \frac{\partial\rho}{\partial t} + \nabla.\mathrm{j} = 0 \tag{1.48}$$



(a) (b)

Figure 1.3: Feyman Diagrams of different couplings in QED, (a) shows a quark radiating a photon and (b) shows pair production from a photon.

1.2.2 The Strong Force

In 1968, Deep Inelastic Scattering (DIS) experiments at SLAC demonstrated that protons have internal structure. It showed that the proton made 3 distinct fundamental particles. Today, we call them quarks. This means a new theory was required to explain a force much stronger than electromagnetism, which would explain the binding of these quarks to form mesons and baryons. A perfect theory was already in place to explain all this: Gell-Mann and Zweig's Quantum Chromodynamics (QCD). It introduces a new charge called colour for quarks. A quark could have one of 6 colours $(r, g, b, \bar{r}, \bar{g}, \bar{b})$, different from the electric charge, and they would be propagated by the gluon field. This is the birth of Quantum Chromodynamics (QCD). Since colour neutrality is necessary, there are no free quarks. They either bind with another quark with the opposite colour ($\bar{r}$, $g\bar{g}$ or $\bar{b}$) and form mesons or bind with 2 other quarks (rgb or $\bar{r}\bar{g}\bar{b}$) to form baryons[2]. What makes it even more different from QED is that gluons, unlike photons, do not need to be colour neutral, meaning gluons can self-interact. QCD is modelled on the $SU(3)$ group; hence, it has 8 vector bosons called gluons and 8 generators $(3^2 - 1)$ given by the Gell-Mann Matrices, denoted by λ_a such that $T^a = \frac{\lambda^a}{2}$. This gives us the complete QCD Lagrangian as:

$$\boxed{\mathcal{L}_{\text{QCD}} = i\bar{\psi}\gamma^\mu\partial_\mu\psi + g\bar{\psi}\gamma^\mu\psi\frac{\lambda_a}{2}A^a_\mu - \frac{1}{4}F^a_{\mu\nu}F^{\mu\nu}_a} \tag{1.49}$$

Recall from Equation 1.36, gluons would have self-interaction. This, combined with the large QCD coupling, yields a large gluon abundance. This makes the analysis in the proton-proton collision experiment trickier. We will address this in later chapters.

1.2.3 Weak Force and Electroweak Unification

The Weak interaction was introduced in nuclear physics to explain the β decay. $n \to p + e^- + \bar{\nu_e}$. Now, with the quark picture, we can reimagine this. A proton (p) is made of 2 u quarks and 1 d quark, and a neutron (n) is made of 2 d quarks

[2]CERN has recently discovered more exotic quark states like pentaquarks and hexaquarks.

and 1 u quark. So beta decay can be seen as a d quark transforming into a u quark, producing an e^- and a $\bar{\nu}_e$ pair. Figure 1.5 gives the Feynman diagram for this process.

In order to model the weak interaction, we can pair fermions into doublets that transform under the weak gauge group. Table 1.2 gives fermions paired into a doublet in the Standard Model.

Table 1.2: Standard Model fermions grouped into electroweak doublets

$$\begin{pmatrix} u \\ d \end{pmatrix}_L \quad \begin{pmatrix} c \\ s \end{pmatrix}_L \quad \begin{pmatrix} t \\ b \end{pmatrix}_L$$
$$\begin{pmatrix} \nu_e \\ e \end{pmatrix}_L \quad \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}_L \quad \begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}_L$$

So to describe the transformation of this doublet, the Weak interaction is modelled by the non-Abelian $SU(2)_L$ group. The number of generator in $SU(2)$ is 3 $(2^2 - 1)$. The L is for left-handed; this indicates that the $SU(2)$ group only transforms left-handed fermions. This was a necessary condition because it was experimentally proven that right-handed neutrinos didn't exist. Neutrinos, being chargeless and colourless, don't interact with photons and gluons. This leaves the

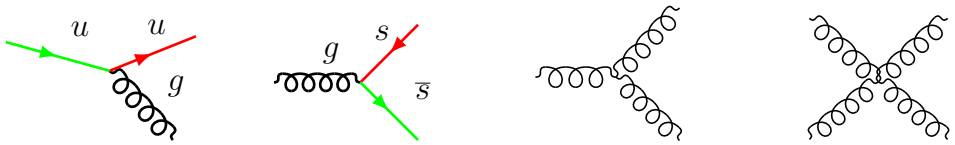


Figure 1.4: Feyman Diagrams of different couplings in QCD, (a) shows a quark radiating gluons, (b) shows pair production from gluons, (c) shows the cubic coupling of gluons and (d) shows the quartic gluon coupling.

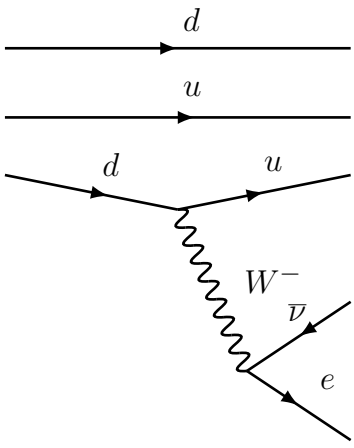


Figure 1.5: Feyman Diagrams of a neutron undergoing β decay to produce proton, electron and neutrino

weak interaction the only gauge field that couples with neutrinos. Hence, the absence of the right-handed neutrinos suggests that the Weak interaction is always left-handed. Let's call the component vector bosons as W^1, W^2 and W^3. Since it's a non-Abelian gauge theory, self-interactions between them are allowed. The generators T^a of the $SU(2)$ group are given by the Pauli matrices ($\frac{\sigma_a}{2}$).

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{1.50}$$

There is an issue with this picture: Notice that the difference in charge between them is 1, which means this action does not conserve electric charge. Meaning the weak decay should be propagated by charged bosons. This is where electroweak unification comes into play. Instead of treating QED and Weak as 2 different Lagrangians, we instead merge them into a single $SU(2)_L \times U(1)_Y$ group action, the suffix Y for $U(1)$ group is for Hypercharge, to distinguish it from $U(1)_{EM}$, which will arise from this after symmetry breaking (We will discuss this in detail in the coming sections). We can thus define the covariant derivative as :

$$\begin{aligned} D_\mu &= (\partial + ig\frac{\sigma_a}{2}W_\mu^a + ig'(I_2)B_\mu) \\ &= \partial + \frac{i}{2}\begin{bmatrix} gW_\mu^3 + g'B_\mu & gW_\mu^1 - igW_\mu^2 \\ gW_\mu^1 + igW_\mu^2 & g'B_\mu - gW_\mu^3 \end{bmatrix} \end{aligned} \tag{1.51}$$

Taking into account the kinetic terms for the bosons, we get the Lagrangian for the electroweak interaction :

$$\begin{aligned} \mathcal{L}_{\text{EW-unbroken}} = i\bar{\psi}\gamma^\mu\partial_\mu\psi + g\bar{\psi}\gamma^\mu\psi\frac{\sigma_a}{2}W_\mu^a + g'\bar{\psi}\gamma^\mu\psi B_\mu \\ - \frac{1}{4}W_{\mu\nu}^a W_a^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} \end{aligned} \tag{1.52}$$

The range of these weak interactions is very small, suggesting that the weak bosons are massive. This creates a unique problem: if we include the mass term in the Lagrangian, it will break gauge invariance, as we have demonstrated below.

Let M be the mass for the field W_μ, so the resulting mass term in the Lagrangian would be

$$\mathcal{L}_{\text{mass}} = \frac{1}{2}M^2 W_\mu W^\mu \tag{1.53}$$

If we apply the gauge transformation to Equation 1.53 we will get :

$$\frac{1}{2}M^2 W_\mu W^\mu \to \frac{1}{2}M^2(W_\mu - \partial_\mu\alpha)(W^\mu - \partial^\mu\alpha) \neq \frac{1}{2}M^2 W_\mu W^\mu \tag{1.54}$$

The gauge invariance issue also effects fermions mass term $m(\bar{\psi}_L\psi_R + \bar{\psi}_R\psi_L)$. Each of those has an L and an R part, which transform differently and are therefore not gauge-invariant.

1.2.4 Brout-Englert-Higgs mechanism

The Brout-Englert-Higgs mechanism (BEH)[30–32] was proposed by Robert Brout and François Englert, Peter Higgs and Gerald Guralnik, Carl R. Hagen, and Tom Kibble to introduce a mass term to massive bosons without breaking Lorentz and gauge invariance. The core idea of this theory was to propose a Goldstone boson, a massless complex scalar field that would spontaneously break the symmetry.

Higgs Potential

The Higgs Potential is modelled as a complex doublet Φ.

$$\Phi = \begin{pmatrix} \phi_+ \\ \phi_0 \end{pmatrix} = \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix} \implies \Phi^\dagger = \begin{pmatrix} \phi_+^* \\ \phi_0^* \end{pmatrix} = \begin{pmatrix} \phi_1 - i\phi_2 \\ \phi_3 - i\phi_4 \end{pmatrix} \tag{1.55}$$

With this, we can construct a gauge-invariant potential $V(\phi)$. To respect gauge invariance, $\Phi \to \Phi' = \Phi e^{i\alpha}$, all odd powers of Φ are discarded, and from renormalisation, all powers greater than 4 are also discarded. That gives us:

$$V(\Phi) = \lambda|\Phi^\dagger\Phi|^2 - \mu^2\Phi^\dagger\Phi \tag{1.56}$$

If we solve for the minima of this potential, we get

$$\Phi^\dagger\Phi = \phi_0^2 + \phi_+^2 = \frac{\mu^2}{2\lambda} \tag{1.57}$$

There are infinitely many solutions to this equation, meaning we can choose any one of them without losing generality. so we choose $\phi_+^2 = 0$ leaving $\phi_0^2 = \frac{\mu^2}{2\lambda}$. Again, without loss of generality, we can choose a real-valued field as ϕ_0 and hence define

$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \nu + H \end{pmatrix} \tag{1.58}$$

where ν is the vacuum expectation value such that $\nu = \frac{\mu}{\sqrt{\lambda}}$ and h is the Higgs fields. Thus, the ground state is :

$$\Phi_\nu = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \nu \end{pmatrix} \tag{1.59}$$

Now if we apply the potential $V(\Phi)$ on the newly defined Φ we get

$$\begin{aligned} V(\Phi) &= \frac{\lambda}{4}(\nu + H)^4 - \frac{\mu^2}{2}(\nu + H)^2 \\ &= \left(\frac{\lambda}{4}\nu^4 - \frac{\mu^2}{2}\nu^2\right) \quad + \nu(\lambda\nu^2 - \mu^2)H + \frac{3\lambda\nu^2 - \mu^2}{2}H^2 + \lambda\nu H^3 + \frac{\lambda}{4}H^4 \end{aligned} \tag{1.60}$$

Substituting $\mu^2 = \lambda\nu^2 = \frac{m_H^2}{2}$.

$$\boxed{V(H) \approx \frac{1}{2}m_h^2 H^2 + \lambda\nu H^3 + \frac{\lambda}{4}H^4} \tag{1.61}$$

Thereby getting the Higgs self-interaction and the Higgs boson mass terms.

In this thesis, we will extensively study the self-interaction of the Higgs boson and develop methods to accurately measure the Higgs Trilinear Coupling. For particle physics, rather than directly measuring a parameter, we often come across the κ framework, where we measure the ratio of the parameter to the expected SM value. In our case, that would be the κ_λ.

$$\kappa_\lambda = \frac{\lambda_{obs}}{\lambda_{SM}} \tag{1.62}$$

Spontaneous Breaking of symmetry

To explain how massive bosons interact with the Higgs, we can apply the Klein-Gordon equation with the Electroweak covariant derivative from Equation 1.51 :

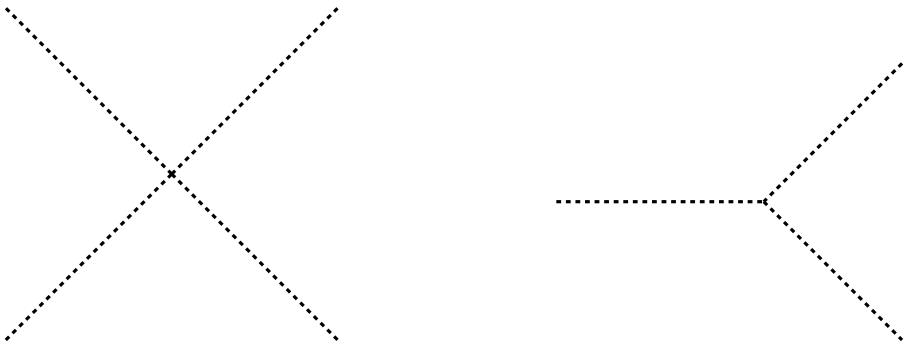
$$\begin{aligned} D_\mu \Phi &= (\partial + ig\frac{\sigma_a}{2}W_\mu^a + ig'(I_2)B_\mu)\begin{bmatrix} 0 \\ \frac{\nu+H}{\sqrt{2}} \end{bmatrix} \\ &= \frac{1}{\sqrt{2}}\begin{bmatrix} 0 \\ \partial^\mu H \end{bmatrix} + \frac{i}{2}\begin{bmatrix} gW_\mu^3 + g'B_\mu & gW_\mu^1 - igW_\mu^2 \\ gW_\mu^1 + igW_\mu^2 & g'YB_\mu - gW_\mu^3 \end{bmatrix}\begin{bmatrix} 0 \\ \frac{\nu+H}{\sqrt{2}} \end{bmatrix} \end{aligned} \tag{1.63}$$

Now we can redefine the 4 bosons as

$$\begin{aligned} A_\mu &= \frac{gW_\mu^3 + g'B_\mu}{\sqrt{g^2+g'^2}} \qquad Z_\mu = \frac{-gW_\mu^3 + g'B_\mu}{\sqrt{g^2+g'^2}} \\ W_\mu^- &= \frac{W_\mu^1 + iW_\mu^2}{\sqrt{2}} \qquad W_\mu^+ = \frac{W_\mu^1 - iW_\mu^2}{\sqrt{2}} \end{aligned} \tag{1.64}$$

We can swap the matrix with the newly defined bosons, and we will get

$$D_\mu \Phi = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 \\ \partial^\mu H \end{bmatrix} + \frac{\nu+H}{\sqrt{2}}\begin{bmatrix} (g\sqrt{2})W_\mu^+ \\ (\sqrt{g^2+g'^2})Z_\mu \end{bmatrix} \tag{1.65}$$



(a) (b)

Figure 1.6: Feyman Diagrams of self couplings of Higgs, (a) Trilinear Coupling $\lambda\nu H^3$ (b) Quartic Coupling $\frac{\lambda}{4}H^4$.

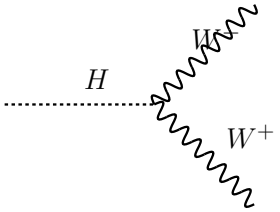
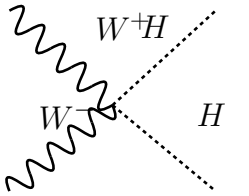
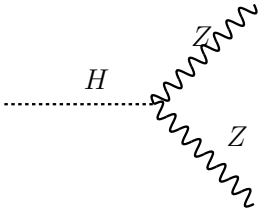
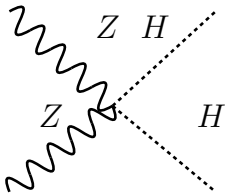
Now we can multiply this by the complex conjugate of the $D_\mu\Phi^\dagger$ to get the interaction of the bosons with Higgs boson and mass terms; $m_W = \frac{g\nu}{2}$ and $m_Z = \frac{\sqrt{g^2+g'^2}\nu}{2}$.

$$D_\mu\Phi^\dagger D_\mu\Phi = \frac{1}{2}\Box H + m_W^2 W_\mu^+ W^{\mu-} + \frac{g^2\nu}{2} H W_\mu^+ W^{\mu-} + \frac{g^2}{4} H^2 W_\mu^+ W^{\mu-} + m_Z^2 Z_\mu Z^\mu + \frac{(g^2+g'^2)\nu}{2} H Z_\mu Z^\mu + \frac{(g^2+g'^2)}{4} H^2 Z_\mu Z^\mu \tag{1.66}$$

Additionally, if we define a right triangle with sides $g\nu/2$ and $g'\nu/2$, we see that the hypotenuse equals m_Z. With that, we can define the Weinberg angle θ_w as

$$cos^2\theta_W = \frac{m_Z^2}{m_W^2} \tag{1.67}$$

Table 1.3: Terms in the Higgs Lagrangian after symmetry breaking

Term	**Description**	**Feynman Diagram**
$\frac{g^2\nu}{2} H W_\mu^+ W^{\mu-}$	HWW interaction (cubic)	
$\frac{g^2}{4} H^2 W_\mu^+ W^{\mu-}$	H^2WW interaction (quartic)	
$\frac{(g^2+g'^2)\nu}{2} H Z_\mu Z^\mu$	HZZ interaction (cubic)	
$\frac{(g^2+g'^2)}{4} H^2 Z_\mu Z^\mu$	H^2ZZ interaction (quartic)	

Yukava potential

A gauge-invariant fermionic mass term can be produced by fermions coupling with the scalar Higgs fields. Recall that in subsection 1.2.3 we addressed that left-handed fermions can be grouped into a doublet while right-handed fermions remain as singlets. From this, we can consider the Higgs field coupling to up (u) and down (d) quarks.

$$\mathcal{L}_{\text{Yukawa}} = -y_d \bar{Q}_L H d_R - y_u \bar{Q}_L \tilde{H} u_R + \text{h.c.} \tag{1.68}$$

where Q_L is the doublet of up (u) and down (d) quarks, $\tilde{H}$ is the Higgs conjugate defined as $\tilde{H} = i\sigma_2 H^*$ and h.c is short for hermitian congugate. We can expand this into its matrix for easier understanding.

$$\begin{aligned}\mathcal{L}_{\text{Yukawa}} &= -y_d \begin{bmatrix}\bar{u}_L & \bar{d}_L\end{bmatrix} \begin{pmatrix}0 \\ \frac{v+H}{\sqrt{2}}\end{pmatrix} d_R - y_u \begin{bmatrix}\bar{u}_L & \bar{d}_L\end{bmatrix} \begin{pmatrix}\frac{v+H}{\sqrt{2}} \\ 0\end{pmatrix} u_R + \text{h.c.} \\ &= \frac{-y_d}{\sqrt{2}} \bar{d}_L (v+H) d_R + \frac{-y_u}{\sqrt{2}} \bar{u}_L (v+H) u_R\end{aligned} \tag{1.69}$$

From this, we can get the mass terms as $\frac{-y_f \nu}{\sqrt{2}}$ and the coupling to the Higgs [33]. When we observe the equation carefully, we will notice that if a fermion doesn't have a right-handed term, then the mass term and the coupling to the Higgs boson cease to exist. This is the case for neutrinos. Unlike other fermions, neutrinos couple only to the weak force; there is no production mode for right-handed neutrinos. Ergo, neutrinos have no mass according to the Standard Model. Many experiments measuring neutrino oscillation have confirmed that neutrinos have a very small mass, hence violating the Standard Model. Models like the See-Saw model predict a very massive right-handed neutrino to explain neutrino mass and neutrino oscillations. So far, there hasn't been any experimental validation of such theories.

Now we can finally put together all of the Standard Model into one langraian, the langraian of the theory of almost everything.

$$\begin{aligned}\mathcal{L}_{\text{SM}} = &-\frac{1}{4} G^a_{\mu\nu} G^{a\mu\nu} - \frac{1}{4} W^i_{\mu\nu} W^{i\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} \\ &+ \sum_f \bar{\psi^f} i\gamma^\mu D_\mu \psi^f + (D_\mu H)^\dagger (D^\mu H) \\ &+ \frac{1}{2} m_h^2 H^2 + \lambda \nu H^3 + \frac{\lambda}{4} H^4 \\ &- \sum_f y_f \bar{\psi^f}_L H \psi^f{}_R + \text{h.c.}\end{aligned} \tag{1.70}$$

1.3 Higgs and Di-Higgs in LHC

In the year 2012, a new massive scalar Boson of mass 125GeV was discovered in the CMS and ATLAS detectors of the Large Hadron Collider of CERN in Geneva, Switzerland [35, 36]. The properties measured of this newly discovered particle were consistent with the Standard Model Higgs boson. This discovery gave experimental confirmation to the BEH mechanism. The discovery method was based on the $H \to ZZ^* \to 4\ell$ and $H \to \gamma\gamma$ decay channels, as these provided the cleanest signal reconstruction.

1.3.1 Higgs Production and Decay

In the previous section, we discussed in detail how the Higgs boson interacts with all massive particles with a coupling proportional to the mass of the fermions and to the square of the mass for bosons. This can be seen from the branching fraction of the Higgs, as shown in Figure 1.9a. Following this, we can now address the most observed production methods and decay channels for the Higgs bosons. Only by measuring these channels can we completely validate the BEH mechanisms. Figure 1.9b summarises the production modes of Higgs bosons in proton-proton

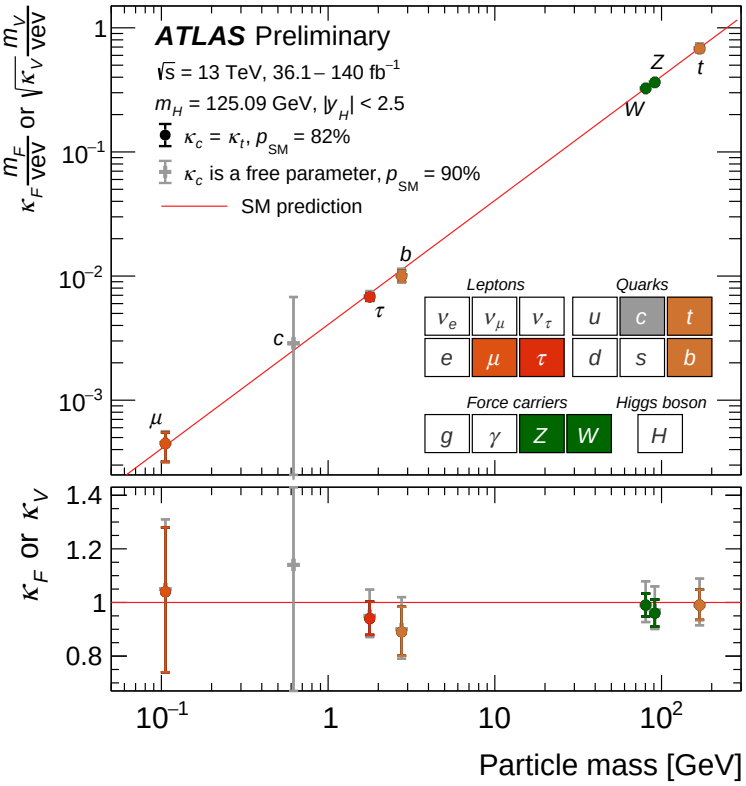
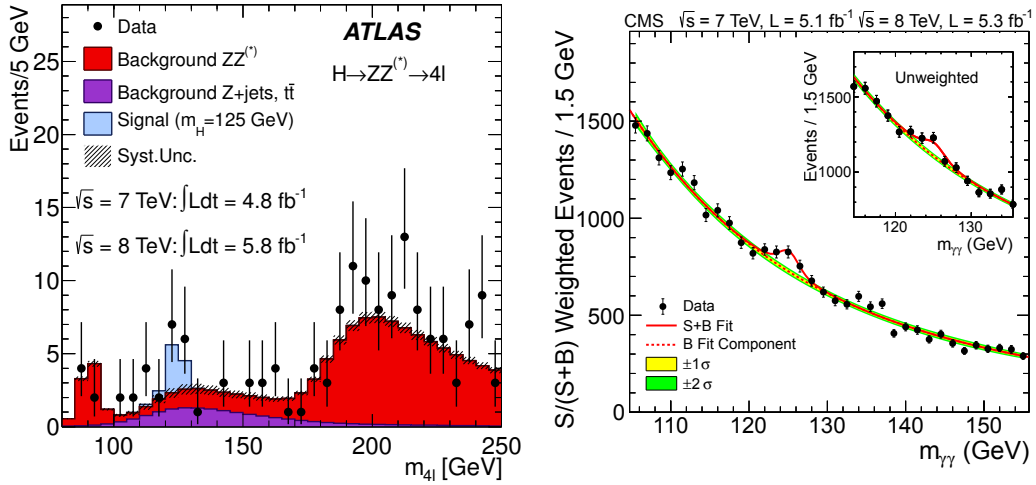


Figure 1.7: Reduced Higgs boson coupling strength modifiers (κ) as a function of their respective particle mass, and their experimental uncertainties[34].

collision experiments, such as those at the LHC.

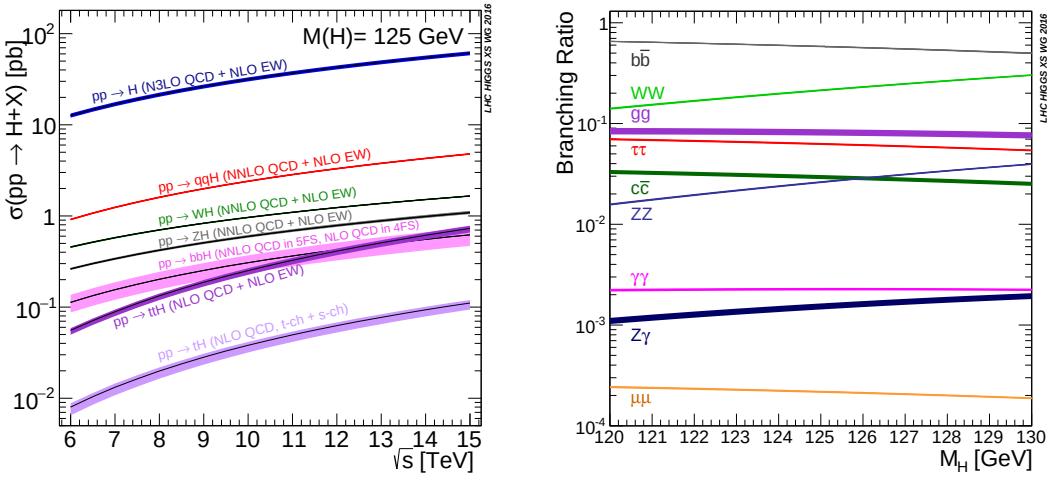
The following are the major production modes of Higgs bosons.

- **gluon- gluon Fusion (ggF):** ggF has the highest cross-section among all the production modes of higgs. This is thanks to the strong coupling to fermions, combined with the large coupling of the Higgs boson to the top quark. The



(a) (b)

Figure 1.8: **Higgs Boson Discovery in ATLAS and CMS**. a) Histogram of 4 lepton invariant mass for Higgs boson decaying to int 4 leptons via a pair of Z bosons and their backgrounds as measured by ATLAS[35]. b)Histogram of $\gamma\gamma$ invariant mass for Higgs bosons decaying to 2 photons (γ) measured by CMS[36]



(a) (b)

Figure 1.9: The production modes of Higgs bosons in proton-proton collision experiments at the LHC as a function of centre-of-mass energy ($\sqrt{s}$) and (b) and branching fraction of Higgs bosons as a function of invariant mass of the Higgs boson (M_H) [37]

Feynman diagram of this is given in Figure 1.10 (a).

- **Vector Boson Fusion (VBF):** In this process, the quarks from the protons radiate vector bosons ($W^{\pm}$, Z^0) which fuse to form a Higgs boson. This channel's signature is two forward jets. The Feynman diagram of this is given in Figure 1.10 (b).

- **Vector boson associated production (VH) (Higgsstrahlung):**. Here, a pair of quarks first creates a vector boson, which then radiates a Higgs boson. The Feynman diagram of this is given in Figure 1.10 (c).

- **Top (bottom) quark associated production (ttH + tH, bbH + bH):** Similar to ggF, this process also leverages the strong coupling and large coupling of the Higgs boson to top quarks, but unlike ggF, this involves creating top (bottom) quarks; hence it is suppressed by phase space. The Feynman diagram of this is given in Figure 1.10 (d,e).

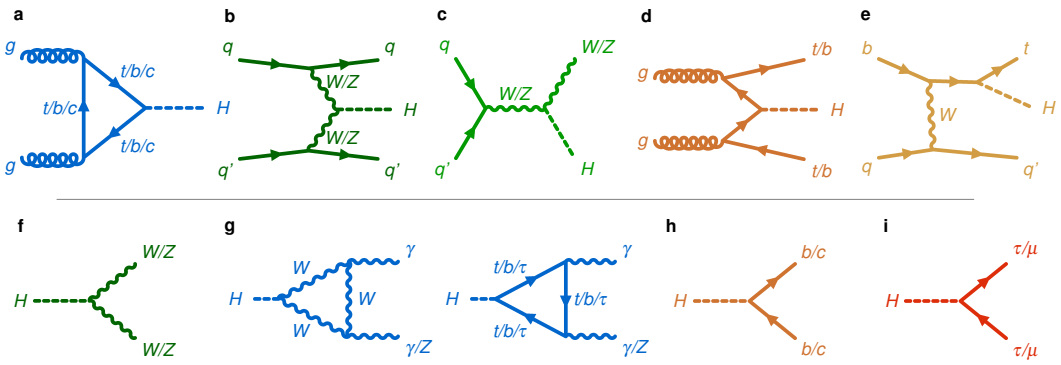


Figure 1.10: Feynman Diagrams of different production modes and decay channels of Higgs, Production: (a) gluon-gluon fusion (ggF), (b) vector boson fusion (VBF), (c) vector boson associated production (VH), (d) top quark associated production (ttH), (e) bottom quark associated production (bbH). Decay: (f)$H \to W/ZW/Z$ (g) $H \to \gamma\gamma$ (h) $H \to b\bar{b}$, and (i) $H \to \tau^+\tau^-$. [6]

In addition to the branching ratio, another factor that helps determine which channels are best to analyse is reconstruction and tagging efficiency. Channels which produce final states as leptons or photons are relatively easier to analyse as they are less affected by the large QCD background found in proton-proton collision experiments. This made $H \to ZZ$ and $H \to \gamma\gamma$ the first to be discovered. Following these were the $H \to W^{\pm}W^{\mp}$ and $H \to \tau^+\tau^-$, both of which suffered from missing energy from neutrinos, requiring the use of multivariate analysis like Boosted Decision Trees (BDT).[3] Following more developments in Jet reconstruction and tagging led to discoveries $H \to b\bar{b}$ and ttH.

1.3.2 From Higgs to Di-Higgs

As in the case of the Single Higgs, Different production modes give us access to different SM couplings. A recurring pattern for Di-Higgs production is the formation of a single Higgs, followed by a trilinear coupling. Equation 1.62 gives the expression of reduced trilinear coupling constant κ_λ in terms of trilinear coupling (λ). The following is the list of productions in descending order of cross-section. Equation 1.62 gives the

- **gluon- gluon Fusion (ggF):** In this production mode, we have the previously mentioned ggF single Higgs decaying to two other Higgs (Figure 1.12a) and the box diagram (Figure 1.12b). But what makes Di-Higgs interesting is that these two processes share the same initial and final states; hence, the observed Di-Higgs state is a superposition of the two diagrams. The core part of this thesis is to understand the consequences of this superposition and to develop methods to overcome the challenge.

- **Vector Boson Fusion (VBF):** In VBF production mode, we have three Di-Higgs final states, which could interfere with each other. The previously mentioned VBF single Higgs decaying to two other Higgs (Figure 1.13a), two vector boson fusions producing each producing one Higgs (Figure 1.13b) and Di-Higgs from the quadrilinear coupling of Higgs boson to vector boson (Figure 1.13c)

- **Top (bottom) quark associated production (ttH + tH, bbH + bH):** In ttHH production mode, we have three Di-Higgs final states. The previously mentioned ttH single Higgs decaying to two other Higgs (Figure 1.14a), and

[3]We will discuss more details of the role of BDTs and NN in LHC analysis in chapter 4.

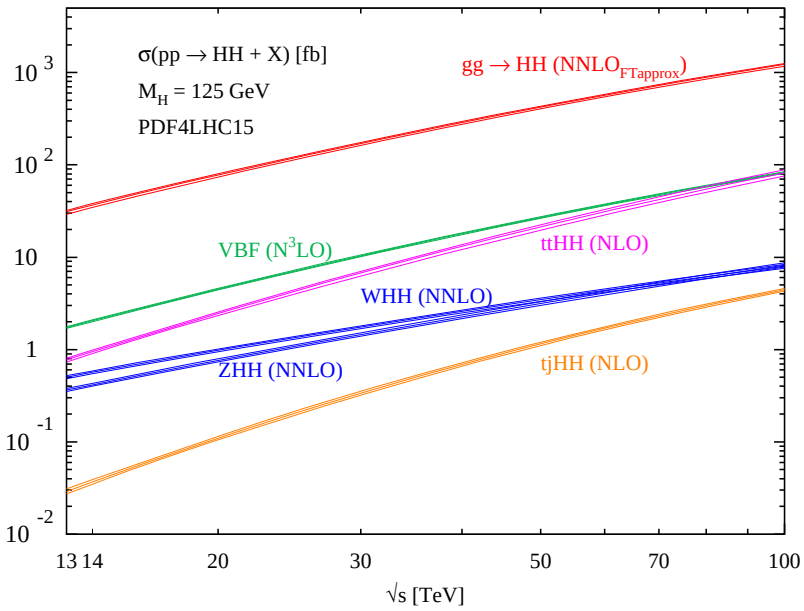


Figure 1.11: $pp \rightarrow HH + X$, production cross-section in various channels as a function of centre-of-mass energy ($\sqrt{s}$)[7].

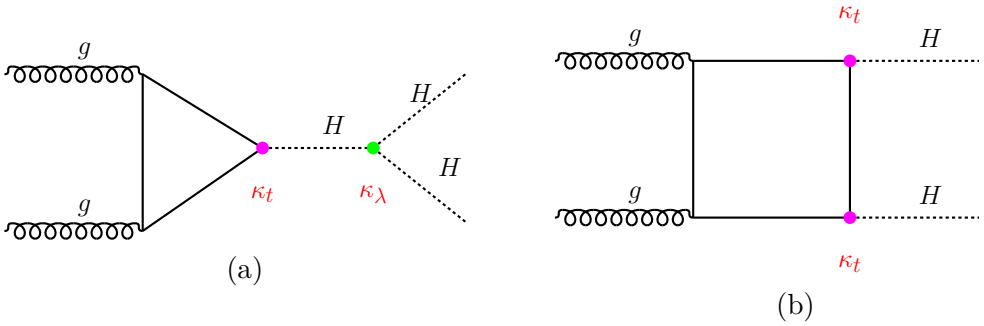


Figure 1.12: Production of Higgs boson pairs through gluon-gluon-fusion (ggF) with trilinear Higgs boson self-coupling (κ_λ (triangle) (a) and involving only top quark loop (box) (b) contributions.

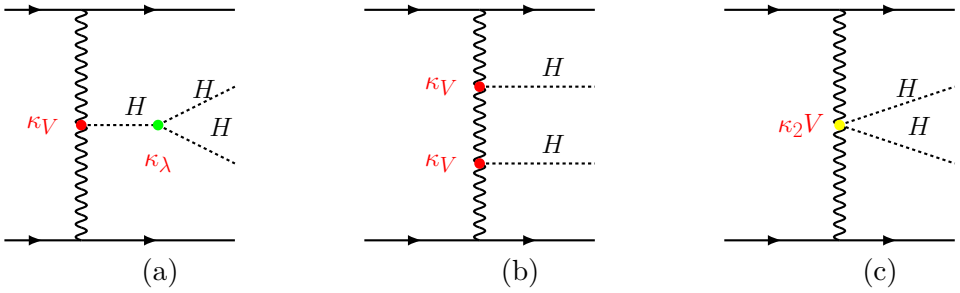


Figure 1.13: Production of Higgs boson pairs through vector-boson-fusion (VBF) with trilinear Higgs boson self-coupling (κ_λ) (a), involving 2 Higgs bosons produced without self-coupling (b) and involving VVHH coupling (κ_{2V}) (c) contributions.

2 diagrams from ($q\bar{q} \rightarrow t\bar{t}$) where the tops radiate one or more single Higgs boson (Figure 1.14b, Figure 1.14c)

- **Vector boson associated production (VH) (Higgsstrahlung):**. Here, a pair of quarks first creates a vector boson, which then radiates a Higgs boson. There are 3 possible diagrams, one involving a Trilinear Higgs coupling (Figure 1.15a), and the other two are similar to the VBF case (Figure 1.15b, Figure 1.15c).

Table 1.4 gives summary for predicted cross-section for center of mass energy $\sqrt{s} = 13$ TeV.

1.3.3 Decay Channels for Di-Higgs Analysis

Figure 1.16 gives the branching fraction for Di-Higgs for different decay channels based on the branching fraction of single Higgs as seen in Figure 1.9b. Four major channels are analysed for Di-Higgs, based on branching fractions and reconstruction and tagging efficiencies.

- $HH \rightarrow b\bar{b}b\bar{b}$: The channel with the highest branching fraction due to the large branching fraction of Higgs to $b\bar{b}$. The major challenge in this channel is tagging the 4 b jets and background rejection.

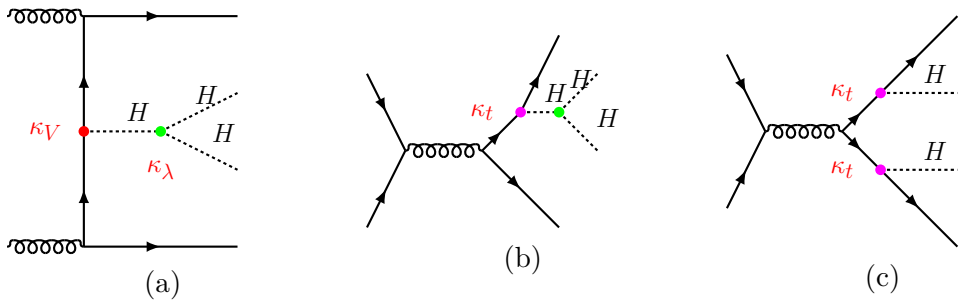


Figure 1.14: Production of Higgs boson pairs in association with a top–antitop quark pair ($pp \to t\bar{t}HH$), including contributions sensitive to the trilinear Higgs boson self-coupling (κ_λ)(a) and (b) and involving 2 Higgs bosons produced without self-coupling (c).

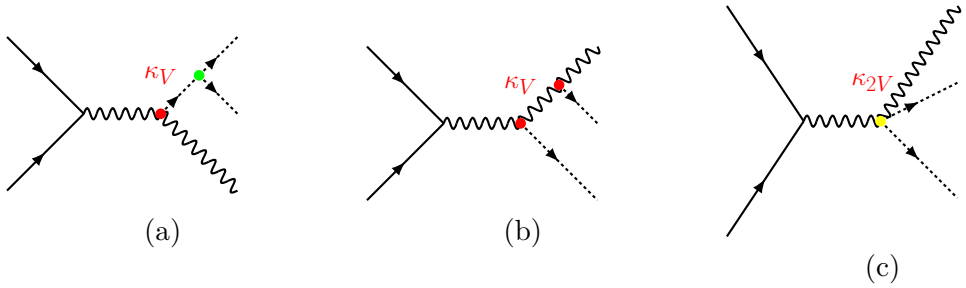


Figure 1.15: Production of Higgs boson pairs in association with a vector boson ($pp \to VHH$, where $V = W, Z$), including contributions sensitive to the trilinear Higgs boson self-coupling (κ_λ)(a) and involving 2 Higgs bosons produced without self-coupling (b) and involving VVHH coupling (κ_{2V}) (c).

- $HH \to b\bar{b}\gamma\gamma$: This channel has the lowest among the four in terms of branching fraction, but it makes up for it in Tagging efficiency and background rejection, due to the clean $\gamma\gamma$ signal.
- Multilepton: Another channel with a clean signal is the Multilepton channel. It has a higher branching fraction than $HH \to b\bar{b}\gamma\gamma$, but because neutrinos are present, energy reconstruction is incomplete, which affects the overall efficiency.
- $HH \to b\bar{b}\tau^+\tau^-$: This channel aims to strike the middle ground between tagging efficiency and branching fraction. This channel also suffers from missing energies, but more sophisticated analysis tools, such as the Graph Attention Network, are used to address this issue. This channel is the focus of this thesis. We will explore this in chapter 6

Figure 1.17 gives the current exclusion limits on Trilinear coupling κ_λ from the latest ATLAS combined analysis[38]. We can see that the two major contributions to this come from $HH \to b\bar{b}\tau^+\tau^-$**:** and $HH \to b\bar{b}\gamma\gamma$**:** Hence, improving in these channels with modern Machine Learning tools like Neural Simulation Based Inference would significantly improve our overall estimate of κ_λ and by extension the

Higgs potential.

Process	σ_{NNLL} [fb]	Theoretical uncertainties
Gluon fusion	$33.49\ ^{+4.3\%}_{-6.0\%}$	$\pm 2.1\%$ (PDF) $\pm 2.3\%$(α_s) $\pm 5.0\%$ (top)
VBF	$1.62\ ^{+2.3\%}_{-2.7\%}$	$\pm 2.3\%$ (PDF + α_s)
$t\bar{t}HH$	$0.772\ ^{+1.7\%}_{-4.5\%}$	$\pm 3.2\%$ (PDF + α_s)
W^+HH	$0.329\ ^{+0.32\%}_{-0.41\%}$	$\pm 2.2\%$ (PDF + α_s)
W^-HH	$0.173\ ^{+1.2\%}_{-1.3\%}$	$\pm 2.8\%$ (PDF + α_s)
ZHH	$0.362\ ^{+3.4\%}_{-2.6\%}$	$\pm 1.9\%$ (PDF + α_s)
$tjHH$	$0.0281\ ^{+5.2\%}_{-3.2\%}$	$\pm 4.5\%$ (PDF + α_s)

Table 1.4: Cross sections and theoretical uncertainties for Higgs pair production processes values taken from "Handbook of LHC Higgs cross sections" [37]

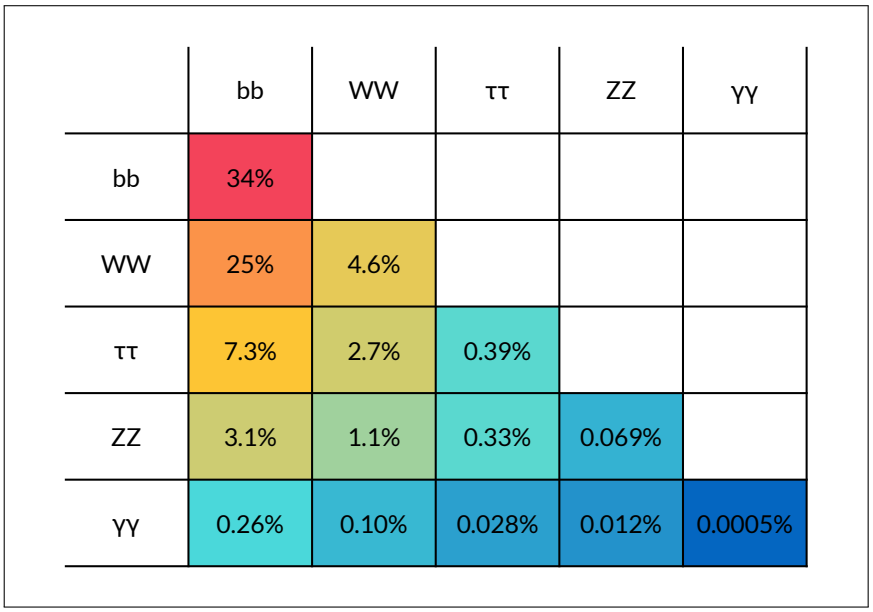


Figure 1.16: Branching fraction for Di-Higgs for different decay channels

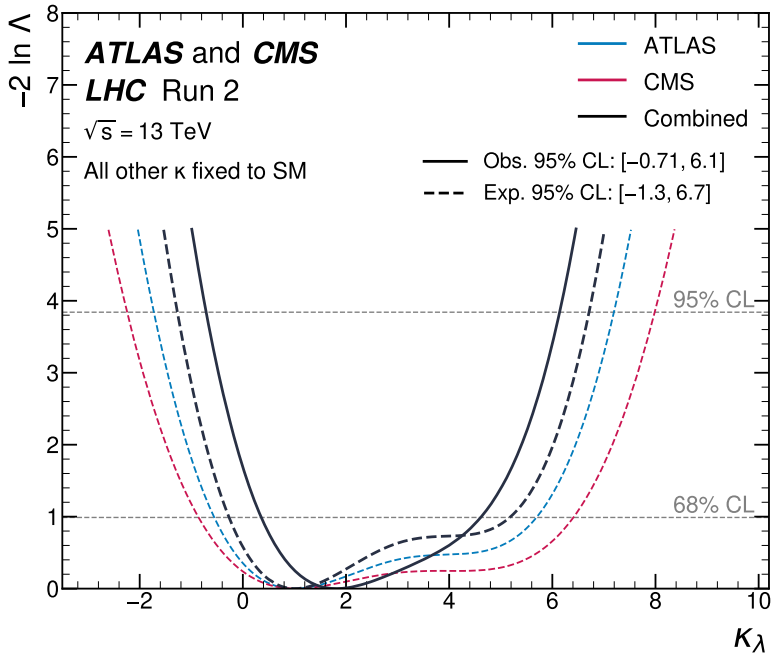


Figure 1.17: Combined ATLAS + CMS κ_λ likilihood scan and their exclusion limits[39]

Chapter 2
Experimental Apparatus

Experiment is the sole judge of scientific "truth." But what is the source of knowledge? Where do the laws that are to be tested come from? Experiment, itself, helps to produce these laws, in the sense that it gives us hints.
- Richard P. Feynman

To properly test a theory as elegant as the Standard Model of particle physics, we require an equally sophisticated apparatus. The Large Hadron Collider and the experiments it hosts are just that. The work described in this thesis was carried out with the ATLAS Experiment, which utilises highly energetic proton beams generated by the Large Hadron Collider. In this chapter, we will briefly explore the workings of the Large Hadron Collider and the ATLAS[40] experiment.

2.1 Large Hadron Collider

The Large Hadron Collider (LHC)[41, 42] is a 27 km-long circular particle accelerator that accelerates protons and ions to nearly the speed of light. Located on the Swiss-French border, this particle accelerator can collide protons at 13.6 TeV, the highest energy achieved so far. The LHC also collides heavy ions, such as Lead and Xenon, as well as light ions, such as Oxygen and Neon. For this thesis, we will focus on the proton-proton collision in the LHC.

2.1.1 Acceleration Process

The journey of the proton in the LHC begins with a bottle of hydrogen gas (H_2). The hydrogen gas is then ionised to H^- (1 proton, 2 electrons) and sent to LINAC4[43], a linear accelerator that accelerates these ions to 160 MeV. The electrons of the H^- are stripped after this, leaving H^+ or just protons. This is followed by the Booster, a circular synchrotron that accelerates protons to 2 GeV. The proton then goes to the Proton Synchrotron (PS), where its energy is boosted to 26 GeV. Lastly, before reaching the LHC, the proton passes through the Super Proton Synchrotron (SPS), which raises the beam to an injection energy of 450 GeV and then injects it into the LHC rings as counter-rotating beams. From here, the LHC further accelerates the beam to the final energy of 6.8 TeV. LHC uses superconducting dipole magnets to accelerate the protons and quadrapole magnets

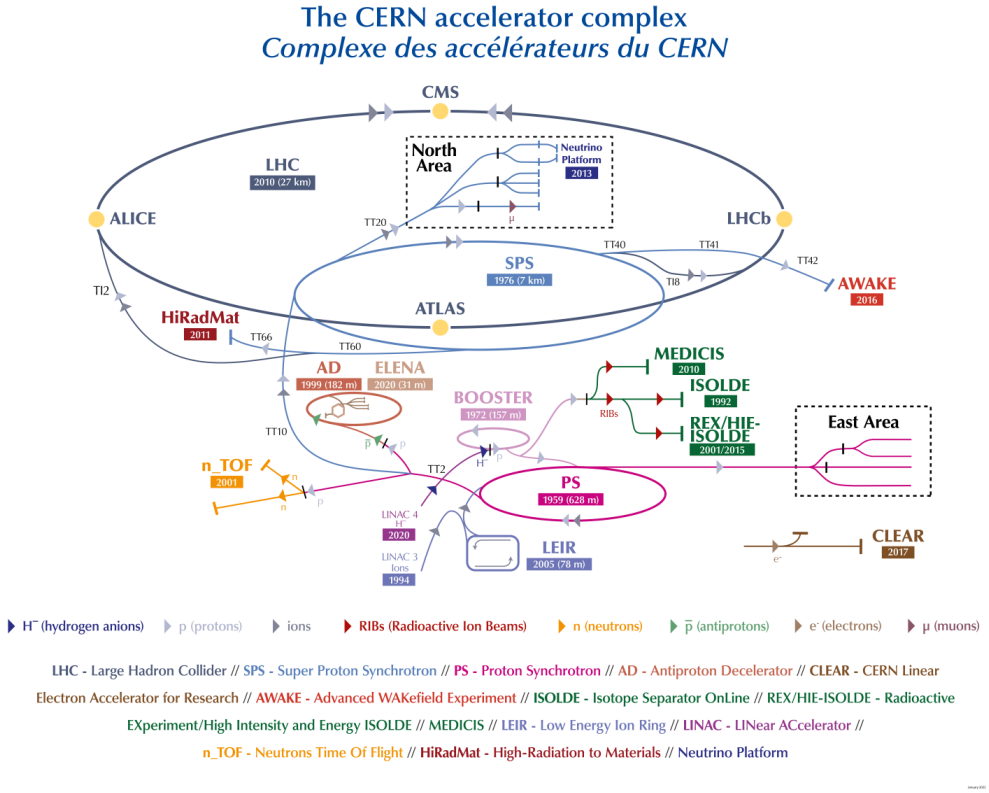


Figure 2.1: CERN Accelerator Complex [44]

to guide and focus the beams. The LHC also uses higher-order multipole magnets to correct the proton beam's path. In proton-proton collisions produced by the LHC, bunch crossings occur every 25 ns, corresponding to a collision frequency of 40 MHz.

2.1.2 Luminosity

Luminosity is the measure of the intensity of the particle collider. Formally, we can define Instantaneous Luminosity ($\mathcal{L}$) as the rate of particle collisions (N) per unit cross-section (σ).

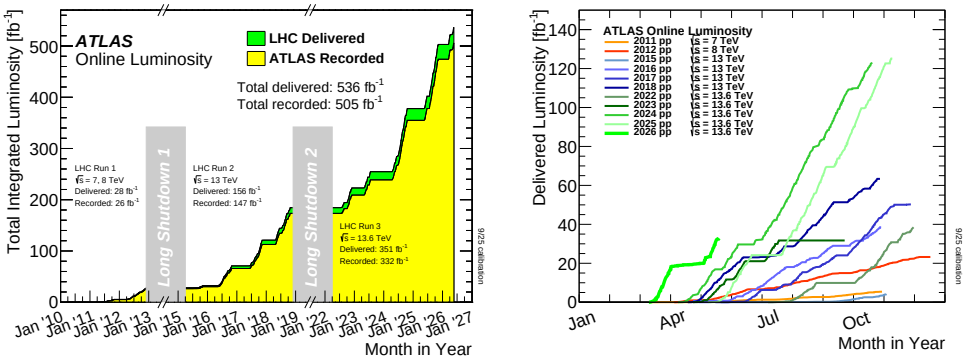
$$\mathcal{L} = \frac{1}{\sigma}\frac{dN}{dt} \tag{2.1}$$

The time integral of the instantaneous luminosity gives the Integrated Luminosity (L). This depends on various accelerator parameters, the number of protons in each bunch (n_1,n_2), N_b the number of bunches in the beam, f_{rev} the revolution frequency of the LHC, the RMS crossection of the colliding bunches (σ_1,σ_1) and $\mathcal{F}$ the geometric factor to account for the non-zero crossing angle between the beams. The general expression for the instantaneous luminosity $\mathcal{L}$ is given by :

$$L = \frac{n_1 n_2 f_{rev} N_b}{4\pi\sigma_1\sigma_2}\mathcal{F} \tag{2.2}$$

Another necessary quantity in the context of luminosity is **pile-up**. Pile-up denoted by μ is defined as the number of collisions in a bunch crossing. Since our

interest is in high-energy collisions or hard processes, a large pile-up would increase contamination, which must be addressed in reconstruction. Due to sustained R&D in accelerator physics over the last decade, the LHC has gradually increased the instantaneous luminosity. The integrated luminosity per year is given in Figure 2.2.



(a) (b)

Figure 2.2: Integrated Luminosity over the years up to the end of Run 3 (2026). (a) Total recorded luminosity as a function of years and (b) Integrated luminosity as a function of months for different data-taking years[45].

2.1.3 Experiments within the LHC

The LHC has 4 interaction points (IP), each hosting major experiments.

- IP1 - (**ATLAS** (**A** **T**oroidal **L**HC **A**pparatu**S**): ATLAS is a general-purpose detector which has a wide range of physics goals like measurement of Higgs properties, Vector-Boson Scattering, Dark matter, SUSY and other Beyond Standard Model physics searches.

- IP2 - **ALICE** (**A** **L**arge **I**on **C**ollider **E**xperiment): ALICE[8] experiment is dedicated to the study of Quark Gluon Plasma (QGP) through heavy ion collision. Recently, the ALICE experiments have also begun studying light-ion collisions, such as those involving oxygen, to better understand nuclear structure.

- IP5 - **CMS** (**C**ompact **M**uon **S**olenoid): CMS[9], just like ATLAS, is a general-purpose detector, but uses very different technologies in its subdetectors. CMS also does similar studies to ATLAS. This enables ATLAS and CMS to validate each other's physics discoveries independently.

- IP8 - **LHCb** (**L**arge **H**adron **C**ollider **b**eauty) : LHCb[10] is forward only detector. LHCb focuses on flavour physics and hence has a dedicated vertex detector, VELO, and a particle-ID detector, the RICH.

2.1.4 LHC Operations

The operational period is divided into Runs. Each Run lasts roughly 2 to 3 years, during which most of the technology used remains fixed, ensuring consistent

physics results. Each Run is followed by a long shutdown for major upgrades and maintenance. LHC operates for about 8 months a year and is shut down for the winter. During these shutdowns, maintenance work is carried out for the experiments and the accelerator facilities themselves. A brief summary of what is achieved or expected to be achieved in each run is given below.

- **Run 1** (2010 - 2012): The very first Run of the LHC, the beam was set to a centre-of-mass energy of 7 TeV. This run had an integrated luminosity of 30fb^{-1} and an average pileup of 21. Run I was marked by the Higgs Boson discovery at the ATLAS and CMS experiments, cross-section measurements for W and Z bosons, and top-quark production.

- **Run 2** (2015 - 2018): Following the upgrades in the shutdown, the LHC had a collision energy of 13TeV with an integrated luminosity of 140fb^{-1} and an average pileup of 34 In Run II, we enter the precision measurement era, with the discovery of $H \to b\bar{b}$, $H \to \tau^+\tau^-$, and $t\bar{t}H$ production at the ATLAS and CMS experiments. The LHCb experiment found hints of lepton-flavour universality violation, and the ALICE experiment showed strong evidence that the quark–gluon plasma behaves as a near-perfect fluid.

- **Run 3** (2022 to present): The centre-of-mass energy was now brought up to 13.6 TeV, and the integrated luminosity as of 2025 is at 56fb^{-1} and an average pileup of 55. Run III is focusing on rare channels like $H \to Z\gamma$, Higgs Pair Production, and EFT interpretations of anomalous couplings.

- **Run 4 and 5** (2030 - 2045): This would be the "High Lumi" phase or the **HL-LHC** It is expected that the integrated luminosity would be close to 3000fb^{-1}, with an average pileup of 200 by the end The phase is expected to shed light on very rare processes and to constrain parameters such as the Higgs self-coupling to very high precision Due to the higher pileup, Major upgrades are required for the experiments The upgrades to the ATLAS detector for HL LHC will be discussed in section 2.3.

Figure 2.3 gives a summary of various runs, including the HL-LHC.

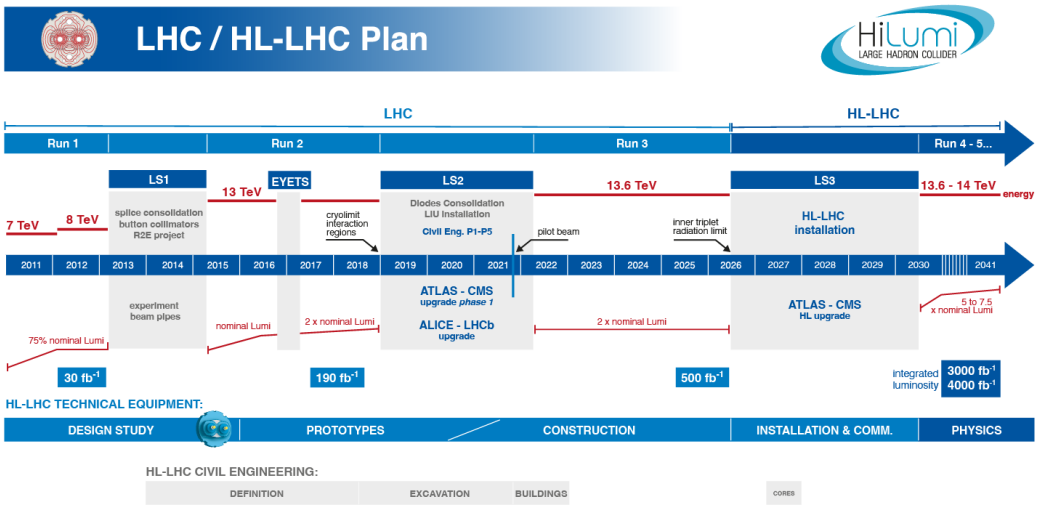


Figure 2.3: Summary of operations and upgrades of LHC and Hi-Lumi LHC[46]

2.2 ATLAS Detector

The purpose of a particle physics detector is to reconstruct the products of the hard process correctly. This is done using multiple subdetectors to identify and study different particle types accurately. It usually includes a tracker to track charged particles and calorimeters to measure particle energy. Sometimes an extra muon detector to identify muons The ATLAS detector also follows this rule It's a general-purpose cylindrical detector covering almost 4π-solid angle[1] The ATLAS detector is also the largest of its kind, spanning 46m in length and 25m in diameter The schematic view of the ATLAS detector is given in Figure 2.4.

2.2.1 Coordinates in Cylindrical detectors

Due to the Cylindrical symmetry of the ATLAS detector and the energy scale of the events measured, the coordinates used by my ATLAS are z, η and ϕ The z and ϕ are the same as the standard cylindrical coordinate system of (z, θ, ϕ), where z is the distance along the beam line and ϕ corresponds to the angle around the beam line Instead of θ, which would be the angle along the plane containing the beam line, we use a quantity called pseudorapidity given by η The relation betwen pseudorapidity η and angle θ is given by:

$$\eta = -\ln\left(\tan\left(\frac{\theta}{2}\right)\right) \tag{2.3}$$

Pseudorapidity of 0 corresponds to the direction orthogonal to the beam pipe and pseudorapidity of $\pm\infty$ corresponds to the directions of the beams The ATLAS detector has a coverage of $|\eta| \leq 4.9$ For ease of construction and maintenance, the

[1]Excluding support structures and the beampipe

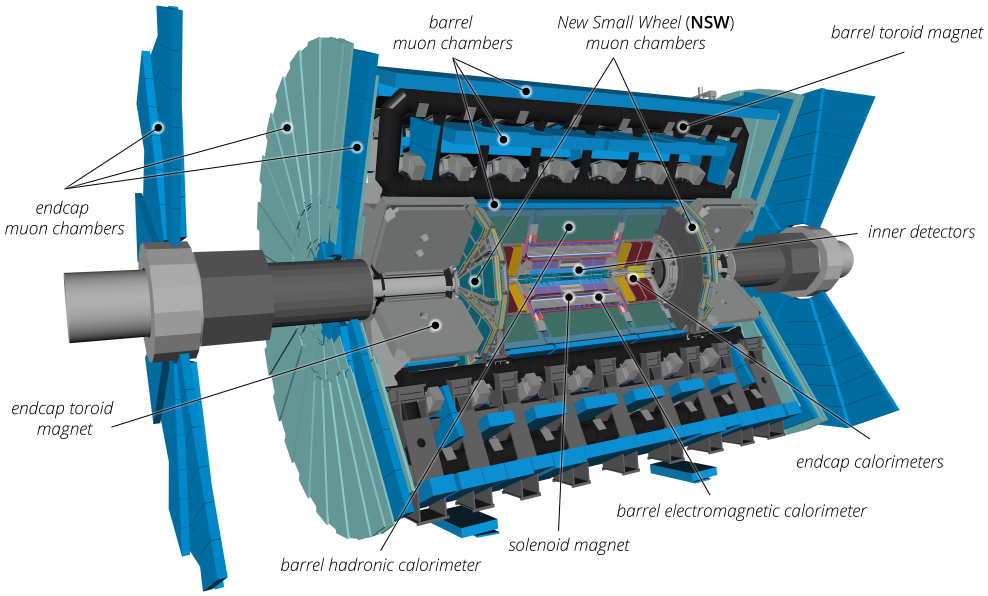


Figure 2.4: The schematic diagram of the ATLAS Detector with Run 3 configuration, indicating the locations of the detector sub-systems[47]

sub-detectors within ATLAS are also divided into endcaps for the forward region (higher $|\eta|$) and a barrel for the central region (lower $|\eta|$. In the following subsections, we will discuss the various sub-detectors, their functions, and their construction.

2.2.2 Inner Detector (ID)

One of the most powerful tools for reconstructing a particle's properties is its track With a particle's track parameters, we can determine properties such as its decay chain and charge Hence, having a good tracking detector and a suitable tracking algorithm is crucial for obtaining accurate measurements in the experiment The Inner Detector is the main tracker for the ATLAS detector The ID is surrounded by a 2T solenoid magnet that curves the path of charged particles based on their momentum and electric charge The ID is 6.2m in length, and 2.1m in diameter and covers a pseudorapidity ($|\eta|$) of 2.5. The components of the ID are described below from the innermost to the outermost

- Insertable B-Layer (IBL): Placed closest to the beam pipe, the Insertable B-Layer (IBL), as its name suggests, helps detect B-mesons. The B-meson doesn't have a sufficiently long lifetime to be detected directly; its decay products must be analysed to detect it. The IBL also play a key role in ambiguity resolution, a pivotal part of track reconstruction. We will discuss this in more detail in chapter 3. The IBL was introduced into the ATLAS detector during the first long shutdown.

- Pixel detector: The pixel detector is found around 3.3 cm from the beamline It is composed of 4 layers of pixels, arranged as concentric cylinders in the barrel region, and 3 disks of pixels at each endcap These pixel cells are made of silicon

doped to function as reverse-biased p-n diodes When charged particles pass through them, they create an electron-hole pair, which leads to a detectable electric signal.

- Semi-Conductor Tracker (SCT): Placed around the Pixel detector is the Semi-Conductor Tracker (SCT). Similar to a Pixel detector, it's also made of silicon but is segmented into strips, meaning each layer provides only 1D information.

- Transition Radiation Tracker (TRT): The last detector before the calorimeters is the Transition Radiation Tracker (TRT) The former 3, TRT is the gasious detector It's composed of 300k $4mm$ dimameter polypropylene tubes filled with $Xe - CO_2 - O_2$ or $Ar - CO_2 - O_2$ gas and has a tungsten wire passing through its centre When charged particles pass through them, they ionise the gas within the tube, creating a detectable electric signal.

Figure 2.6 and Figure 2.5 shows the schematic diagrams of the Inner Dectector (ID). The inner detector is scheduled to be replaced at the end of 2026 This will be replaced by a fully silicon detector called the Inner Tracker (ITk)[48] This is in preparation for HL-LHC, which will have 4 times larger pile-up than current runs, hence requiring more precise detectors See subsection 2.3.1 for more information The ITk will also include a new tracking algorithm based on the ACTS framework, which is elaborated in detail in chapter 3.

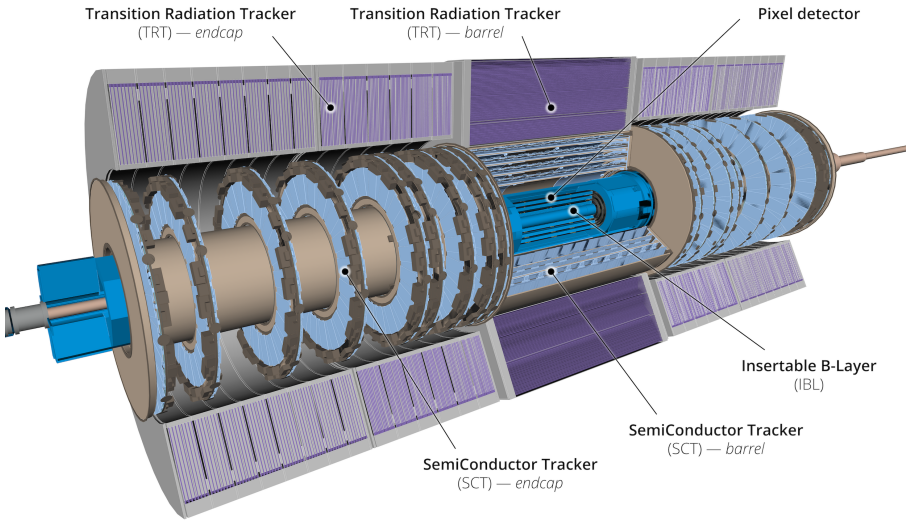


Figure 2.5: Schematic diagram of the ATLAS Inner Detector.[47]

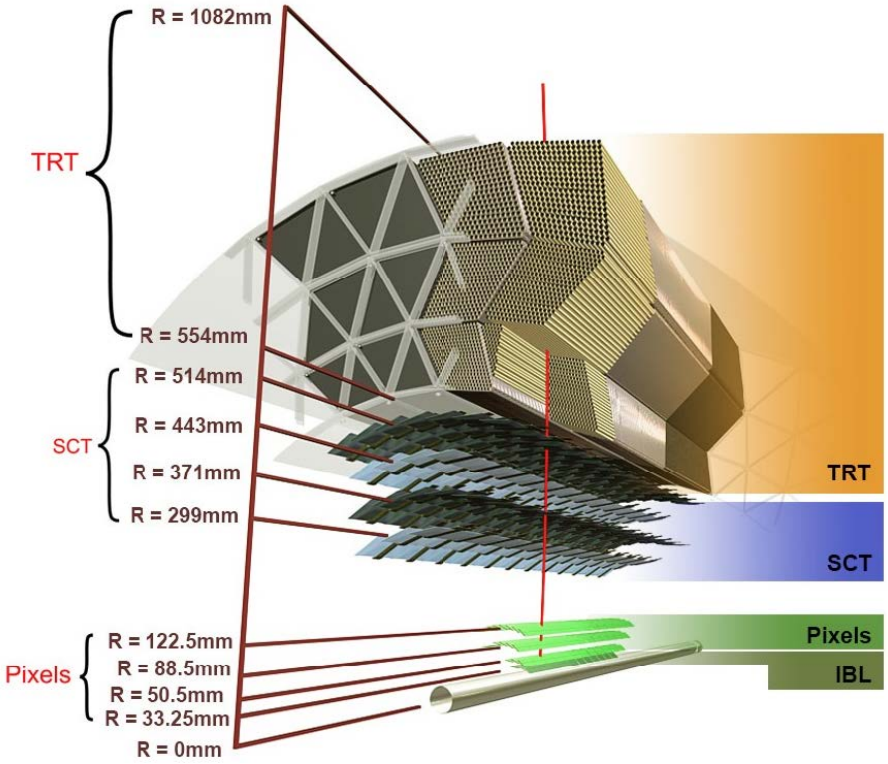


Figure 2.6: Cross-sectional diagram of the ATLAS Inner Detector.[47]

2.2.3 Calorimeters

One of the most essential pieces of information in a HEP experiment is the energy of the produced particle This is the job of the calorimeters The energy resolution $\frac{\sigma(E)}{E}$ of a calorimeter is the quadratic sum of 3 terms: the stochastic term $\frac{S}{\sqrt{E}}$, which depends on the sampling and intrinsic fluctuations of the shower, the Noise term $\frac{N}{E}$ which models the noise from hardware and the environment and the constand term C which introduced by imperfect cell intercalibration and residual noncompensation

$$\frac{\sigma(E)}{E} = \frac{S}{\sqrt{E}} \oplus \frac{N}{E} \oplus C \tag{2.4}$$

In ATLAS, we employ 3 types of Sampling calorimeters, meaning it alternates between active(areas that detect energy) and passive (areas that induce showering) areas: the **Electromagnetic Calorimeter (ECAL)** to reconstruct electrons, photons, and some light hadrons, and the **Hadronic Calorimeter (HCAL)** for reconstructing hadrons The ECal and HCal cover $|\eta| < 3.2$, hence an additional sampling calorimeter is used for the region between $3.1 < |\eta| < 4.9$ called the Forward Calorimeter (FCal). The FCal measures both the electromagnetic and hadronic particles. We will briefly discuss the workings of these calorimeters in the following subsections.

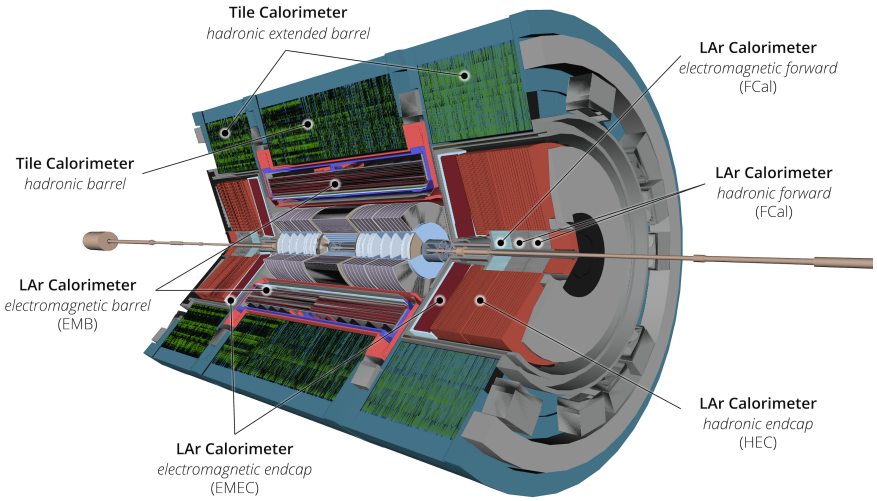


Figure 2.7: Schematic diagram of Liquid Argon Calorimeter.[47]

Electromagnetic Calorimeter (Ecal)

The ECal of the ATLAS detector is called the **Liquid Argon Calorimeter (LAr)** The active area in the LAr detector is filled with Argon ($^{39.9}_{18}$Ar) gas, which is cooled to $89K$, and the passive areas are made of Lead ($^{207.2}_{82}$Pb). The energy carried by electrons and photons is distributed through electromagnetic showers, which are then collected by electrodes in the active region. To maximise contact between these particles and the passive area, ATLAS engineered the lead layers to have an Accordion geometry. Similar to the Inner Detectors, the calorimeters are also structurally divided into barrel and endcap regions. The barrel of the ECal covers $|\eta|1.475 <$ and the endcap covers $1.375 < |\eta| < 3.2$. Figure 2.8 is the schematic diagram of a section of the LAr calorimeter showing the aforementioned accordion shape.

Similar to the Inner Detectors, the calorimeters are also structurally divided into barrel and endcap regions. The barrel of the ECal covers a pseudorapidity range $|\eta|1.475 <$, and the endcap covers a pseudorapidity range $1.375 < |\eta| < 3.2$. The Overall Energy resolution of the Ecal is given by :

$$\frac{\sigma(E)}{E} = \frac{0.1}{\sqrt{E}} \oplus \frac{0.001}{E} \oplus 0.007 \tag{2.5}$$

Hadronic Calorimeter (Hcal)

The Hadronic Calorimeter of the ATLAS detector surrounds the ECal The HCal is structurally divided into 3 regions: the large central barrel covering a pseudorapidity range $|\eta| < 1$, an Extended Barrel covering $1.7 < |\eta| < 3.2$ and the endcap covering $1.7 < |\eta| < 3.2$, but unlike ECal, it uses different technologies between the barrel regions and the endcap. The barrel region is called the Tile Calorimeter as it uses plastic scintillating tiles as its active material, is connected to photomultipliers,

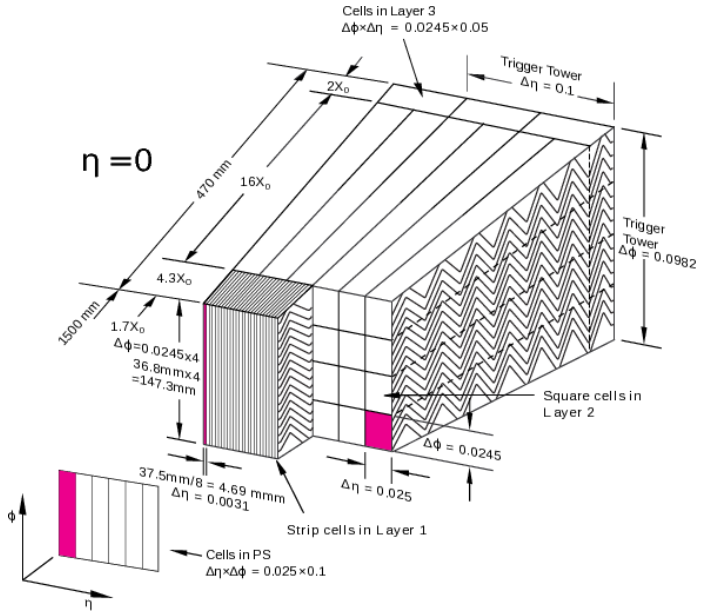


Figure 2.8: Section of the Liquid Argon Calorimeter, detailing the accordion structure[49].

and uses steel as its passive material or absorber The endcap uses the same LAr technology as the ECal, except that the absorber material is Copper ($^{62.9}_{29}$Cu) rather than lead Energy resolution of the Hcal is given by :

$$\frac{\sigma(E)}{E} = \frac{0.5}{\sqrt{E}} \oplus \frac{0.01}{E} \oplus 0.034 \tag{2.6}$$

Forward Calorimeter (Fcal)

The Forward Calorimeter, located close to the beam pipe, covers a pseudorapidity range of $3.1 < |\eta| < 4.9$ It uses technology similar to LAr, but with different materials for the absorbing layers It has one layer with a Copper absorber and 2 layers with a tungsten ($^{183.8}_{74}$W) absorber.

2.2.4 Muon Spectrometer (MS)

Muons are very interesting particles from the point of detection, as they pass through the HCal and ECal with minimal interaction This means a dedicated system is required to reconstruct muons The Muon Spectrometer is a system of multiple subdetectors that together perform tracking and calorimetry The system is placed inside a 4T Toroidal magnetic field, which bends the muons' trajectories This helps properly measure the particle's 4-momentum Figure 2.9 is a schematic diagram of the ATLAS Muon Spectrometer There are 4 main components within the MS:

- Resistive Plate Chambers (RPC): Is a gaseous detector that is composed of 2 resistive bakelite plates filled with a mixture of various gases between them A large potential is applied between the thesis plates; thus, when an electron passes through it, it initiates an electron avalanche Located in the Barrel

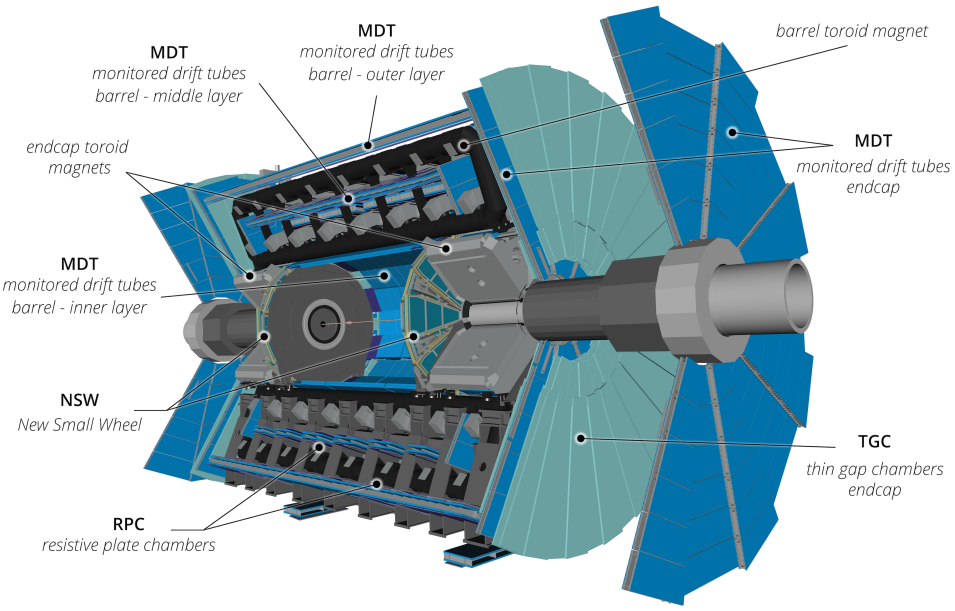


Figure 2.9: Schematic diagram of the Muon spectrometer.[47]

region, this covers a pseudorapidity range of $|\eta| < 1.05$ It's mainly used as a trigger due to its fast response.

- Thin Gap Chambers (TGC): The TGC has a similar function to that of RPC and provides Trigger information, but uses the multi-wire proportional chamber technology. It's located in the endcaps and hence covers a pseudorapidity range of $1.05 < |\eta| < 2.4$

- Monitored Drift Tubes (MDT) chamber: MDTs are drift tubes filled with $93\%Ar$ and $3\%CO_2$. They are located in both the barrel and the endcap.

- New Small Wheel (NSW): The NSW was installed prior to Run 3 to replace Cathode Strip Chambers (CSC). These gaseous detector technologies, such as micromegas and TGC, help with tracking and triggering. They cover a pseudorapidity range of $1.3 < |\eta| < 2.7$. They are designed for High luminosity conditions expected at HL-LHC.

2.2.5 Data Acquisition in the ATLAS Experiment

Storing the complete detector readout for every bunch crossing is technically and economically infeasible due to the enormous volume of data generated Consequently, the ATLAS data acquisition (DAQ) system is designed to efficiently collect detector signals, perform online event selection, and record only the most relevant events for offline analysis The DAQ system integrates detector readout electronics, trigger systems, event building infrastructure, and data storage services Detector information is first processed by the trigger system, which rapidly identifies potentially interesting events Accepted events are assembled into complete detector records and

transferred to permanent storage facilities, where they are made available for offline reconstruction and physics analysis The following sections provide an overview of the trigger architecture and the data storage framework employed by the ATLAS experiment.

Triggers

The trigger system is a fundamental component of the ATLAS data acquisition chain Its primary purpose is to reduce the incoming event rate from the 40 MHz bunch-crossing frequency to a manageable rate that can be recorded and processed offline This reduction is achieved through a sequence of increasingly sophisticated selection stages that identify signatures associated with potentially interesting physics processes Since Run 3, ATLAS has employed a two-level trigger architecture comprising a hardware-based Level-1 (L1) trigger and a software-based High-Level Trigger (HLT) The L1 trigger uses coarse-granularity information from the calorimeter and muon systems to make rapid decisions within a fixed latency, reducing the event rate to approximately 100 kHz Events accepted by L1 are subsequently processed by the HLT, which performs more detailed reconstruction using full detector information The HLT further reduces the event rate to approximately 1-2 kHz for permanent storage. This hierarchical design enables ATLAS to retain events of potential physics interest while operating within the constraints of available computing and storage resources.

Storage and Backup

Events accepted by the trigger system are transferred to the ATLAS computing infrastructure for permanent storage, reconstruction, and analysis. Raw detector data are initially recorded at CERN and subsequently distributed through the Worldwide LHC Computing Grid (WLCG), a global network of computing centres organised in a tiered architecture. The Tier-0 centre at CERN performs prompt data reconstruction and archives the raw data. Processed datasets are then distributed to Tier-1 and Tier-2 computing centres worldwide, which provide large-scale storage, reprocessing capabilities, and computational resources for physics analyses. Multiple replicas of both raw and reconstructed datasets are maintained across the grid to ensure data integrity, reliability, and long-term accessibility. This distributed storage model provides redundancy against hardware failures and enables efficient access to data by the international ATLAS collaboration.

Offline Track reconstruction

An essential tool for reconstructing a particle's properties is its trajectory. With a particle's track parameters, we can determine properties such as its decay chain and charge. Hence, having a good tracking detector and a suitable tracking algorithm is crucial for obtaining accurate measurements in the experiment. The main track reconstruction algorithm for ATLAS is part of the Athena Reconstruction Framework. We will dive deeper into the in chapter 3.

2.3 High-Luminosity LHC and ATLAS Upgrades

The High-Luminosity Large Hadron Collider (HL-LHC)[50] is the next major upgrade of the LHC, designed to significantly extend its physics reach by increasing the integrated luminosity delivered to the experiments. Beginning operation after Long Shutdown 3 (LS3), the HL-LHC is expected to provide approximately 3 ab^{-1} of proton–proton collision data at a centre-of-mass energy of $\sqrt{s} = 14$ TeV, corresponding to about an order of magnitude more data than the original LHC programme. To achieve this increase in luminosity, the HL-LHC will operate with instantaneous luminosities of up to 7.5×10^{34} $\text{cm}^{-2}\text{s}^{-1}$, resulting in an average of around 200 simultaneous proton–proton interactions (pile-up) per bunch crossing. While this large dataset will enable precision measurements of Standard Model processes and improve sensitivity to rare phenomena and physics beyond the Standard Model, it also poses significant challenges for event reconstruction, tracking, and object identification.

To maintain and enhance its performance under these demanding conditions, the ATLAS detector will undergo an extensive Phase-II upgrade. Two of the key detector upgrades relevant to this work are the replacement of the Inner Detector with the all-silicon Inner Tracker (ITk)[48] and the installation of the High Granularity Timing Detector (HGTD)[51], which provides precise timing information to mitigate the effects of pile-up.

2.3.1 Inner Tracker

The Inner Tracker (ITk) is the new all-silicon tracking detector that will replace the current Inner Detector (ID) of the ATLAS experiment for operation during the High-Luminosity LHC (HL-LHC). The existing Inner Detector was designed for the luminosity and radiation environment of the original LHC programme and would experience significant performance degradation under the HL-LHC operating conditions. The ITk has therefore been designed to provide efficient and precise charged-particle tracking at an average pile-up of 200, while withstanding the substantially increased radiation levels expected over the detector's lifetime.

Unlike the current Inner Detector, which consists of silicon pixel detectors, silicon strip detectors (SCT), and a transition radiation tracker (TRT), the ITk is composed entirely of silicon sensors. The detector comprises an inner Pixel Detector surrounded by an outer Strip Detector, providing nearly hermetic tracking coverage over the pseudorapidity range $|\eta| < 4.0$. The Pixel Detector consists of five barrel layers together with multiple endcap disks on either side of the interaction point. The innermost layers employ finely segmented pixel sensors to maintain low occupancy even in the dense HL-LHC environment, enabling precise measurements of track parameters and primary and secondary vertices. These measurements are crucial for identifying displaced decay vertices and for flavour-tagging algorithms used to identify jets originating from b- and c-quarks.Surrounding the pixel detector, the Strip Detector consists of four barrel layers and six endcap disks on each side. Double-sided strip modules provide stereo measurements, enabling reconstruction of three-dimensional points while extending the tracking volume to large radii. Figure 2.10 gives the Schematic diagram of the ITk

The ITk is expected to achieve significantly improved tracking performance com-

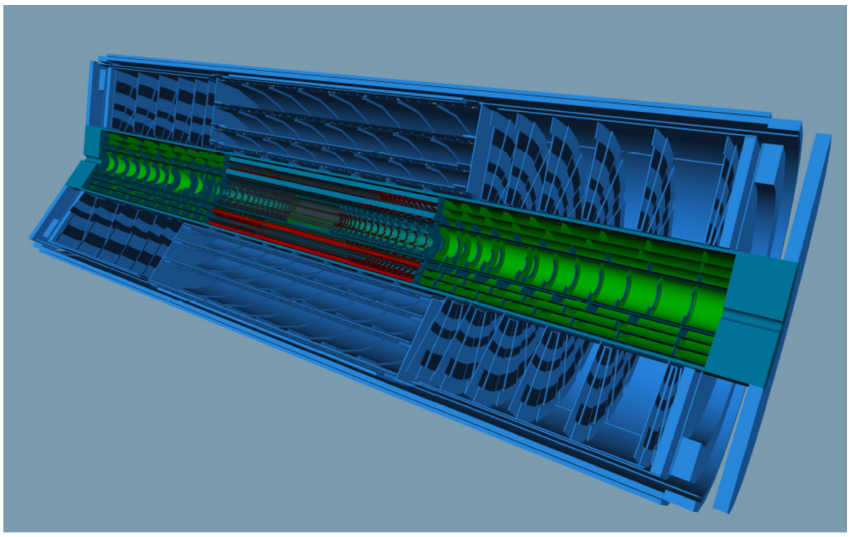


Figure 2.10: Schematic diagram of the ITk detector. The two pixel endcaps are highlighted in green[48]

pared with the current Inner Detector despite the challenging HL-LHC conditions. Furthermore, the extended tracking acceptance up to $|\eta| = 4.0$ substantially enhances the reconstruction of forward objects, benefiting analyses involving vector boson fusion, boosted objects, and other processes with forward signatures. The new ITk geometry will also utilise the ACTS tracking framework, which we will discuss in detail in chapter 3.

2.3.2 High Granularity Timing Detector

The High Granularity Timing Detector (HGTD) is a new detector system that will be installed in the forward region of the ATLAS experiment as part of the Phase-II upgrade for HL-LHC operation. Its primary purpose is to mitigate the effects of the extreme pile-up environment expected at the HL-LHC by providing precise timing measurements for charged particles. The HGTD is located between the Inner Tracker (ITk) and the endcap calorimeters, covering the pseudorapidity range $2.4 < |\eta| < 4.0$. It consists of four active silicon sensor layers arranged in two double-sided modules on each side of the interaction point. The detector employs Low-Gain Avalanche Detectors (LGADs), which provide moderate internal signal amplification while maintaining excellent timing performance and radiation hardness. Each LGAD sensor is segmented into fine pads with an area of approximately 1.3×1.3 mm^2 in the innermost region, limiting occupancy even in the highest pile-up conditions. Figure 2.11 shows an illustration of the HGTD detector.

The sensors are designed to achieve a time resolution of approximately 30–50 ps per track throughout the HL-LHC lifetime, despite exposure to radiation levels approaching 2.5×10^{15} n_{eq}/cm^2. The precise timing information provided by the HGTD effectively adds a fourth dimension to track reconstruction. The HGTD is expected to improve the performance of object reconstruction in the forward region, including jet reconstruction, flavour tagging, missing transverse momentum measurement,

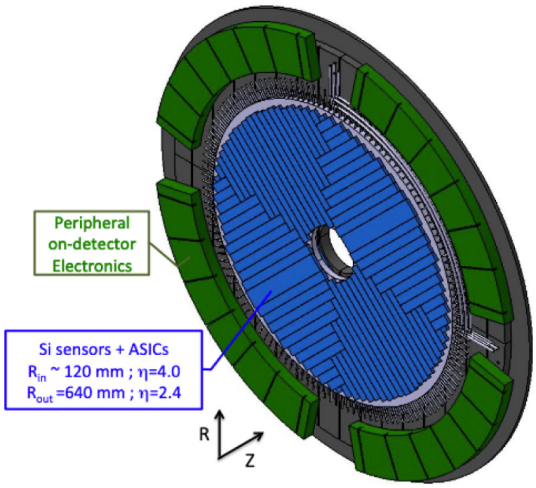


Figure 2.11: Illustration of the HGTD detector. The blue region represents the active region, and the green region represents the detector's off-detector electronics.[51]

and the identification of isolated leptons and photons. These improvements are essential for preserving the physics performance of ATLAS throughout the HL-LHC programme.

Chapter 3
ACTS Score-Based Ambiguity Solver and its Implementation in Athena

With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.
— John von Neumann

Researchers at ATLAS began preparing for the High Luminosity phase of the LHC starting in 2012. This brought new challenges, such as a higher pile-up from 50 to 200, necessitating an upgrade to the Athena tracking framework. The task was rather challenging, as parts of the original tracking algorithm were written almost a decade ago and did not utilise the latest C++ version or conventions. In parallel with the development of Athena Tracking, a ground-up approach to creating experiment-independent open-source tracking software called ACTS[52] had begun. ACTS stands for **A**CTS **C**ommon **T**racking **S**oftware. The plan so far is to use ACTS as the tracking algorithm for the ITk Detector in ATLAS for the high-luminosity phase. To achieve this, we would like most of the Athena-Tracking framework's current features to be available in the ACTS framework. One such feature is the Score-Based Ambiguity resolution in Athena. The key difference between the ACTS and the Athena frameworks is that ACTS, at its core, is designed to be flexible and detector-independent, whereas Athena, being the legacy tracking software of the ATLAS experiment, has hard-coded parts for the ATLAS tracking geometry. Hence, the main challenge of the qualification project is to make a detector-independent version of the legacy Ambiguity solver without sacrificing quality.

3.1 Tracking in ATLAS and ACTS

From a purely software point of view, Track reconstruction, or Tracking, can be seen as the process of generating track parameters from a collection of detector responses. The tracking algorithms in both ACTS and Athena follow the same 4 steps.

chapters/ACTS/

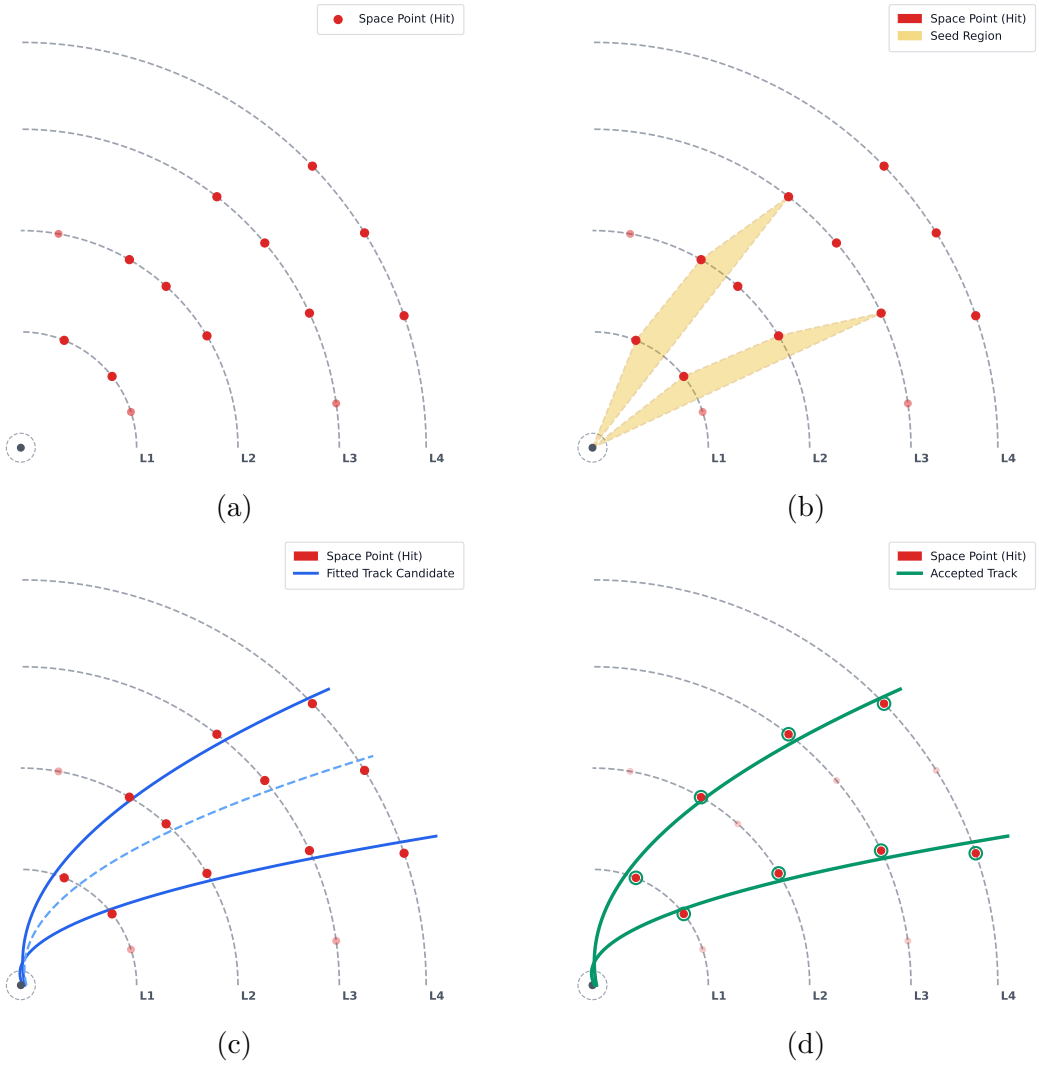


Figure 3.1: Four main stages of tracking reconstruction in ACTS: (a) *Space Point Formation*, (3D space point coordinates as red filled circles, lying on concentric detector layers grey dashed arcs, L1–L4); (b) Seeding, (seeds highlighted as translucent yellow regions); (c) Track Finding & Fitting, (candidate track trajectories shown as blue curves); and (d) Ambiguity Resolution,

3.1.1 Space Point Formation and Clustering

The track reconstruction process starts by converting the raw detector inputs into data. In subsection 2.3.1, we learned that the Inner Tracker in ATLAS is divided into multiple layers of silicon trackers. Each segment in those layers can be identified by its local coordinate l_i. In the clustering phase, it's easier to work with local coordinates than with global coordinates with the full detector. The clustering algorithm uses the Connected Component Analysis (CCA) to group segments into clusters. Detectors can give 2 kinds of signals: binary, indicating whether the detector segment is fired; for these, the 2D local position vector $\vec{r}$ is given by

$$\vec{r} = \frac{1}{N} \sum_{i=1}^{N} l_i \tag{3.1}$$

The second type measures the amount of charge q_i in a detector segment; for these, the position vector $\vec{r}$ is given by

$$\vec{r} = \frac{1}{\sum_{i=1}^{N} q_i} \sum_{i=1}^{N} q_i l_i \tag{3.2}$$

Space Point Formation is the process of converting position vectors $\vec{r}$ to the 3D global coordinates using a transformation function, which depends on the position and orientation of a layer with respect to the global frame or reference. Figure 3.1(a) illustrates the space point formation. The points where clusters are found are called **Hits**, and the points where the cluster is expected to be but is not are called **Holes**. Additionally, some hits will be marked as outliers after track fitting if they are significantly away from the fitted track.

3.1.2 Track Seeding

Particle trajectories exhibit certain symmetries, and we can use group-correlated points before the Track-finding stage to improve track-finding performance. Seeding is the process of creating collections of measurements through which a track could be created, due to the 2T solenoid magnet around the Inner Tracker. We can assume a homogeneous magnetic field along the z-axis. This means the tracks will be straight lines in the rz-plane and circular in the transverse plane. We can design algorithms to leverage these correlations and generate seeds for building the tracks. Figure 3.1(b) illustrates the seedind process. Every three points that can potentially form a track are selected as a seed, as we need at least three points to define a circle in the transverse plane.

3.1.3 Track Finding and Track Fitting

Track Finding is the process of finding the set of hits that corresponds to a specific particle trajectory. In proton-proton experiments like ATLAS, the number of true tracks could easily exceed 2000, making it very difficult to find good tracks. Figure 3.1(c) illustrates the Track Finding and Track Fitting process. The first main difference (algorithmically) between Athena and ACTS comes in the track-finding and fitting algorithm. Athena uses the **Global Chi-squared Fitter(GX2F)** fitter, which uses χ^2 between the fit and hits to assess good and bad fits; on the other hand, ACTS uses **Combinatorial Kalman Fitter**(**CKF**)[1]. The Combinatorial Kalman Fitter is a tracking algorithm based on the Kalman Filter formalism. The Kalman Filter (KF) uses the current state $\vec{x}_k$ and the state before it to predict the state at $\vec{x}_{k+1}$. The predicted state is then compared to the hits at $k+1$ to update the state. The combinatorial part arises because there may be multiple hits at $k+1$, and we

[1]Due to bremsstrahlung radiation in electrons, electron trajectories behave differently; hence both ACTS and ATLAS use **Gaussian Sum Fitter(GSF)**[53] for electron tracks.

need to consider all of them, leading to a combinatorial explosion in the number of possible tracks.

3.1.4 Ambiguity Resolution

Due to the CKF's approach of finding every possible track, it tends to generate multiple tracks that share many hits but differ only in fit quality. We call these tracks duplicate tracks. As a result of this approach, CKF has more duplicate tracks than the Global Chi-squared Fitter. Along with duplicates, we can also get Fake tracks, which are formed from incorrect combinations of unrelated clusters. Ambiguity resolution is the process of resolving track ambiguities to identify good tracks and distinguish them from duplicates and fake tracks. Figure 3.1(d) illustrates the ambiguity resolution process.

The base ambiguity solver in ACTS is the **Greedy Solver**. The greedy solver is a greedy algorithm designed solely to remove any shared hits from tracks. It compares every pair of tracks that share at least three hits and removes the track with the poorer χ^2. The key part is that after each iteration, it recomputes the number of shared hits in each track. By starting this process from the tracks with the highest shared hits, it's specifically designed to target duplicate tracks of lower quality. The major drawback of this method is the computation cost. The greedy solver has complexity of $\mathcal{O}(N^2)$ where N is the number of tracks. Athena, on the other hand, uses a **Score-Based Ambiguity Solver**. This algorithm uses scores and cuts to filter out fake and duplicate tracks. The objective of this project is to develop a score-based ambiguity Solver in the ACTS Framework, implement it in Athena for the new ITk detector geometry, and compare the performance between the native Athena implementation and the ACTS Framework's score-based ambiguity solver.

3.2 ACTS Score-Based Ambiguity Solver

A Score-Based Ambiguity Solver filters tracks by 3 processes. **Track Scoring**: which scores each track based on some parameters and weights, **Track filtering**: Which removes tracks based on some parameters and weights and **Track cleaning**: Which removes shared hits from tracks based on certain parameters. These 3 steps help us discriminate good tracks from duplicates and thus resolve the ambiguity. To bring this algorithm into the ACTS Framework, which is detector-independent, we need to find creative ways to inform it without sacrificing flexibility. This is achieved through the initialisation process, which uses configuration files that can be customised to meet the needs of the experiments.

Initialisation and Detector Configuration

In the ACTS framework, the whole detector is divided into volumes. Therefore, each hit, hole or outlier has a volume ID associated with it. In experiments such as ATLAS, not all volumes behave uniformly; their properties and responses vary depending on the materials or technologies used. For example, in ATLAS ITk, we have pixels and strips. These groups of volumes will be called detectors, and for the score-based ambiguity solver, we can configure eta-based cuts and weights

for each detector. The solver then uses the configuration at the track-scoring and track-cleaning stages. To enable easy modification and customisation, users can pass the configuration via a JSON file. We have provided a helper function to address this. Initialisation is the step that prepares the ambiguity solver by setting up the global and detector configurations and computing prerequisite variables, such as the number of hits, holes, and outliers per detector, as well as the number of hits shared with other tracks.

3.2.1 Implementation of Eta Based Cuts

Since the final ranking is based on the `trackScore`, **Cuts** are implemented by setting the `trackScore` to zero. The cuts applied are **eta-based**, meaning the user can define eta bins in the detector config and set cuts for each bin. The cuts are defined as arrays in the JSON config. There are 4 η based cuts implemented in ACTS standalone:

- `minHitsPerEta`: Minimum number of hits a track needs to have per η bin.
- `maxHolesPerEta`: Maximum number of holes a track can have per η bin.
- `maxOutliersPerEta`: Maximum number of outliers a track can have per η bin.
- `maxSharedHitsPerEta`: Maximum number of hits a track is allowed to share with another track per η bin.

3.2.2 Track Scoring for Ambiguity Solving

The **Track score** is derived from properties such as p_T, χ^2 and the number of hits, holes and outliers in each detector. There are 2 scoring types available to users: Ambiguity-Score and Simple-Score.

Ambiguity-Score

The Ambiguity-score is a product of three terms (see Equation 3.3). The first term is the p_T-dependent term, so higher-p_T tracks receive a higher score. The next term is the χ^2 dependent term. So, the lower the χ^2, the better the score, and for $\chi^2 = 0$, the χ^2 dependent term would be 1. The last term is based on the number of hits and holes in each detector.

$$S = (\log(p_T) - 1)\frac{1}{\log(10 + 10 * \frac{\chi^2}{nDoF})} \prod_{i=1}^{N_{detectors}} \left(w_i^{\text{hits}} * w_i^{\text{holes}}\right) \tag{3.3}$$

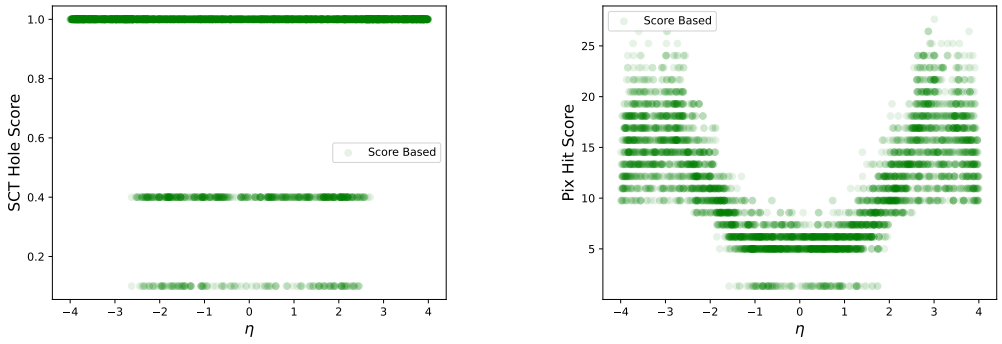
w_i^{hits} in Equation 3.3 depends on the number of hits (n_i^{hits}). The values of these weights are taken from an array ($f_{\text{hits},i}$) at indexes equal to the number of hits. For example, if a track has 5 hits, the 6th value of the factor array is taken as the weight. The array length is set to the expected maximum number of hits; for tracks with more hits than the array size, a different mechanism is used. Equation 3.4 gives the exact expression for the hit weights. The array (f_i^{hits}) is the ratio of *goodHits* and *fakeHits* vectors set from the detector config.

$$w_i^{\text{hits}} = \begin{cases} f_i^{\text{hits}}[n_i^{\text{hits}}], & \text{if } n_i^{\text{hits}} \leq n_{\text{max}}^{\text{hits}}, \\ \left(n_i^{\text{hits}} - n_{\text{max}}^{\text{hits}} + 1\right) \cdot f_i^{\text{hits}}\left[n_{\text{max}}^{\text{hits}}\right], & \text{if } n_i^{\text{hits}} > n_{\text{max}}^{\text{hits}} \end{cases} \tag{3.4}$$

Similarly, these weights (w_i^{holes}) are also taken from an array (f_i^{holes}) at indexes equal to the number of holes. For tracks with holes larger than the array, the mechanism is different from that for hits. Equation 3.5 gives the exact expression for the hole weights. The array (f_i^{holes}) is the ratio of *goodHoles* and *fakeHoles* vectors, also set from the detector config.

$$w_i^{\text{holes}} = \begin{cases} f_i^{\text{holes}}[n_i^{\text{holes}}], & \text{if } n_i^{\text{holes}} \leq n_{\text{max}}^{\text{holes}}, \\ \dfrac{f_i^{\text{holes}}\left[n_{\text{max}}^{\text{holes}}\right]}{n_i^{\text{hole}} - n_{\text{max}}^{\text{holes}} + 1}, & \text{if } n_i^{\text{holes}} > n_{\text{max}}^{\text{holes}}. \end{cases} \tag{3.5}$$

The 2-case structure for w_i^{hits} and w_i^{holes} lets us linearly scale the score of the tracks up or down if it exceeds the maximum number of hits or holes. Figure 3.2 shows an example of the weights for the pixel detector. We can see the discrete nature of the weights. In this example, $n_{\text{holes,max}}$ is 3 (hence 3 lines) and $n_{\text{hits,max}}$ is 26.



(a) (b)

Figure 3.2: Examples of individual scores vs η for 2 events for the ATLAS ITk geometry. a) is the score weight for the SCT hole score, and b) is the score weight for the pixel hits score.

Simple-Score

The simpleScore, like the ambiguityScore, also has 3 terms, but the key difference is that the terms are **added** together rather than multiplied. The first term is a fixed number. This is set at 100. The second term is the χ^2 dependent term. So, the lower the χ^2, the better the score, and for $\chi^2 = 0$, the χ^2 dependent term would be 1. The last term is based on the number of hits, holes and outliers on a track in a detector. Equation 3.6 gives the simple scoring function.

$$S = 100 + \frac{1}{\log\left(10 + 10 \cdot \frac{\chi^2}{n_{\mathrm{DoF}}}\right)} + \sum_{i=1}^{N_{\mathrm{detectors}}} \left(n_i^{\mathrm{hits}} \cdot W_i^{\mathrm{hits}} + n_i^{\mathrm{holes}} \cdot W_i^{\mathrm{holes}} + n_i^{\mathrm{outliers}} \cdot W_i^{\mathrm{outliers}}\right) \tag{3.6}$$

Here $N_{detector}$ is the number of detectors (pixels, strips, etc.). W is the score weight, which is configured from the detectorConfig.

3.2.3 Tracking Cleaning

The cleaning step is the most important step for ambiguity solvers. It is in this step that we remove unwanted tracks. The cleaning step is performed by calling `getCleanedOutTracks()` iteratively from the `solveAmbiguity()` function. For each track, the `getCleanedOutTracks()` returns a boolean; true if the track is acceptable and false otherwise. For each track, `getCleanedOutTracks()` loops over all hits and decides whether to count a hit based on the number of tracks the hits are shared with and the `trackScore`. If the hit is not shared with any other track, it is accepted without condition. For shared hits, we accept them if the track's score exceeds the minimum score for shared tracks (set in the global config), provided that the total number of accepted shared hits does not exceed the maximum limit (set in the global config). Tracks with the number of accepted hits less than the configured minimum number (`minUnsharedHits`) are rejected.

3.2.4 Customisation of the Score-Based Solver

Since ACTS is designed to be used with various tracking detectors and needs, the Score-Based Solver can be customised in two ways. **Global configurations** which are instructions that are applied to all tracks. and Optional Functions, which let users define the cuts, scores and hit selection. The Score-Based Solver can be customised in two different ways. **Global configurations** which are instructions that are applied to all tracks. We limit the number of global configurations to the absolute essential one.

- Minimum acceptable `trackScore` - used in the Finalisation step
- Minimum score for shared track - used in the track cleaning stage
- Maximum shared hits in a track - used in the track cleaning stage
- Minimum unshared hits in a track - used in the track cleaning stage
- Ambiguity scoring flag - used in track scoring

Optional Cuts and Scores

The default cuts and scoring in the ACTS standalone have only a basic set of cuts and score modifiers. Users can define custom cuts and scores using this structure for more flexibility. This is done with the `Optionals` structure. This

structure contains 3 vectors: 2 of type `OptionalScoreModifier`; used for adding scores into ambiguityScore (optionals.score) and simpleScore(optionals.weights), and `OptionalFilter`; used for adding optional cuts.

- The `OptionalScoreModifier` type is a std::function which takes track as input and a floating point number (score) by reference and returns void. Users can define functions with the same set of parameters and create custom score-modifying functions.
- The `OptionalFilter` type is a std::function that takes a track as an input parameter and returns a boolean; true if the track is rejected and false otherwise. Users can define functions with the same parameters and make custom cuts.

Optional Hit Selection

The purpose of the optional hit selector is similar to that of the optional cuts and the optional scores. It enables customising hit selection. Hence, the optional hit selector is defined in the same optionals structure as other optionals. The optional hit selector is called from the `getCleanedoutTracks()` function. Users can use the custom function to change `TrackStateTypes` from say `RejectedHit` to `SharedHit` or `OtherTrackStateType` to `outlier`. This will then affect the hit selection. Adding Optional Hit Selection to the solver follows the same mechanism as the other optional features; the function should have the parameters of type `TrackProxy`, `TrackStateProxy`, and `TrackStateTypes` and return void.

3.3 Integrating the ACTS Score-Based Solver into Athena

In section 3.2, we have seen how the ACTS Score-Based Solver is designed to be customisable. Implementing the ACTS Score-Based Solver in Athena involves 2 steps. First, port the configuration for each score modifier and cut from its Athena counterpart into the `DetectorConfig` file; second, add any missing score modifiers and cuts from the ACTS standalone as optional features.

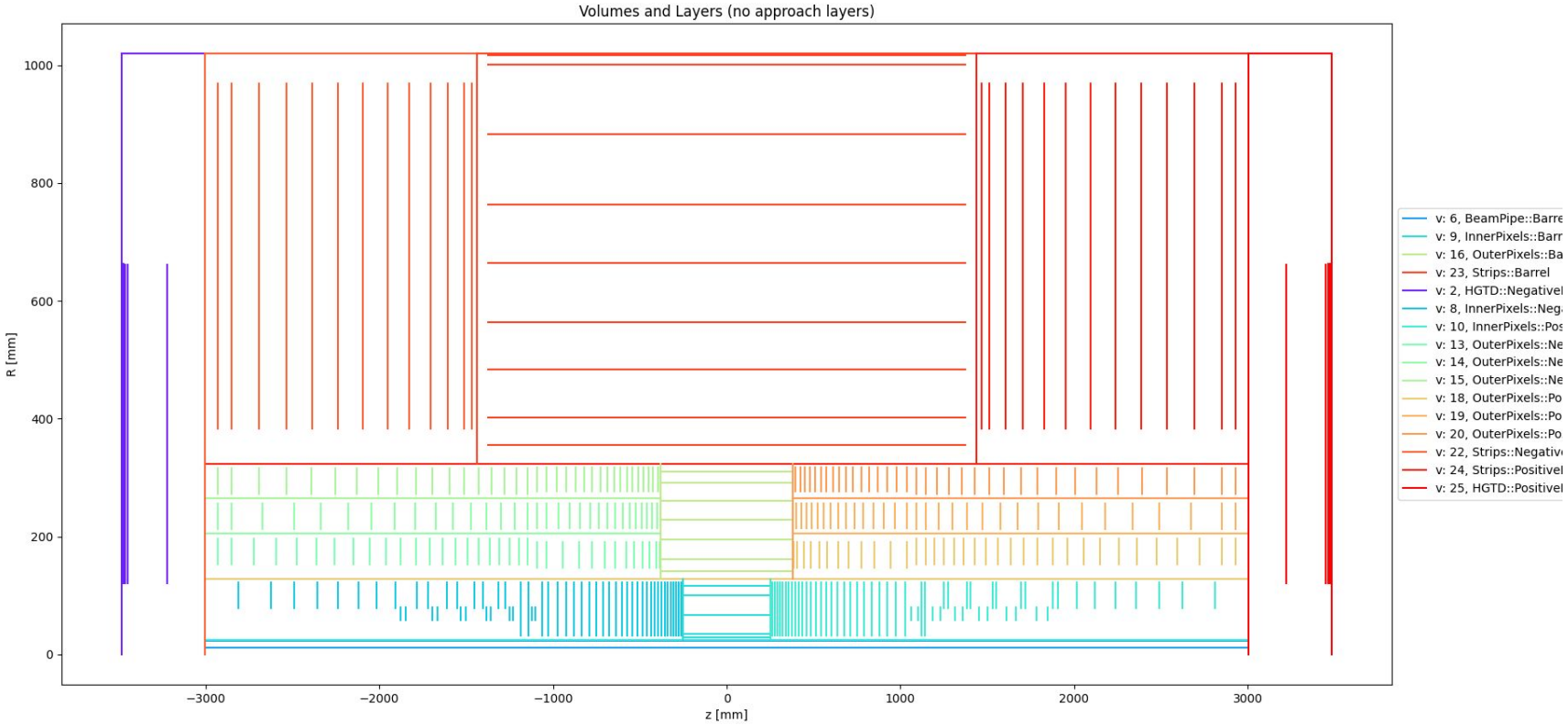


Figure 3.3: Layer geometry for ATLAS ITk Detector

3.3.1 Ambiguity Configuration for ITk Detector

Figure 3.3 gives the volume map of the ITk Detector. The detector consists of Inner Pixels and Silicon Strips, which need a slightly different configuration for the Score-Based solver. Table 3.1 lists the global configurations used in the inner pixels and strips of the ITk tracker. [2]

Parameter	**Inner Pixel**			**Strips**		
hitsScoreWeight	20			10		
holesScoreWeight	-10			-5		
outliersScoreWeight	-2			-2		
otherScoreWeight	0			0		
maxHolesPerEta	2			2		
maxOutliersPerEta	2			2		
sharedHitsFlag	false			false		
	η **bins**					
$\lvert\eta\rvert$ range	0.0 - 2.0	2.0 - 2.6	2.6 - 4.0	0.0 - 2.0	2.0 - 2.6	2.6 - 4.0
$n_{\text{min}}^{\text{hits}}(\eta)$	2	2	2	1	0	0
$n_{\text{max}}^{\text{SharedHits}}(\eta)$	8	10	20	10	10	10
$n_{\text{max}}^{\text{holes}}(\eta)$	2	2	2	2	2	2
$n_{\text{max}}^{\text{outlier}}(\eta)$	2	2	2	2	2	2

Table 3.1: Table of InnerPixels configuration

In subsection 3.2.2 we discussed the good and bad hits or holes, whose ratio defined $f_i^{\text{hits/holes}}$. Figure 3.4 gives the f_i^{hits} for Inner Pixels and f_i^{holes} and for Inner Pixels and ITk Strips. For ITk strips have no hit-based scaling, meaning $f_i^{\text{holes}} = 1, \forall i..$ [3]

3.3.2 Optional cuts and scores in Athena

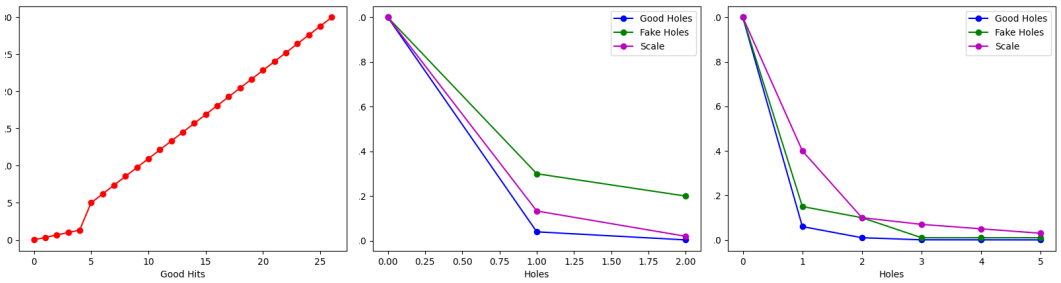
The ACTS standalone has only a basic set of cuts and score modifiers, and all the Athena cuts and scoring algorithms can not be implemented that way. Hence, we use the optional features to implement the missing cuts and score modifiers. These are optional score modifiers in the Athena implementation.

1. Silicon Strips Double-Hole Score

2. Contributing Pixel Layers Score

[2] The `DetectorConfig` file for the ACTS Score-Based solver implementation in Athena can be found in ActsAmbiguityConfig.json

[3] For Inner Pixel f_i^{hits}, good and bad hits are not shown separately, as bad hits are defined to be an array of ones, meaning $f_i^{\text{hits}} = \text{goodHits}$.



(a) (b) (c)

Figure 3.4: Scaling factor arrays; scaling factor vs hits/holes/outlier count. a) f_i^{hits} for Inner Pixels, b) f_i^{holes} for Inner Pixels and c) f_i^{holes} for Silicon Strips

3. Innermost Pixel Layer Hits Score

4. Silicon Hits Score

Apart from score modifiers, we also have an optional cut function that adds eta-based cuts missing in the ACTS standalone. These extra cuts and score modifiers often use variables that are es not present in the ACTS standalone tracks, and we have to add them to the tracks. These extra variables are added as extra columns to the track and can be accessed using hash strings. The extra variables are computed in the `addSummaryInformation()` function in the *ScoreBasedSolverCutsImpl.h* file. But such an operation is better moved to track fitting.

Silicon Strips Double-Hole Score

Double holes are a very strong sign that the track is fake, as the likelihood of two holes appearing back-to-back from a legitimate track is low. Hence, we use a factor array $f[n]$ to reduce the track's score.

$$w = \begin{cases} f[n], & \text{if } n \leq n_{\max}, \\ \dfrac{f\left[n_{\max}\right]}{n - n_{\max} + 1}, & \text{if } n > n_{\max}. \end{cases} \tag{3.7}$$

The factor array $f[i]$ is defined as the ratio of the good Double Holes vector to fake Double Holes, and these vectors are hardcoded in the code.

Contributing Pixel Layers Score

The pixel detector is further divided into layers based on their radial distance from the beam pipe. The number of contributing pixel layers is the number of layers in the pixel module that have a hit for the given track. The existing score is multiplied by a weight term w dependent on the number of contributing pixel layers in a track. See Equation 3.8.

$$w = \begin{cases} f[n], & \text{if } n \leq n_{\max}, \\ (n - n_{\max} + 1) \cdot f\left[n_{\max}\right], & \text{if } n > n_{\max}. \end{cases} \tag{3.8}$$

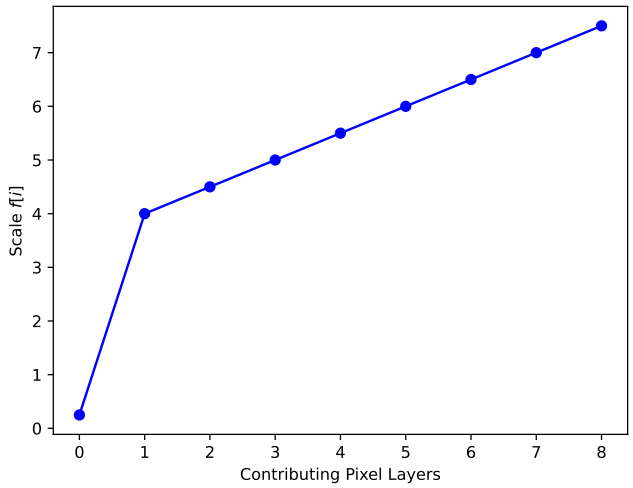


Figure 3.5: Scaling factor vs number of innermost pixel layer hits

Here n is the number of contributing pixel layers, n_{max} is the maximum expected number of contributing pixel layers, and f is the factor array given by Equation 3.9. The current values of n_{max} is set at 16, Score_{max} at 30.0 and Score_{min} at 5.0.

$$f[i] = \begin{cases} 0.01 + (i \cdot 0.33), & \text{if } i < \text{Score}_{min} \\ \text{Score}_{max} + \left((i - \text{Score}_{min}) \cdot \dfrac{\text{Score}_{max} - \text{Score}_{min}}{\text{n}_{max} - \text{Score}_{min}} \right), & \text{if } i \geq \text{Score}_{min} \end{cases} \tag{3.9}$$

Innermost Pixel Layer Hits Score

In the Innermost Pixel Layer Hits Score, the existing score is multiplied by a weight term w dependent on the number of hits on the innermost layer (sometimes called the B layer). The weight term w is computed similarly to the Contributing Pixel Layers Score. Equation 3.10 gives the description for the weight term w.

$$w = \begin{cases} f[n], & \text{if } n \leq n_{\text{max}}, \\ (n - n_{\text{max}} + 1) \cdot f\left[n_{\text{max}}\right], & \text{if } n > n_{\text{max}}. \end{cases} \tag{3.10}$$

Unlike the Contributing Pixel Layers Score, the factor array $f[i]$ is preset. Figure 3.5

Optional Eta based cuts

The eta-based cuts not implemented in the standalone ACTS version can be added via the Optional Function. In defining these functions, we specify the eta-based binning and the cut thresholds for each eta bin. The following are the 5 η-based cuts added in the Athena Implementation.

1. Maximum η (absolute) for a track

2. Maximum Z Impact for a track

3. Maximum Primary Impact (D0) for a track

4. Minimum number of silicon hits

5. Maximum number of double holes

3.4 Results

The results are generated by comparing ACTS Greedy, ACTS Score-Based and Athena solvers. The results are obtained by using $t\bar{t}$ events at 200 pile-up. The same RDO file was used for all 3 solvers for consistency. Due to technical difficulties, we couldn't convert Athena tracks to ACTS tracks. Hence, for ACTS Greedy and ACTS Score-Based, we used the ACTS tracks, and for Athena Solver, we used the Athena tracks. Once we resolve the issue with track conversion from Athena to ACTS, we can make a much fairer comparison between the 3 solvers.

3.4.1 Results from scoring

The primary goal of the work is to mimic the Athena Solver; hence, the first check is whether the track scores produced by the ACTS Score-Based Solver match those of the Athena Solver. Figure 3.6 is the Overall score distribution for the Score-Based Solver compared to the Overall score distribution of the Athena Solver. To make a fair comparison, we removed the **nGangedPixels Score** from the Athena Solver for these studies, as it is not yet available in ACTS Tracks. At first glance, the overall shape and scale are the same for both solvers.

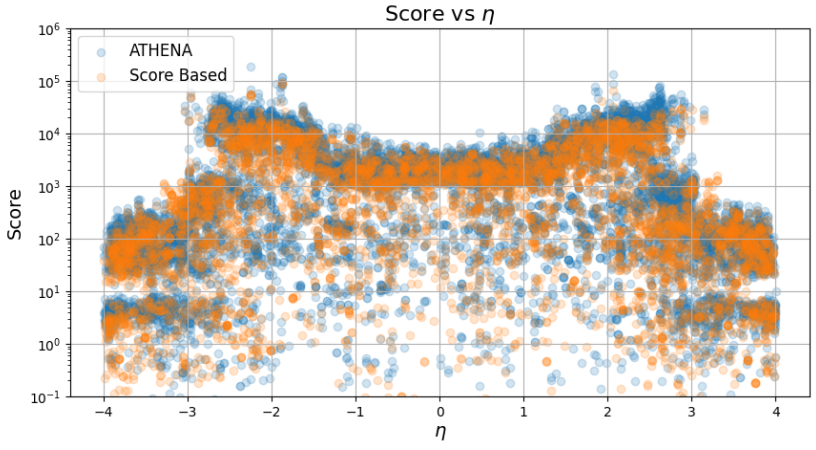


Figure 3.6: Comparing between Athena Solver score and Score Based Solver score with respect to η

However, we can notice small differences in the scores, necessitating further probing and checking the individual sub-scores that make up the overall score to

ensure that every part of the scorer functions correctly. Figure 3.7 is the one-to-one comparison of the score-based solver scores to its Athena counterpart.

Figure 3.7 shows that **only** the χ^2 Score has a mismatch. This is because for Athena tracks, the plot also includes pattern tracks, and for pattern tracks, the χ^2 score is always 1 in Athena Solver. Figure 3.7(i) has a mismatch. We can see that the Athena Contributing Pixel Layers Score is shifted up compared to the Score-Based Solver score. This difference is because the number of Contributing Pixel Layers is off between the ACTS track and the Athena tracks. This might be due

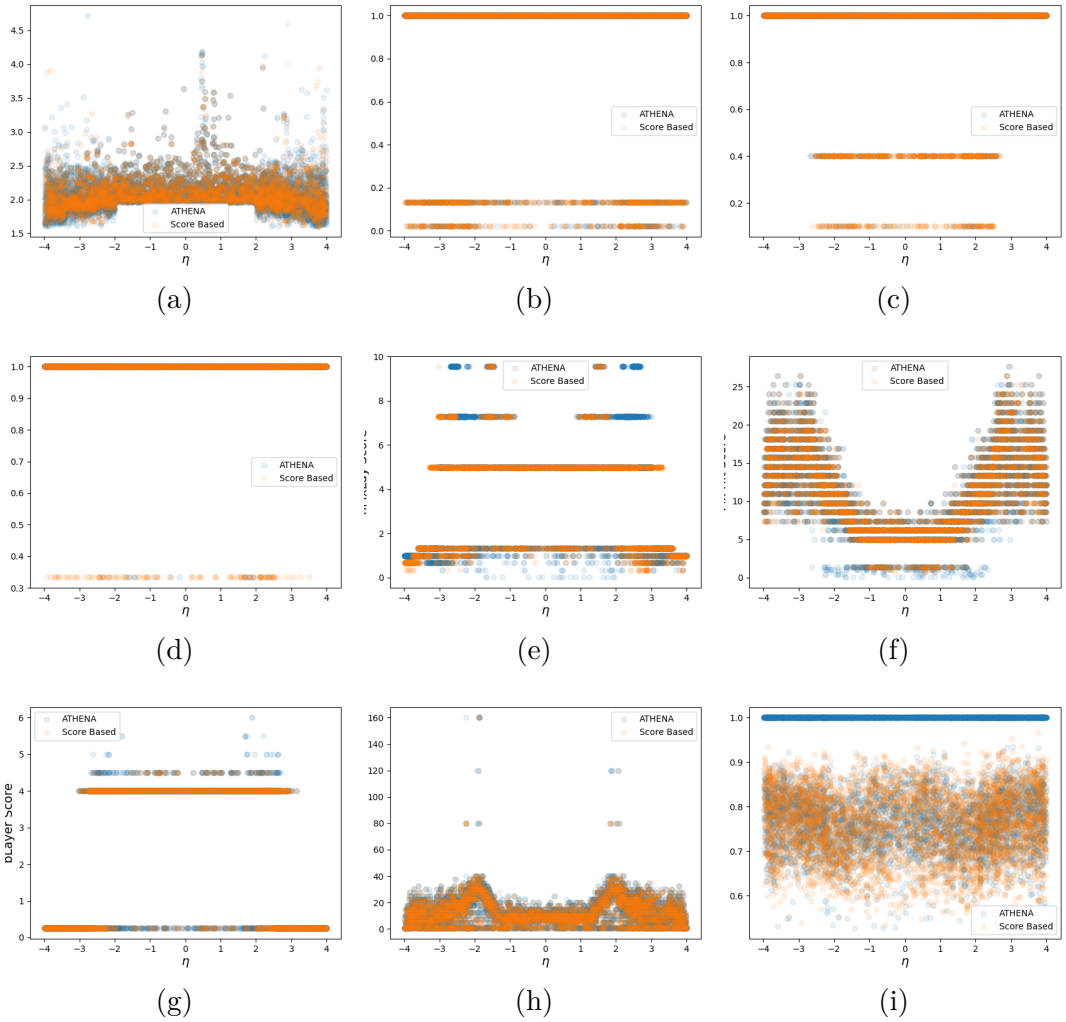
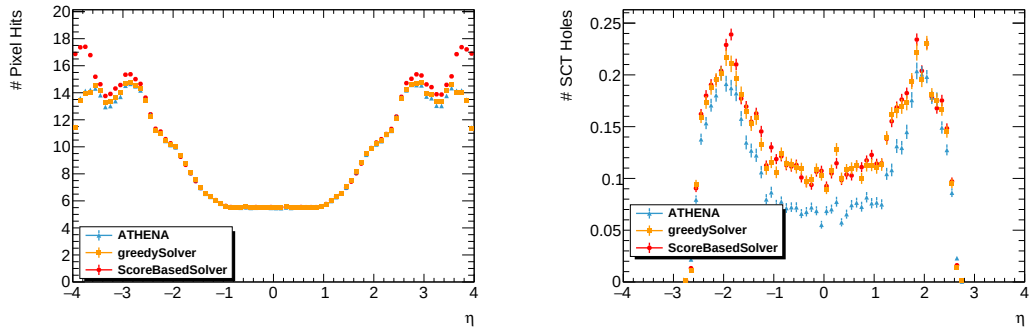


Figure 3.7: Comparing between Athena Solver score and Score Based Solver score concerning eta η. a) Score based on p_T b) Score based on number of pixel holes c) Score based on number of SCT holes d) Score based on number of SCT double holes e) Score based on number of Contributing Pixel Layers f) Score based on number of pixel hits g) Score based on the number of hits on the innermost layer (b layer) h) Score based on number of silicon hits i) Score based on the χ^2

to differences in the definitions of layers between ACTS and Athena. The Athena implementation needs further study to implement this score properly.

3.4.2 Results from primary comparison

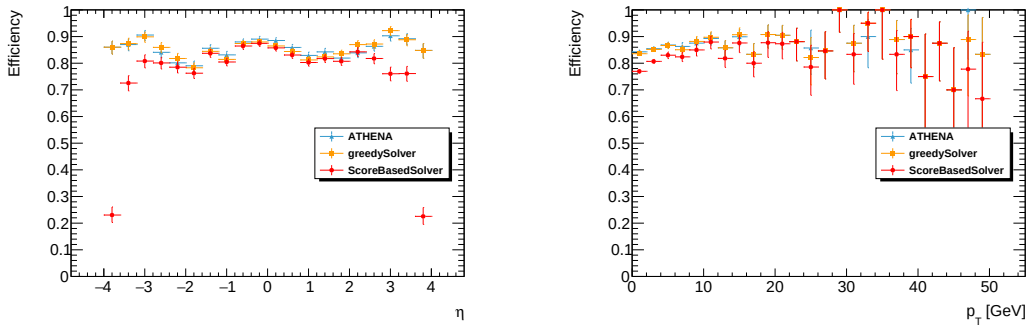
The primary comparison revealed that track parameters, such as the number of Pixel hits and SCT holes, are similar, suggesting low Fake Track rates. (see Figure 3.8). However, this is not the best way to check the solver's performance.



(a) (b)

Figure 3.8: Comparison of track parameters between Athena Solver, ACTS Greedy Solver, and ACTS Score-Based Solver a) number of hits on Pixel detector vs η b) number of holes on Strip detector vs η

A better comparison would be to check the Efficiency and Fake rate. Efficiency is the rate at which we correctly identify the truth track. Ideally, we would expect this number to be one. Figure 3.9 compares the efficiency of ACTS Score-Based Solver to Athena Solver and ACTS Greedy Solver. On average, efficiency seems good, but there is room for improvement.



(a) (b)

Figure 3.9: Comparison of efficiency between Athena Solver, ACTS Greedy Solver, and ACTS Score-Based Solver a) efficiency vs η b) efficiency vs p_T

The fake rate is the fraction of fake tracks that are misclassified as good tracks. The better the solver, the lower the fake rate. Ideally, the fake rate should be zero. Figure 3.10 compares the fake rate of ACTS Score-Based Solver to Athena Solver

and ACTS Greedy Solver. Overall, the Fake rate from the ACTS Score-Based Solver is in good agreement with the other 2 Solvers.

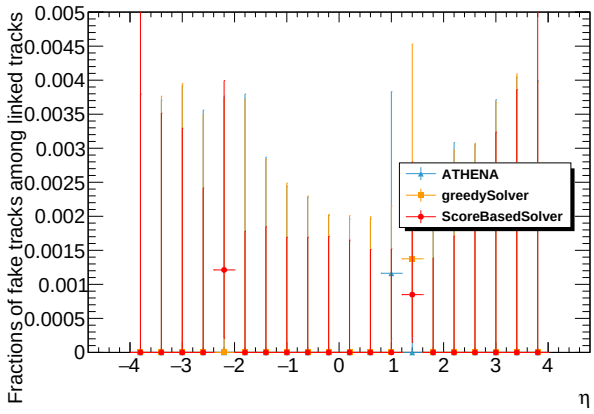
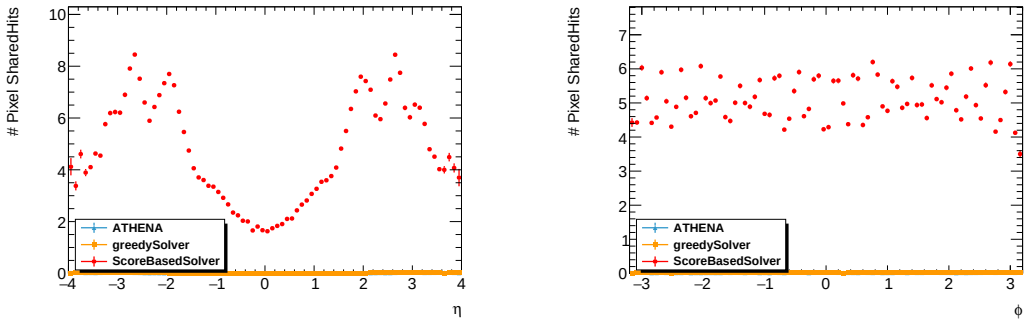


Figure 3.10: Comparison of Fake rate Vs η between Athena Solver, ACTS Greedy Solver and ACTS Score-Based Solver

Once we have good Efficiency and Fake rate, the next question is about the duplicate rate. Unfortunately, we could not perform a direct comparison plot. But we notice that the number of tracks after the ACTS Score-Based Solver is **40%** more than that of both ACTS Greedy and Athena Solver. Given that we have similar efficiency and a fake rate to those of the other solvers, these extra tracks must be duplicates. This is evident when we compare the number of shared hits per track. Figure 3.11 gives the shared hit comparison between Athena Solver, ACTS Greedy Solver, and ACTS Score-Based Solver. We can clearly see that the Score-Based solver has more shared hits than the others. (More on this is in section 3.5.)



(a) (b)

Figure 3.11: Comparison of number of shared pixel hits between Athena Solver, ACTS Greedy Solver, and ACTS Score-Based Solver: a) Number of shared hits Vs η; b) Number of shared hits vs p_T.

3.5 Discussion and Conclusion

The comparative performance in Athena ITk Geometry is promising, as the performance matches that of the Greedy Solver and the Athena Solver in terms of Efficiency and Fake rate. That said, compared with the ACTS Greedy Solver, the ACTS Score-Based Solver has a higher average duplicate rate. If we try to avoid it by raising the minimum scores or tightening the cuts, we end up lowering the efficiency. The reason we do not see this with the Athena solver is that we are not testing it with ACTS tracks, which, by design, had more duplicate tracks. The reason for that was the difference in seeding procedure. The Athena seeding has an additional seed filter that removes seeds if the same set of hits is present in another seed. This can reduce the number of duplicate tracks before the ambiguity solver, thereby making the ambiguity solver's job easier. After my project ended, the ACTS seeding also added a seed filter, and the newer results show a reduction in duplicates from 40% to around 2%. The 2% duplicate rate is not unexpected, as there are small differences and shortcomings in the ACTS Score-Based Solver when compared to the Athena implementation, such as:

1. How shared hits are handled. Unlike Athena Solver, the ACTS Score-Based Solver counts how many hits can be accepted and then decides to accept or reject a track based on that. The Athena Solver instead removes the bad hits altogether from the tracks, then refits the track with the remaining hits. This compromise was made to avoid having to recreate the track collection. In the future, we could modify the ACTS Score-Based Solver such that it builds a new track collection with shared hits removed from the tracks.

2. All the selections in Athena Solver are **Not** Available in the current implementation. Some of them are variables from Calorimeter, like η_{WidthEm}, ϕ_{WidthEm} and Max $R_{\Phi,\text{ImpEM}}$. Others are variables calculated from other algorithms or detectors, such as:

    - Cluster splitting and Split probability
    - Dead sensor
    - Number of Ganged Pixels

Once these changes are added to ActsTrK, we could add or modify optional cuts and optional scores to bring the Score-based Solver much closer to the Athena Solver. Alternatively, rather than porting the configuration from Athena, we might be warranted in using an optimiser to find a better configuration and adding optional score modifier functions that could distinguish between good and duplicate tracks. A simple example would be to use the number of shared hits on a track to score it.

Conclusion

A new Score-Based Ambiguity Solver, inspired by the Athena Solver, has been developed within the ACTS Framework. This solver, which incorporates optional cuts and scores, demonstrates promising performance in the Athena ITk Geometry; challenges such as efficiency loss at high eta and issues with cuts and scores remain

to be addressed. Future improvements could include optimising detector configuration files using dedicated algorithms and exploring additional cuts and scores to enhance performance. ACTS now has 3 different Ambiguity solvers: Greedy, Score-Based, and ML. A potential next step is to develop a flexible solver that combines the strengths of these methods, paving the way for even more robust ambiguity resolution. The successful implementation of the Score-Based Solver marks a significant advancement, opening new avenues for innovation and improvement in track reconstruction for the ATLAS experiment and beyond.

Chapter 4
FAIR Universe Higgs Uncertainty Challenge

When a measure becomes a target,
it ceases to be a good measure.
- Charles Goodhart

Experiments in particle physics are essentially very elaborate counting experiments. We quantify discovery by comparing the number of observed events to those under the null hypothesis ($\mathcal{H}_0 = P(n|B)$) or background only, and the new-physics hypothesis ($\mathcal{H}_1 = P(n|S+B)$) or signal + background. And testing if this excess is statistically significant. In practice, we parameterise this with the signal strength μ. The signal strength μ is defined as the ratio of the observed number of signal events (S_{obs}) to the expected number of signal events (S_{exp}) in the standard model. The observed number of events $n_{obs} = \nu(\mu) = \mu S_{exp} + B_{exp}$

$$\mu = \frac{S_{obs}}{S_{exp}} \tag{4.1}$$

Since any dataset is effectively a random sample, the estimated value of μ inherently has an uncertainty around it. We can quantify this with 1σ and 2σ confidence intervals (CI). If the estimated signal strength μ is **significantly** different from the null hypothesis ($\mu = 0$), then we can conclude that new physics has been discovered. The statistical discovery significance (Z) of the estimated μ parameter is given by

$$Z = \sqrt{2\left[(s+b)\ln\left(1+\frac{s}{b}\right) - s\right]} \tag{4.2}$$

Z qualifies how many standard deviations separate the null hypothesis from the new physics one. In particle physics, we use the standard of at least 3σ to declare evidence and at least 5σ to declare discovery. Given the extreme rarity of the events we observe, the chance of discovering new physics is very low **unless** we find a region of phase space with a higher signal density. In the very early days of the LHC, the most prominent way to identify this region was to construct specialised variables that separated the signal from the background, and then use them to define the region of interest. This is often known as the cut-and-count approach. One can only build histograms on these variables to utilise them to improve Z; we will discuss these in section 4.5.

The 2012 Higgs discovery became a gateway for analysing rare and more complex Higgs Boson decay channels, like the $H \to \tau^+\tau^-$. Unlike the previous channels, these channels contained neutrinos that the ATLAS detector cannot directly detect. This means that standard cut-based and simple multivariate analyses didn't provide sufficient background rejection. This called for machine learning algorithms trained to maximise the background rejection rate, starting with simple decision trees. Though this improved measurement precision, there was still much more to be gained from machine learning algorithms. Thus motivated the creation of the Higgs Boson Machine Learning Challenge (HiggsML [14, 15]), which at the time was the most popular Kaggle challenge, attracting 1785 teams. This challenge has significantly heightened interest in applying Machine Learning (ML) techniques within High-Energy Physics (HEP) and, conversely, has exposed physics issues to the ML community. The challenge used the 2014 ATLAS open data [54] for the $pp \to \tau^+\tau^-$ process. The participant's objective was to create an ML classifier that would maximise the discovery significance Z for the Higgs Boson ($H \to \tau^+\tau^-$).

4.1 Towards HiggsML Uncertainty Challenge

The HiggsML Challenge opened the door to extensive research in machine learning and particle physics and motivated additional challenges in ML for particle physics. The dataset itself was widely used in studies of newer methods for years after the challenge concluded. But in the last decade, new methods that require training and validation on larger datasets have begun to emerge. At the same time, there was a need for newer metrics to assess ML models for particle physics, and one of the challenges was their ability to handle systematic uncertainty.

4.1.1 Systematics Uncertainty in Particle Physics

In particle physics, systematic uncertainty is the uncertainty arising from incomplete knowledge of the experiment, the detector, or the theory. This is practically visible as a shift in the distributions of various event features in the data compared to the simulation. Systematic uncertainty was already a complex issue even before the introduction of ML, but with ML algorithms, because training is done on pure datasets (datasets with no systematics), the effect on the final results is more pronounced when an ML discriminant is used. The traditional approach is to estimate the systematic uncertainty in parameter estimation using shifted datasets and propagate it to the final error prediction. However, this does not address the fundamental issue of biased ML models. We will discuss more about classical uncertainty quantification in subsection 4.5.2.

4.1.2 Motivation for an Uncertainty Challenge

In the past 10 years, advanced efforts that integrate uncertainties into the ML training include approaches that explicitly depend on nuisance parameters [55–64], that are insensitive to nuisance parameters [65–82], that use downstream test statistics in the initial training [83–93], and that use Bayesian neural networks for estimating uncertainties [94–96]. Many of these topics were covered in recent forward-

looking review-type articles in Refs. [97, 98]. Unfortunately, many of these works are published with different datasets and problem settings, making comparisons between methods challenging.

This motivated the creation of a publicly available challenge with a large, open dataset focused on uncertainty quantification: the **Fair Universe HiggsML Uncertainty Challenge** [4], the topic of this chapter. The challenge evaluates uncertainty-aware models using a simulated dataset, paired with a biasing script that emulates systematic uncertainty. The participant's objective this time was to develop an algorithm to estimate the signal strength μ of the Higgs boson signal along with its 1 σ confidence interval prediction. This means that to obtain a good estimate of the signal strength, a robust background-rejection algorithm might be required. The project was funded primarily by the US Department of Energy. Through Chalearn, the project funded my PhD, and I served as the lead developer for the Fair Universe HiggsML Uncertainty Challenge, contributing heavily to data generation, baseline method design, the implementation of the competition protocol, and post-challenge analysis.

4.2 Challenge Dataset

The challenge dataset is a simulated particle physics dataset, built specifically for this challenge. This dataset was designed to provide a sufficiently large sample to both train uncertainty-aware models and obtain a proper estimate of the confidence interval around the predicted parameter of interest. To achieve this goal, we took a *fast and cheap* approach rather than a *robust and slow* one. The dataset [99] was created with two widely-used simulation tools, Pythia 8.2 [100] and Delphes 3.5.0 [101]. This is less accurate than GEANT4, which ATLAS or CMS uses. However, the loss in precision was compensated by a gain in speed, as Delphes simulations are at least 1000 times faster than their GEANT4 counterparts. The simulated dataset considered mimicked the process $pp \to \tau^+\tau^-$ (see Figure 4.1), with one τ decaying hadronically and the other leptonically. The signal was again $H \to \tau^+\tau^-$ and 3 background were chosen ($Z \to \tau^+\tau^-$), $t\bar{t} \to \tau^+\tau^-$ + jets and $VV \to \tau^+\tau^-$ + jets. Table A.1 gives the crossection used for the production of the dataset. The event weights were normalised to reflect an integrated luminosity of 10 fb^{-1}. After production and post-processing, the Events were divided into 2 sets: a Private Test Set, used for testing participants' models, and a Public Training Set, available to the public for training, validation, and testing their models.

4.2.1 Events Features

We have organised the dataset into a tabular format, with each row corresponding to a collision event. The features were selected to match the ATLAS Open dataset used for the 2014 HiggsML Challenge[1], which itself comes from the 2015 ATLAS $H \to \tau^+\tau^-$ discovery paper[102]. There are 16 Primary features (starting with PRI), corresponding to 4 momenta for taus and jets Table 4.1.

[1] $m_{\tau\tau}^{\text{MMC}}$ was removed from the FAIR Universe dataset due to its high computational cost.

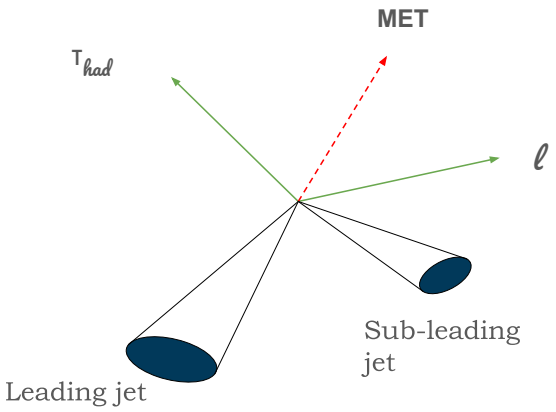


Figure 4.1: Diagram of the particles in the final state chosen: one lepton, one tau hadron, up to two jets, and the missing transverse momentum vector

Table 4.1: List of event features used in the analysis.

Feature	**Description**	**variable name**
$p_{\mathrm{T}}^{\tau_{\mathrm{had}}}$	Transverse momentum of the τ_{had}	PRI_had_pt
$\eta^{\tau_{\mathrm{had}}}$	Pseudorapidity of the τ_{had}	PRI_had_eta
$\phi^{\tau_{\mathrm{had}}}$	Azimuthal angle of the τ_{had}	PRI_had_phi
p_{T}^{ℓ}	Transverse momentum of the lepton	PRI_lep_pt
η^{ℓ}	Pseudorapidity of the lepton	PRI_lep_eta
ϕ^{ℓ}	Azimuthal angle of the lepton	PRI_lep_phi
$p_{\mathrm{T}}^{\mathrm{miss}}$	Missing transverse momentum	PRI_met
ϕ^{miss}	Azimuthal angle of missing transverse momentum	PRI_met_phi
N_{jets}	Number of reconstructed jets	PRI_jet_num
$p_{\mathrm{T}}^{\mathrm{jet_leading}}$	Transverse momentum of leading jet	PRI_jet_leading_pt
$\eta^{\mathrm{jet_leading}}$	Pseudorapidity of leading jet	PRI_jet_leading_eta
$\phi^{\mathrm{jet_leading}}$	Azimuthal angle of leading jet	PRI_jet_leading_phi
$p_{\mathrm{T}}^{\mathrm{jet_subleading}}$	Transverse momentum of subleading jet	PRI_jet_subleading_pt
$\eta^{\mathrm{jet_subleading}}$	Pseudorapidity of subleading jet	PRI_jet_subleading_eta
$\phi^{\mathrm{jet_subleading}}$	Azimuthal angle of subleading jet	PRI_jet_subleading_phi
$\mathrm{Sum}(p_{\mathrm{T}}^{\mathrm{jet}})$	Sum of transverse momenta of all jets	PRI_jet_all_pt

Along with the Primary, we have also released a set of Derived features derived from the Primary to improve sensitivity in signal-background classification. Table 4.2 gives a brief summary of the derived features; More details on the Derived variables are provided in Appendix A.

Table 4.2: List of event features used in the analysis.

Symbol	**Description**	Variable
$m_{\mathrm{T}}(\ell, p_{\mathrm{T}}^{\mathrm{miss}})$	Transverse mass of lepton and met	DER_mass_transverse_met_lep
m_{vis}	Visible invariant mass of τ_{had} and ℓ	DER_mass_vis
$p_{\mathrm{T}}^{\mathrm{H}}$	Vector sum of $\tau_{\mathrm{had}}, \ell, p_{\mathrm{T}}^{\mathrm{miss}}$	DER_pt_h
$\Delta\eta^{\mathrm{jet-jet}}$	Pseudorapidity separation of jets	DER_deltaeta_jet_jet
$m^{\mathrm{jet-jet}}$	Invariant mass of two jets	DER_mass_jet_jet
$\mathrm{Prod}(\eta^{\mathrm{jet-jet}})$	Product of jet pseudorapidities	DER_prodeta_jet_jet
$\Delta R(\tau_{\mathrm{had}}, \ell)$	Angular separation	DER_deltar_had_lep
$p_{\mathrm{T}}^{\mathrm{tot}}$	Vector sum of transverse momentum	DER_pt_tot
$\mathrm{ScalarSum}(p_{\mathrm{T}}^{\mathrm{jet}})$	Scalar sum of transverse momenta	DER_sum_pt
$p_{\mathrm{T}}^{\ell}/p_{\mathrm{T}}^{\tau_{\mathrm{had}}}$	Ratio of transverse momenta	DER_pt_ratio_lep_tau
C_{ϕ}^{miss}	Azimuthal centrality of $p_{\mathrm{T}}^{\mathrm{miss}}$	DER_met_phi_centrality
C_{η}^{ℓ}	Lepton pseudorapidity centrality	DER_lep_eta_centrality

4.2.2 Event Selection and Data Processing

After the production with Pythia and Delphes, a series of preselections was applied to filter out events. First, in this challenge, we focus only on Events in which taus decay into exactly one leptonic tau and one hadronic tau. So, the first step is to count the number of tau hadrons and select only the events that have exactly one leptonic tau and one hadronic tau. Additionally, since the Higgs boson is chargeless, the sum of charges in the two taus should be equal to zero. Hadronic tau (and also the jets) can only be identified in the detector above a certain transverse momentum threshold, so that the raw dataset $p_{\mathrm{T}}^{\tau_{\mathrm{had}}}$, $p_{\mathrm{T}}^{\mathrm{jet_leading}}$ and $p_{\mathrm{T}}^{\mathrm{jet_subleading}}$ have clear thresholds. Figure 4.2 shows the *Cutflow* (progression of the number of events selected as different event selections are applied). When applying systematics shifts (see section 4.3), these thresholds are adjusted so that, if nothing else is done, the threshold positions would be an obvious indication of the nuisance parameter values. To alleviate this, "high thresholds" have been defined, which should be systematically applied after calculating the biased parameters.

The dataset is publicly available on the Zenodo platform [99]. The data is stored as a 16 GB tabular Parquet file [103] and accompanied by a Croissant JSON

Criteria	Low thresholds	High thresholds
N_ℓ	1	
$N_{\tau_{\text{had-vis}}}$	1	
Charge	Opposite charge	
p_{T}^{ℓ} [GeV]	> 20	
$p_{\text{T}}^{\tau_{\text{had-vis}}}$ [GeV]	> 23	> 26
$p_{\text{T}}^{\text{jets}}$ [GeV]	> 23	> 26

Table 4.3: Event selection criteria for the $\tau_\mu\tau_{\text{had}}$ and $\tau_e\tau_{\text{had}}$ channels.

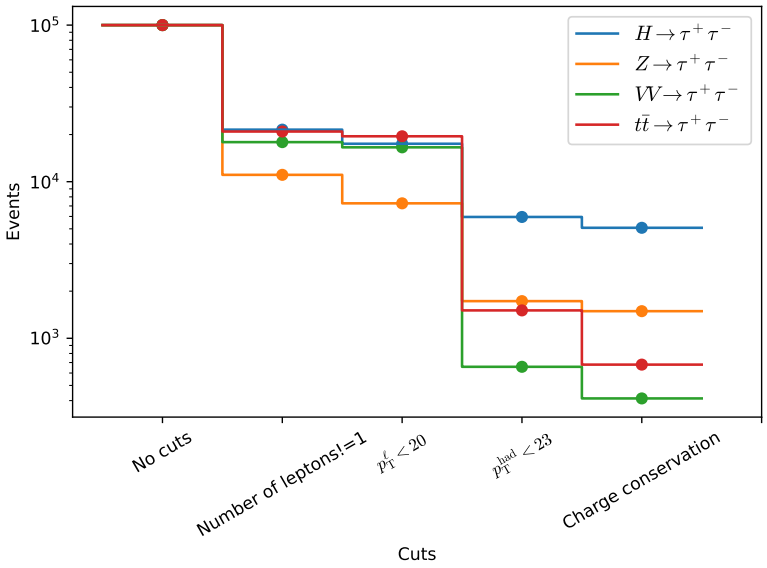


Figure 4.2: *CutFlow*: Progression of the number of events selected as different event selections are applied

metadata file. The dataset paper was selected in the Dataset and Benchmark track of NeurIPS 2025[4]. The dataset comprises 220 million simulated proton-proton collision events and is weighted to represent two weeks of LHC data-taking. Table 4.4 gives the summary of the publicly released data. A separate 120M i.i.d dataset has been used for the final results in section 3.4 and is kept private for future over-training checks. The full dataset was produced with the Perlmutter supercomputer at NERSC[104] consuming 1.8 million CPU hours.

4.3 Implementation of Systematic Biases

The participant's models will be tested on a shifted dataset by introducing six parameterised distortions as a function of six corresponding *Nuisance Parameters*(systematic biases). To help participants prepare for this, we also provide a shift-

Process	Number Generated	LHC Events
$H \to \tau^+\tau^-$	52,101,127	1,015
$Z \to \tau^+\tau^-$	221,724,480	1,002,395
$VV \to \tau^+\tau^- + \mathbf{jets}$	2,105,415	3,783
$t\bar{t} \to \tau^+\tau^- + \mathbf{jets}$	12,073,068	44,190

Table 4.4: Summary of the publicly released data

ing function that transforms the datasets for a given set of nuisance parameter values. There are 6 different types of systematics, three feature-distorting systematics: α_{tes}, α_{jes}, and α_{soft} and three normalisation systematics: $\alpha_{\text{ttbar_scale}}$, $\alpha_{\text{diboson_scale}}$ and $\alpha_{\text{bkg_scale}}$, which change the numbers of each background event-category or weights.

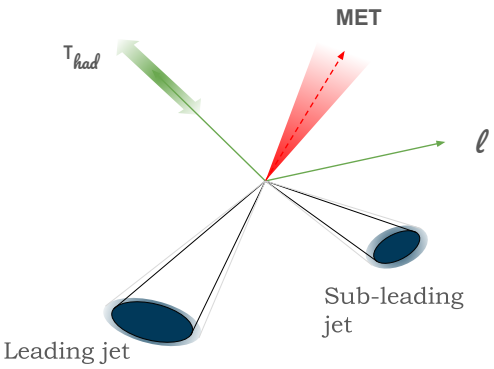


Figure 4.3: The figure shows the shift on different particles in an event due to systematic uncertainty. In green are the τ particles, the one decaying hadronically can have a shift due to α_{tes} uncertainty, represented by the shaded arrows. If the event has jets, the jets can shift due to α_{jes} uncertainty represented as the shaded region.

Table 4.5 lists the different Nuisance Parameters with their Gaussian distribution and the range to which they are clipped. The Gaussian distributions parameterise our ignorance of the exact value of the biases. We think their value is 1 (or zero for α_{soft}). In contrast, their real values differ slightly, as parameterised by their width, thereby biasing our measurement by an unknown amount that can be simulated.

4.3.1 Feature Distortion Systematic biases

To detail the impact of the systematics, we first need to describe how the 4-momenta of the final-state particles can be reconstructed from the primary features. The four parameters (P_x,P_y,P_z, E) of the four-vector of each particle in the final state can be reconstructed from the primary features as follows[2]:

[2]The Missing ET vector is, by definition, in the transverse plane, hence the z component is zero

Variable	Mean	Sigma	Range
Tau-hadron Energy Scale α_{tes}	1.	0.01	[0.9, 1.1]
Jet Energy Scale α_{jes}	1.	0.01	[0.9, 1.1]
Soft Missing Transverse Energy α_{soft}	0.	1.	[0., 5.]
$t\bar{t}$ Normalisation $\alpha_{\text{ttbar_scale}}$	1.	0.02	[0.8, 1.2]
Di-boson Normalisation $\alpha_{\text{diboson_scale}}$	1.	0.25	[0., 2.]
Total Normalisation $\alpha_{\text{bkg_scale}}$	1.	0.001	[0.99, +1.01]

Table 4.5: List of six systematic bias Nuisance Parameters defined in the challenge, with the mean and sigma of their Gaussian (Log-normal for α_{soft}) distribution and their range. The corresponding α is set to the Mean value whenever a systematic bias is switched off.

$$\mathbf{P} = \begin{pmatrix} p_{\text{T}} * \cos(\phi) \\ p_{\text{T}} * \sin(\phi) \\ p_{\text{T}} * \sinh(\eta) \\ p_{\text{T}} * \cosh(\eta) \end{pmatrix} \quad , \quad \mathbf{P}_{\text{met}} = \begin{pmatrix} p_{\text{T}}^{\text{miss}} \cos(\phi_{\text{met}}) \\ p_{\text{T}}^{\text{miss}} \sin(\phi_{\text{met}}) \\ 0 \\ p_{\text{T}}^{\text{miss}} \end{pmatrix} \tag{4.3}$$

Tau-hadron Energy Scale (TES)

α_{tes} is meant to describe the fact that the detector is not calibrated correctly for the measurement of the hadron momentum, meaning when the detector reports a momentum $\mathbf{P}_{\text{had}}$ it really is :

$$\mathbf{P}_{\text{had}}^{\text{biased}} = \alpha_{\text{tes}} \cdot \mathbf{P}_{\text{had}} \tag{4.4}$$

Jet Energy Scale (TES)

Similar to TES, α_{jes} is meant to describe the fact that the detector is not calibrated correctly for the measurement of the jet momenta, meaning when the detector reports a momentum $\mathbf{P}_{jet}$ it really is :

$$\begin{aligned} \mathbf{P}_{\text{leading jet}}^{\text{biased}} &= \alpha_{\text{jes}} \cdot \mathbf{P}_{\text{leading jet}} \\ \mathbf{P}_{\text{subleading jet}}^{\text{biased}} &= \alpha_{\text{jes}} \cdot \mathbf{P}_{\text{leading jet}} \\ \text{Sum}({p_{\text{T}}}_{\text{biased}}^{\text{jet}}) &= \alpha_{\text{jes}} \cdot \text{Sum}(p_{\text{T}}^{\text{jet}}) \end{aligned} \tag{4.5}$$

If the number of jets is three or more, the impact of α_{jes} on missing ET cannot be calculated, given that detailed information on the additional jets (beyond two) is not available; this is a legitimate approximation as the total jet transverse momentum would be, in most cases, dominated by the first two leading. α_{tes} and α_{jes} also have an impact on $\mathbf{P}_{\text{MET}}$: $\mathbf{P}_{\text{MET}}$ is obtained from the opposite of the sum of all visible objects in the event, so that changing one of the visible objects (like $\mathbf{P}_{\text{had}}$, $\mathbf{P}_{\text{leading jet}}$ or $\mathbf{P}_{\text{subleading jet}}$) has a correlated impact on $\mathbf{P}_{\text{MET}}$ (this calculation is performed on the first two coordinates and E_{MET} is recalculated from their modulus):

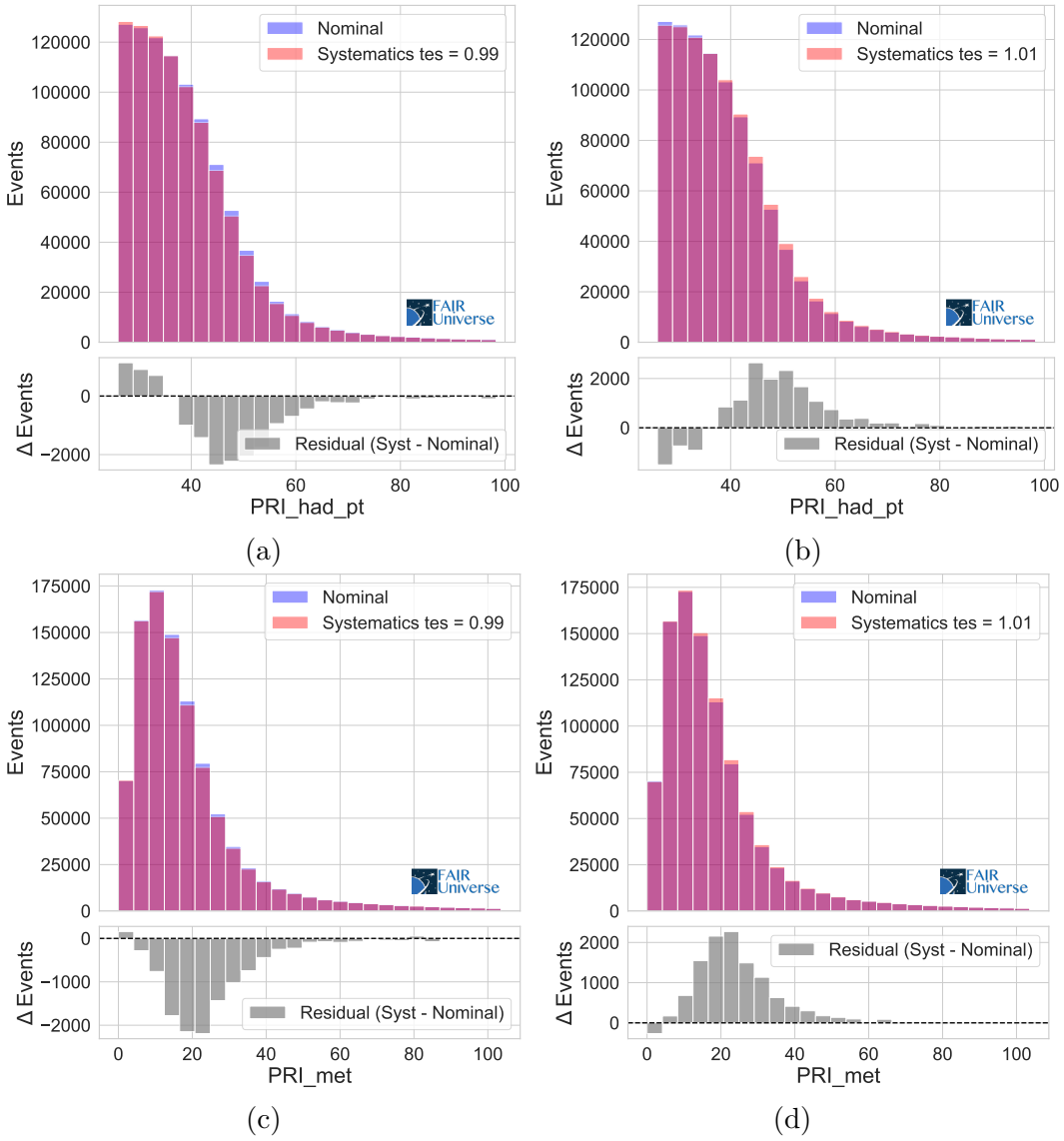


Figure 4.4: Shift in the $p_{\mathrm{T}}^{\tau_{\mathrm{had}}}$ distribution due to Tau Energy Scale Systematics; (a) Down shift of 1% and (b) Up shift of 1% and shift in the $p_{\mathrm{T}}^{\mathrm{miss}}$ distribution due to Tau Energy Scale Systematics; (c) Down shift of 1% and (d) Up shift of 1%

$$\begin{aligned}\mathbf{P}_{\mathrm{MET}}^{biased} = \mathbf{P}_{\mathrm{MET}} &+ (1-\alpha_{\mathrm{tes}})\mathbf{P}_{\mathrm{had}} \\ &+ (1-\alpha_{\mathrm{jes}})\mathbf{P}_{\text{leading jet}} \\ &+ (1-\alpha_{\mathrm{jes}})\mathbf{P}_{\text{subleading jet}}\end{aligned} \tag{4.6}$$

Soft Missing Transverse Energy (Soft MET)

α_{soft} has a different role; it introduces additional noise into the measurement of the missing ET vector, which is not present in the simulation. A random 2D vector of norm $ET_{\mathrm{soft}} = \mathrm{Lognormal}(\alpha_{\mathrm{soft}})$ is added to $\mathbf{P}_{\mathrm{MET}}$ (with different values event by

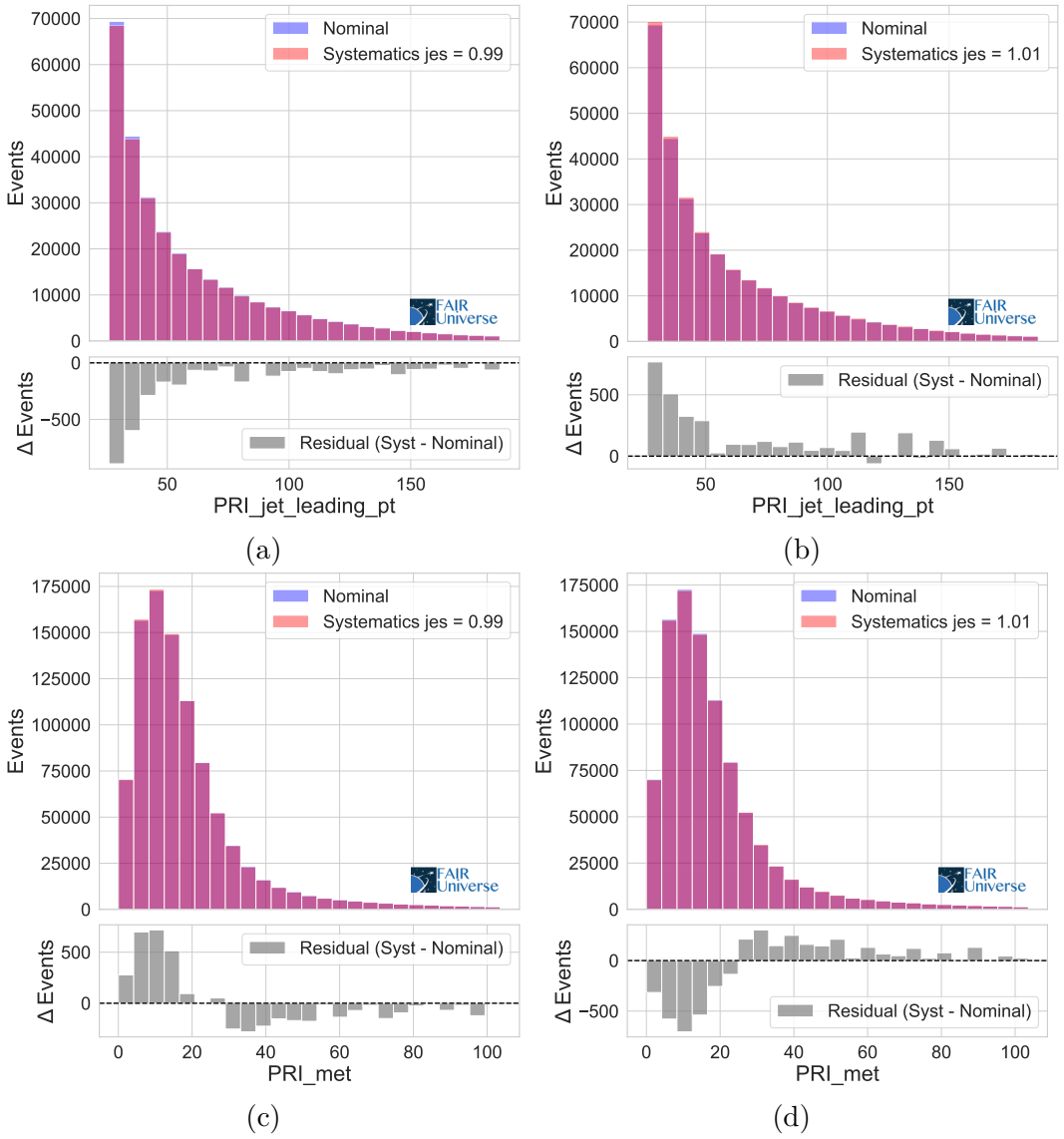


Figure 4.5: Shift in the $p_{\mathrm{T}}^{\mathrm{jet_leading}}$ distribution due to Jet Energy Scale Systematics; (a) Down shift of 1% and (b) Up shift of 1% and sift in the $p_{\mathrm{T}}^{\mathrm{miss}}$ distribution due to Jet Energy Scale Systematics; (a) Down shift of 1% and (b) Up shift of 1%

event, by contrast with α_{soft}, which has a fixed value for a given pseudo-experiment) (this calculation is performed on the first two coordinates and E_{MET} is recalculated from their modulus):

$$\mathbf{P}_{\mathrm{MET}}^{biased} = \mathbf{P}_{\mathrm{MET}} + \begin{pmatrix} Gauss(0, ET_{\mathrm{soft}}) \\ Gauss(0, ET_{\mathrm{soft}}) \\ 0 \\ 0 \end{pmatrix}$$

Derived features are also impacted if they depend on these Primary features. Thus, for each of α_{tes}, α_{jes} and α_{soft}, different features are impacted in a correlated

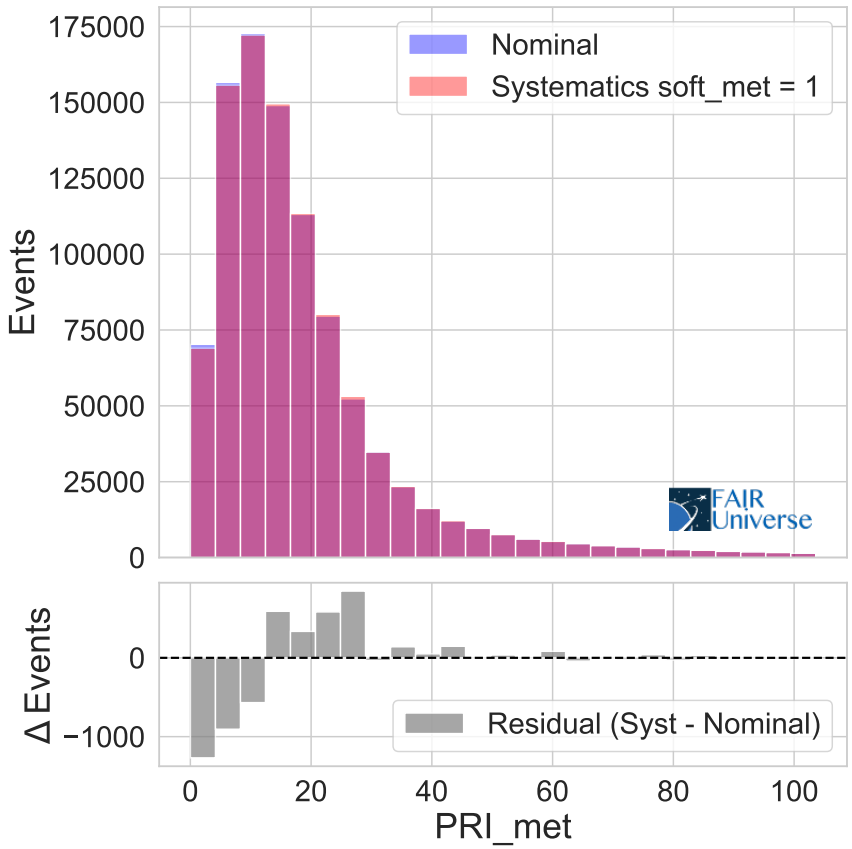


Figure 4.6: Shift in the $p_{\mathrm{T}}^{\mathrm{miss}}$ distribution due to soft-met systematics with 1Gev shift to ET_{soft}

way.

4.3.2 Normalisation Systematics

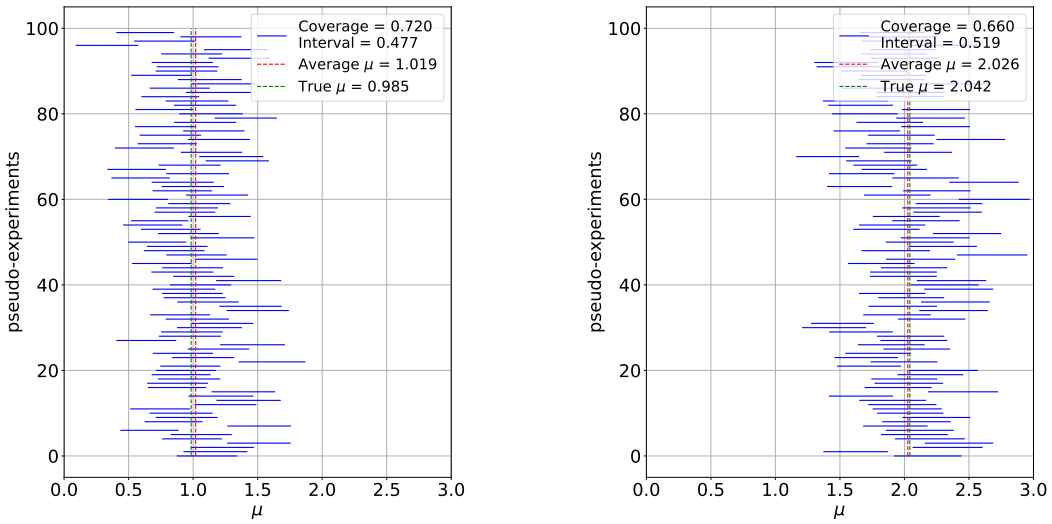
So $\alpha_{\mathrm{bkg_scale}}$ only affects the overall background level, leaving the background distributions unchanged. $\alpha_{\mathrm{ttbar_scale}}$ and $\alpha_{\mathrm{diboson_scale}}$ affect only the proportions of the smaller backgrounds, thereby distorting the overall background distribution.

$$w^{\mathrm{bias}} = \begin{cases} \alpha_{\mathrm{bkg_scale}} \times w, & \text{if } \texttt{DetailedLabel} = \texttt{"ztautau"}, \\ \alpha_{\mathrm{bkg_scale}} \times \alpha_{\mathrm{ttbar_scale}} \times w, & \text{if } \texttt{DetailedLabel} = \texttt{"ttbar"}, \\ \alpha_{\mathrm{bkg_scale}} \times \alpha_{\mathrm{diboson_scale}} \times w, & \text{if } \texttt{DetailedLabel} = \texttt{"diboson"}. \end{cases}$$

4.4 Metric and Evaluations

As this challenge focuses on uncertainties, the participant's model will be tested on a dataset with systematics and nuisance parameters with unknown values. We also want to gauge participants' ability to predict their models' uncertainty correctly. So a simple Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)

against optimal values of μ could gauge the accuracy of the model, but not the 1-sigma Confidence Interval (CI) for μ predicted. The solution we opted for was to validate the accuracy of the CI using pseudo-experiments. We test the participants' model multiple times using bootstrap datasets, each with random values for the nuisance parameters and the signal strength μ. For each pseudo-experiment i, the model returns the 16th and 84th quantiles, $\mu_{16,i}$ and $\mu_{84,i}$, respectively, which bound the CI. Figure 4.7 gives examples of the prediction from a good model on different values of μ_{true}.



(a) (b)

Figure 4.7: *Coverage plot for 2 different* μ_{true}: all the predicted intervals (blue lines) for each pseudo experiment generated for a given μ_{true} (vertical dotted line). The coverage (here $70 \pm 5\%$) is determined by the fraction of time the horizontal blue lines intersect the vertical line.

Notice that the μ_{true} falls within the predicted CIs (blue lines) around 68% of the time, which is expected as they are 1-sigma CIs. The challenge of designing a good metric is that, ultimately, we need models with good coverage and a small Confidence interval. Unlike accuracy, we can't use RMSE or MAE against the optimal CI value (68.27%), since overcoverage is less concerning than undercoverage. So metric based on two criteria: **Average Interval Width w** and **Coverage** c is constructed, here is the breakdown:

- Average Interval Width, the average of the predicted widths of each pseudo-experiment (small the better) is computed as

$$\mathbf{w} = \frac{1}{N_{test}} \sum_{i=1}^{N} |\mu_{84,i} - \mu_{16,i}|$$

- **Coverage**, the frequency with which μ_{truth} is covered by the CI (the closer to the standard 68.27% probability, the better) is computed as

$$c = \frac{1}{N_{test}} \sum_{i=1}^{N} 1 \text{ if } \mu_{\text{true},i} \in [\mu_{16,i}, \mu_{84,i}]$$

- Since we only have a finite number of pseudo-experiments (N_{test}), we introduce σ_{68}, the binomial statistical error on c as

$$\sigma_{68} = \sqrt{\frac{(1 - 0.6827)0.6827}{N_{\text{test}}}}$$

- A penalising function $f(c)$ has been defined to penalise the departure of c from the expected 68.27%.

$$f(c) = \begin{cases} 1 + \left| \dfrac{c - (0.6827 - 2\sigma_{68})}{\sigma_{68}} \right|^4, & \text{if } c < 0.6827 - 2\sigma_{68}, \\ 1, & \text{if } 0.6827 - 2\sigma_{68} \leq c \leq 0.6827 + 2\sigma_{68}, . \\ 1 + \left| \dfrac{c - (0.6827 + 2\sigma_{68})}{\sigma_{68}} \right|^3, & \text{if } c > 0.6827 + 2\sigma_{68}. \end{cases} \tag{4.7}$$

  We opted for an asymmetric penalty function because, within the High Energy Physics (HEP) field, overestimating uncertainty is deemed more acceptable than underestimating it [105, 106]. Hence, coverage exceeding 68.27% incurs a lower penalty than coverage below 68.27%.

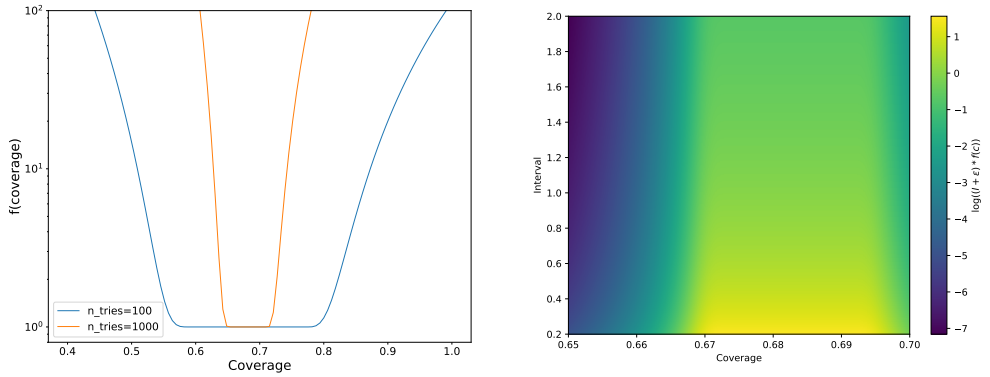
- Finally, we add $\epsilon = 10^{-2}$ as a regularisation term to guard against submissions that report unrealistically narrow CIs and get the **Quantile Score** used to rank participants:

$$\boxed{\text{score} = -\ln(\mathbf{w} + \epsilon) f(c))} \tag{4.8}$$

4.5 Benchmark models and Starting Kit

The **Starting Kit** Repository contains a series of Jupyter notebooks, each containing different benchmark models that can quantify systematic uncertainty to a good extent. Along with the code to run the competition locally, the starting kit also includes code for installing necessary packages, loading and visualising data, and training and evaluating a model using the metrics described in section 4.4. In LHC, we analyse a very large number of events in search of a very rare process, which essentially gives us a Binomial distribution with a very large number of trials (n) and a very small probability of success (p). The Poisson limit theorem states that such processes can be modelled using a Poisson distribution. The probability of observing n events with an expected rate of ν is given by

$$P(n|\mu) = \frac{\nu(\mu)^n}{n!} e^{-\nu(\mu)} = \tag{4.9}$$



(a) (b)

Figure 4.8: (a) Coverage Penalty function and (b) distribution of quantile score with respect to interval and coverage.

Given a dataset $\mathcal{D}$, we can rewrite $P(n|\nu(\mu))$ as the likelihood $\mathcal{L}(\mu)$.

$$\mathcal{L}(\mu|\mathcal{D}) = \frac{\nu(\mu)^n \cdot e^{-\nu(\mu)}}{n!} \prod_i^n p(x_i|\mu) \tag{4.10}$$

In order to utilise the per-event part of the likelihood ($p(x_i|\mu)$) we need to have a precise estimate of the prior probability p. In the absence of such a perfect estimate for the prior, we ignore them in our likelihood estimate. This will reduce the statistical strength of the parameter estimation. From the Neyman-Pearson lemma, we know that the best test statistic for a parameter of interest is obtained by the ratio (λ_μ) of the likelihood under different hypotheses. The expression for the likelihood ratio is given by:

$$\frac{\mathcal{L}(\mu)}{\mathcal{L}(1)} = \left(\frac{\nu(\mu)}{\nu(1)}\right)^n e^{-(\nu(\mu)-\nu(1))} \tag{4.11}$$

To simplify the calculation, we can compute the two times negative log of the likelihood ratio, thus giving the expression of t_μ

$$t_\mu = -2\log\left(\frac{\mathcal{L}(\mu)}{\mathcal{L}(1)}\right) = -2n\log\left(\frac{\nu(\mu)}{\nu(1)}\right) + (\nu(\mu) - \nu(1)) \tag{4.12}$$

By maximising the likelihood ratio or minimising the negative log-likelihood ratio (NLL), we can find the most probable value of μ for the given dataset. If the estimated signal strength μ is **significantly** different from the null hypothesis $\mu = 0$, then we can conclude that new physics has been discovered. The key issue with this method is that, given the extreme rarity of the events we observe, the chance of discovering new physics is very low unless we find a region of phase space with a higher signal density. In the very early days of the LHC, the most prominent way to find this region was to use the per-event information x_i to construct specialised variables that separated the signal from the background. This is often known as the cut-and-count approach. The specialised variable was later replaced by scores

from machine-learning classifiers. For the machine-learning discriminant, we used an XGBoost [107] Binary classifier, with Higgs Boson events as the signal.

4.5.1 Summary Histogram with Machine Learning Discriminants

The advantage of using the cut-and-count method is that we obtain a region with very high signal density, enabling precise measurements. But it comes at the cost of losing information outside the region of interest. This is solved by using a binned discriminant. Instead of choosing a region above a threshold, the whole discriminant is binned, and the likelihood function now depends on the product of each bin.

$$\mathcal{L}(\mu, \theta) = \prod_{k=1}^{N_{\text{bins}}} \frac{(\mu S_k + B_k)^{n_k} e^{-(\mu S_k + B_k)}}{n_k!} \tag{4.13}$$

where n_k is the number of events in bin k and S_k and B_k are respectively the expected number of signal and background in bin k. We can now perform the same hypothesis test we saw in Equation 4.12 but on each bin, and we get;

$$t_\mu = -2 \log \left(\frac{\mathcal{L}(\mu)}{\mathcal{L}(1)} \right) = 2 \sum_{k=1}^{N_{\text{bins}}} -n_k \log(\mu S_k + B_k) + (\mu S_k + B_k) + \text{const} \tag{4.14}$$

4.5.2 Quantifying Systematic Uncertainity

When unknown biases occur, the prediction on the amount of background and signal events per bin will be wrong, biasing the estimation of μ. The simplest solution is to build a statistical model of the systematics and use it in the parameter estimation.

Parameter Estimation with only one systematics

In chapter 4, we addressed the shifting script provided with the data, which enables us to train our classifier with varying systematics. For the simplest case of a single systematic, we can map the shift in each bin from the nominal ($\theta = 1$) to an up- (($\theta > 1$)) or down-shifted ($0 < \theta < 1$) dataset. From these points, we can construct a polynomial interpolation curve to estimate the signal ($S_k(\theta)$) and background ($B_k(\theta)$) in each bin. Figure 4.9 shows one such fit curve for the 24th bin.

We add a Gaussian constant $\pi(\theta)$ to constrain the distribution of θ. Finally, the minimisation function $\mathcal{L}_{\text{full}}(\mu, \theta)$ now regresses both μ and θ, thus making the model less susceptible to systematic bias.

$$\mathcal{L}_{\text{full}}(\mu, \theta) = \left[\prod_{i=1}^{N_{\text{bins}}} \frac{(\mu B_k(\theta) + B_k(\theta))^{n_k} e^{-(\mu B_k(\theta) + B_k(\theta))}}{n_k!} \right] \cdot \pi(\theta) \tag{4.15}$$

Parameter Estimation With Multiple Systematics

We could continue the process across multiple nuisance parameters, but we quickly run into the issue of correlated uncertainty. If we assume that all nuisance

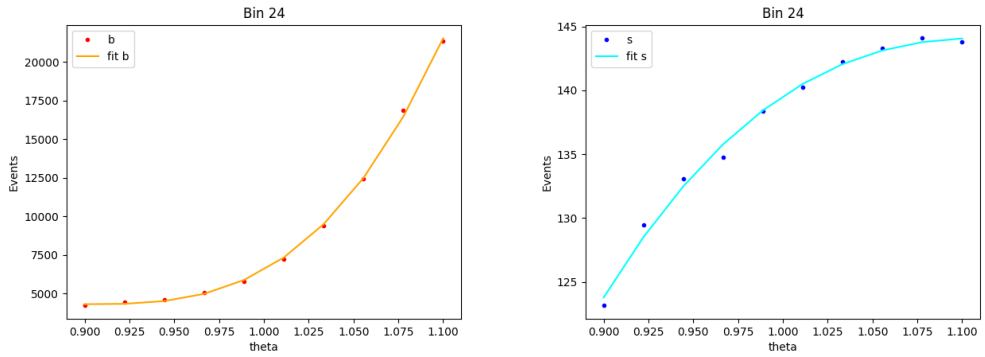


Figure 4.9: Interpolation curve to estimate the signal and background in the 24th bin

parameters are correlated, we will need to perform N-dimensional fits and generate samples for every combination of NP values. In many LHC analyses, we have close to 100 nuisance parameters, making this almost impossible. The approach, more common in ATLAS and other LHC analyses, and implemented in the HistFactory[108] framework, uses only 2 shifted datasets for each NP: 1 σ up (η^+) and 1 σ down (η^-). This is because, unlike FAIR Universe, real ATLAS or CMS GEANT4 simulations are a lot more costly. Secondly, it assumes that the NPs are uncorrelated, allowing us to add the effects of each NP, simplifying the calculations. Since the systematics are uncorrelated, we can build an interpolation function $W(\alpha_m)$ where each NP m. The Hist-Factory Approach uses a combination of a 6th-degree polynomial interpolation function and exponential extrapolation functions. For $(-1 > \alpha_m > 1)$, it follows the polynomial function. Beyond it, it follows the exponential function. Equation 4.16 gives the interpolation function for a sample X at bin i.

$$W_k^X(\alpha_m) = \begin{cases} (\eta^+)^{\alpha_m}, & \text{if } \alpha_m > 1 \\ 1 + \sum_{n=1}^{6} c_n \alpha^n{}_m, & \text{if } -1 \leq \alpha_m \leq 1 \\ (\eta^-)^{-\alpha_m}, & \text{if } \alpha_m < -1 \end{cases} \tag{4.16}$$

Using a 6th-degree polynomial would require 7 constraints. We obtain 3 directly from the 3 points; then we can apply the differentiability and continuity conditions at the 2 boundaries of the polynomial, yielding the 4 additional constraints required for interpolation. Figure 4.10, gives the interpolation function for $\eta^- = 1.1$ and $\eta^+ = 1.5$

Finally, we can merge the individual systematics m with $W_k^X(\vec{\alpha}) = \prod_m W_k^X(\alpha_m)$ and with $\nu_k(\vec{\alpha}) = \mu S_k W_k^S(\vec{\alpha}) + B_k W_k^B(\vec{\alpha})$ we get

$$\mathcal{L}_{\text{full}}(\mu, \alpha) = \left[\prod_{i=1}^{N_{\text{bins}}} \frac{\nu_k(\vec{\alpha})^{n_k} e^{-\nu_k(\vec{\alpha})}}{n_k!} \right] \cdot \prod_m \pi_m(\alpha_m) \tag{4.17}$$

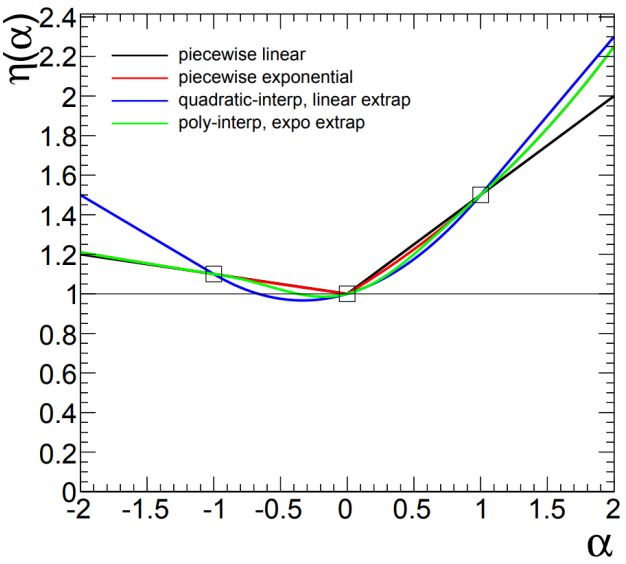


Figure 4.10: Interpolation function for $\eta^- = 1.1$ and $\eta^+ = 1.5$ [108]

4.5.3 Profile Likelihood for Uncertainty Estimation

If our interest were only in finding the best point estimate, maximising the joint likelihood (or minimising its negative log-likelihood) would suffice. However, a point estimate alone provides no quantitative measure of uncertainty. To obtain a rigorous uncertainty estimate that naturally incorporates both statistical fluctuations and systematic variations, we perform a *profile likelihood analysis*. We define the profile likelihood ratio test statistic q_μ as:

$$q_\mu = -2 \ln \left(\frac{\mathcal{L}(\mu, \hat{\hat{\alpha}}(\mu))}{\mathcal{L}(\hat{\mu}, \hat{\alpha})} \right) \tag{4.18}$$

Here, $(\hat{\mu}, \hat{\alpha}) = \text{argmax}_{\mu,\alpha}\, \mathcal{L}(\mu, \alpha)$ represents the global Maximum Likelihood Estimators (MLE) across all parameters simultaneously and $\hat{\hat{\alpha}}(\mu) = \text{argmax}_{\alpha}\, \mathcal{L}(\mu, \alpha)$ is the *conditional maximum likelihood estimator* of the nuisance parameters α, evaluated independently for each fixed value of the signal strength μ. Rather than fixing α to nominal values, profiling allows α to re-optimise to its conditional best-fit value $\hat{\hat{\alpha}}(\mu)$ at every candidate value of μ. When evaluating a signal strength μ away from the global fit $\hat{\mu}$, the nuisance parameters adjust dynamically to absorb a portion of the discrepancy between model predictions and observed data. This conditional adaptability ensures that:

$$\mathcal{L}(\mu, \hat{\hat{\alpha}}(\mu)) \geq \mathcal{L}(\mu, \alpha_0) \tag{4.19}$$

where α_0 represents frozen nuisance parameters. Consequently, the likelihood ratio drops more slowly, which **flattens the profile curve** of q_μ. This broadening of the test-statistic curve directly translates the uncertainty in α into a wider confidence interval for μ.

By Wilk's Theorem, the test statistic q_μ asymptotically follows a chi-squared distribution with 1 degree of freedom ($q_\mu \sim \chi^2_1$). The combined statistical and

systematic 68.27% (1σ) central confidence interval $\text{CI}_{68\%} = [\mu_{16}, \mu_{84}]$ is obtained by finding the roots where the test statistic equals unity:

$$\text{CI}_{68\%} = [\mu_{16}, \mu_{84}] \quad \text{where} \quad q_{\mu_{16}} = q_{\mu_{84}} = 1 \tag{4.20}$$

4.6 Competition Overview

The competition was hosted on the Codabench platform, which allowed participants to submit their models as code. This allowed us to repeatedly re-evaluate the same model using many pseudo-experiments. To safely test ML models repeatedly, the Perlmutter supercomputer[104] served as the competition's backend. Each model, hence, ran on a node with 128 GB of RAM, 2 CPUs, and 4 A100 GPUs, which were state-of-the-art in AI at the time. The competition ran from September to April and received roughly 300 submissions from 180 participants. The competition was conducted in 2 phases: a Public Phase on Codabench, which featured a live benchmark that allowed participants to submit their models and receive results within a day. In this phase, we used 10 sets, each with 100 pseudo-experiments. To ensure efficient resource use, each participant's model inference was executed within a 20-second time limit on the Purlmuter GPU Node. The values of μ_{truth} were randomly generated to avoid participants gaming the system. The best submissions from this phase were moved to the Final Phase of the competition.

4.6.1 The Final Phase

At the end of the competition, a clear trio was at the top of the public leaderboard: **HEPHY** with a quantile score of 0.878, followed by **IBRAHIME** (0.823) and **HZUME** (0.179). For the Final Phase, participants were asked to nominate their best model, which was then re-evaluated privately on a new dataset (i.i.d. with the original one). To ensure fairness, we froze all the seeds within the competition algorithm, thereby ensuring that all participants were evaluated under identical conditions (same nuisance parameters, same μs, same pseudo-experiments). The submissions were evaluated across 1000 sets, each with 100 pseudo-experiments, to minimise statistical variance. Figure 4.11 shows the results for all trials for the trio. The $\text{CI}_{68\%}$ width is seen falling at large values of μ: this is due to the clipping of the Confidence Interval to a maximum value of 3 (which was not done in Figure 4.7), which was the maximum value in this competition. Such clipping would be meaningless in the context of a real physics measurement where μ is truly unknown. This is the only "hack" identified as specific to the competition context. When merging all trials, the scores obtained by the top trio are: **HEPHY** -0.582, Ibrahim: -0.576, and **HZUME** -2.16.

Final Ranking

- 1st tie: **HEPHY** with *"Unbinned inclusive crosssection measurements with machine-learned systematic uncertainties"* [109] (Lisa Benato, Cristina Giordano, Claudius Krause, Ang Li, Robert Schöfbeck, Maryam Shooshtari, Dennis Schwarz, Daohan Wang) from Marietta Blau Institute for Particle Physics

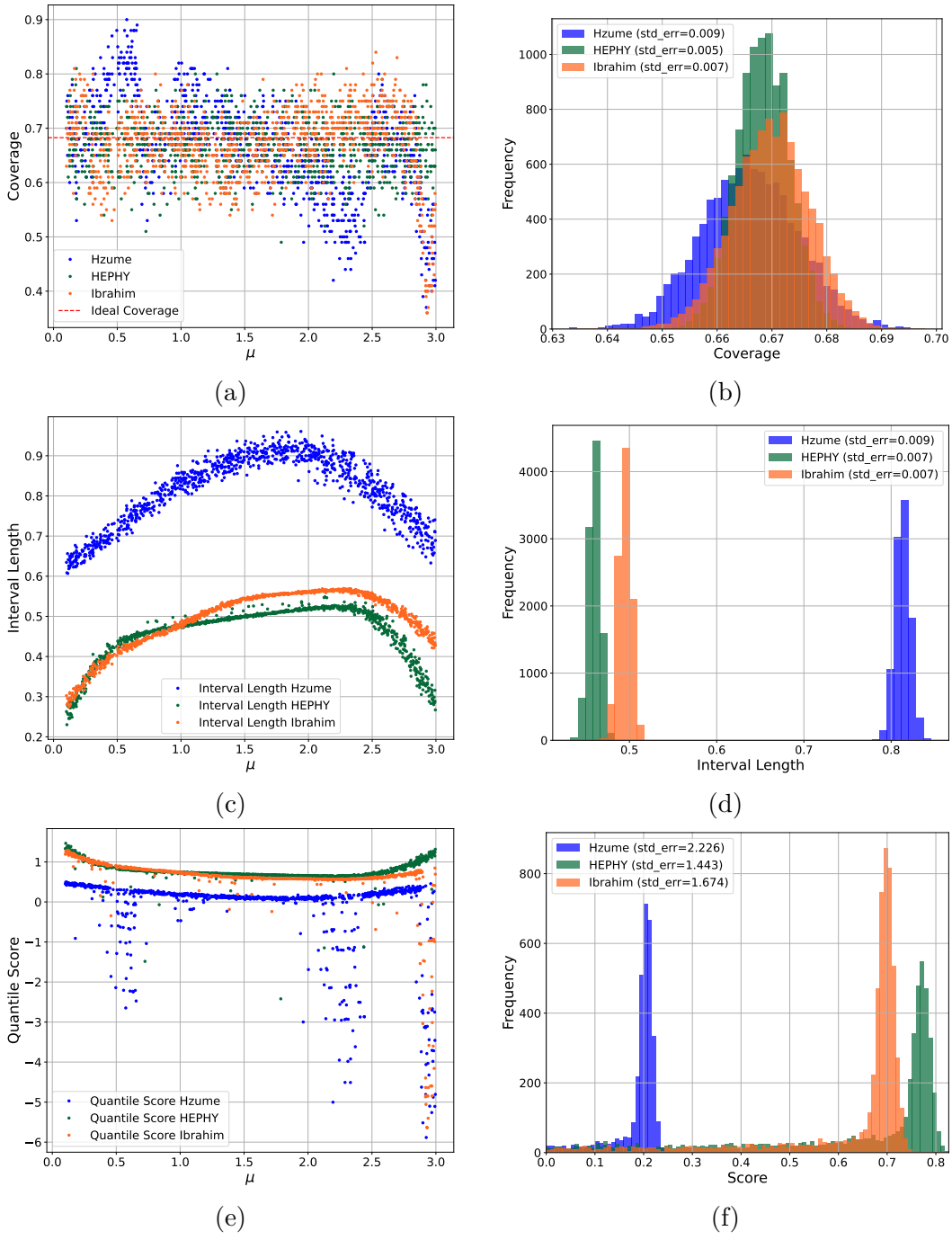


Figure 4.11: Comparative study of the three finalists (blue for **HZUME**, green for **HEPHY** and orange for **IBRAHIME**) with 1000 trials of 100 pseudo-experiments (see section 4.4). 4.11a the coverage from each trial, 4.11c the average $\mathrm{CI}_{68\%}$ width and 4.11e the quantile score and 4.11b gives the distribution of the coverage, 4.11d gives the distribution of the $\mathrm{CI}_{68\%}$ width and 4.11f gives the distribution of the final score from the trials.

(**MBI**), Vienna[3].

- 1st tie **IBRAHIME** (Ibrahim Elsharkawy) with *"Contrastive Normalizing Flows for Uncertainty-Aware Parameter Estimation"* [110] (Ibrahim Elsharkawy) from University of Illinois at Urbana-Champaign, USA wins $2000.
- 3rd : **HZUME** (Hashizume Yota) with *"Decision-Tree Aggregated Features and Hybrid Bin-Classifier / Quantile-Regressor"* from Kyoto University, Japan wins $500

A closer look at the winning solutions reveals that they use **Simulation-Based Inference** at their core. SBI is a **Machine Learning** method that aims to learn the N-dimensional shape of the event distributions and use that to infer the parameter of interest directly. Hence, will discuss SBI and the 2 First-prize solutions in the next chapter (subsection 5.2.1 and subsection 5.2.2).

4.7 Conclusions and Outlook

The competition brought together cutting-edge infrastructure for AI training and inference, large datasets, and a standardised scoring method for uncertainty computation. The dataset is permanently stored in the Zenodo platform [99] and publicly available, making it a potential standard benchmark for uncertainty quantification in HEP. Owing to its ease of use, the dataset has also been adopted for machine learning teaching exercises at CentraleSupélec in France. Beyond this specific metric, we expect that this unique large dataset equipped with a biasing script will be the basis of future studies, for example: (i) the precise parametrisation of density and density ratios over several order of magnitudes which is fundamental to precision physics (ii) development of morphing/Optimal Transport techniques to parameterise multidimensional non-parametric biases (iii) the same studies but with a focus on learning with a limited number of instances. The competition concluded with 2 winning solutions from **HEPHY** and **IBRAHIME**, both competitive yet different; however, their results are not highly correlated, suggesting the optimum has not yet been reached. We believe that submissions from the FAIR Universe challenge can push the boundaries of Uncertainty-Aware Artificial Intelligence in the coming years, within and outside the HEP community. With the growing size of datasets in high-energy physics, the sophistication of tools, and the precision requirements to explore new phenomena, uncertainty quantification will be an essential part of machine learning in the future.

[3]Formerly Institute for High Energy Physics (HEPHY)

Chapter 5
Simulation-Based Inference with FAIR Universe Dataset

Essentially, all models are wrong,
But some are useful.
- George E. P. Box

The success of the binned likelihood is that ML classifiers are designed to increase the entropy or information content in each bin, which in turn maximises the precision of the parameter estimation. However, this is mathematically only true for single binary classification. If we have more categories than just the signal background, a single classifier score does not provide sufficient statistics. The go-to solution in these cases is to use multi-dimensional binning with various classifier scores. But even in those cases, we would often need large bins to avoid empty or negative bins in our computation. Additionally, processes with large destructive interference, such as off-shell Higgs production, Di-Higgs, EFTs, etc., tend to have a larger fraction of events with negative weights, further worsening the issue. An alternative solution is Simulation-Based Inference. **Simulation-Based Inference**(**SBI**) is a machine learning method that aims to learn the N-dimensional shape of the event distributions and use that to infer the parameter of interest directly. As we discussed in section 4.5, the likelihood ($\mathcal{L}$) for a parameter of interest given a dataset can be expressed as

$$\mathcal{L}(\mu|\mathcal{D}) = \frac{\nu(\mu)^n \cdot e^{-\nu(\mu)}}{n!} \prod_i^n p(x_i|\mu) \tag{5.1}$$

From the Neyman-Pearson lemma, we know that the best test statistic for a parameter of interest is obtained by the ratio (λ_μ) of the likelihood under different hypotheses.

$$\lambda_\mu = \frac{\mathcal{L}(\mu|\mathcal{D})}{\mathcal{L}(1|\mathcal{D})} = \frac{\nu(\mu)^n \cdot e^{-\nu(\mu)}}{\nu(1)^n \cdot e^{-\nu(1)}} \prod_i^n \frac{p(x_i|\mu)}{p(x_i|1)} \tag{5.2}$$

The essence of SBI is to utilise machine learning algorithms that take in high-level information about an event and derive an approximate likelihood $p(x_i|\mu)$, for any parameter of interest (μ). To achieve this, one approach is to train a parameterised neural network that takes both the event and the parameters of interest as input. Unfortunately, when we start addressing systematic uncertainty, starts to become

harder to calibrate for all the parameters. A more practical approach is the **Semi-Parametric Simulation-Based Inference**.

5.1 Semi-Parametric Simulation-Based Inference

The purpose of the Semi-Parametric Simulation-Based Inference is to separate the parameters (μ in our case) from the probability of each process per event.

$$p(x_i|\mu) = \frac{1}{\nu(\mu)} \sum_{X}^{C} f_X(\mu) \cdot \nu_X \cdot p_X(x_i) \tag{5.3}$$

Here, X is any process from C, the set of all processes. The $f_X(\mu)$ is the dependence of the parameter of interest for different processes X. The theory description gives $f_X(\mu)$ for each process. Lastly, we have the yield terms ν_X for each process X and $\nu(\mu)$ is the overall yield for a given value of μ to normalise $p(x_i|\mu)$. Thus, making it easier to estimate the overall likelihood than to train parametric neural networks, which are hard to train for realistic physics analyses. In Semi-Parametric SBI, we can often decorrelate nuisance parameters (NPs) from the main model, enabling us to scale SBI to a larger number of NPs.

5.1.1 Approximating probability ratios with Classifiers

There are two main approaches to determine the $p_X(x_i)$ for any event x_i: **Neural Likelihood Estimation**(**NLE**) and **Neural Density Estimation**(**NDE**). Though NLE has found some success in toy studies, most full LHC analyses follow a variation of NDE, as is the case for this thesis. The 2015 *"Approximating Likelihood Ratios with Calibrated Discriminative Classifiers"*[55] paper proposes a more effective method, which is to instead approximate ratios of probabilities with the help of the score of a binary classifier. This approach was successfully implemented by the ATLAS Collaboration on real data, where they reanalysed the off-shell Higgs Boson signal strength measurement for the Run-2 data[1].

Many loss functions can achieve this goal; [111] gives a summary of many such functions. We can mathematically demonstrate this convergence for the case of the binary cross-entropy (BCE) ($\mathcal{L}[s(x_i)]$);

$$\mathcal{L}[s(x)] = [y \log s(x) + (1-y) \log(1-s(x))] \tag{5.4}$$

When the optimal model that minimises the BCE loss ($\mathcal{L}[s(x)]$) we get

$$\begin{aligned} \frac{\partial}{\partial s} E[\mathcal{L}[s(x)]|x] &= -\frac{p(x|\mathcal{H}_1) \cdot \nu(\mathcal{H}_1)}{s(x)} + \frac{p(x|\mathcal{H}_0) \cdot \nu(\mathcal{H}_0)}{1-s(x)} = 0 \\ \implies s(x_i) &= \frac{p(x|\mathcal{H}_1) \cdot \nu(\mathcal{H}_1)}{p(x|\mathcal{H}_0) \cdot \nu(\mathcal{H}_0) + p(x|\mathcal{H}_1) \cdot \nu(\mathcal{H}_1)} \end{aligned} \tag{5.5}$$

Where $\mathcal{H}_0$ is the target process and $\mathcal{H}_1$ is the reference hypothesis. When training, if one uses a balanced dataset (where the sum of weights for both classes is the same, which implies $\nu(\mathcal{H}_0) = \nu(\mathcal{H}_1)$), we could rewrite Equation 5.6 as

$$s(x_i) = \frac{p(x_i|\mathcal{H}_1)}{p(x_i|\mathcal{H}_0) + p(x_i|\mathcal{H}_1)} \implies \frac{p(x_i|\mathcal{H}_0)}{p(x_i|\mathcal{H}_1)} = \frac{s(x_i)}{1 - s(x_i)} \tag{5.6}$$

5.1.2 Validation of Density ratio Estimators

The theoretical guarantee that the ML algorithm's density-ratio approximation is good enough is subject to practical limitations. The distribution of the reference sample, the phase space coverage of the reference and the target sample, the number of events for training, etc. An essential part of the SBI workflow is validating the ML classifiers. Traditional metrics like AUC and ROC, which we discussed earlier, measure a classifier's separation power but not its ability to estimate the density ratio accurately. Calibration curves and reweighting plots achieve this.

Calibration curves

The calibration curve is a direct test of Equation 5.6. For this, we separately build histograms of the target ($p(x_i|\mathcal{H}_0)$)and reference samples ($p(x_i|\mathcal{H}_1)$) over the score distribution, and combine them as per Equation 5.6. If the ML classifier is well-trained, then the height of each bin in the combined histogram will match the bin edge, meaning the bin at score = 0.4 will have a frequency of 0.4. Figure 5.1 shows the calibration curves for the S vd B Network trained with the BDT and TabICL classifiers as examples of good and bad calibration curves[1]. The difference is better seen in the residue plot below.

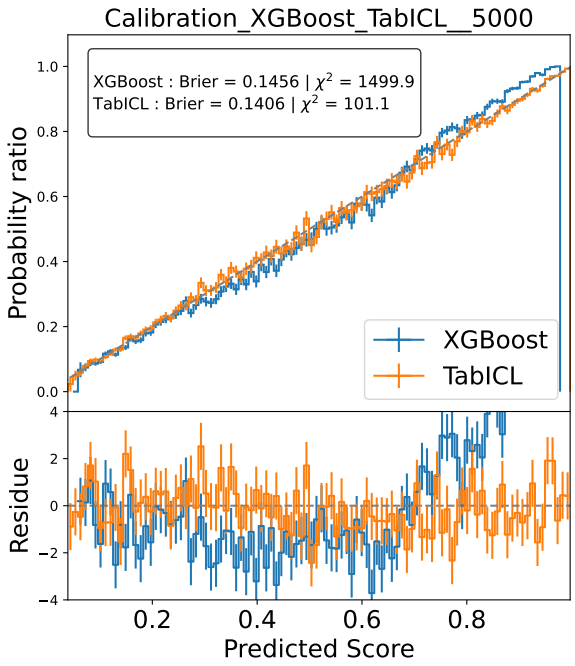


Figure 5.1: Calibration curves for two classifiers, TabICL in orange (example of good calibration) and BDT in blue (example of bad calibration)

[1]The full analysis for Tabular Foundational Models like TabICL is in Appendix B.

Reweighting curves

An indirect method to measure the accuracy of the predicted density ratios $\left(\frac{p_X(x_i)}{p_{ref}(x_i)}\right)$ is by reweighting the samples of the one-hypothesis distribution (e.g. $\mathcal{H}_1$) with the predicted density ratio. If the network is perfectly trained, then the reweighted distribution will match the other hypothesis distribution (e.g. $\mathcal{H}_0$).

$$w_i^{\mathcal{H}_0} \simeq w_i^{\text{rw}} = w_i^{\mathcal{H}_1} \left(\frac{p(x_i|\mathcal{H}_0)}{p(x_i|\mathcal{H}_1)}\right) \tag{5.7}$$

Figure 5.2 gives the reweighted plots for different variables. One can see that the signal-reweighted histogram (in orange) is almost perfectly aligned with the background (in green). This tells us whether a classifier correctly predicts the density ratio, but only from the perspective of the chosen variable. For a more thorough verification, one must perform this test across various features, including those not used for classifier training.

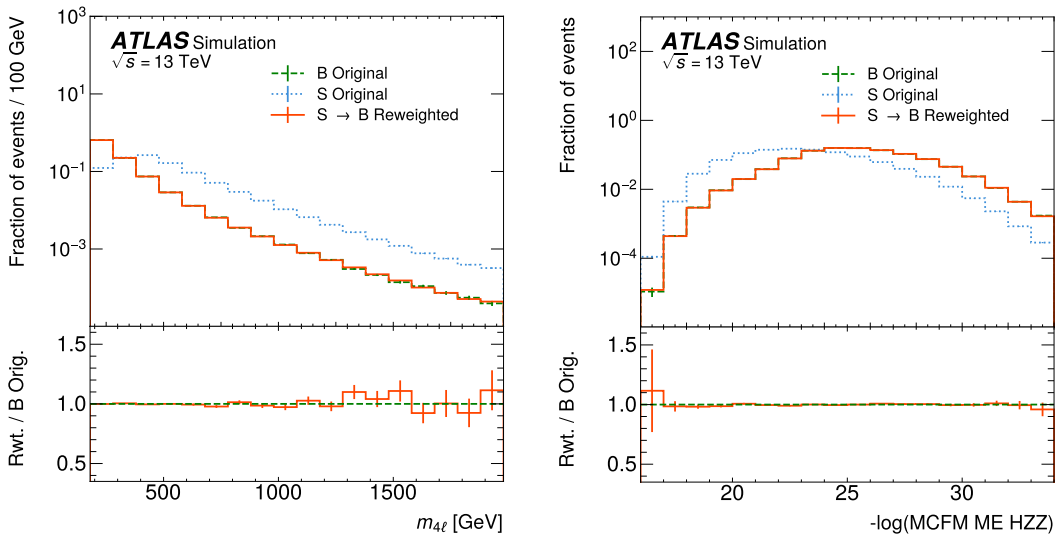


Figure 5.2: Reweighting curves curve for the Signal (S) vs Background (B Network trained for the ATLAS off-shell Higgs Boson [2]. Histogram of $m_{4\ell}$ variable (left) and $-\log$(MC FM ME HZZ)(right) with S (dotted blue line), B (dotted green line) and the S Reweighted to match B (solid orange line).

5.2 Winning Submissions to FAIR Universe challenge

As mentioned in the previous chapter, the two First Prize winners both used different variants of simulation-based inference. Here, we will briefly discuss these methods, how they incorporate systematic uncertainty, and whether there is room to merge the best parts of both.

5.2.1 `HEPHY`: Unbinned inclusive cross-section measurements with machine-

learned systematic uncertainties

The HEPHY team uses simulation-based inference (SBI) to construct a flexible, unbinned, and refinable surrogate of the extended likelihood [112] that captures the full high-dimensional event information for inference of the signal strength μ and the nuisance parameters $\vec{\alpha}$ via a multiclassifier and parametric regressors [109].[2] In order to properly constrain nuisance parameters, the HEPHY team split them into five each sensitive to different physics processes. Table 5.1 gives a summary of the event selection they used. In these five, 2 are signal regions(SR), and they are both analysed with unbinned (UB) analysis, and among the 3 Control regions (CR) the `highMT-VBFJet` is the only one analysed in unbinned analysis.

Table 5.1: Summary of the event selections. The "UB" label indicates whether the selection is included in the unbinned analysis.

Region	**Requirements**	**Type**
`lowMT-VBFJet`	$p_{\mathrm{T}}^{\mathrm{jet_leading}} > 50$ GeV $p_{\mathrm{T}}^{\mathrm{jet_subleading}} > 30$ GeV $m_{\mathrm{T}} \leq 70$ GeV	UB, SR
`highMT-VBFJet`	$p_{\mathrm{T}}^{\mathrm{jet_leading}} > 50$ GeV $p_{\mathrm{T}}^{\mathrm{jet_subleading}} > 30$ GeV $m_{\mathrm{T}} > 70$ GeV	UB, CR
`lowMT-noVBFJet-ptH100`	$p_{\mathrm{T}}^{\mathrm{H}} > 100$ GeV $m_{\mathrm{T}} \leq 70$ GeV veto on `VBFJet`	UB, SR
`lowMT-noVBFJet-ptH0to100`	$p_{\mathrm{T}}^{\mathrm{H}} \leq 100$ GeV $m_{\mathrm{T}} \leq 70$ GeV veto on `VBFJet`	CR
`highMT-noVBFJet`	$m_{\mathrm{T}} > 70$ GeV veto on `VBFJet`	CR

Statistical Overview

HEPHY's method utilises a variant of Semi-Parametric Neural Density Estimation that employs a multi-classifier rather than a binary one. This multi-classifier model directly predicts the probability $p_X(x_i)$ for all samples (X), which will give us the per-event probability ratio required for Equation 5.2. The equation for the probability ratio is given by:

$$\frac{p(x_i|\mu)}{p(x_i|1)} = \frac{1}{\nu(\mu)} \sum_{X}^{C} f_X(\mu) \cdot \nu_X \cdot \frac{p_X(x_i)}{p(x_i|1)} \tag{5.8}$$

The upside is that it reduces the number of models we need to train. But directly predicting probabilities may lead to poorly calibrated probabilities. To compensate

[2]The notion used here follows the ATLAS NSBI convention, hence slightly different from the one seen in the original paper.

for potential miscalibration, the HEPHY team used a dedicated isotonic regression (IREG) to adjust the classifier outputs. First, they independently calibrated each multiclassifier output $p_X^*(x_i)$ using one-vs-rest binary isotonic regression. Then they retained the calibrated output for the signal-vs-background class ($p = \mathrm{H}$) and rescaled the calibrated outputs of the three background classes so that the sum of $p_X(x_i)$ is unity. Figure 5.3 gives the calibration curves before and after the Isotonic regression step.

$$p_X(x) = \begin{cases} \mathrm{IREG}(p_H^*(x_i)), & \text{if } X = H \\ \frac{\mathrm{IREG}(p_X^*(x_i))(1-\mathrm{IREG}(p_H^*(x_i)))}{\sum_{q=\mathrm{Z},t\bar{t},\mathrm{VV}} \mathrm{IREG}(p_q^*(x_i))}, & \text{if } X \neq H. \end{cases} \tag{5.9}$$

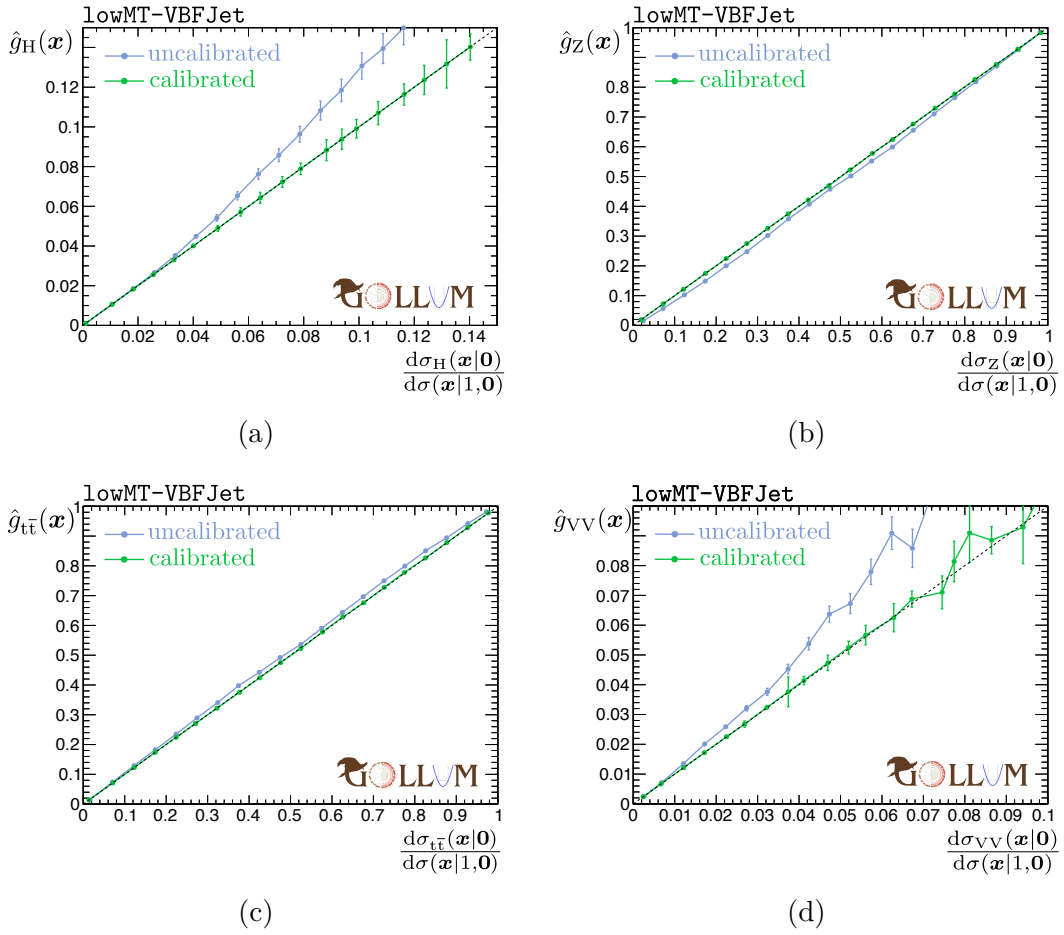


Figure 5.3: Calibration curve from different processes in the lowMT VBF jets region [109]. (a)$H \to \tau^+\tau^-$ process, (b) $Z \to \tau^+\tau^-$ process (c)$t\bar{t} \to \tau^+\tau^-$ process and (d) $VV \to \tau^+\tau^-$ process.

Binned and Unbinned Parametric Systematic Uncertainty

To address feature-distortion systematics, combine accounts for possible correlations using a multi-dimensional polynomial function. In the binned regions

(`lowMT-noVBFJet-ptH100` and `highMT-noVBFJet`) , the $\vec{\alpha}$ dependent shift in bin-yield (ν_k) is modelled by the following.

$$\begin{aligned}\alpha_A\hat{\Delta}_{k,A} = &\alpha_{\text{tes}}\hat{\Delta}_{k,\text{tes}} + \alpha_{\text{jes}}\hat{\Delta}_{k,\text{jes}} + \alpha_{\text{soft}}\hat{\Delta}_{k,\text{soft}} \\ &+ \alpha^2_{\text{tes}}\hat{\Delta}_{k,\text{tes,tes}} + \alpha^2_{\text{jes}}\hat{\Delta}_{k,\text{jes,jes}} \\ &+ \alpha^2_{\text{soft}}\hat{\Delta}_{k,\text{soft,soft}} + \alpha_{\text{tes}}\alpha_{\text{jes}}\hat{\Delta}_{k,\text{tes,jes}} \\ &+ \alpha_{\text{tes}}\alpha_{\text{soft}}\hat{\Delta}_{k,\text{tes,soft}} + \alpha_{\text{soft}}\alpha_{\text{jes}}\hat{\Delta}_{k,\text{soft,jes}}\end{aligned} \tag{5.10}$$

They fit the value of $\alpha_A\hat{\Delta}_{k,A}$ for $\vec{\alpha} \in \mathcal{V}$, the set of NP vectors, for each bin by minimising the χ^2,

$$\chi^2 = \sum_{\vec{\alpha}\in\mathcal{V}} \left(\log\left(\frac{\nu_k(\vec{\alpha})}{\nu_k(0)}\right) - \alpha_A\hat{\Delta}_{k,A} \right)^2 \tag{5.11}$$

The fitted function can be plugged into Equation 4.17 and regressed for μ and $\vec{\alpha}$. To perform the same regression in the unbinned analysis, the HEPHY team used a single multi-classifier-like network for each unbinned region and sample (12 in total). The model has 2 hidden layers with 128 nodes each, takes 28 features as input, and has an output layer with 9 nodes for 9 coefficients ($\Delta_A(x_i)$) of the polynomial function,

$$\begin{aligned}\alpha_A\hat{\Delta}_A(x) = &\alpha_{\text{tes}}\hat{\Delta}_{\text{tes}}(x) + \alpha_{\text{jes}}\hat{\Delta}_{\text{jes}}(x) + \alpha_{\text{soft}}\hat{\Delta}_{\text{soft}}(x) \\ &+ \alpha^2_{\text{tes}}\hat{\Delta}_{\text{tes,tes}}(x) + \alpha^2_{\text{jes}}\hat{\Delta}_{\text{jes,jes}}(x) \\ &+ \alpha^2_{\text{soft}}\hat{\Delta}_{\text{soft,soft}}(x) + \alpha_{\text{tes}}\alpha_{\text{jes}}\hat{\Delta}_{\text{tes,jes}}(x) \\ &+ \alpha_{\text{tes}}\alpha_{\text{soft}}\hat{\Delta}_{\text{tes,soft}}(x) + \alpha_{\text{soft}}\alpha_{\text{jes}}\hat{\Delta}_{\text{soft,jes}}(x)\end{aligned} \tag{5.12}$$

In order to learn the polynomial coefficients, they trained the multiclassifier with a custom loss function,

$$\begin{aligned}L[\hat{\Delta}_A] = &\sum_{\vec{\alpha}\in\mathcal{V}} \Bigg[\int p(x|0)\,\text{Soft}^+(\alpha_A\hat{\Delta}_A(x)) \\ &\qquad + \int p(x|\vec{\alpha})\,\text{Soft}^+(-\alpha_A\hat{\Delta}_A(x)) \Bigg] \\ \approx &\sum_{\vec{\alpha}\in\mathcal{V}} \Bigg[\sum_{\{x_i,w_i\}\in\mathcal{D}_0} w_i\,\text{Soft}^+(\alpha_A\hat{\Delta}_A(x_i)) \\ &\qquad + \sum_{\{x_i,w_i\}\in\mathcal{D}_{\vec{\alpha}}} w_i\,\text{Soft}^+(-\alpha_A\hat{\Delta}_A(x_i)) \Bigg]\end{aligned} \tag{5.13}$$

where $\text{Soft}^+(x) = \log(1+\exp(x))$ and the last approximation defines the empirical version of the loss function. We emphasise that for each term in the sum over $\mathcal{V}$,

the sample $\mathcal{D}_0$ in the first term is the same. The functional minimum of Eq. 5.13 is given by

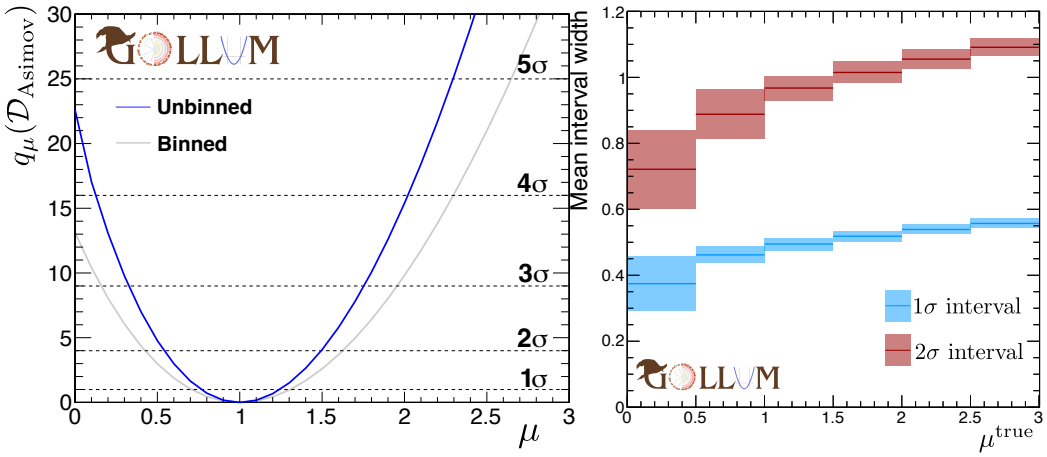
$$g_X(x|\vec{\alpha}) = \exp\left(\alpha_A \hat{\Delta}_A(x)\right) \simeq \frac{p_X(x|\vec{\alpha})}{p_X(x|\vec{0})} \tag{5.14}$$

Finally, we combine everything into a single density ratio and perform the profiling as we discussed before

$$\frac{p(x_i|\mu,\vec{\alpha})}{p(x_i|1,0)} = \frac{1}{\nu(\mu,\vec{\alpha})} \sum_X^C f_X(\mu) \cdot g_X(x|\vec{\alpha}) \cdot \frac{p_X(x_i|0)}{p(x_i|1,\vec{0})} \tag{5.15}$$

Results and Discussions

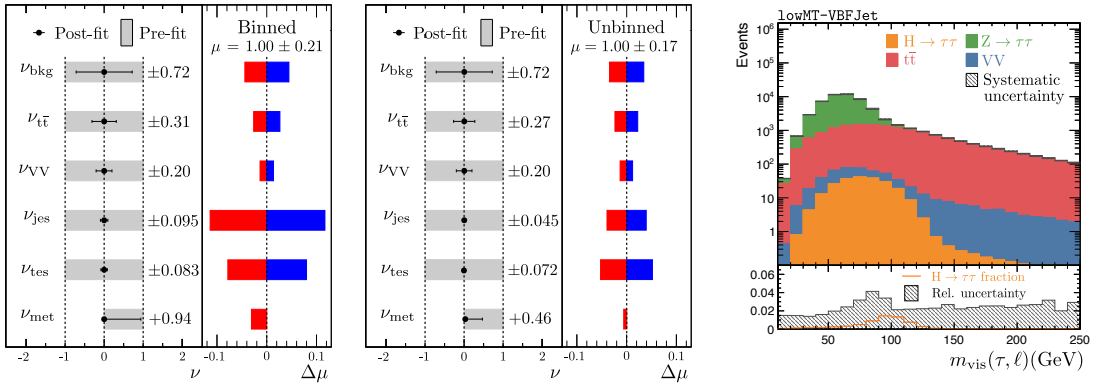
By utilising SBI, HEPHY has shown to improve 1σ bounds by 20% compared to the binned case Figure 5.4(a). Since the challenge design ensured that all μ were within the interval $[0, 3]$, the team implemented clipping for the intervals at 0 and 3, which is visible in Figure 5.4(b).



(a) (b)

Figure 5.4: Results from HEPHY's Submission [109]. Likelihood Curve from binned/unbinned inference (a), Average $\text{CI}_{68\%}$ and $\text{CI}_{95\%}$ Widths vs μ predicted (b)

They used a multiclassifier to best estimate the probability ratios and also to fit the per-event systematics variation, which has given a better estimate on systematics Figure 5.5(a),(b). Using a multiclassifier rather than 1 network per sample for probability estimation could be a strategy to adopt if there aren't enough samples to train certain backgrounds. Figure 5.5(c) gives the post-fit histogram on m_{vis}.



(a) (b) (c)

Figure 5.5: Pre-fit and post-fit Plots from HEPHY's Submission [109]. Pre-fit and post-fit nuisance parameter values for the binned analysis (a) and the unbinned analysis (b) in the Asimov fit to the nominal simulation and (c) Pre- and post-fit distributions in the `lowMT-VBFJet` region for m_{vis} observable.

5.2.2 `IBRAHIM`: Contrastive Normalizing Flows for Uncertainty-Aware Parameter Estimation

Ibrahim's CNF methods [110] take a different approach from HEPHY's, and even from the ATLAS method we will see later. Instead of modelling the systematic shift to bins or densities as discussed before, they plan to mitigate the shift altogether by using Normalising flows to create an abstract variable $\Phi[x_i, \vec{\alpha}]\mathbb{R}^{\mathrm{>}} \to \mathbb{R}$ such that $\Phi(x_i|\vec{\alpha}) \simeq \Phi(x_i|\vec{0})$. This would mean we can express the likelihood ratio as;

$$\frac{p(x_i|\mu, \vec{\alpha})}{p(x_i|1, \vec{0})} \to \frac{p(x_i|\mu, \Phi[x_i, \vec{\alpha}])}{p(x_i|1, \Phi[x_i, \vec{0}])} \tag{5.16}$$

For this to work, the abstract variable $\Phi[x_i, \vec{\alpha}]$ should have sufficient information to discriminate between the target and reference in the traditional sense of SBI. This is achieved through the Contrastive Normalising Flows.

Brief introduction to Normalising Flow

Normalising Flows (NF) are a bijective function($f_\theta(x) = z$) that maps an n-dimensional distribution ($\mathcal{D}$) from a Gaussian ($\mathcal{N}(0, I)$) of the same dimension.

$$X \sim p_{\text{data}} \quad \Longrightarrow \quad f_\theta(X) \sim \mathcal{N}(0, I) \tag{5.17}$$

for the mapping $z = f_\theta(x)$ we can write the probability distribution as

$$p_\theta(x) = p_z(z).|\det J_{f_\theta}(x)| = p_z(f_\theta(x))|\det J_{f_\theta}(x)|$$

We can use negative log likelihood as the loss function to tune NF ($f_\theta(x)$)

$$\mathcal{L} = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} -\log(p_\theta(x)) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} -\Big[\log p_z(f_\theta(x)) + \log|\det J_{f_\theta}(x)|\Big] \tag{5.18}$$

Finally, we can define $\Phi[x_i]$ as

$$\Phi[x_i] = \frac{p_\theta(x)}{1 + p_\theta(x)} \tag{5.19}$$

In practice, Normalising Flows learn a transformation that maps a simple base distribution (typically a Gaussian) to the data distribution. Instead of learning this mapping in a single step, the transformation is decomposed into a sequence of simpler, invertible functions. Each function incrementally warps the space, making the overall transformation easier to model and train. This composition allows the model to efficiently represent highly complex distributions while still enabling exact likelihood computation through the change-of-variables formula.

Contrastive Normalising Flow for HEP

The distribution we are training on is composed of various physics processes (Signal and Background). We can train the NFs separately to learn the signal and background, where the loss is given by

$$\mathcal{L} = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} -\Big[\log p_\theta^{(s)}(x) + c \cdot \log p_\theta^{(b)}(x) \Big] \tag{5.20}$$

With this, we can adjust how signal-like or background-like the distribution we want the NF to learn using the contrastive variable c. Similar to the methods discussed before, the events have been divided into 3 categories, based on the number of jets. The 3 regions are 0-jet, 1-jet, and 2-jet. They used 1-jet and 2-jet events to train the CNFs. For each region, they trained 4 networks, 2 with $c = 2$, and 2 with $c = 0.5$.

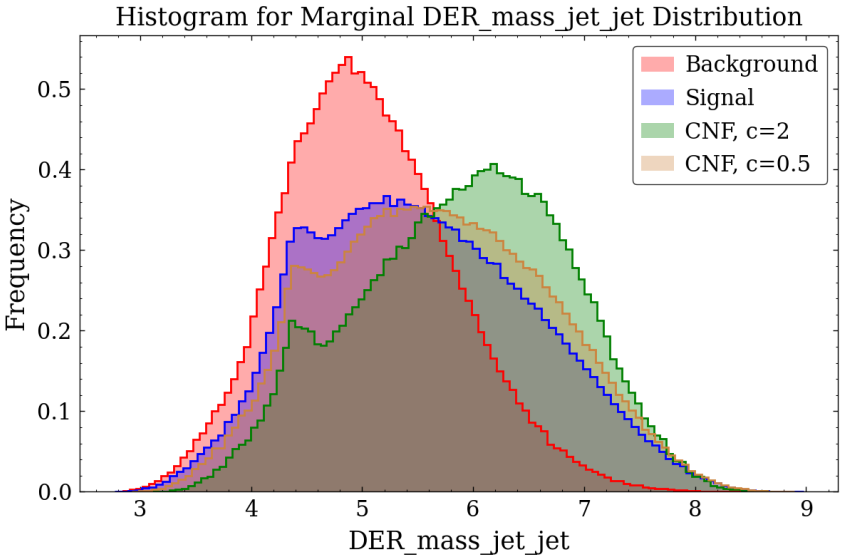


Figure 5.6: Different learned distributions here shown as histogram of $m^{\text{jet}-\text{jet}}$ from CNFs trained with varying values of c.

Each of the CNFs was trained with augmented datasets with various systematics or NP values ($\vec{\alpha}$). This means the final values of abstract variables ($\Phi^s_{c=0.5}[x_i]$,

$\Phi^b_{c=0.5}[x_i]$, $\Phi^s_{c=2}[x_i]$ and $\Phi^s_{c=2}[x_i]$) will be independent of NP ($\vec{\alpha}$). Figure 5.6 different learned distribution with varying values of c. Figure 5.7 shows the flow chart of the full analysis. In addition to the event features, Ibrahime has used the abstract variable *Phi* as the input to the NN classifier to do the density ratio estimation (As mentioned in Equation 5.16). Finally, the NNClassifier score was binned and, using maximum likelihood estimation (Equation 4.14), the values of μ and $\text{CI}_{68\%}$ were determined.

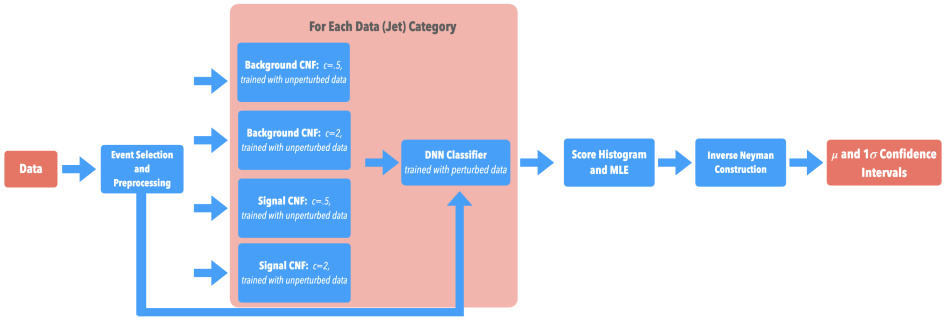
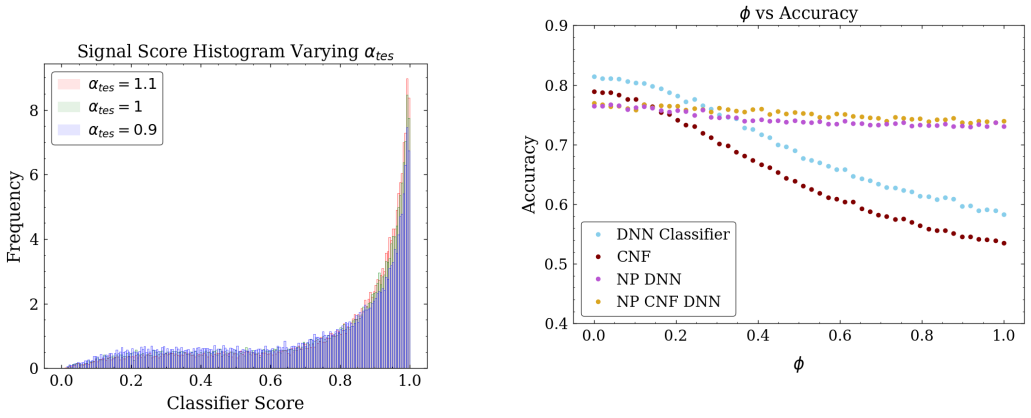


Figure 5.7: Flow -chart of full CNF-based Classifier pipeline from data to final μ predictions.[110]

Results and Discussion

Figure 5.8(a) gives the change in histograms of the classifier score of signal score with different values of NP α_{tes}. This shows the stability of the DNN score thanks to using Φ variables. Figure 5.8(b) further shows why combining it with the DNN will get more stability, systematics uncertainty-based shift (parametrised by ϕ)analyses.



(a) (b)

Figure 5.8: Change in DNN score histogram with various values of α_{tes} (a) and evolution of accuracy for different approaches as a function of nuisance parameter ϕ

The major challenge in using this method is that, unlike the FAIR Universe framework, in a realistic LHC analysis, it is far more costly to create shifted datasets,

so only datasets with shifts of 1σ up or down per nuisance parameter are typically available. Also, NFs are very resource-intensive, and with the CNF approach, we would need to retrain the entire CNF system if new systematics are to be implemented. CNF also doesn't fully address the case of unknown unknowns, as the assumption at the end is that the DNN score is uncertainty-aware and hence won't shift if **known** systematics are present.

5.3 ATLAS SBI Workflow with FAIR Universe HiggsML Challenge

The 2024 ATLAS off-shell Higgs to 4 lepton Paper[1, 2] proved the applicability of SBI to real particle physics data. We present here the implementation of the same SBI approach for the FAIR Universe HiggsML dataset using the **Iris-HEP** SBI toolkit [113]. As in the Winning submissions, to achieve a stable, well-calibrated classifier, we partition the phase space into regions and fit only parts of it with SBI, and the rest with analyses using a binned discriminant. Unlike HEPHY, which used a series of selections on transverse mass and momentum, we use a multiclassifier-based discriminant ($\mathrm{NN}_{\mathrm{presel}}$) to achieve this goal.

$$\mathrm{NN}_{\mathrm{presel}}(x) = \log\left[\frac{\sum_{sig} P_{sig}(x)}{\sum_{bkg} P_{bkg}(x)}\right] \tag{5.21}$$

Using the $\mathrm{NN}_{\mathrm{presel}}(x)$, they divided the events into 3 regions,

- **Control Region** (**CR**): For events with $\mathrm{NN}_{\mathrm{presel}}(x) < -1$, these events will be analysed using a binned histogram on the $\mathrm{NN}_{\mathrm{presel}}(x)$.
- **Signal Region- Binned** (**BSR**): For events with $\mathrm{NN}_{\mathrm{presel}}(x) > 4.5$, these events will be analysed using a binned histogram on the $\mathrm{NN}_{\mathrm{presel}}(x)$.
- **Signal Region- Unbinned** (**UB, SR**): For events with $-1 < \mathrm{NN}_{\mathrm{presel}}(x) < 4.5$, these events will be analysed using the SBI approach we discussed above.

With a large enough dataset, we can train an ML classifier between the reference and the target sample and generate the density ratio $\left(\frac{p_X(x_i)}{p_{ref}(x_i)}\right)$.

$$\frac{p(x_i|\mu)}{p_{ref}(x_i)} = \frac{1}{\nu(\mu)} \sum_{X}^{C} f_X(\mu) \cdot \nu_X \cdot \frac{p_X(x_i)}{p_{ref}(x_i)} \tag{5.22}$$

Finally, we can represent $\frac{p(x_i|\mu)}{p(x_i|1)}$ as $\frac{p(x_i|\mu)}{p_{ref}(x_i)} \cdot \frac{p_{ref}(x_i)}{p(x_i|1)}$, making it possible to use the density ratio for computation in place of probability.

Recall that the dataset includes 4 processes: Higgs, $t\bar{t}$, Z and VV(di-boson), but the dataset has a disproportionately low number of VV process events; hence, for the initial trials, we will remove the $VV \to \tau\tau$ category. For FAIR Universe data, μ linearly depends on the Higgs process and no other process, meaning $f_H(\mu) = \mu$ and $f_{t\bar{t}}(\mu) = f_Z(\mu) = 1$. Combining this with Equation 5.22 we get a complete expression for the probability ratio for $\frac{p(x_i|\mu)}{p_{\mathrm{ref}}(x_i)}$ for the FAIR Universe dataset as

$$\frac{p(x_i|\mu)}{p_{\text{ref}}(x_i)} = \frac{1}{\nu(\mu)}\left[\mu \cdot \nu_H \cdot \frac{p_H(x_i)}{p_{\text{ref}}(x_i)} + \nu_Z \cdot \frac{p_Z(x_i)}{p_{\text{ref}}(x_i)} + \nu_{t\bar{t}} \cdot \frac{p_{t\bar{t}}(x_i)}{p_{\text{ref}}(x_i)}\right] \tag{5.23}$$

Figure 5.9 gives the reweighted plots for $H \to \tau^+\tau^-$ and $H \to \tau^+\tau^-$ for $\Delta R(\tau_{\text{had}}, \ell)$ and $\log(m_{\text{vis}})$.

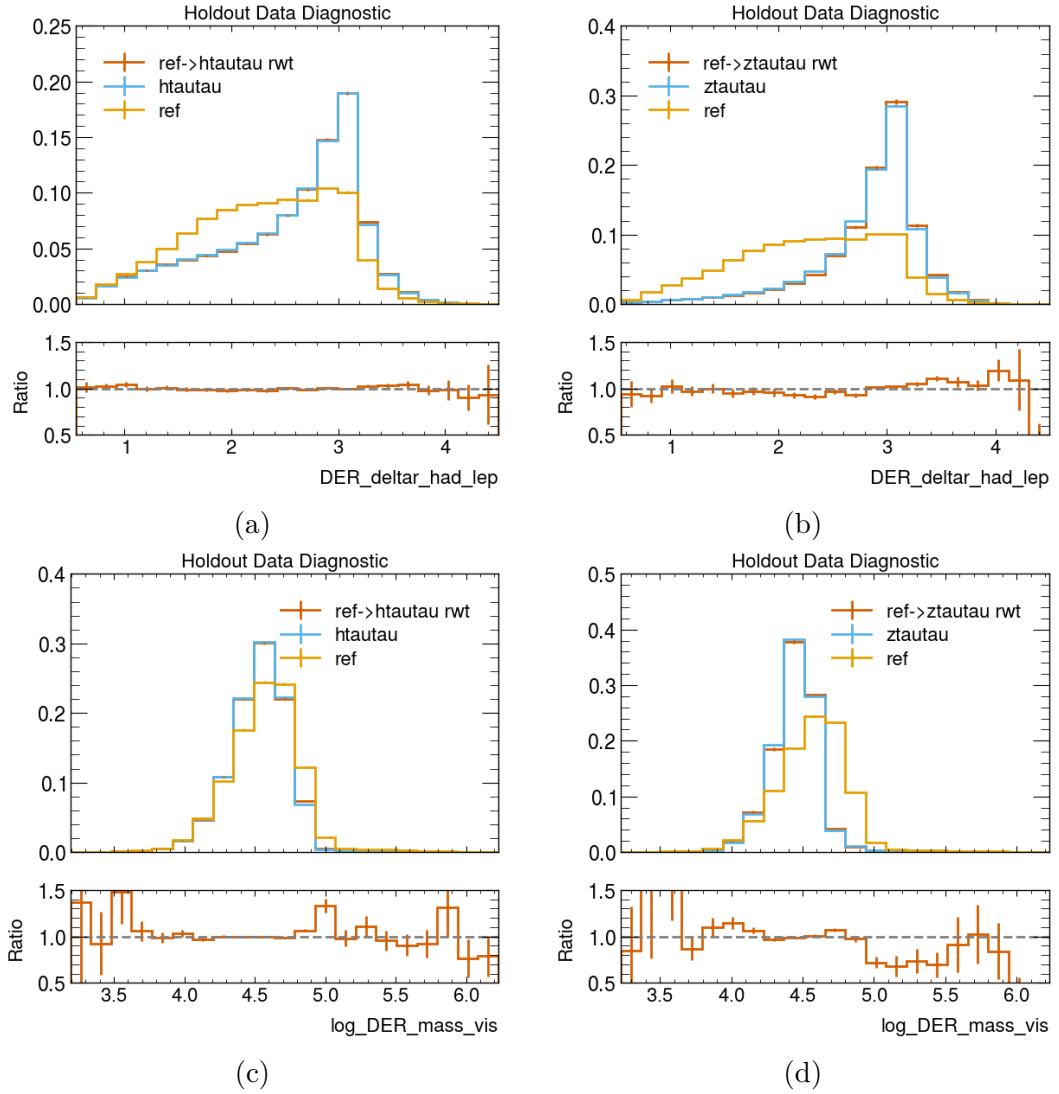


Figure 5.9: Reweighted plots for $H \to \tau^+\tau^-$ (a,c) and $Z \to \tau^+\tau^-$ (b,d) for $\Delta R(\tau_{\text{had}}, \ell)$ and $\log(m_{\text{vis}})$

Figure 5.10 gives the Calibration curved for various classifiers for $H \to \tau^+\tau^-$, $Z \to \tau^+\tau^-$ and $\bar{t}t \to \tau^+\tau^-$.

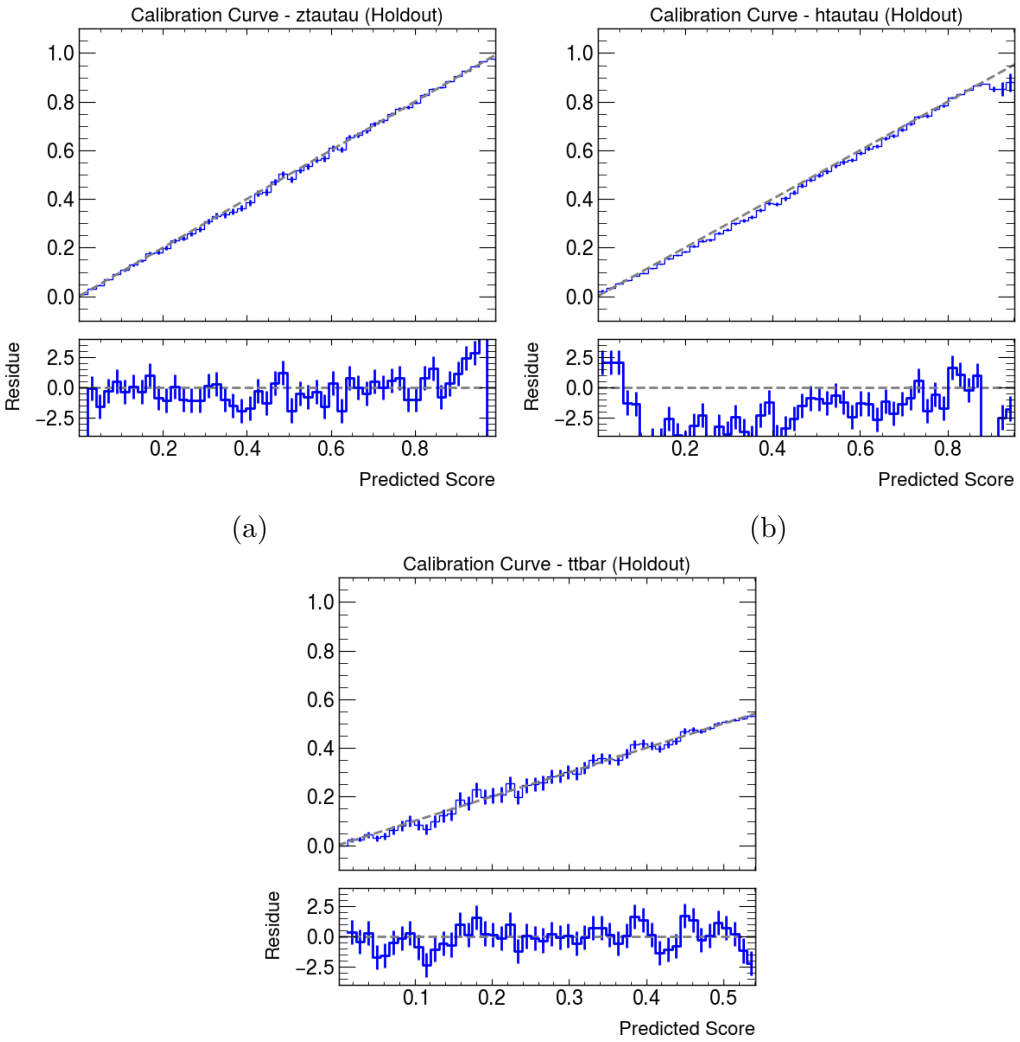


Figure 5.10: Calibration curves for various classifiers for (a) $H \to \tau^+\tau^-$, (b) $Z \to \tau^+\tau^-$ and (c)$\bar{t}t \to \tau^+\tau^-$

5.3.1 Modelling systematics in Semi-Parametric SBI

In Equation 4.5.2, we discussed that there are 2 kinds of systematics, the first is normalisation systematics, which can be modelled by floating the appropriate samples and the constraint function π. The second is feature-distorting systematics, which shift the yield in each bin in the feature histograms. Since SBI fundamentally learns N-dimensional shape information, we can utilise it to learn shape changes due to systematics. We can thus expand Equation 5.3 to account for μ and α as :

$$\frac{p(x_i|\mu,\vec{\alpha})}{p_{ref}(x_i)} = \frac{1}{\nu(\mu,\vec{\alpha})} \sum_{X}^{samples} f_X(\mu) \cdot \nu_X(\vec{\alpha}) \cdot \frac{p_X(x_i,\vec{\alpha})}{p_{ref}(x_i)} \tag{5.24}$$

This gives us two $\vec{\alpha}$ depended variables $\nu_X(\vec{\alpha})$ and $p_X(x_i,\vec{\alpha})$. In both cases, we will use the HistFactory approach described in Equation 4.5.2 and assume that the

Nuisance Parameters are uncorrelated and can be factorised and written as:

$$\frac{p(x_i|\mu,\vec{\alpha})}{p_{ref}(x_i)} = \frac{1}{\nu(\mu,\vec{\alpha})} \sum_{X}^{samples} f_X(\mu) \frac{p_X(x_i)}{p_{ref}(x_i)} \prod_{m}^{N_{syst}} \cdot \nu_X(\alpha_m) \cdot \frac{p_X(x_i|\alpha_m)}{p_X(x_i)} \tag{5.25}$$

$\nu_X(\alpha_m)$ is exactly the same as shifting bin yield, and hence we can utilise Equation 4.16 to make that interpolation, giving :

$$G_X(\alpha_m) = \begin{cases} \left(\frac{\nu_X(\alpha_m^+)}{\nu_X}\right)^{\alpha_m}, & \text{if } \alpha_m > 1 \\ 1 + \sum_{n=1}^{6} c_n \alpha^n{}_k, & \text{if } -1 \leq \alpha_m \leq 1 \\ \left(\frac{\nu_X(\alpha_m^-)}{\nu_X}\right)^{-\alpha_m}, & \text{if } \alpha_m < -1 \end{cases} \tag{5.26}$$

we will use a similar interpolation as Equation 4.16 but between $\frac{p_X(x_i|\alpha_m^+)}{p_X(x_i)}$and $\frac{p_X(x_i|\alpha_m^-)}{p_X(x_i)}$ to get $\frac{p_X(x_i|\alpha_m)}{p_X(x_i)}$ or $g_X(x_i|\alpha_m)$. The 1-up and 1-down ratios will be approximated with the help of ML classifiers, just as in the case of $\frac{p_X(x_i)}{p_{ref}(x_i)}$.

$$g_X(x_i|\alpha_m) = \begin{cases} \left(\frac{p_X(x_i|\alpha_m^+)}{p_X(x_i)}\right)^{\alpha_m}, & \text{if } \alpha_m > 1 \\ 1 + \sum_{n=1}^{6} c_n \alpha^n{}_k, & \text{if } -1 \leq \alpha_m \leq 1 \\ \left(\frac{p_X(x_i|\alpha_m^-)}{p_X(x_i)}\right)^{-\alpha_m}, & \text{if } \alpha_m < -1 \end{cases} \tag{5.27}$$

Finally, we can now put this all together to create the joint conditional probability as

$$\frac{p(x_i|\mu,\alpha)}{p_{ref}(x_i)} = \frac{1}{\nu(\mu,\alpha)} \sum_{X}^{C} f_X(\mu) \cdot \nu_X \cdot \frac{p_X(x_i)}{p_{ref}(x_i)} \prod_{m}^{N_{syst}} G_X(\alpha_m) g_X(x_i|\alpha_m) \tag{5.28}$$

This joint conditional probability can then be used to create the joint likelihood function $\mathcal{L}(\mu,\alpha)$, which can then be profiled as seen in subsection 4.5.3. Figure 5.11 gives the likelihood curves and the profile likelihood curve for NSBI and Histogram-based methods[3] (Blue), for Statistics Only and Stat + syst.

5.3.2 Discussion and conclusion

We discussed various methods for performing unbinned systematics analysis using machine-learning measurements. The FAIR Universe Dataset and Benchmark

[3]The Histogram method is based on the preselection score (NN_{presel}), hence it need not be the best distribution for Binned analysis

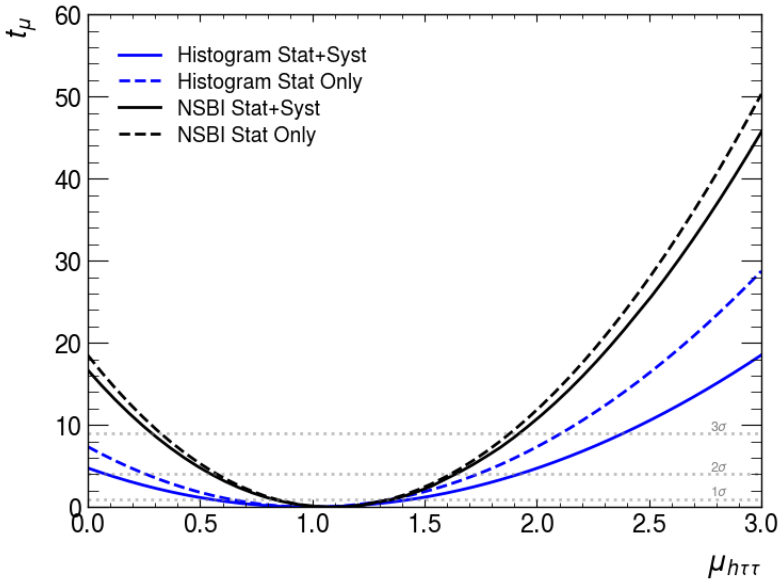


Figure 5.11: The Likelihood curves to scan μ, to compare between Neural Simulation Based Inference (NSBI)(Black) and Histogram-based methods (Blue), for Statistics Only (dashed lines) and Stat + syst (solid lines)

provides us with a uniform platform, where we have full control over systematics and signal strength, allowing us to estimate the pros and cons of various approaches before implementing them in formal LHC analyses. The benchmark is still simpler than the many LHC analyses we will use SBI in; for instance, the POI is linear and depends on only a single sample, the Higgs, and we have limited systematics whose behaviour is perfectly understood, unlike in real work cases where we depend on other measurements to determine the distributions of the Nuisance Parameters. Taking all this into consideration, the safest bet for the Di-Higgs analysis moving forward is the ATLAS NSBI method, which uses N different ML classifiers in place of a single multiclassifier, allowing us to refine each classifier individually and making classifier tuning easier. Additionally, this means we can use different classifiers (Mixed Classifier Approach) to handle different samples or different systematics. This is critical in regimes with a limited number of training samples. We will explore this in detail in chapter 7, where we apply Semi-Parameteric SBI to the $HH \to b\bar{b}\tau^+\tau^-$.

Chapter 6
Study of Higgs pair production in $b\bar{b}\ \tau^+\tau^-$ final state in ATLAS

If I have seen further it is by standing on the shoulders of giants. – Sir Isaac Newton

---

In subsection 1.3.2 we introduced the production and decay of Di-Higgs. In this chapter, we will explore various steps taken for the experimental analysis of Higgs pair production in the $bb\tau^+\tau^-$ final state. The work presented in this chapter is taken from the recent public analysis by the $HH \to b\bar{b}\tau^+\tau^-$ analysis group[3]. We have used this as the baseline for building the SBI analysis, which we will discuss in chapter 7. Although they analysed both lep-had and had-had channels for Run 2 and partial Run 3 data, for this thesis, we will be focusing only on partial Run 3 data: ggF $HH \to bb\tau^+\tau^-$, where both taus decay hadronically. Since this thesis is the first to utilise **SBI** for the analysis of $HH \to bb\tau^+\tau^-$, we began with this channel to understand the problem and set the course for the full SBI analysis in all other channels.

6.1 Theoretical context

As far as this thesis is concerned, we will be inferring 2 parameters: the signal strength of the Di-Higgs production μ and the Higgs self-coupling constant λ. For simplicity, we instead determine a related quantity called κ_λ, which is the ratio of the observed value of the λ constant to the expected value in the standard model; hence, by definition, the Standard Model value of κ_λ is one.

$$\kappa_\lambda = \frac{\lambda_{\text{observed}}}{\lambda_{\text{SM}}} \tag{6.1}$$

The amplitude for the ggF production of Di-Higgs ($\mathcal{M}_{\kappa_\lambda}$) at leading order is given by

$$\mathcal{M}_{\kappa_\lambda} = \kappa_\lambda \cdot \mathcal{M}_{HH_\Delta} + \mathcal{M}_{HH_\square} \tag{6.2}$$

The total cross-section for the production of Di-Higgs (σ_{κ_λ}) is hence given by

$$\begin{aligned} \sigma_{\kappa_\lambda} &\propto \mathcal{M}^*_{\kappa_\lambda}\mathcal{M}_{\kappa_\lambda} \\ &= \kappa_\lambda^2\, \mathcal{M}^*_{HH_\Delta}\mathcal{M}_{HH_\Delta} + \mathcal{M}^*_{HH_\square}\mathcal{M}_{HH_\square} \\ &\quad + \kappa_\lambda\Big(\mathcal{M}^*_{HH_\Delta}\mathcal{M}_{HH_\square} + \mathcal{M}^*_{HH_\square}\mathcal{M}_{HH_\Delta}\Big) \end{aligned} \tag{6.3}$$

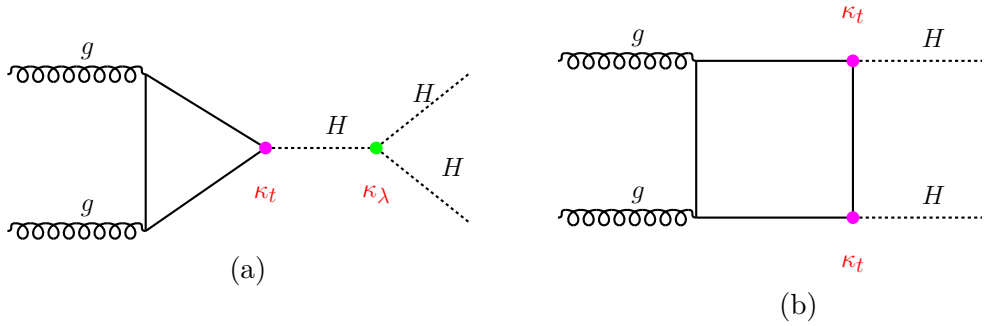


Figure 6.1: Production of Higgs boson pairs through gluon-gluon-fusion (ggF) with trilinear Higgs boson self-coupling (κ_λ (triangle) (a) and involving only top quark loop (box) (b) contributions.

The $(\mathcal{M}^*_{HH_\Delta}\mathcal{M}_{HH_\square} + \mathcal{M}^*_{HH_\square}\mathcal{M}_{HH_\Delta})$ is the interference term, as it originated from the quantum interference between the box and the Triangle process. Hence, we can rewrite the cross-section equation as :

$$\sigma_{\kappa_\lambda} = \kappa_\lambda^2 \cdot \sigma_{HH_\Delta} + \kappa_\lambda \cdot \sigma_{HH_{\text{interference}}} + \sigma_{HH_\square}$$

Instead of working with σ_{HH_Δ}, $\sigma_{HH_{\text{interference}}}$ and $\sigma_{HH_\square}$, which could be difficult as the interference terms are riddled with negative weighted events, a more commonly used approach is to rewrite the σ_{κ_λ} linear combination of samples of different κ_λ values. Technically, we can take any three values of κ_λ to build the kl framework. The $\{\kappa_{\lambda 0}, \kappa_{\lambda 1}, \kappa_{\lambda 5}\}$ basis is taken as $\kappa_{\lambda 1}$ is used for signal strength studies and $\{\kappa_{\lambda 0}, \kappa_{\lambda 5}\}$ as 2 BSM case. $\kappa_{\lambda 0}$ is the BSM case with no self-interaction, but $\kappa_{\lambda 5}$ is an arbitary choice, There are analysis which use $\kappa_{\lambda 10}$ for in thier basis. The three benchmark cross sections with $\{\kappa_{\lambda 0}, \kappa_{\lambda 1}, \kappa_{\lambda 5}\}$ basis can be written in matrix form as

$$\begin{pmatrix} \sigma_{\kappa_{\lambda 1}} \\ \sigma_{\kappa_{\lambda 5}} \\ \sigma_{\kappa_{\lambda 0}} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1 & 1 \\ 25 & 5 & 1 \\ 0 & 0 & 1 \end{pmatrix}}_{A} \begin{pmatrix} \sigma_{HH_\Delta} \\ \sigma_{HH_{\text{interference}}} \\ \sigma_{HH_\square} \end{pmatrix}. \tag{6.4}$$

Inverting the system gives

$$\begin{pmatrix} \sigma_{HH_\Delta} \\ \sigma_{HH_{\text{interference}}} \\ \sigma_{HH_\square} \end{pmatrix} = A^{-1} \begin{pmatrix} \sigma_{\kappa_{\lambda 1}} \\ \sigma_{\kappa_{\lambda 5}} \\ \sigma_{\kappa_{\lambda 0}} \end{pmatrix}, \qquad A^{-1} = \begin{pmatrix} \frac{1}{20} & -\frac{1}{20} & 0 \\ -\frac{5}{4} & \frac{1}{4} & 1 \\ 0 & 0 & 1 \end{pmatrix}. \tag{6.5}$$

Using this we can express σ_{κ_λ} as

$$\sigma_{\kappa_\lambda} = \sigma_{\kappa_{\lambda 0}} \left(\frac{\kappa_\lambda^2 - 6\kappa_\lambda + 5}{5} \right) + \sigma_{\kappa_{\lambda 1}} \left(\frac{-\kappa_\lambda^2 + 5\kappa_\lambda}{4} \right) + \sigma_{\kappa_{\lambda 5}} \left(\frac{\kappa_\lambda^2 - \kappa_\lambda}{20} \right) \tag{6.6}$$

We will utilise this formulation when building a statistical model for both binned and unbinned (SBI) cases.

6.2 Data and Monte Carlo samples

As we discussed in the previous chapter, Monte Carlo (MC) simulations are essential for building both the statistical and ML models. Unlike a FAIR Universe Dataset, which used Pythia for all processes in event generation, when we build MC for realistic analyses, we use different tools for different tasks within the event-generation process. Table 6.1 summarises the various Matrix element (ME) generators, Parton Distribution Functions (PDF), Parton Shower (PS), and underlying event (UE) tunes for the Run 3 had-had MC samples. For the detector simulation, Background samples were generated with a full ATLAS Geant4 simulation, and for the background, ATLAS Fastsim was used, which relies on a parameterisation of the calorimeter response. The Partial Run 3 contained data from pp collisions at centre of mass energy $s = \sqrt{13.6}$ TeV and corresponded to an integrated luminosity of 56fb^{-1}. The simulated events are reconstructed and analysed using the same software that processes real data. To improve the agreement between simulation and data, corrections are applied to account for differences in trigger, reconstruction, identification, and isolation efficiencies, as well as in the energy scale and resolution of jets, electrons, muons, and hadronically decaying taus.

Table 6.1: Summary of the nominal HH signal and background event samples used to analyse the Run 3 dataset.[3]

Process	ME Generator	PDF set	Parton Shower	Tune
Signal				
$gg \to HH$ (ggF)	POWHEG BOX v2	PDF4LHC21	PYTHIA 8	A14 NNPDF2.3
$qq \to qqHH$ (VBF)	MG5_aMC 2.7.3	NNPDF3.0NLO	PYTHIA 8	A14
Top-quark				
$t\bar{t}$	POWHEG BOX v2	NNPDF3.0NLO	PYTHIA 8	A14
t-channel	POWHEG BOX v2	NNPDF3.0NLO	PYTHIA 8	A14
s-channel	POWHEG BOX v2	NNPDF3.0NLO	PYTHIA 8	A14
Wt	POWHEG BOX v2	NNPDF3.0NLO	PYTHIA 8	A14 NNPDF2.3
$t\bar{t}W$	Sherpa 2.2.14	NNPDF3.0NLO	Sherpa 2.2.14	–
$t\bar{t}Z$	MG5_aMC @ NLO 2.8.1	NNPDF3.0NLO	PYTHIA 8	A14
W/Z+jets	Sherpa 2.2.14	NNPDF3.0NNLO	Sherpa 2.2.14	–
WW, WZ, ZZ	Sherpa 2.2.14/16	NNPDF3.0NNLO	Sherpa 2.2.14/16	–
Single Higgs boson				
ggF	POWHEG BOX v2	PDF4LHC15NLO	PYTHIA 8	A14 NNPDF2.3
VBF	POWHEG BOX v2	PDF4LHC15NLO	PYTHIA 8	A14 NNPDF2.3
$qq \to WH$	POWHEG BOX v2	PDF4LHC15NLO	PYTHIA 8	A14 NNPDF2.3
$qq \to ZH$	POWHEG BOX v2	PDF4LHC15NLO	PYTHIA 8	A14 NNPDF2.3
$gg \to ZH$	POWHEG BOX v2	PDF4LHC15NLO	PYTHIA 8	A14 NNPDF2.3
ttH	POWHEG BOX v2	PDF4LHC21*	PYTHIA 8	A14

6.3 Object reconstruction and Tagging

The $b\bar{b}\tau^+\tau^-$ had-had channel requires 4 tagged jets, 2 B-tagged and 2 tau-tagged. So for this channel, object reconstruction is essentially jet reconstruction.

6.3.1 Jet reconstruction

Jets are reconstructed from particle-flow objects[114]. A particle-flow object is a combination of tracks and calorimeter cells corresponding to a jet. The jets are selected with the anti-kt algorithm [115, 116] with $R = 0.4$. The Jet Calibration uses the procedures from Ref[117]. For Jets in the forwards region ($2.5 \leq |\eta| \leq 4.5$) must have $p_T > 30$ and the ones in central region ($|\eta| < 2.5$) must have $p_T > 20$. Additionally, they should also follow the **Tight** forward jet-vertex-tagging (fJVT)[118, 119] and jet-vertex-tagging (fJVT)[120] respectively.

6.3.2 b-Jet Tagging

ATLAS utilised the GN2 algorithm [121], a Graph Attention-based algorithm for flavour tagging. In previous chapters, we discussed BDTs and DNNs as ML Classifiers, but they have a fundamental limitation that Graph-based ML algorithms can overcome. BDTs and DNNs often work with tables; ergo, they use a fixed set of features as input. In particle physics data, each event can have a different number of constituents. The traditional way to handle them is to summarise them to a fixed number of features. Graph-Based Algorithms can naturally handle events with varying numbers of constituent particles and can also learn relationships among these constituents or objects. Transformers are one more step further from this. In GNNs, we provide an adjacency matrix; in Transformers, we start with a fully connected graph, and the algorithm learns the adjacency matrix via self-attention. A single attention layer is often insufficient to capture all properties and connections. So most algorithms, including Large Language Models (LLMs), use multi-head attention, which is effectively an ensemble of attention layers that learn different information. GN2 was trained on simulated jets, which are labelled as b-jets by matching them to a B-hadron with $p_{\mathrm{T}} \geq 5$ GeV within a cone of $\Delta R = 0.3$ around the jet axis. In addition to the classification task, it's trained on auxiliary tasks such as track-vertex matching and physics-process prediction. The GN2 algorithm uses charged-track information to classify b-jets, c-jets, and light jets. The b-tagging is restricted to the central jets, and a loose working point that satisfies 85% b-tagging efficiency was chosen. This corresponds to a c-jet (light-flavour jet) misidentification rate of approximately 0.2 (0.01). The MVA also utilises a transformer-based classifier, thereby taking object-level information into account, which includes b-tagging information. To provide a detailed b-tagging information for the MVA classifier, jets are classified into mutually exclusive categories based on the most stringent efficiency working point they satisfy. Efficiency is defined as the following categories: 85–77%, 77–70%, 70–65%, and below 65%. For b and c-jets, the tagging efficiencies and misidentification rates are calibrated with $t\bar{t}$ events and for light-quark and gluon jets with Z+jets events [122–124]. Tagging efficiency scale factors are derived using simulated $t\bar{t}$ events and calibrated using established methods [122–124].

6.3.3 τ_{had} Jet Tagging

The visible component of hadronically decaying τ leptons is reconstructed from jets seeded by topological clusters of calorimeter energy deposits, following the procedure described in Ref. [125]. Tracks associated with each τ_{had} candidate are classified by a recurrent neural network (RNN) into four categories: τ tracks, conversion

tracks, isolation tracks, and fake tracks. In more recent developments, RNNs are somewhat obsolete architectures, but they were the first to be able to handle a variable number of objects. The modern approach to handling a variable number of objects is to use a transformer-like GN2 for b-tagging. The number of tracks identified as τ tracks determines the τ_{had} track multiplicity. The τ_{had} energy is reconstructed from a weighted combination of the jet energy and information from the τ_{had} substructure. A boosted regression tree is used to derive the final energy calibration, which is subsequently corrected in situ using $Z \to \tau\tau$ events to reproduce the energy response observed in data. The τ_{had} candidates are required to satisfy $p_{\text{T}} > 20\text{GeV}$ and $|\eta| < 2.5$, excluding the calorimeter transition region $1.37 < |\eta| < 1.52$. They must have one or three associated tracks and a unit electric charge. To suppress jets misidentified as τ_{had} candidates, signal candidates are required to pass the Loose working point of the RNN-based identification algorithm [126], which provides identification efficiencies of approximately 85% for one-track and 75% for three-track τ_{had} candidates. In the $\tau_{\text{lep}}\tau_{\text{had}}$ channel, τ_{had} candidates are additionally required to satisfy the Loose working point of the RNN-based electron veto (eVeto) [125] to suppress electrons misidentified as one-track τ_{had} candidates. This working point has an efficiency of approximately 95% for genuine one-track τ_{had} candidates. The eVeto requirement is applied only in analysis regions where it provides an overall improvement in sensitivity, thereby minimising unnecessary signal losses. Anti-τ_{had} candidates, used to estimate backgrounds containing fake τ_{had} candidates, satisfy the same reconstruction and selection requirements as signal τ_{had} candidates but fail the Loose RNN-based τ_{had} identification requirement. In regions where this background-estimation method is employed, at most one anti-τ_{had} candidate is selected per event.

6.4 Event Selection Analysis Region

6.5 Multivariate Analysis

One key innovation in Multivariate Analysis (MVA) is the use of Transformer-based classifiers, similar to those we discussed in subsection 6.3.2. Unlike for b-tagging, here objects are tagged jets as a whole. The Transformer is used both for region selection and for constructing the discriminants in each region.

6.5.1 Network Architecture

Figure 6.2 gives the general architecture of the GNN used for this analysis. There are 4 main stages in this architecture.

- **Object-level feature engineering:** Each event has object features (per particle) and global features (per event). Instead of treating them separately, the architecture copies over the global feature to each object to create the **Combined Inputs**
- **Latent Space representation:** The combined inputs are then passed through Dense Deepset layers, which use ReLU activation[127] to create a

Table 6.2: Cutflow for the SM ggF HH signal in the hadhad channel in Run 3. The cuts are applied sequentially.[3]

Cut	Yield	Step Efficiency	Cumulative Efficiency
No cut	56.4586	100.000%	100.000%
$N_{e/\mu} = 0$	54.7659	97.002%	97.002%
$N_{\tau} = 2$	19.0235	34.736%	33.695%
Pass trigger	12.9745	68.202%	22.981%
OS	11.9880	92.397%	21.233%
> 40 GeV	11.8685	99.004%	21.022%
$N_{\text{central jets}} \geq 2$	11.4078	96.118%	20.206%
$N_{b\text{-jets}} \geq 1$	10.8891	95.454%	19.287%
$N_{b\text{-jets}} \geq 2$	7.4949	68.829%	13.275%
$N_{b\text{-jets}} = 2$	6.9865	93.217%	12.375%
Trigger thresholds	4.8567	69.515%	8.602%
Passes GRL for category and bucket	4.6912	96.594%	8.309%
> 60 GeV	4.6329	98.756%	8.206%
$m_{bb} > 40$ GeV	4.6202	99.726%	8.183%
$m_{bb} < 210$ GeV	4.5305	98.060%	8.025%
2tag	4.4526	98.280%	7.887%
Loose–loose (LL) τ had-ID	3.5914	80.658%	6.361%
ggF high-m_{HH} SR	2.7429	76.373%	4.858%
ggF low-m_{HH} SR	0.7789	21.689%	1.380%
VBF SR	0.0696	1.938%	0.123%

256-dimensional latent space object representation.

- **Transformer Encoder:** The latent space objects then pass through 4 multi-head attention layers, each with 4 heads, to create a **Conditional Representation** which then passes through a global attention pooling to get a 126-dimensional summary representation (each event is represented by a 128D vector)

- **Classifiers:** The summary representation is used as input for 4 different classifiers, 2 main classifiers, for events with $m_{HH} > 350$ GeV, a binary classifier to classify between ggF and VBS Signal and for the rest of the events ($m_{HH} < 350$ GeV), a multiclassifier to classify SM ggF signal, SM VBF signal, $\kappa_{\lambda 5}$ ggF signal and background. Scores of these classifiers are used to build the discriminant in those regions. The 2 auxiliary classifiers are tasked with object-level top parentage and Higgs boson parentage. They help the GNN as a whole train better.

6.5.2 Training Features and Strategy

The Training of the Transformer utilises a 5-fold cross-validation strategy. Figure 6.3 shows the different validation folds and their respective train, validation, and

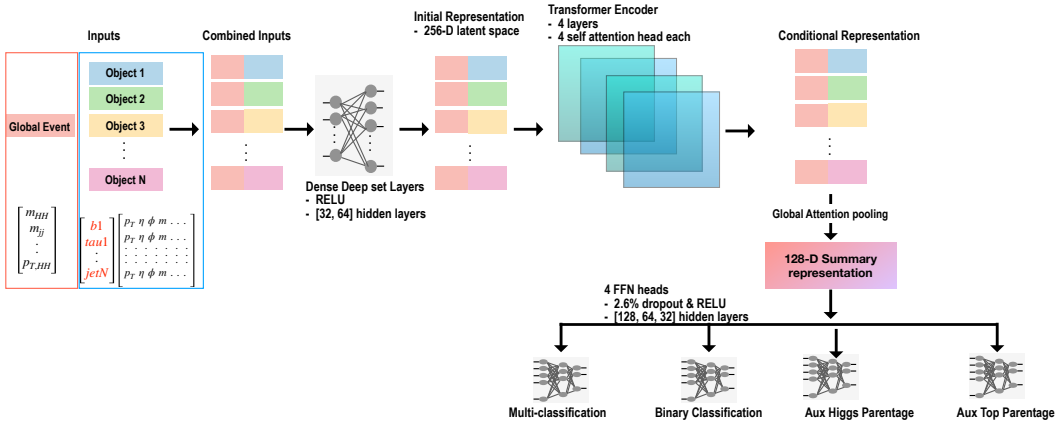


Figure 6.2: General Architecture for Graph Neural Networks (Transformer) used in this analysis.[3]

test sets.

The summary of input features used in the MVA. Event-level features are listed in the upper portion of the table, separated by channel, and object-level features common to all objects in the graph are listed in the lower portion Table 6.3(top) and object-level feature Table 6.3(bottom).

6.5.3 Event Categorisation

Event categorisation uses the scores from the 2 classifiers we discussed in subsection 6.5.1. From the multiclassifier heads, we can get p_{ggF} and p_{VBF}, the probabilities of the ggF and VBF signals, respectively. Using this, we can define ggF–VBF categorisation score S_{cat}, defined as

$$S_{\text{cat}} = \frac{p_{\text{ggF}}}{p_{\text{ggF}} + p_{\text{VBF}}}, \tag{6.7}$$

Events with $S_{\text{cat}} < 0.1$ are assigned to the SR VBF signal region, where the discriminant ($S_{\text{SR VBF}}$) is defined by

$$S_{\text{SR VBF}} = \frac{2\,p_{\text{VBF}}}{p_{\text{VBF}} + p_{\text{B}}} - 1. \tag{6.8}$$

For the remaining events, divide them into 2, SR Hi signal region with $m_{HH} > 350$ GeV and the rest $m_{HH} < 350$ GeV to SR Lo. The SR Hi signal region has a discriminant ($S_{\text{SR Hi}}$) defined by the predictions from the binary classifier $p_{HH,\text{ Hi}}$ and $p_{\text{B, Hi}}$, as

$$S_{\text{SR Hi}} = \frac{2\,p_{HH,\text{ Hi}}}{p_{HH,\text{ Hi}} + p_{\text{B, Hi}}} - 1, \tag{6.9}$$

Table 6.3: Summary of input features used in the MVA. Event-level features are listed in the upper portion of the table, separated by channel, and object-level features common to all objects in the graph are listed in the lower portion. Each object additionally carries a validity flag, and only valid objects are passed to the network.[3]

Variable	**Description**
Global event-level features	
m_{HH}	Invariant mass of the reconstructed HH system using the di-τ four-momentum after the MMC algorithm
m_{bb}	Invariant mass of the b-jet pair
ΔR_{bb}	ΔR between the two b-jets
m_{jj}	Invariant mass of the two leading VBF jets
ΔR_{jj}	ΔR between the two leading VBF jets
$\eta_{\text{VBF } j_0} \cdot \eta_{\text{VBF } j_1}$	Product of pseudorapidities of the two leading VBF jets
$\Delta R_{\tau\tau}$	ΔR between the two τ_{had} candidates
Object-level features (per-object inputs to the transformer network)	
η	Pseudorapidity; set to 0 for $\mathbf{E}_{\text{T}}^{\text{miss}}$
ϕ	Azimuthal angle
m	Mass; set to 0 for $\mathbf{E}_{\text{T}}^{\text{miss}}$, leptons, and taus
p_{T}	Transverse momentum
PCBT	GN2 pseudo-continuous b-tagging score (bins: 85%–77%–70%–60%) ; set to -2 for non-b-jet objects
isTau	Boolean flag: object is a τ_{had} candidate
isJet	Boolean flag: object is a non-b-tagged jet
isBjet	Boolean flag: object is a b-tagged jet
isMET	Boolean flag: object is $\mathbf{E}_{\text{T}}^{\text{miss}}$
isMMC	Boolean flag: object is the MMC di-τ system
isLepton	Boolean flag: object is a lepton (only in $\tau_{\text{lep}}\tau_{\text{had}}$ channel)



Figure 6.3: Cross Validation datasets for this Analysis, For each Network, 3/5th are used for training (green), 1/5th for Validation (Yellow) and 1/5th for Testing (Orange)

The SR Lo signal region has a discriminant ($S_{\text{SR Lo}}$) given by

$$S_{\text{SR Lo}} = \frac{2\,p_{\text{BSM}}}{p_{\text{BSM}} + p_{\text{B}}} - 1. \tag{6.10}$$

where p_{BSM} and p_{B} are the probabilities of ggF $\kappa_{\lambda 5}$ and background samples repectively. Figure 6.4 gives the full flow chart of the Event selection.

6.6 Systematic uncertainties

6.6.1 Experimental Uncertainities

Experimental uncertainties include contributions from the integrated luminosity measurement, pileup modelling, and the reconstruction and identification of detector objects. These uncertainties are evaluated using the tools and recommendations provided by the ATLAS Combined Performance (CP) groups. Systematic uncertainties associated with detector objects are available at different levels of granularity through various reduction schemes. As experimental uncertainties have a limited impact on the overall sensitivity of the analysis, the most aggressive reduction schemes are adopted wherever possible, as summarised in Table Table 6.4. An exception is made for the JES/JER uncertainties, for which the reduction scheme recommended for combined analyses is used. The relative uncertainty on the integrated luminosity is 2.0% for the Run 3 (2022–2023) dataset [128, 129].

6.6.2 Theoretical modelling uncertainties

Normalisation and Acceptance Uncertainties

Systematic uncertainties in background and signal predictions are categorised into overall normalisation, relative acceptance, and shape uncertainties. Overall

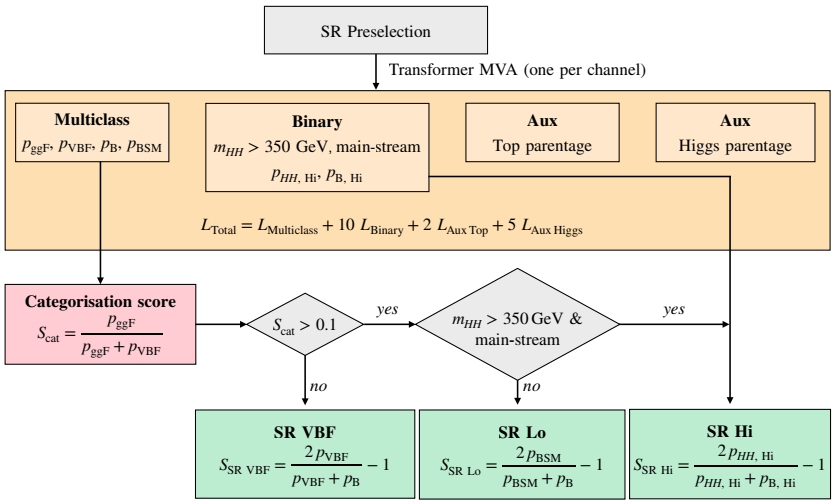


Figure 6.4: Flow chart of the event categorisation from the MVA analysis.[3]

Table 6.4: Overview of experimental systematic uncertainty reduction schemes used in the analysis.

Source	Reduction scheme
EGamma	Total reduction scheme (limited sensitivity)
Muon	Tightest reduction scheme (limited sensitivity)
Tau	Standard
FTAG	No reduction available, thus the full set of uncertainties is used
JES	Category-based reduction
JER	Full set of uncertainties is used

cross-section uncertainties are assigned to minor background processes, including s- and t-channel single top, Z+light-flavour, W+jets, WW, and WZ, as detailed in Table 6.5. Conversely, the normalisations for $t\bar{t}$, tW, and Z+HF are unconstrained and left as free parameters in the likelihood fit to data, with relative acceptance uncertainties evaluated relative to the Z+HF control region. For single-Higgs backgrounds, theoretical uncertainties from QCD scale variations and PDF sets are included for the ttH and ZH channels. A 100% normalisation uncertainty is assigned to minor $H \to \tau\tau$ processes (ggF, VBF, and WH) produced with heavy-flavour jets, where ggF+HF represents the leading background modelling uncertainty. Finally, signal cross-section and $H \to b\bar{b}/\tau\tau$ branching ratio uncertainties follow theoretical recommendations, with acceptance uncertainties evaluated from scale variations, PDF propagation, and parton-shower comparisons.

Shape and Specific Modelling Uncertainties

Unless explicitly stated otherwise, acceptance shape and normalisation uncertainties are treated as uncorrelated across regions. For $t\bar{t}$ and tW modelling, matrix-

Table 6.5: Cross-section uncertainties used for the background processes whose overall normalisation is not determined from data in the fit.[3]

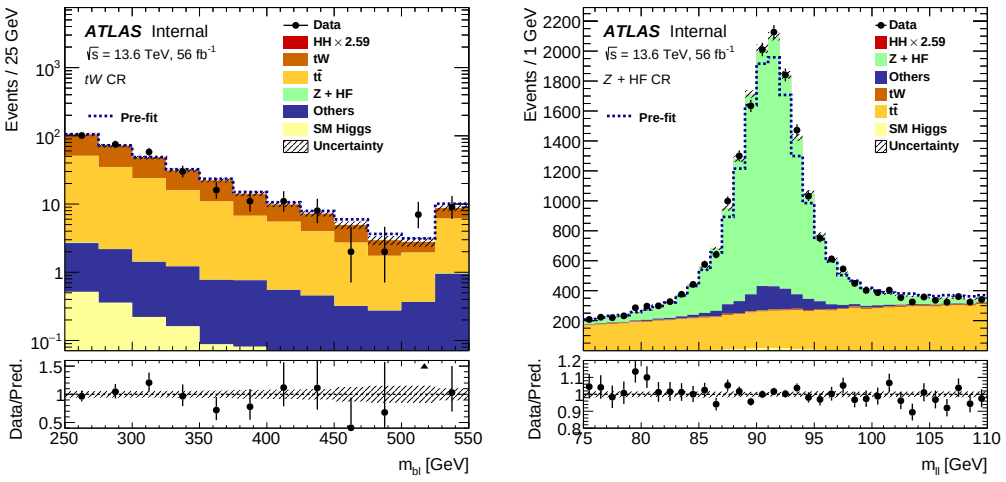
Sample	Value	Comment
Z+light flavour	5%	
W+jets	5%	
single top s- and t-channel	4.1%	s/t average
tW channel	5.3%	
diboson	6%	
ttH	$^{+5.8\%}_{-9.2\%}$ (scale) $\pm 3.6\%$ (PDF+α_S)	YR4
$qqZH$	$^{+0.5\%}_{-0.6\%}$ (scale) $\pm 1.9\%$ (PDF+α_S)	YR4
$ggZH$	$^{+25\%}_{-19\%}$ (scale) $\pm 2.4\%$ (PDF+α_S)	YR4
WH	$^{+0.5\%}_{-0.7\%}$ (scale) $\pm 1.9\%$ (PDF+α_S)	YR4
ggF $H \to \tau\tau$	$\pm 3.9\%$ (scale) $\pm 3.2\%$ (PDF+α_S)	YR4
VBF $H \to \tau\tau$	$^{+0.4\%}_{-0.3\%}$ (scale) $\pm 2.1\%$ (PDF+α_S)	YR4

element matching (h_{damp}, $p_{\text{T,hard}}$), parton-shower choices (Herwig 7 vs. Pythia 8), scale variations, PDFs, and initial- and final-state radiation are evaluated using alternative samples or internal weight variations. Additional specific uncertainties account for the top-quark decay recoil scheme in $t\bar{t}$, while tW–$t\bar{t}$ interference is evaluated by comparing diagram removal with diagram subtraction, reduced by a factor of two to avoid overestimating the uncertainty. For Z+HF processes, generator uncertainties are evaluated by comparing nominal Sherpa with MadGraph5_aMC@NLO + Pythia 8 using FxFx merging, which directly impacts the classifier score shape. Dedicated Z+HF uncertainties also cover heavy-flavour composition differences (Z+$bc/bl/cc$ relative to Z+bb) and Z boson p_{T} mismodelling derived from data. Lastly, classifier score shapes for both signal and single-Higgs processes (ttH and ZH) incorporate scale variations, PDF propagation, and comparisons between the generator and shower in the $\tau_{\text{had}}\tau_{\text{had}}$ signal regions.

6.7 Results and discussion

The analysis uses binned likelihood estimation to infer the 2 parameters of interest, signal strength μ_{HH} and Higgs self-coupling constant κ_λ. For μ_{HH}, the likelihood construction is similar to the example from the fair universe (section 4.5). In the $\tau_{\text{had}}\tau_{\text{had}}$ channel, we use five regions, two as Control Regions and 3 as Signal Regions. The two control regions, Z + HF CR and tW CR, the $m_{\ell\ell}$ and $m_{b\ell}$ variables, respectively, were used as the discriminants. Figure 6.5 shows fitted variables along with the background contributions after the fit, shown as filled histograms. The size of the combined statistical and systematic uncertainties for the sum of the

fitted signal and background is indicated by the hatched band.



(a) (b)

Figure 6.5: Post-fit distributions of fit variables in the control regions for Run3 $\tau_{\text{had}}\tau_{\text{had}}$ channel. The background contributions after the fit are shown as filled histograms. The size of the combined statistical and systematic uncertainties for the sum of the fitted signal and background is indicated by the hatched band. (a) Z+HF CR with $m_{\ell\ell}$ discriminant and tW CR with $m_{b\ell}$ discriminant.[3]

For the three Signal regions (SR ggF high-mass, SR ggF low-mass, and SR VBF) and the two control regions, the discriminants described in subsection 6.5.3 are used. Figure 6.6 shows fitted variables along with the background contributions after the fit, shown as filled histograms. The size of the combined statistical and systematic uncertainties for the sum of the fitted signal and background is indicated by the hatched band.

For κ_λ, we will use the parametric form we described in Equation 6.6, with the full negative loglikelihood given as:

$$t_\mu = -\log\left(\frac{\mathcal{L}(\kappa_\lambda)}{\mathcal{L}(1)}\right) = \sum_{k=1}^{N_{\text{bins}}} -n_k \log(\nu(\kappa_\lambda)_k) + (\nu(\kappa_\lambda)_k) + \text{const} \tag{6.11a}$$

$$\begin{aligned}\nu(\kappa_\lambda)_k = &\left(\frac{\kappa_\lambda^2 - 6\kappa_\lambda + 5}{5}\right)\cdot \nu_k^{\kappa_{\lambda 0}} + \left(\frac{-\kappa_\lambda^2 + 5\kappa_\lambda}{4}\right)\cdot \nu_k^{\kappa_{\lambda 1}} \\ &+ \left(\frac{\kappa_\lambda^2 - \kappa_\lambda}{20}\right)\cdot \nu_k^{\kappa_{\lambda 5}} + \sum_{\text{x} \in \text{ backgrounds}} \nu_k^{\text{x}}\end{aligned} \tag{6.11b}$$

Figure 6.7 gives the expected and the observed negative loglikelihood curve described in Equation 6.11b for $\tau_{\text{lep}}\tau_{\text{had}}$ and $\tau_{\text{had}}\tau_{\text{had}}$ channels for both Run 2 and Run 3.

When we combined the results from Figure 6.7, we obtained the final unblinded results shown in Figure 6.8.

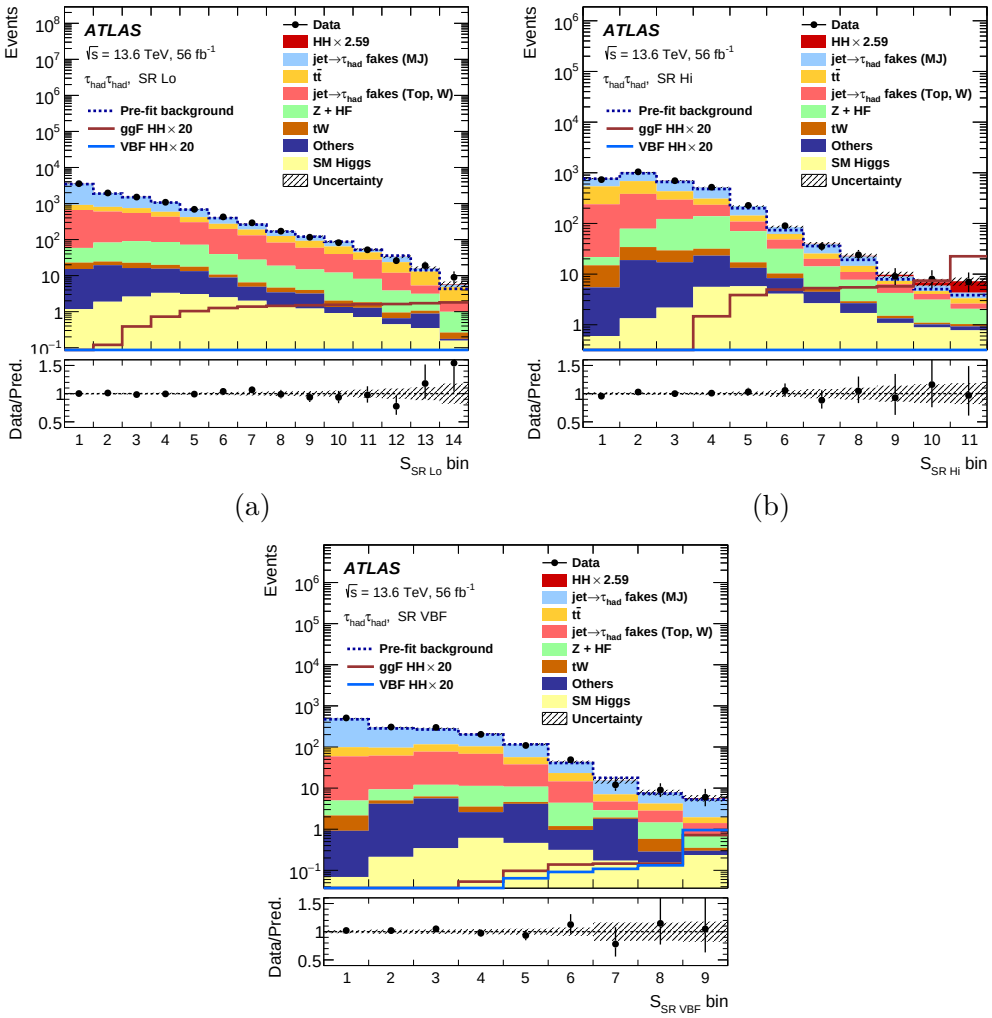
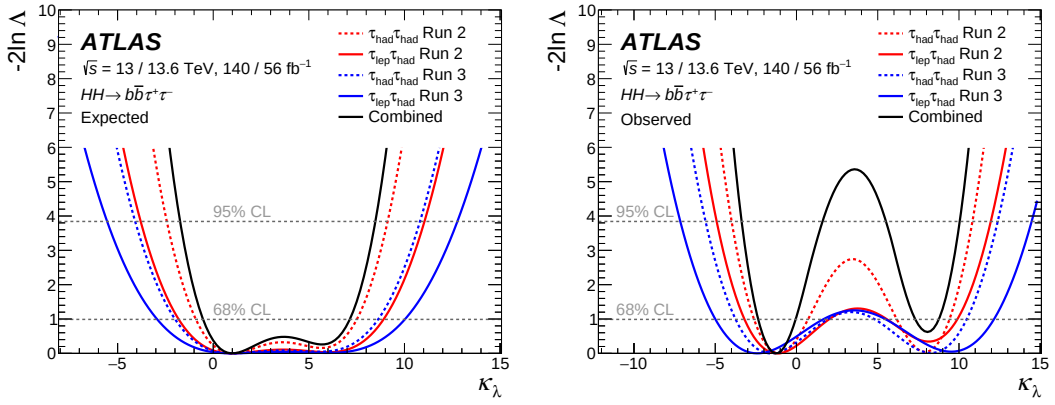


Figure 6.6: Post-fit distributions of respective GNN discriminants in the signal regions for Run3 $\tau_{\text{had}}\tau_{\text{had}}$ channel. The background contributions after the fit are shown as filled histograms. The size of the combined statistical and systematic uncertainties for the sum of the fitted signal and background is indicated by the hatched band. (a) ggF low-mass SR, (b) ggF high-mass SR and (c) VBF SR.[3]

The κ_λ scans shows in Figure 6.8 shows that we can set a $\text{CL}_{95\%}$ interval of $\kappa_\lambda \in [-3.4, 1.6] \cup [5.5, 10.1]$ and the best fit value is $\kappa_\lambda = 2.6$, 2.6 times the Standard Model prediction. The exact positions of the best fit and the distance between the 2 minima of the NLL can be attributed to shape changes across different κ_λ hypotheses. Though the current analysis improved the κ_λ interval by around 15% compared to the past by utilising tools like GNNs and improved tagging, we still have a long way to go before we can achieve statistically significant results for new discoveries. In section 2.3 we discussed how HL-LHC can open doors to new discoveries by providing more luminosity and, by extension, statistics. As we mentioned, the κ_λ parameter estimation strongly relies on the shapes of different hypotheses,



(a) (b)

Figure 6.7: Negative loglikelihood curve for κ_λ scan for various channels for expected (a) and the observed (b). The dashed lines correspond to $\tau_{\text{had}}\tau_{\text{had}}$ and the solid lines to $\tau_{\text{lep}}\tau_{\text{had}}$ and Run 2is in blue and Run 3 in red. Finally, the combined results are shown in black.[11]

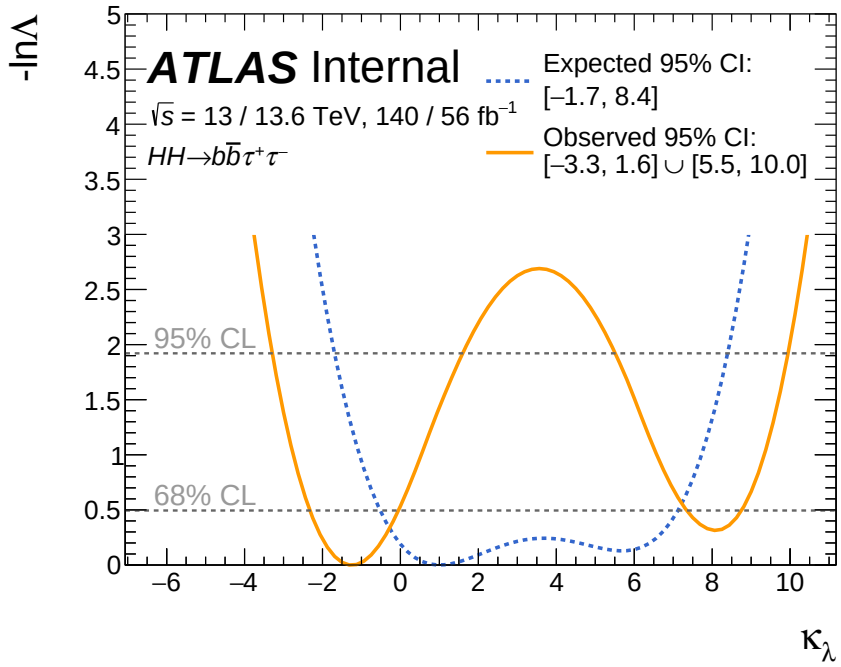


Figure 6.8: Negative loglikelihood curve for κ_λ scan for all channels Run 2+3 for expected (blue dashed line), and the observed (orange solid line).[3]

and an analysis that captures the n-dimensional shapes of the events could unlock additional information and hence provide better limits on κ_λ. In the next chapter (chapter 7) we will discuss on how to build this using Simulation-Based Inference.

Chapter 7
Simulation-based Inference for Di-Higgs Analysis

The journey of a thousand miles begins with a single step.
— Lao Tzu

Simulation-Based Inference full potential can be unlocked when applied to problems with quantum interference. The 2024 ATLAS off-shell Higgs to 4 lepton Paper[1, 2] proved the applicability of SBI to real particle physics data. In that work, the interference occurs between the ggF Higgs and the ($qqZZ$) processes, with the latter involving loop contributions. A similar phenomenon occurs in Di-Higgs production, where quantum interference arises between the triangle and box loop amplitudes. In this chapter, we will expand on what was discussed in chapter 5, apply it to the Higgs pair-production task, and discuss the proof of concept for such an analysis, the expected challenges, and the roadmap for a full analysis. We plan to follow the standard analysis addressed in chapter 6 as closely as possible to have a reliable comparison of gains with SBI. Hence, for the dataset preparation, we used the same Monte Carlo samples, triggers, and jet tagging as the Binned analysis. Finally, we re-implemented the Binned analysis using the same Transformer-based Classifier within our framework to compare the two analyses.

As we discussed in chapter 5, there are many possible routes to take when applying SBI to an analysis. The one we took was the Semi-Parametric SBI using Neural Density Ratio Estimation. The main argument for attempting this as the first step was its successful use in the off-shell Higgs paper, which helped gain acceptance within the ATLAS Collaboration. Regardless, certain critical reasons prevented us from pursuing the winning solutions in the FAIR Universe Challenge. Both Ibrahim's and HEPHY's methods relied on numerous systematics-shifted simulations to train the model. In ATLAS analyses, we are always limited to only 2 systematic samples, apart from the nominal sample, one up-shifted and the other down-shifted by 1σ, which makes it difficult to implement those methods.

7.1 Semi-Parametric SBI for Di-Higgs

We are addressing two parameter estimation methods in this analysis. First is the signal strength μ, and the other is the Higgs self-coupling constant λ through the κ_λ parameter. The signal strength μ measurement is similar to the FAIR Universe

case, with the only difference coming from the samples themselves; hence, the general expression for the density ratio is given by:

$$\frac{p(x_i|\mu)}{p_{\text{ref}}(x_i)} = \frac{1}{\nu(\mu)}\left[\mu \cdot \nu_{\kappa_{\lambda 1}} \cdot \frac{p_{\kappa_{\lambda 1}}(x_i)}{p_{\text{ref}}(x_i)} + \sum_{\text{B} \in \text{ bkgs}} \nu_{\text{B}} \cdot \frac{p_{\text{B}}(x_i)}{p_{\text{ref}}(x_i)}\right] \tag{7.1}$$

In the FAIR Universe dataset, nearly 20 processes were grouped into four categories. Similarly, we can group various physics processes involved in this channel. In traditional binned analysis, the groups depend on calibration and systematics. In the case of SBI, there is one more point to keep in mind: the phase-space distribution of the different processes and the reference process. This is important because they determine how well we can train our density-estimating classifiers. Appendix C presents the grouping of samples into different categories, and Table 7.1 gives the summary of each category along with their numbers of samples (N), total yield, average weight ($\bar{w}$), and effectiveness (E_{ff}). The two metric which we need to pay attention to when building the machine learning classifier are the number of samples and effectiveness. The number of samples indicates which ML algorithms can be trained; for example, DNNs or Transformers would need far more events to achieve stable results than BDTs, and effectiveness reflects the relative spread of the weights. The exact expression for effectiveness is given by

$$E_{ff} = \frac{1}{N} \cdot \frac{\sum w_i}{\sqrt{\sum w_i^2}} \tag{7.2}$$

where w is the per-event weight. If the effectiveness is too low, it means the spread is too large, which could hinder a binary classifier from properly learning the density ratios.

Table 7.1: Summary Statistics by Di-Higgs analysis samples

Category	**Yield**	**Number of Samples**	**Average Weight**	**Effectiveness (E_{ff})**
$HH \to b\bar{b}\tau^+\tau^-$, $\kappa_\lambda = 1$	3.999 ± 0.022	50204	0.022	0.644
$HH \to b\bar{b}\tau^+\tau^-$, $\kappa_\lambda = 0$	7.386 ± 0.043	42585	0.043	0.707
$HH \to b\bar{b}\tau^+\tau^-$, $\kappa_\lambda = 5$	5.123 ± 0.038	22266	0.038	0.810
Diboson	42.725 ± 0.697	121216	0.016	0.031
Single Higgs	68.874 ± 0.423	214677	0.001	0.123
Rest (Z+ jets and tops)	1600.185 ± 21.134	120791	0.031	0.047

Taking the categories from Table 7.1, we can write the full expression of the probability ratio $\frac{p(x_i|\mu)}{p_{\text{ref}}(x_i)}$ as :

$$\frac{p(x_i|\mu)}{p_{\text{ref}}(x_i)} = \frac{1}{\nu(\mu)}\left[\mu \cdot \nu_{\kappa_{\lambda 1}} \cdot \frac{p_{\kappa_{\lambda 1}}(x_i)}{p_{\text{ref}}(x_i)} + \nu_H \cdot \frac{p_H(x_i)}{p_{\text{ref}}(x_i)} + \nu_{VV} \cdot \frac{p_{VV}(x_i)}{p_{\text{ref}}(x_i)} + \nu_{\text{Rest}} \cdot \frac{p_{\text{Rest}}(x_i)}{p_{\text{ref}}(x_i)}\right] \tag{7.3}$$

From here we can take the κ_λ formulation we discussed in Equation 6.6 and get a complete expression for the probability ratio for $\frac{p(x_i|\kappa_\lambda)}{p_{\text{ref}}(x_i)}$ as :

$$\begin{aligned}\frac{p(x_i|\kappa_\lambda)}{p_{\text{ref}}(x_i)} = \frac{1}{\nu(\kappa_\lambda)}\Bigg[& \left(\frac{\kappa_\lambda^2 - 6\kappa_\lambda + 5}{5}\right) \cdot \quad \nu_{\kappa_{\lambda 0}} \cdot \frac{p_{\kappa_{\lambda 0}}(x_i)}{p_{\text{ref}}(x_i)} \\ & + \left(\frac{-\kappa_\lambda^2 + 5\kappa_\lambda}{4}\right) \cdot \quad \nu_{\kappa_{\lambda 1}} \cdot \frac{p_{\kappa_{\lambda 1}}(x_i)}{p_{\text{ref}}(x_i)} \\ & + \left(\frac{\kappa_\lambda^2 - \kappa_\lambda}{20}\right) \cdot \qquad \nu_{\kappa_{\lambda 5}} \cdot \frac{p_{\kappa_{\lambda 5}}(x_i)}{p_{\text{ref}}(x_i)} \\ & + \nu_H \cdot \frac{p_H(x_i)}{p_{\text{ref}}(x_i)} + \qquad \nu_{VV} \cdot \frac{p_{VV}(x_i)}{p_{\text{ref}}(x_i)} \quad + \nu_{\text{Rest}} \cdot \frac{p_{\text{Rest}}(x_i)}{p_{\text{ref}}(x_i)}\Bigg]\end{aligned} \tag{7.4}$$

As we discussed in section 5.3, we can compute these density ratios ($\frac{p_X(x_i)}{p_{\text{ref}}(x_i)}$) with ML Classifiers.

7.1.1 Preselection

In chapter chapter 5, we discussed how both **HEPHY** and **ATLAS** were used in the phase-space to define regions, and how the signal-rich region was utilised for the unbinned analysis. The primary purpose of selection is to identify the signal-rich region for unbinned analysis. Additionally, we have to consider the reference sample when selecting the preselection region to avoid unsupported events in the ML training. Our preselection strategy is a combination of both. Similar to ATLAS, we have a discriminant based on a multiclassifier score. We experimented with various multiclassifier architectures, including BDTs, DNNs, and GATs, and ultimately converged on DNNs for the final results. The features and hyperparameters used for the preselection DNN are given in Table 7.2

The architecture for the Preselection Network follows a simple DNN with 4 dense layers of 1024 nodes, followed by 1 dense layer of 128 nodes, both with swish activation, and finally concluding at a 6-node output layer with softmax activation to give the probability for each class (P_{x}). The expression for the preselection discriminant ($\texttt{score}_{\texttt{presel}}$) is given by :

$$\texttt{score}_{\texttt{presel}} = \ln\left(\frac{\sum\limits_{i\in \text{ signal processes}} P_i}{\sum\limits_{j\in \text{ Background processes}} P_j}\right) \tag{7.5}$$

where the signal processes were all processes excluding the Rest (top and $Z+$ jets)) and the background processes were top and $Z+$jets categories. Figure 7.1

gives the distribution of the score discriminant with and without considering the weights. The 3 backgrounds are diboson, single Higgs, and Rest (top and $Z+$Jets). We present the histograms with and without event weights to illustrate the difference between the MC sample size and the yield. We see that it's predominantly single Higgs, which has the largest number of MC samples. The reason we consider both is to assess the number of training events in each sample and the impact of those samples, given their weights.

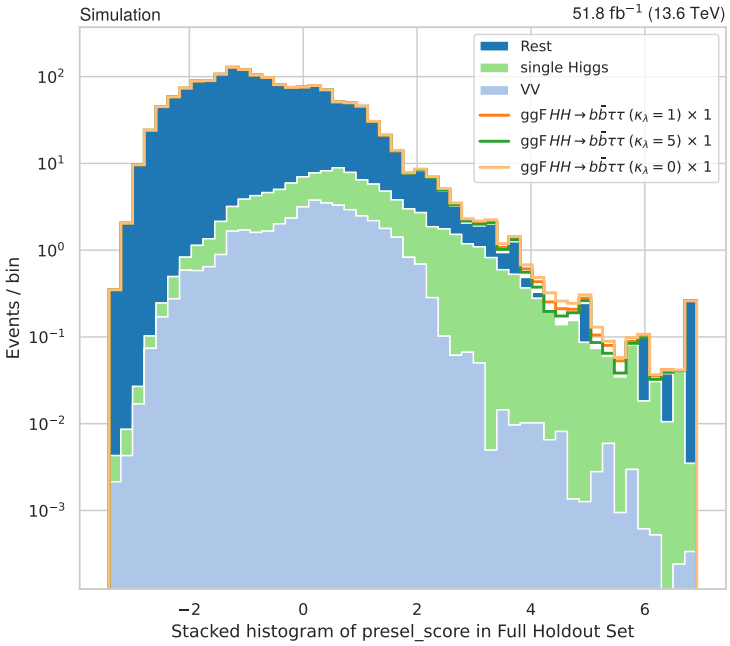
To further improve preselection, we can also use additional cuts on features such as mass and ΔR to select the signal-rich region for the unbinned analysis. Table 7.3 gives the cuts applied to select the unbinned signal region.

7.2 Choice of Reference sample in SBI

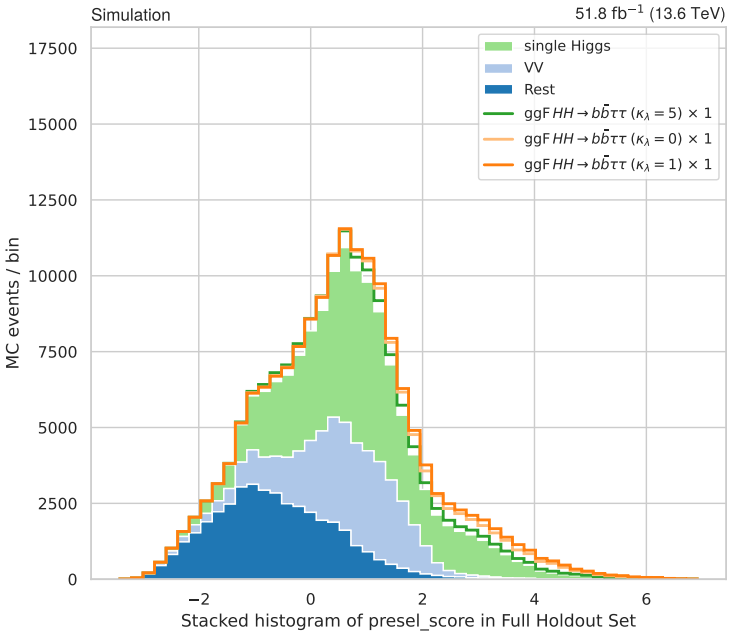
In section 5.3, we addressed how we can construct the probability ratio from the density ratio estimated by the ML algorithms. A crucial step in this process is choosing the reference function. Ideally, under the assumption of infinite data, any reference sample could be used for NSBI analysis. Unfortunately, we have a rather limited number of events; hence, a poorly chosen reference sample can degrade the quality of ML training. There is no straightforward answer to the question of the ideal reference; we have several intuitions, but no clear metric for making that choice. The simplest verification is to build histograms of different variables in different ways and check whether the target distribution falls within the reference distribution. A more accurate question would be for every bin in the target distribution, does the reference sample have enough events? But one-dimensional verification is not ideal, since we are trying to estimate the performance of the reference samples in a multivariate classifier; a one-dimensional approach is insufficient. The next approach was to use a pair plot to visualise the 2D coverage across different variables. Figure 7.2 shows the pair plots for a few key features: m_{HH}, m_{bb}, $m_{\tau\tau}^{\text{MMC}}$, and $E_{\text{T}}^{\text{miss}}$ used to train the model. There are 2 distributions

Table 7.2: Neural Network Multiclassifier training parameters for Preselection Classifier

Parameter	**Value**
Training parameters	
Loss	categorical crossentropy
optimizer	Adam
Learning rate	10^{-4}
Batch size	64
Epochs	20
Global features	m_{HH}, m_{BB}, MET, ΔR_{BB}, m_{MMC}, $\Delta R_{\tau\tau}$, $\Delta R_{b0\tau}$, $\Delta R_{b1\tau}$, n_{jets}



(a)



(b)

Figure 7.1: Histogram of preselection discriminant. There are 6 samples in this; the 3 backgrounds are diboson (light blue), single Higgs (light green), and Rest (top and Z + Jets) (dark blue). The HH signals are individually stacked on top of the background. a) is with event weights b) is without event weights

shown, one for $HH \to b\bar{b}\tau^+\tau^-$ and the other for single Higgs. The single Higgs is shown as the reference distribution ultimately chosen.

Table 7.3: Selection cuts applied to select the Unbinned signal region

Variable	Lower bound	Upper bound
$\texttt{score}_{\texttt{presel}}$	-1.0	5
m_{HH}	60	1000
m_{bb}	30	160
mmc_m	60	160
ΔR_{BB}	—	3.5
$\Delta R_{\tau\tau}$	—	3.5
$\Delta R_{b0,\tau}$	—	3.5
$\Delta R_{b1,\tau}$	—	3.5

7.2.1 Gaussian Kernel Density Estimation

To quantify the concept of overlap, we can extend this intuition into a full metric using Gaussian Kernel Density Estimation (KDE) [130]. The idea is first to construct a 3σ boundary for both the target and reference distributions. And then divide the n-dimensional volume of the region of the target inside the reference

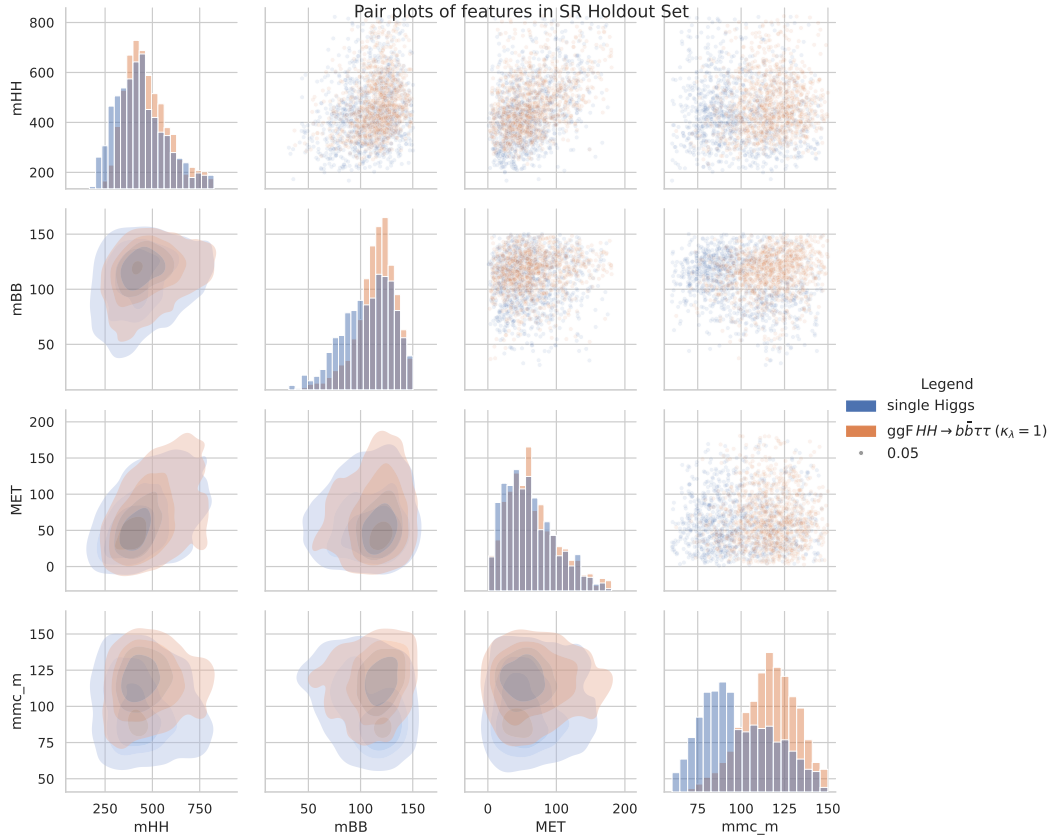


Figure 7.2: Pair plot for m_{HH}, m_{bb} and $m_{\tau\tau}^{\text{MMC}}$ and $E_{\text{T}}^{\text{miss}}$ used in the signal region on $HH \to b\bar{b}\tau^+\tau^-$, $\kappa_\lambda = 1$ and single H.

by the volume of the target. We introduce the Ordered Overlap metric ($\mathcal{O}_{S\to R}$), which measures the extent to which a set of Samples S is contained within the high-density region of a reference set R. Unlike symmetric similarity measures, Ordered Overlap is directional, as the density estimate is constructed only from the reference distribution R. The probability density of R is estimated using kernel density estimation (KDE):

$$\hat{p}_R(x) = \frac{1}{Nh^d} \sum_{i=1}^{N} K\left(\frac{x - r_i}{h}\right), \tag{7.6}$$

where $K(\cdot)$ denotes the kernel function, h is the bandwidth parameter, N is the number of reference samples, and d is the dimensionality of the sample space. The density threshold (τ) is obtained from the empirical distribution of KDE values evaluated at the reference samples:

$$\tau = Q_\alpha\left(\{\hat{p}_R(r_i)\}_{i=1}^{N}\right), \tag{7.7}$$

where $Q_\alpha()$ denotes the α-quantile operator. The corresponding high-density region of the reference distribution is defined as

$$\Omega_R = \left\{x \in \mathbb{R}^d : \hat{p}_R(x) > \tau\right\}, \tag{7.8}$$

where $\hat{p}_R(x)$ is the kernel density estimate of the reference distribution and τ is the selected density threshold. The overlap region between the target set S and the high-density region of the reference distribution is denoted as

$$\mathcal{O}_{S,R} = S \cap \Omega_R \tag{7.9}$$

Equivalently, the overlapping samples can be expressed as

$$\mathcal{O}_{S,R} = \{s_j \in S : \hat{p}_R(s_j) > \tau\}. \tag{7.10}$$

The Ordered Overlap score ($\mathcal{O}_{S\to R}$) is then defined as the fraction of samples from S that belong to the overlap region:

$$\mathcal{O}_{S\to R} = \frac{|\mathcal{O}_{S,R}|}{|S|} = \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}\left[s_j \in \mathcal{O}_{S,R}\right] \tag{7.11}$$

where $|\mathcal{O}_{S,R}|$ denotes the number of samples in the overlap region, and $|S| = M$ is the total number of samples in the target set. Figure 7.3 gives the Overlap score ($\mathcal{O}_{S\to R}$) for different samples and references.

From $\mathcal{O}_{S\to R}$ in Figure 7.3, we can rule out any Di-Higgs sample and di-boson samples which have a poor score in many regions. That leaves out the single Higgs and the Rest(Z + jets and tops). From Table 7.1, we can see that Rest(Z + jets and tops) have too low effectiveness and too few training events, making it a poor choice. Thus, we finally chose a single Higgs sample as our reference sample for this analysis.

7.2.2 Uniform Reference Distribution

From the overlap metric and the effectiveness and number of events argument, one could think that a uniform distribution of the number of features would satisfy all these conditions and serve as a reference distribution. Hence, we tested the theory with BDT and NN. We generated a dataset such that for each feature, 10M events were generated from a uniform distribution, with a range $[\mu - 6\sigma^2 \ , \ \mu + 6\sigma^2]$. The range was chosen to maintain uniformity while minimising the number of samples lost in preselection. Figure 7.4 gives the ROC curve and calibration curve for the DNN classifier between $HH \to b\bar{b}\tau^+\tau^-, \kappa_\lambda = 1$ and uniform distribution.

The ROC and AUC are near perfect, but as we discussed in chapter 5, good classifiers need not be good density estimators, and this is evident in Figure 7.4(b), where we see that despite a very good ROC, the calibration is subpar. This is understood as the distributions between the uniform and any physics process being sufficiently different that a DNN can easily learn them; most scores are either 0 or 1, and the calibration is noticeably good in these regions. This is not enough; we also need the calibration to be good in the central part of the score distribution.

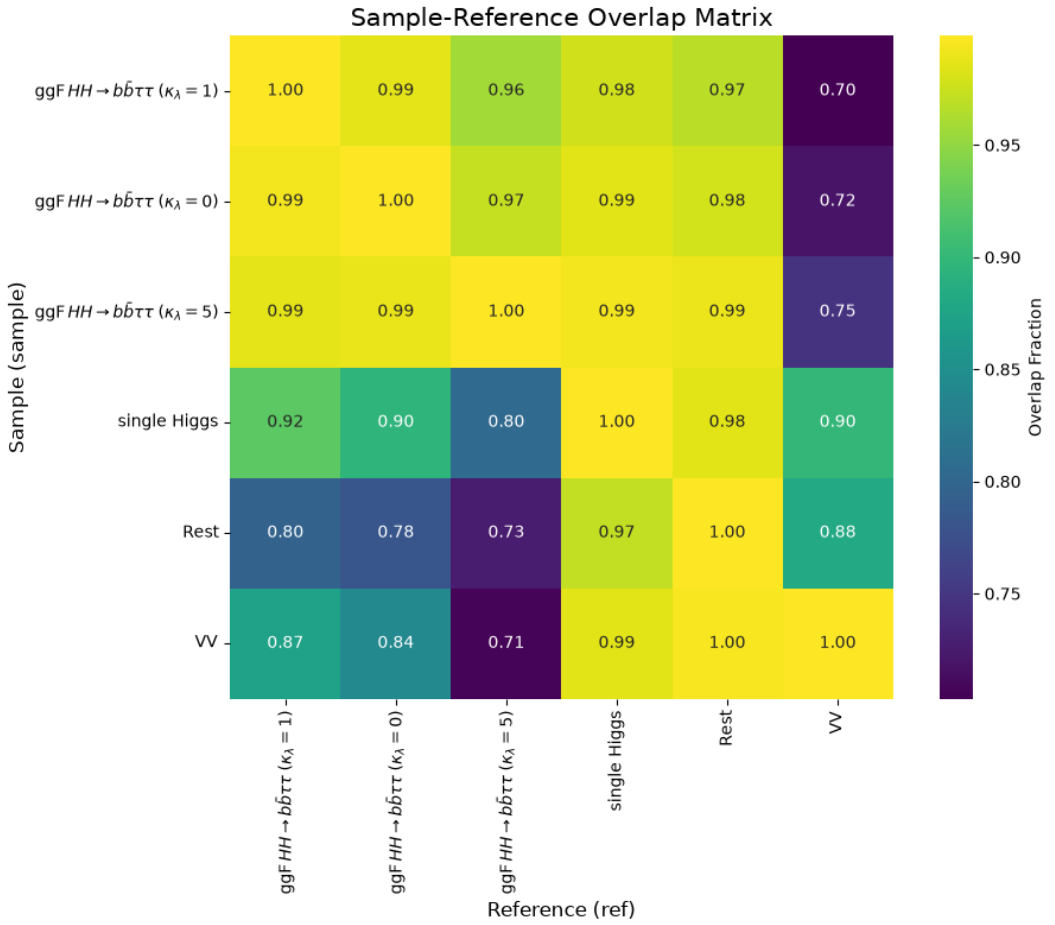
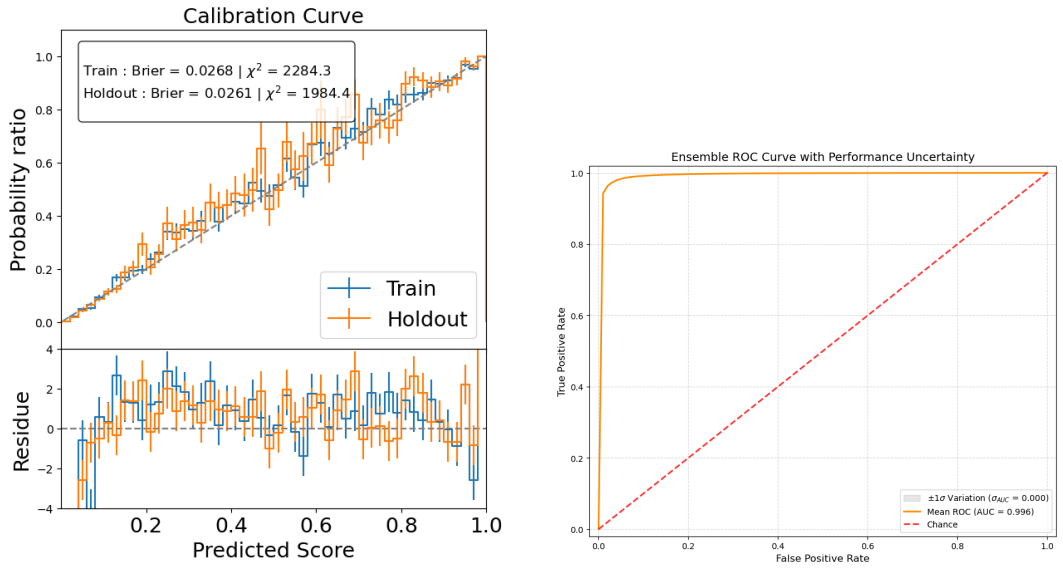


Figure 7.3: Ordered Overlap $\mathcal{O}_{S\to R}$ between 6 different sample to find best covering references



(a) (b)

Figure 7.4: Calibration Curve (a) and ROC curve for DNN trained to classify $HH \to b\bar{b}\tau^+\tau^-, \kappa_\lambda = 1$ and uniform distribution

That being said, there are limitations to our studies, such as the preselection we used being the same as that used when a single Higgs is the reference. We also haven't studied other distributions, and our current implementation assumes that all features are perfectly uncorrelated, which is not physically realistic.

7.3 Density Ratio Estimation Classifiers

All SBI methods discussed in chapter 5 rely on some form of dense neural network for density ratio estimation. This is not a hard-and-fast rule; technically, any classifier with the right loss function can perform this task. The ATLAS Off-shell Higgs to 4 lepton paper [1, 2], used around 100 DNNs to get stable classifier scores. This comes at the cost of computing resources and requires larger simulated samples. In our case, we have 2 major differences from the off-shell Higgs boson analysis or FAIR Universes. The first is that we have more final-state objects, which gives us the advantage of using Graph-Based algorithms like Transformers, as discussed in subsection 6.5.1. The other is the limited number of training events, which prompts us towards data-frugal architectures like BDTs. A detailed study of data-frugal methods and Tabular Foundation Models is presented in Appendix B. So we decided to compare DNN, BDT, and the Graph Attention Network (GAT) head-to-head. For this, it is necessary to keep as many external factors as possible, the main one being input features. In subsection 6.5.1, we discussed how the dataset has object-level and global (event-level) features. But DNN and BDT can only work with event-level features. To make a fair comparison between GAT and BDTs, we flattened the object-level features for the DNN and BDTs so that the same information is used

across all three models. Additionally, we replace the ϕ variable with $\cos(\phi)$ and $\sin(\phi)$ wherever applicable to improve training, as angles often scale poorly.

Table 7.4: General parameters and training features used for the DNN, BDT, and GAT classifiers.

Parameter	**Value**
Additional options	
Calibration	False
Use trigonometric conversion	True
Use particle features	True
Global features	m_{HH}, m_{BB}, MET, ΔR_{BB}, m_{MMC}, $\Delta R_{\tau\tau}$, $\Delta R_{b0\tau}$, $\Delta R_{b1\tau}$, n_{jets}
Object fields	η, ϕ, m, p_T, $\text{quant}_{\text{GN2}}$, isTau, isNu, isBjet, isJet, isMET, isMMC

7.3.1 Dense Neural Network

Dense Neural Network is the standard ML Classifier for SBI and has been used in many applications, including the 3 we discussed in chapter 5. For the DNN benchmark, we utilised a simple NN with 3 dense layers of 1024 nodes each, combined with a swish activation. The final layer was a single node with a sigmoid activation function, and combining it with the BCE loss provides the right conditions to infer the density ratio from the score. Table 7.5 gives the full set of hyperparameters and training features used for the DNN.

7.3.2 Boosted Decision Trees

Boosted Decision Trees (BDTs) are often seen as a faster but less accurate alternative to DNNs and other neural networks. Compared to Neural Networks, BDTs are often more stable and thus don't require a large ensemble to achieve the stability we need for density ratio estimation. For this thesis, we explored various BDT algorithms before selecting the classifier. The 4 we considered are CatBoost[131], XGBoost[107] and the Scikit-learn Hist Gradient Boost Method. One key point to consider here is that, unlike FAIR Universe datasets. The event weights in Di-Higgs samples can be negative due to subtraction terms introduced by matching procedures in particle physics simulations. The event weights for Di-Higgs samples can be negative since negative weights are how we emulate the effects of destructive interference in particle physics simulations. Unfortunately, CatBoost and XGBoost cannot handle negative weights in any form, leaving only Scikit-learn Hist Gradient Boost Method. Thus, Scikit-learn Hist Gradient Boost Method was chosen as the preferred BDT tool and Table 7.6 lists the architecture and training features used for the BDT.

Table 7.5: Neural network architecture and training parameters.

Parameter	**Value**
Architecture	
Number of layers	3
Layer type	Dense
Nodes per layer	1024
Activation function	Swish
Training parameters	
Number of models (N_{models})	10
Learning rate	10^{-4}
Batch size	4096
Epochs	100
Early stopping patience	30
Early stopping mode	Minimize validation loss
Learning rate reduction factor	0.1
Learning rate reduction patience	10
Minimum learning rate	10^{-8}

Table 7.6: Hyperparameters used for the BDT training.

Parameter	**Value**
Learning rate	0.08
Maximum iterations	10000
Maximum depth	10
Maximum bins	255
Minimum samples per leaf	30
Early stopping	True
Iterations without improvement	20
Tolerance	10^{-4}
Other parameters	
Calibration	False
Use trigonometric conversion	True
Use particle features	False
Train without weights	False
Number of models (N_{models})	100

7.3.3 Graph Attention Transformer

Witnessing the success of Graph Attention-based Networks in the standard analysis. It was reasonable to benchmark its performance against SBI. Here, we have used a much simpler model that utilises object-level information without the risk

of overfitting. Unlike the histogram-based approach, overtraining is fatal for SBI, as it can introduce bias into our final parameter estimates. Figure 7.5 gives the architecture of the GNN used.

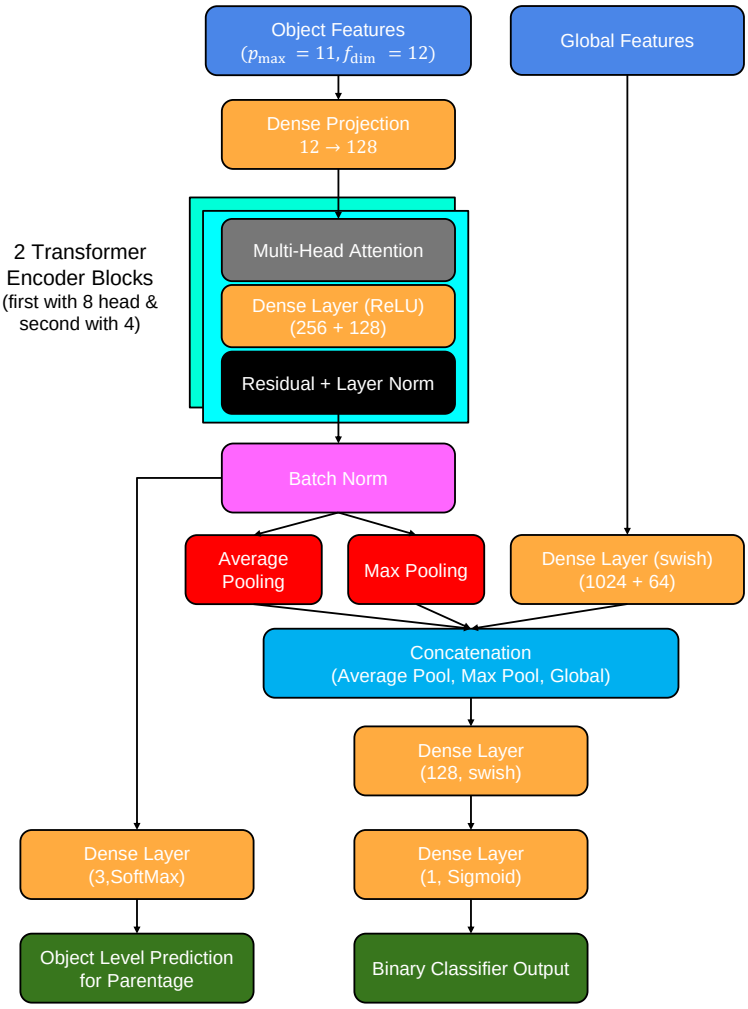


Figure 7.5: General architecture for Graph Attention Transformer used for SBI analysis.

The object input is passed through a single dense layer to project it to 128 dimensions, which is used as input by the transformer blocks. The 2 Transformer blocks consist of a multi-head attention layer (with 8 heads in the first and 4 in the second) that learns the relationships between objects, followed by a dense and normalising layer. A batch norm is placed at the end of the 2 attention blocks. As for Global (event-level) features, they pass through 2 dense layers before being merged with the outputs of the transformer blocks. We use 2 pooling methods to compress the transformer block's outputs: average pooling and max pooling. The

concatenation layers combine the results from the 2 different pooling layers and merge them with the information from the global features. The rest of the network is a simple MLP that leads to the final node, which predicts probabilities. As an auxiliary task, we also add a parentage classifier that takes information directly after the batch norm and uses it to compute object-level parentage. Table 7.7 gives the hyperparameters for the GAT classifier.

Table 7.7: Hyperparameters used for the GAT training.

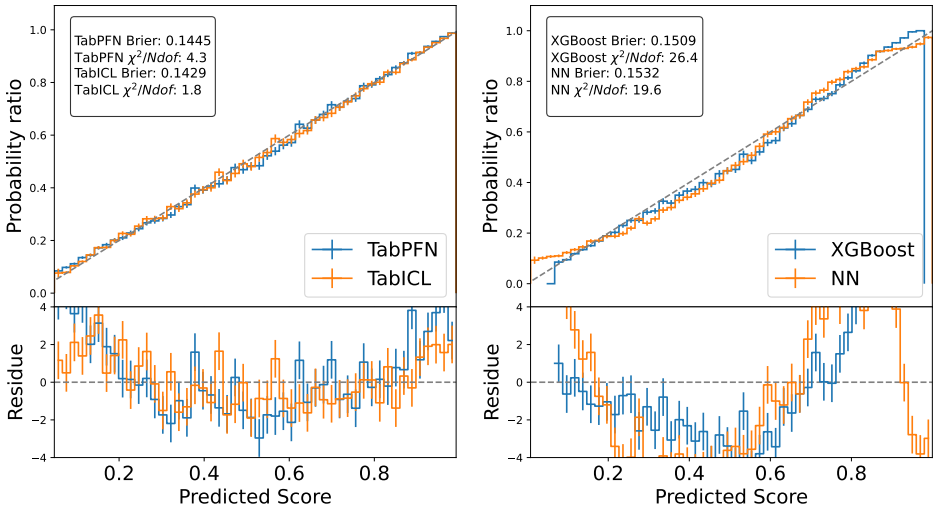
Parameter	**Value**
Number of models (N_{models})	20
Learning rate	10^{-4}
Batch size	2096
Epochs	100
Early stopping patience	30
Early stopping mode	Minimize validation loss
Learning rate reduction factor	0.1
Learning rate reduction patience	10
Minimum learning rate	10^{-8}
Other parameters	
Calibration	False
Use trigonometric conversion	True
Train without weights	False

Apart from the reduced complexity, there are 2 main differences between our architecture and the GNN used in the standard analysis. First, they combine global features into each object, whereas we treat them separately; second, they use global attention pooling rather than average or max pooling. Both these changes are worth testing in the SBI case, but weren't implemented due to time constraints.

7.3.4 Tabular Foundation models for HEP

In Appendix B, we discuss in detail the advantages of using Tabular Foundation over traditional ML algorithms such as DNNs or BDTs when we have a limited number of training events, as in the case with tops and Z + jets in our analysis. In Figure 7.6, we present the calibration curves for 2 Tabular Foundation models, TabICL and TabPFN (left), and compare them with the BDT and DNN (right) when trained on the same 2000 events and tested on 800k events.

The major drawback we identified in this study is that there is no straightforward way to incorporate event weights into either of the two tabular foundation models. This is even more critical when negative weights are involved, and hence, utilising the currently available TFMs for SBI is not possible. But if this issue is resolved or we adopt other foundational models in particle physics, we might be able to address concerns about the lack of training events.

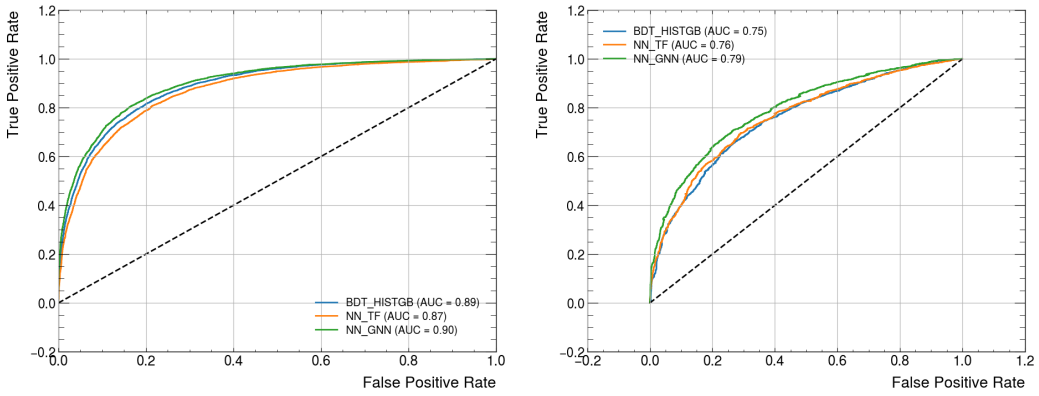


(a) (b)

Figure 7.6: Calibration curves for 2 Tabular Foundation models, TabICL and TabPFN (a), and compare them with the BDT and DNN (b) when trained with the same 2000 events and tested with 800k events.

7.3.5 Validation of classifiers

The primary comparison is the ROC curve, which provides insight into the information each classifier extracts from different samples. Figure 7.7 gives the ROC curve and the AUC score for the 3 different models that we are comparing for 3 different samples. These 3 are chosen for comparison as all the HH samples have very similar behaviour.

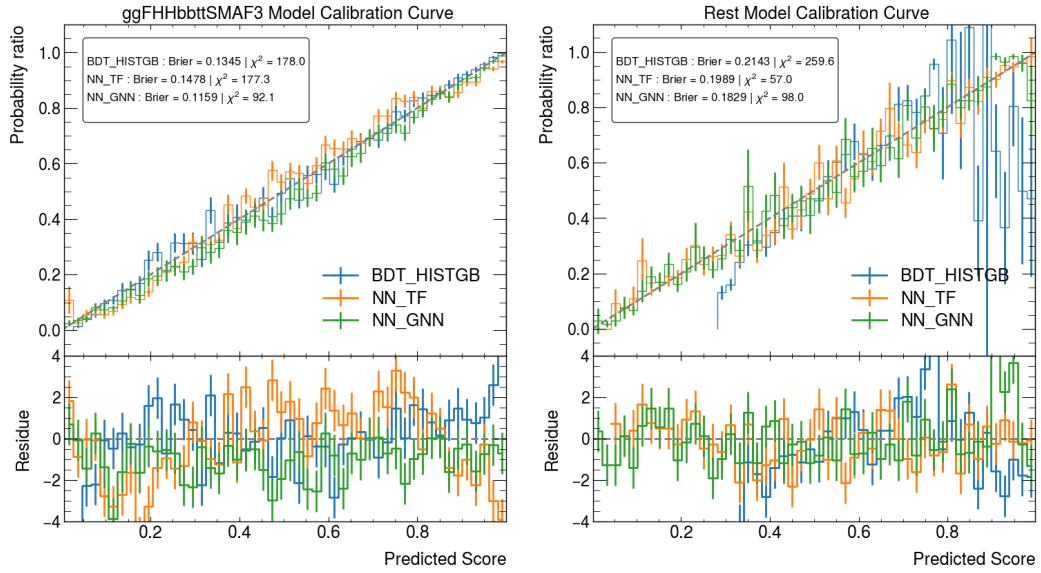


(a) (b)

Figure 7.7: Comparing receiver operating characteristic Curve using different Binary Classifiers Graph Attention Transformers (GAT) in green, Dense Neural Networks (DNN) in orange and Boosted Decision Tree (BDT) in blue, on different samples (a)SM $HH \to b\bar{b}\tau^+\tau^-$, (b) Rest ($Z$+ jets and tops)

From the ROC curves, we can see that GAT has an advantage. As we discussed

in subsection 5.1.2, Good classifiers need not be good density estimators. We have 2 main tools for validating density ratios: calibration curves and Reweighting plots. Figure 7.8 gives the calibration curves, along with the Brier score and χ^2 deviation.

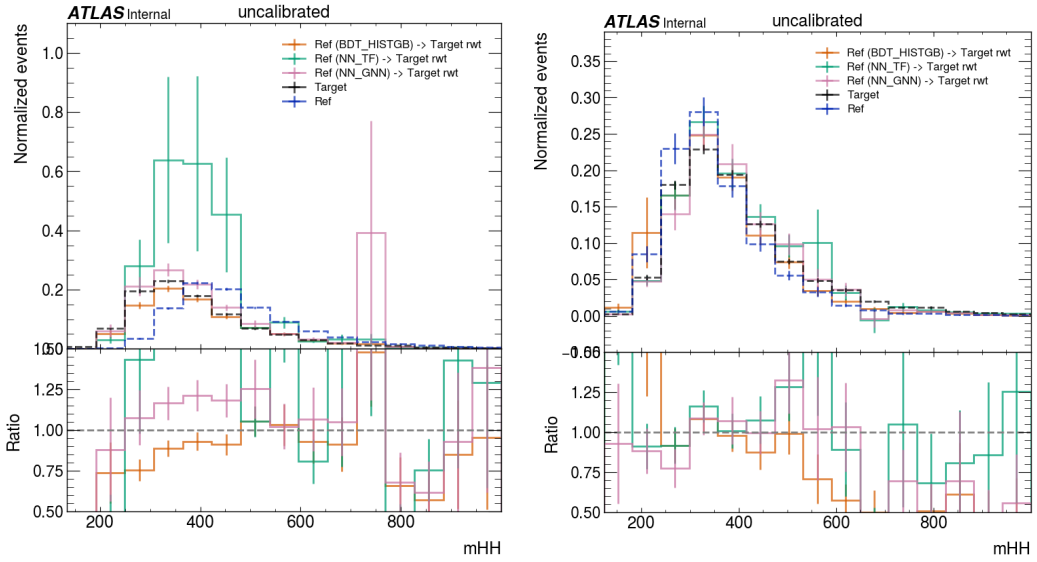


(a) (b)

Figure 7.8: Comparing calibration curves using different Binary Classifiers Graph Attention Transformers (GAT) in green, Dense Neural Networks (DNN) in orange and Boosted Decision Tree (BDT) in blue, on different samples (a)SM $HH \to b\bar{b}\tau^+\tau^-$, (b) Rest ($Z+$ jets and tops)

Figure 7.9 gives reweighting plots on the m_{HH} variable. Here, the reference distribution (single Higgs) is reweighted by the computed density ratio, which ideally should match the target distribution.

Though the calibration and ROC curves give us confidence in using GAT across all samples, the reweighting plots show a different story. It seems that GAT performs poorly in specific regions of the phase space that are under-represented. The only place where GAT has a clear advantage was with the Rest sample (Figure 7.9(c)). But since the yield or average weight in this sample is much higher than the rest, any mis-calibration will have a drastic effect. So ultimately, we opted for a mixed approach: BDT for all samples except the *Rest* sample (top and $Z+$jets), for which we used GAT.



(a) (b)

Figure 7.9: Comparing reweighting plots of Di-Higgs Invariant mass (m_{HH}) using different Binary Classifiers; BDT (orange), DNN (green) and GAT (pink), reference distribution (dashed blue) and target distribution in the dashed black line. On different samples (a)SM $HH \to b\bar{b}\tau^+\tau^-$, (b) Rest ($Z+$ jets and tops)

7.4 Results

Before we estimate κ_λ, we can first infer signal strength μ for SM ($\kappa_\lambda = 1$) and the 2 BSM case ($\kappa_\lambda = 0$ and $\kappa_\lambda = 5$). We performed a negative log-likelihood scan on μ as described in Equation 7.3 and inferred the signal strength μ. To demonstrate the statistical power of SBI relative to the Histogram method, we have also performed a binned analysis of the GNN discriminant developed for the standard analysis (subsection 6.5.1). In the previous section, we addressed the poor calibration of the Rest sample despite our best efforts.

Finally, Figure 7.10 shows the Comparison of likelihood curves for stat-only results, where only Di-Higgs, single higgs, and diboson samples are considered between the SBI method and the GNN Histogram, along with its $CL_{68\%}$ as a shaded region. From this, we demonstrate that the SBI method has greater statistical power than the histogram method for the set of backgrounds currently considered.

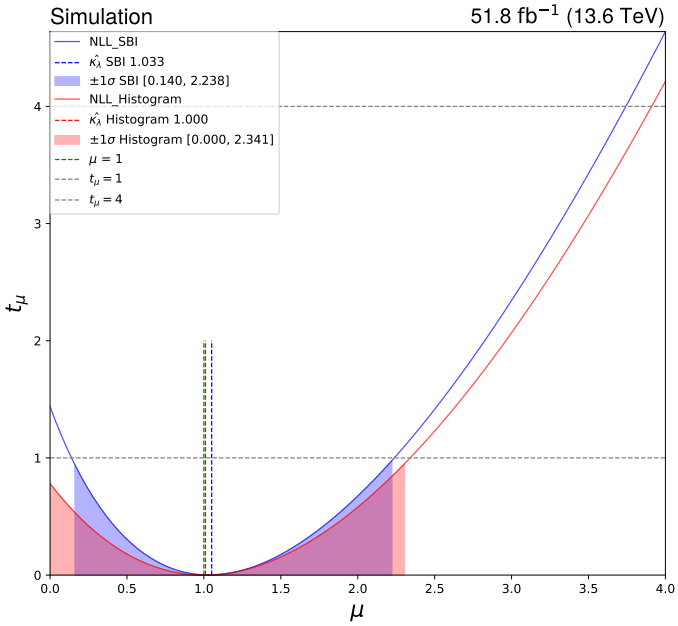


Figure 7.10: Comparison of likelihood curves which scan over signal strength μ for stat-only results, where **only** Di-Higgs, single higgs, and diboson samples are considered. Likelihood curve of the SBI method in blue and Likelihood curve of the Histogram method in red. The green dotted vertical shows the true μ, and the blue and red dotted vertical lines show the predicted μ for the SBI and Histogram methods, respectively.

We repeated this study for both the BSM cases ($\kappa_\lambda = 0$ and $\kappa_\lambda = 5$); Figure 7.11 shows the NLLs for $\kappa_\lambda = 0$ and $\kappa_\lambda = 5$ for both the SBI and Histogram methods. Apart from demonstrating statistical power, we can also utilise this as an additional study to understand network biases before combining them for κ_λ studies.

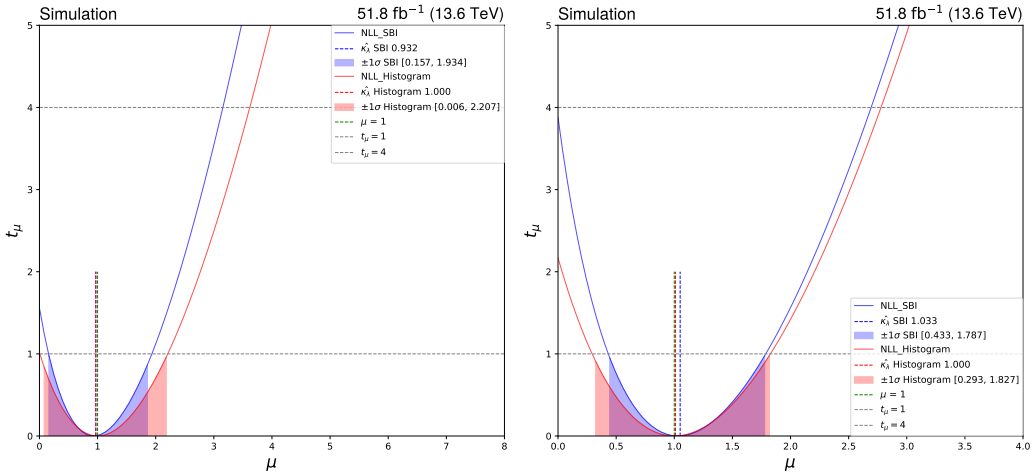


Figure 7.11: Comparison of Likelihood Curves which scan over signal strength μ for stat-only results for BSM samples (a) for $\kappa_\lambda = 5$ and (b) for $\kappa_\lambda = 0$. The legend follows the same structure as Figure 7.10.

The final step in the analysis is to construct the Negative log likelihood as described in Equation 7.4 and infer the κ_λ parameter. As in the μ case, we show results only with a single Higgs and Diboson background. But also studied the effects of different samples by utilising pseudo-universes in which we assume the specific backgrounds are absent. section D.2 gives the full set of plots for all the different combinations.

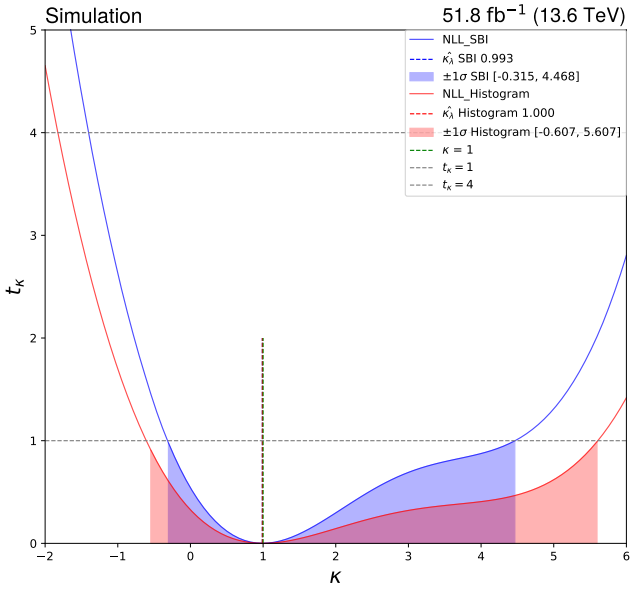


Figure 7.12: NLL scan over Higgs self coupling κ_λ for stat-only results with $\kappa_{\lambda true} = 1$, where **only** Di-Higgs, single higgs, and diboson samples are considered. The legend follows the same structure as Figure 7.10.

The expected limits in this restricted set of backgrounds from the SBI method are $\kappa_\lambda \in [-0.315, 4.468]$, and from the Histogram method are $\kappa_\lambda \in [-0.607, 5.607]$. This gives around 23% smaller $\text{CL}_{68\%}$ interval. This massive gain is thanks to the second minima rising, which could be attributed to the SBI's ability to learn more nD-shape information. With the help of the polynomial expansion from Equation 6.6, we can construct datasets of any κ_λ value and examine the performance of both models in different truth κ_λ values. Figure 7.13 gives the NLL plot comparison as we saw in Figure 7.12 but for $\kappa_\lambda \in \{0, 1, 2, 3, 4, 5\}$, We can put togather the $\text{CL}_{68\%}$ from each of theis scans and visualise them togather with Figure 7.14.

7.5 Discussion and Future Steps

From Figure 7.14, we see that SBI yields 20-30% smaller $\text{CI}_{68\%}$ limits than the histogram counterpart. This is, of course, true only in the regime where tops and Z+jets do not exist, which is unphysical. The major restriction when using Z or tops is that it has a very low number of MC samples relative to its yield, meaning any miscalibration will have a grave effect on the final NLL and by extension the κ_λ or μ estimation. Despite this major limitation, the study still shows that SBI methods have provided greater statistical power than the histogram method. This means we should generate more samples for this decay channel, both for background samples like Top and Z, and for Di-Higgs and single Higgs samples. Giving us possibilities to fully leverage the GAT's potential rather than relying on BDTs. The other approach might be to look for data-frugal methods, such as the Tabular foundational models we discussed in subsection 7.3.4 and in more detail in Appendix B. The last two pieces required are the estimation of systematic uncertainty and Neyman Construction. In section 6.6, we discussed systematic uncertainties in this analysis channel. Most of the major uncertainties pertain to the top and Z + jets samples. Once we have the stats-only Likelihood for Top and Z, we can pursue the procedure prescribed in subsection 5.3.1 to build density ratios for shape systematics and perform a fully unbinned estimation of κ_λ and its limits. The parameter of interest is subject to a boundary in the allowed parameter space, violating the regularity conditions required for Wilks' theorem to hold. Consequently, the standard asymptotic approximation cannot be reliably used to derive confidence intervals, and Neyman construction is employed to obtain statistically well-calibrated limits. Furthermore, similar to how pseudo-experiments were used to validate participants' models in the FAIR Universe challenge, Neyman construction provides a framework for assessing the coverage and reliability of the confidence intervals derived using SBI methods. The chapter presents the work done to train and validate classifiers, introduces innovative methods to test the pipeline even when we have few poorly calibrated networks, and finally constructs the pipeline for statistical-only results, which gives us a glimpse of the gains we can achieve when this analysis is completed.

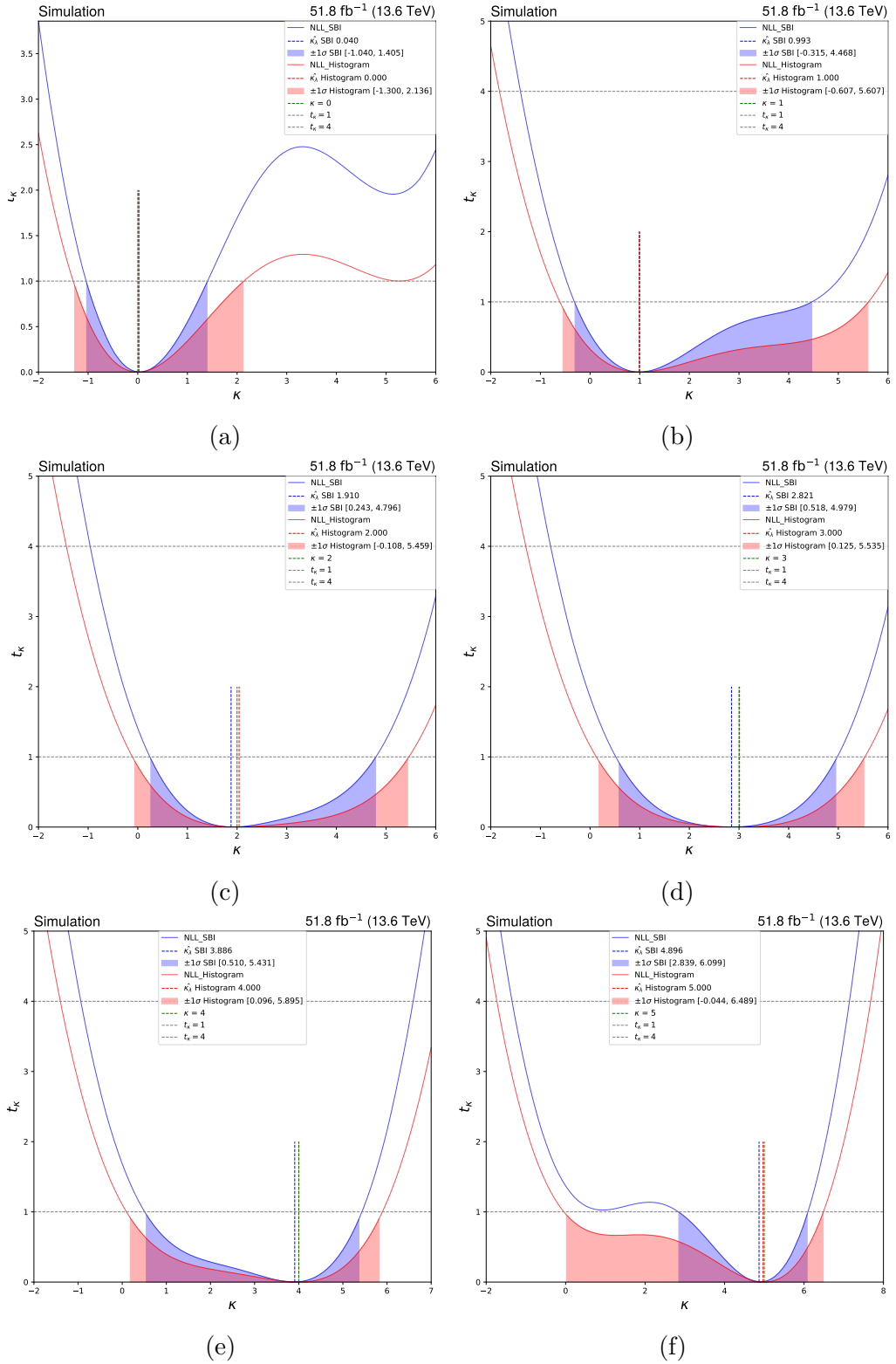


Figure 7.13: NLL scan over Higgs self coupling κ_λ for stat-only results, where **only** Di-Higgs, single higgs, and diboson samples are considered.

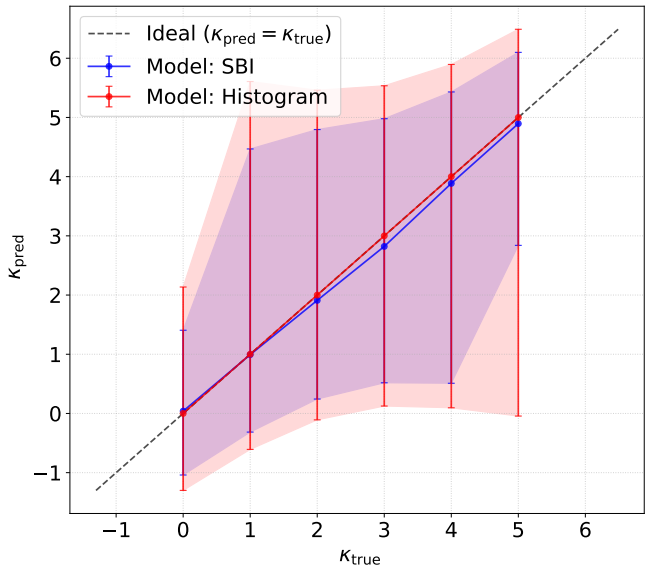


Figure 7.14: Precdicted κ_λ values for various truth values of κ_λ and their $\mathbf{CL}_{68\%}$ intervals, for SBI (Blue) and Histogram-based (Red)

Conclusion

Since the 2010s, the development of machine learning and artificial intelligence has accelerated rapidly. In the span of 15 years, we went from multi-variate analysis based on boosted decision trees to full-fledged transformers, and the application of such modern methods has been an integral part of particle physics research. The question this thesis seeks to answer is how we can properly benchmark these developments for particle physics research so that we can both compare models fairly and gain insight into new developments in ML for HEP. The first part of the thesis, the **FAIR Universe HiggsML Uncertainty Challenge**[4], directly addresses this. Building a benchmark that is complex enough to represent the physics we want these models to tackle, while also being tunable, maintainable, and controllable, so that a multitude of studies can be carried out with its help. Three key points to highlight. First is the **FAIR**ness. The data and its generation code are available for anyone to use, modify, and reproduce. This has enabled the FAIR Universe HiggsML dataset[12] to be used for studies different from the challenge, such as benchmarking New models like Tabular foundation models or teaching purposes. The second is the **Open-source** and **Open-endedness**: the challenge code is also available to the public, so it's possible to extend the benchmark for more systematic analysis, modify the existing ones or change the correlation between the systematics. And lastly, a reasonable **Long-term benchmark**. The two winning submissions are excellent solutions that could be applied to realistic HEP analysis. Still, as developments in physics and AI progress, we will need a long-term benchmark to quantify and rank models, and this challenge serves that role very well.

The second part of the thesis was also a benchmarking exercise, but of a different kind. In the FAIR Universe case, we were trying to create a FAIR, accessible version of a realistic HEP problem. In SBI for Di-Higgs, we were doing the opposite. How to build an exhaustive test bench to analyse, compare, and establish procedures to build Simulation-Based Inference for Di-Higgs analysis in the $b\bar{b}$ $\tau^+\tau^-$ channel in the full complexity of an ATLAS analysis. The one-word answer we reached was compartmentalisation. We treated each sample set separately. This is required because the quality of the density-ratio estimation varies across samples. The first part of using calibration curves and reweighting plots was already achieved and demonstrated in previous works. and the second is to compare performance against a selected set of samples or backgrounds. This helps us directly diagnose which networks or BDTs were causing bias. This also helps us assess the prospects of the analysis before we invest in additional montecarlo samples. With the help of these

pseudo universes, we were able to show that the current SBI method can have a gain of approximately 20% when top and Z + jets are ignored. This gives us prospects or a roadmap for the future.

Numerous steps are still missing in this analysis. One of the major constraints we have right now is an insufficient number of Monte Carlo samples, especially for the top and Z + jets samples. More studies into data-frugal Machine learning are required to tackle this issue. We addressed this by conducting studies on the Tabular foundation model using the FAIR Universe dataset, although the results are promising in the absence of event weights. Solving it for weighted datasets is critical before applying it in SBI. Once the statistics-only analysis is complete, we can start building density estimators to account for systematic shifts and move from statistics-only to a full stat+syst analysis. And the last step would be to perform the Neyman construction to obtain the 1 and 2 σ limits for the κ_λ parameter. As the saying goes, the journey of a thousand miles begins with a single step. I believe the thesis took initial steps toward building Simulation-Based inference for the Di-Higgs analysis in the $b\bar{b}$ $\tau^+\tau^-$ channel.

Bibliography

[1] ATLAS Collaboration, *Measurement of off-shell higgs boson production in the $h^* \to zz \to 4\ell$ decay channel using a neural simulation-based inference technique in 13tev pp collisions with the atlas detector*, Reports on Progress in Physics **88** (may, 2025) 057803. https://doi.org/10.1088/1361-6633/adcd9a.

[2] ATLAS Collaboration, *An implementation of neural simulation-based inference for parameter estimation in atlas*, Reports on Progress in Physics **88** (may, 2025) 067801. https://doi.org/10.1088/1361-6633/add370.

[3] ATLAS Collaboration, "Improved analysis of non-resonant higgs boson pair production in the $b\bar{b}\tau^+\tau^-$ final state with 196 fb^{-1} of data collected at $\sqrt{s} =$ 13 tev and 13.6 tev with the atlas detector." Url=https://arxiv.org/abs/2607.26879, 2026.

[4] W. Bhimji, R. Chakkappai, P.-W. Chang, Y.-T. Chou, S. Diefenbacher, J. Dudley, I. Elsharkawy, S. Farrell, A. Ghosh, C. Giordano, I. Guyon, C. Harris, Y. Hashizume, S.-C. Hsu, E. E. Khoda, C. Krause, A. Li, B. Nachman, D. Rousseau, R. Schöfbeck, M. Shooshtari, D. Schwarz, I. Ullah, D. Wang, and Y. Zhang, *FAIR universe higgsML uncertainty dataset and competition*, in *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2026. https://openreview.net/forum?id=F48NfH4NFE.

[5] Wikipedia contributors, "Standard model of elementary particles." https://en.wikipedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg, 2025. Wikimedia Commons.

[6] ATLAS Collaboration, *A detailed map of higgs boson interactions by the atlas experiment ten years after the discovery*, Nature **607** (July, 2022) 52–59. http://dx.doi.org/10.1038/s41586-022-04893-w.

[7] J. Alison *et al.*, *Higgs boson potential at colliders: Status and perspectives*, Rev. Phys. **5** (2020) 100045, arXiv:1910.00012 [hep-ph].

[8] ALICE Collaboration, *The alice experiment at the cern lhc*, Journal of Instrumentation **3** (Aug., 2008) S08002–S08002. http://dx.doi.org/10.1088/1748-0221/3/08/S08002.

[9] CMS Collaboration, *The cms experiment at the cern lhc*, Journal of Instrumentation **3** (Aug., 2008) S08004–S08004. http://dx.doi.org/10.1088/1748-0221/3/08/S08004.

[10] LHCb Collaboration, *The lhcb detector at the lhc*, Journal of Instrumentation **3** (Aug., 2008) S08005–S08005. http://dx.doi.org/10.1088/1748-0221/3/08/S08005.

[11] ATLAS Collaboration, *Improved analysis of non-resonant Higgs boson pair production in the $b\bar{b}\tau^+\tau^-$ final state with* 196 fb^{-1} *of data collected at* $\sqrt{s} = 13$ *TeV and* 13.6 *TeV with the ATLAS detector*, tech. rep., CERN, Geneva, 2026. https://cds.cern.ch/record/2964646.

[12] W. Bhimji, P. Calafiura, R. Chakkappai, P.-W. Chang, Y.-T. Chou, S. Diefenbacher, J. Dudley, S. Farrell, A. Ghosh, I. Guyon, C. Harris, S.-C. Hsu, E. E. Khoda, B. Nachman, P. Nugent, D. Rousseau, B. Thorne, I. Ullah, and Y. Zhang, "Fair universe dataset." https://github.com/FAIR-Universe/FAIR_Universe_dataset, 2025.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, *Imagenet: A large-scale hierarchical image database*, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009.

[14] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyon, B. Kégl, and D. Rousseau, *The Higgs boson machine learning challenge*, in *Proceedings of the NIPS 2014 Workshop on High-energy Physics and Machine Learning*, G. Cowan, C. Germain, I. Guyon, B. Kégl, and D. Rousseau, eds. PMLR, Montreal, Canada, 13 Dec, 2015. http://proceedings.mlr.press/v42/cowa14.html.

[15] "Higgs boson machine learning challenge." https://www.kaggle.com/c/higgs-boson, 2014.

[16] A. Duncan and M. Janssen, *Heisenberg's umdeutung paper*, in *Constructing Quantum Mechanics Volume Two: The Arch, 1923-1927*. Oxford University Press, 08, 2023. https://academic.oup.com/book/0/chapter/420743155/chapter-pdf/52065501/oso-9780198883906-chapter-4.pdf. https://doi.org/10.1093/oso/9780198883906.003.0004.

[17] E. Schrödinger, *Quantisierung als eigenwertproblem*, Annalen der Physik **384** (1926) 361, https://onlinelibrary.wiley.com/doi/pdf/10.1002/andp.19263840404. https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.19263840404.

[18] W. Pauli, *Über den einfluß der geschwindigkeitsabhängigkeit der elektronenmasse auf den zeemaneffekt*, Zeitschrift für Physik **31** (Feb., 1925) 373–385. http://dx.doi.org/10.1007/BF02980592.

[19] Bose, *Plancks gesetz und lichtquantenhypothese*, Zeitschrift für Physik **26** (Dec., 1924) 178–181. http://dx.doi.org/10.1007/BF01327326.

[20] E. Fermi, *Zur quantelung des idealen einatomigen gases*, Zeitschrift für Physik **36** (Nov., 1926) 902–912. http://dx.doi.org/10.1007/BF01400221.

[21] P. A. M. Dirac, *On the theory of quantum mechanics*, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character **112** (10, 1926) 661, https://royalsocietypublishing.org/rspa/article-pdf/112/762/661/24516/rspa.1926.0133.pdf. https://doi.org/10.1098/rspa.1926.0133.

[22] O. Klein, *Quantentheorie und fünfdimensionale relativitätstheorie*, Zeitschrift für Physik **37** (Dec., 1926) 895–906. http://dx.doi.org/10.1007/BF01397481.

[23] W. Gordon, *Der comptoneffekt nach der schrödingerischen theorie*, Zeitschrift für Physik **40** (Jan., 1926) 117–133. http://dx.doi.org/10.1007/BF01390840.

[24] P. A. M. Dirac, *The quantum theory of the electron*, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character **117** (02, 1928) 610, https://royalsocietypublishing.org/rspa/article-pdf/117/778/610/25048/rspa.1928.0023.pdf. https://doi.org/10.1098/rspa.1928.0023.

[25] E. Noether, *Invariante variationsprobleme*, Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse **1918** (1918) 235. http://eudml.org/doc/59024.

[26] A. Proca, *Sur la théorie ondulatoire des électrons positifs et négatifs*, Journal de Physique et le Radium **7** (1936) 347–353. http://dx.doi.org/10.1051/jphysrad:0193600708034700.

[27] C. N. Yang and R. L. Mills, *Conservation of isotopic spin and isotopic gauge invariance*, Phys. Rev. **96** (Oct, 1954) 191. https://link.aps.org/doi/10.1103/PhysRev.96.191.

[28] M. Gell-Mann, *A schematic model of baryons and mesons*, Physics Letters **8** (1964) 214. https://www.sciencedirect.com/science/article/pii/S0031916364920013.

[29] G. Zweig, *An SU_3 model for strong interaction symmetry and its breaking*, . https://cds.cern.ch/record/570209.

[30] F. Englert and R. Brout, *Broken symmetry and the mass of gauge vector mesons*, Phys. Rev. Lett. **13** (Aug, 1964) 321. https://link.aps.org/doi/10.1103/PhysRevLett.13.321.

[31] P. W. Higgs, *Broken symmetries and the masses of gauge bosons*, Phys. Rev. Lett. **13** (Oct, 1964) 508. https://link.aps.org/doi/10.1103/PhysRevLett.13.508.

[32] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, *Global conservation laws and massless particles*, Phys. Rev. Lett. **13** (Nov, 1964) 585. https://link.aps.org/doi/10.1103/PhysRevLett.13.585.

[33] H. YUKAWA, *On the interaction of elementary particles. i*, Proceedings of the Physico-Mathematical Society of Japan. 3rd Series **17** (1935) 48.

[34] ATLAS Collaboration, *Combined measurements of Higgs boson production and decay at $\sqrt{s} = 13$ TeV using up to 140 fb^{-1} of data collected by the ATLAS Experiment*, tech. rep., CERN, Geneva, 2025. https://cds.cern.ch/record/2937634. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2025-006.

[35] ATLAS Collaboration, *Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc*, Physics Letters B **716** (2012) 1. https://www.sciencedirect.com/science/article/pii/S037026931200857X.

[36] CMS Collaboration, *Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc*, Physics Letters B **716** (2012) 30. https://www.sciencedirect.com/science/article/pii/S0370269312008581.

[37] LHC Higgs Cross Section Working Group, S. Heinemeyer, C. Mariotti, G. Passarino, and R. e. a. Tanaka, *Handbook of LHC Higgs Cross Sections: 3. Higgs Properties*. CERN Yellow Reports: Monographs. CERN, 2013. https://cds.cern.ch/record/1559921. Comments: 404 pages, 139 figures, to be submitted to CERN Report. Working Group web page: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CrossSections.

[38] ATLAS Collaboration, *Combination of searches for higgs boson pair production in pp collisions at $\sqrt{s} = 13$ TeV with the atlas detector*, Phys. Rev. Lett. **133** (Sep, 2024) 101801. https://link.aps.org/doi/10.1103/PhysRevLett.133.101801.

[39] The ATLAS and The CMS Collaborations, *Combination of atlas and cms searches for higgs boson pair production at $\sqrt{s} = 13$ tev*, arXiv e-prints (2026) , arXiv:2602.23991 [hep-ex]. https://arxiv.org/abs/2602.23991.

[40] T. A. Collaboration, *The atlas experiment at the cern large hadron collider*, Journal of Instrumentation **3** (Aug., 2008) S08003–S08003. http://dx.doi.org/10.1088/1748-0221/3/08/S08003.

[41] O. S. Brüning, P. Collier, P. Lebrun, S. Myers, R. Ostojic, J. Poole, and P. Proudlock, *LHC Design Report*. CERN Yellow Reports: Monographs. CERN, Geneva, 2004. https://cds.cern.ch/record/782076.

[42] L. Evans and P. Bryant, *Lhc machine*, Journal of Instrumentation **3** (Aug., 2008) S08001–S08001. http://dx.doi.org/10.1088/1748-0221/3/08/S08001.

[43] M. Vretenar, J. Vollaire, R. Scrivens, C. Rossi, F. Roncarolo, S. Ramberger, U. Raich, B. Puccio, D. Nisbet, R. Mompo, S. Mathot, C. Martin, L. A. Lopez-Hernandez, A. Lombardi, J. Lettry, J. B. Lallement, I. Kozsar, J. Hansen, F. Gerigk, A. Funken, J. F. Fuchs, N. Dos Santos, M. Calviani, M. Buzio, O. Brunner, Y. Body, P. Baudrenghien, J. Bauche, and T. Zickler, *Linac4 design report*, vol. 6 of *CERN Yellow Reports: Monographs*. CERN, Geneva, 2020. https://cds.cern.ch/record/2736208.

[44] F. Landua, *The CERN accelerator complex layout in 2022*, . https://cds.cern.ch/record/2813716. General Photo.

[45] ATLAS Collaboration, "Public atlas luminosity results for run-3 of the lhc." https://twiki.cern.ch/twiki/bin/viewauth/AtlasPublic/LuminosityPublicResultsRun3, July, 2026. 2026-07-30.

[46] CERN, "Lhc/hl-lhc plan: Updated status of the lhc baseline programme." https://hilumilhc.web.cern.ch/sites/default/files/2025-01/HL-LHC_Plan_January2025.pdf, Jan., 2025. https://hilumilhc.web.cern.ch/sites/default/files/2025-01/HL-LHC_Plan_January2025.pdf. High-Luminosity Large Hadron Collider (HL-LHC) Project, accessed: 2026-07-31.

[47] ATLAS Collaboration, *The atlas experiment at the cern large hadron collider: a description of the detector configuration for run 3*, Journal of Instrumentation **19** (may, 2024) P05063. https://doi.org/10.1088/1748-0221/19/05/P05063.

[48] ATLAS Collaboration, *Technical Design Report for the ATLAS Inner Tracker Pixel Detector*, tech. rep., CERN, Geneva, 2017. https://cds.cern.ch/record/2285585.

[49] ATLAS Collaboration, *ATLAS detector and physics performance*. Technical design report. ATLAS. CERN, Geneva, 1999. https://cds.cern.ch/record/391176.

[50] CERN, "Cern yellow reports: Monographs, vol 4 (2017): High-luminosity large hadron collider (hl-lhc) technical design report v. 0.1." https://e-publishing.cern.ch/index.php/CYRM/issue/view/40, 2017. https://e-publishing.cern.ch/index.php/CYRM/issue/view/40.

[51] M. Casado, *A high-granularity timing detector for the atlas phase-ii upgrade*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **1032** (2022) 166628. https://www.sciencedirect.com/science/article/pii/S0168900222002005.

[52] A. Xiaocong, A. Corentin, C. Noemi, C. Angéla, E. Markus, E. Irina, F. Ralf, G. Louis-Guillaume, G. Rocky, G. Paul, G. Hadrien, G. H. M., G. Christian, H. Julia, H. Benjamin, K. Moritz, K. Fabian, K. Bernadette, K. Attila, L. Robert, L. Charles, M. Georgiana, M. Edward, N. Joana, O. J. D., R. David, S. Andreas, S. Bastian, T. Lauren, Y. Tomohiro, Y. Beomki, and Z. Jin, *A common tracking software project*, Computing and Software for Big Science **6** (2022) 8. https://doi.org/10.1007/s41781-021-00078-8.

[53] W. Adam, R. Fruhwirth, A. Strandlie, and T. Todorov, *Reconstruction of electrons with the Gaussian sum filter in the CMS tracker at LHC*, eConf **C0303241** (2003) TULT009, arXiv:physics/0306087.

[54] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyon, B. Kegl, and D. Rousseau, *Learning to discover: the higgs boson machine learning challenge - documentation*, . http://opendata.cern.ch/record/329.

[55] K. Cranmer, J. Pavez, and G. Louppe, *Approximating Likelihood Ratios with Calibrated Discriminative Classifiers*, arXiv e-prints (June, 2015) arXiv:1506.02169, arXiv:1506.02169 [stat.AP].

[56] P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, and D. Whiteson, *Parameterized neural networks for high-energy physics*, The European Physical Journal C **76** (Apr., 2016) . http://dx.doi.org/10.1140/epjc/s10052-016-4099-4.

[57] J. Brehmer, F. Kling, I. Espejo, and K. Cranmer, *MadMiner: Machine learning-based inference for particle physics*, Comput. Softw. Big Sci. **4** (2020) 3, arXiv:1907.10621 [hep-ph].

[58] J. Brehmer, G. Louppe, J. Pavez, and K. Cranmer, *Mining gold from implicit models to improve likelihood-free inference*, Proc. Nat. Acad. Sci. (2020) 201915980, arXiv:1805.12244 [stat.ML].

[59] J. Brehmer, K. Cranmer, G. Louppe, and J. Pavez, *Constraining Effective Field Theories with Machine Learning*, arXiv:1805.00013 [hep-ph].

[60] J. Brehmer, K. Cranmer, G. Louppe, and J. Pavez, *A Guide to Constraining Effective Field Theories with Machine Learning*, arXiv:1805.00020 [hep-ph].

[61] B. Nachman, *A guide for deploying Deep Learning in LHC searches: How to achieve optimality and account for uncertainty*, arXiv:1909.03081 [hep-ph].

[62] A. Ghosh, B. Nachman, and D. Whiteson, *Uncertainty-aware machine learning for high energy physics*, Phys. Rev. D **104** (2021) 056026, arXiv:2105.08742 [physics.data-an].

[63] F. Rozet and G. Louppe, *Arbitrary Marginal Neural Ratio Estimation for Simulation-based Inference*, in . 10, 2021. arXiv:2110.00449 [cs.LG].

[64] ATLAS Collaboration, *An implementation of neural simulation-based inference for parameter estimation in ATLAS*, arXiv:2412.01600 [hep-ex].

[65] A. Blance, M. Spannowsky, and P. Waite, *Adversarially-trained autoencoders for robust unsupervised new physics searches*, JHEP **10** (2019) 047, arXiv:1905.10384 [hep-ph].

[66] C. Englert, P. Galler, P. Harris, and M. Spannowsky, *Machine Learning Uncertainties with Adversarial Neural Networks*, Eur. Phys. J. **C79** (2019) 4, arXiv:1807.08763 [hep-ph].

[67] G. Louppe, M. Kagan, and K. Cranmer, *Learning to Pivot with Adversarial Networks*, arXiv:1611.01046 [stat.ME].

[68] J. Dolen, P. Harris, S. Marzani, S. Rappoccio, and N. Tran, *Thinking outside the ROCs: Designing Decorrelated Taggers (DDT) for jet substructure*, JHEP **05** (2016) 156, arXiv:1603.00027 [hep-ph].

[69] I. Moult, B. Nachman, and D. Neill, *Convolved Substructure: Analytically Decorrelating Jet Substructure Observables*, JHEP **05** (2018) 002, arXiv:1710.06859 [hep-ph].

[70] J. Stevens and M. Williams, *uBoost: A boosting method for producing uniform selection efficiencies from multivariate classifiers*, JINST **8** (2013) P12013, arXiv:1305.7248 [nucl-ex].

[71] C. Shimmin, P. Sadowski, P. Baldi, E. Weik, D. Whiteson, E. Goul, and A. Søgaard, *Decorrelated Jet Substructure Tagging using Adversarial Neural Networks*, arXiv:1703.03507 [hep-ex].

[72] L. Bradshaw, R. K. Mishra, A. Mitridate, and B. Ostdiek, *Mass Agnostic Jet Taggers*, arXiv:1908.08959 [hep-ph].

[73] ATLAS Collaboration, *Performance of mass-decorrelated jet substructure observables for hadronic two-body decay tagging in ATLAS*, ATL-PHYS-PUB-2018-014 (2018) . `http://cds.cern.ch/record/2630973`.

[74] G. Kasieczka and D. Shih, *DisCo Fever: Robust Networks Through Distance Correlation*, arXiv:2001.05310 [hep-ph].

[75] S. Wunsch, S. Jörger, R. Wolf, and G. Quast, *Reducing the dependence of the neural network function to systematic uncertainties in the input space*, Comput. Softw. Big Sci. **4** (2020) 5, arXiv:1907.11674 [physics.data-an].

[76] A. Rogozhnikov, A. Bukva, V. V. Gligorov, A. Ustyuzhanin, and M. Williams, *New approaches for boosting to uniformity*, JINST **10** (2015) T03002, arXiv:1410.4140 [hep-ex].

[77] CMS Collaboration, *A deep neural network to search for new long-lived particles decaying to jets*, Machine Learning: Science and Technology (2020) , 1912.12238.

[78] J. M. Clavijo, P. Glaysher, and J. M. Katzy, *Adversarial domain adaptation to reduce sample bias of a high energy physics classifier*, arXiv:2005.00568 [stat.ML].

[79] G. Kasieczka, B. Nachman, M. D. Schwartz, and D. Shih, *ABCDisCo: Automating the ABCD Method with Machine Learning*, arXiv:2007.14400 [hep-ph].

[80] O. Kitouni, B. Nachman, C. Weisser, and M. Williams, *Enhancing searches for resonances with machine learning and moment decomposition*, arXiv:2010.09745 [hep-ph].

[81] V. Estrade, C. Germain, I. Guyon, and D. Rousseau, *Systematic aware learning - A case study in High Energy Physics*, EPJ Web Conf. **214** (2019) 06024.

[82] A. Ghosh and B. Nachman, *A cautionary tale of decorrelating theory uncertainties*, Eur. Phys. J. C **82** (2022) 46, arXiv:2109.08159 [hep-ph].

[83] S. Wunsch, S. Jörger, R. Wolf, and G. Quast, *Optimal Statistical Inference in the Presence of Systematic Uncertainties Using Neural Network Optimization Based on Binned Poisson Likelihoods with Nuisance Parameters*, Comput. Softw. Big Sci. **5** (2021) 4, arXiv:2003.07186 [physics.data-an].

[84] CMS Collaboration, *Development of systematic uncertainty-aware neural network trainings for binned-likelihood analyses at the LHC*, arXiv:2502.13047 [hep-ex].

[85] L. Heinrich, *Learning Optimal Test Statistics in the Presence of Nuisance Parameters*, arXiv:2203.13079 [stat.ME].

[86] A. Elwood, D. Krücker, and M. Shchedrolosiev, *Direct optimization of the discovery significance in machine learning for new physics searches in particle colliders*, J. Phys. Conf. Ser. **1525** (2020) 012110.

[87] L.-G. Xia, *QBDT, a new boosting decision tree method with systematical uncertainties into training for High Energy Physics*, Nucl. Instrum. Meth. **A930** (2019) 15, arXiv:1810.08387 [physics.data-an].

[88] P. De Castro and T. Dorigo, *INFERNO: Inference-Aware Neural Optimisation*, Comput. Phys. Commun. **244** (2019) 170, arXiv:1806.04743 [stat.ML].

[89] T. Charnock, G. Lavaux, and B. D. Wandelt, *Automatic physical inference with information maximizing neural networks*, Physical Review D **97** (Apr, 2018) . http://dx.doi.org/10.1103/PhysRevD.97.083004.

[90] J. Alsing and B. Wandelt, *Nuisance hardened data compression for fast likelihood-free inference*, Mon. Not. Roy. Astron. Soc. **488** (2019) 5093, arXiv:1903.01473 [astro-ph.CO].

[91] N. Simpson and L. Heinrich, *neos: End-to-End-Optimised Summary Statistics for High Energy Physics*, J. Phys. Conf. Ser. **2438** (2023) 012105, arXiv:2203.05570 [physics.data-an].

[92] P. Feichtinger *et al.*, *Punzi-loss: a non-differentiable metric approximation for sensitivity optimisation in the search for new particles*, Eur. Phys. J. C **82** (2022) 121, arXiv:2110.00810 [hep-ex].

[93] L. Layer, T. Dorigo, and G. Strong, *Application of Inferno to a Top Pair Cross Section Measurement with CMS Open Data*, arXiv:2301.10358 [hep-ex].

[94] S. Bollweg, M. Haußmann, G. Kasieczka, M. Luchmann, T. Plehn, and J. Thompson, *Deep-Learning Jets with Uncertainties and More*, SciPost Phys. **8** (2020) 006, arXiv:1904.10004 [hep-ph].

[95] J. Y. Araz and M. Spannowsky, *Combine and Conquer: Event Reconstruction with Bayesian Ensemble Neural Networks*, JHEP **04** (2021) 296, arXiv:2102.01078 [hep-ph].

[96] M. Bellagente, M. Haußmann, M. Luchmann, and T. Plehn, *Understanding Event-Generation Networks via Uncertainties*, arXiv:2104.04543 [hep-ph].

[97] T. Dorigo and P. De Castro Manzano, *Dealing with Nuisance Parameters using Machine Learning in High Energy Physics: a Review*, arXiv:2007.09121 [stat.ML].

[98] T. Y. Chen, B. Dey, A. Ghosh, M. Kagan, B. Nord, and N. Ramachandra, *Interpretable Uncertainty Quantification in AI for HEP*, in *Snowmass 2021*. 8, 2022. arXiv:2208.03284 [hep-ex].

[99] W. Bhimji, P. Calafiura, R. Chakkappai, P.-W. Chang, Y.-T. Chou, S. Diefenbacher, J. Dudley, S. Farrell, A. Ghosh, I. Guyon, C. Harris, S.-C.

Hsu, K. Elham E, B. Nachman, P. Nugent, D. Rousseau, B. Thorne, I. Ullah, and Y. Zhang, "Fair universe - higgsml uncertainty challenge public dataset." https://zenodo.org/doi/10.5281/zenodo.15131565, 2025.

[100] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, *An introduction to PYTHIA 8.2*, Comput. Phys. Commun. **191** (2015) 159, arXiv:1410.3012 [hep-ph].

[101] DELPHES 3, J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi, *DELPHES 3, A modular framework for fast simulation of a generic collider experiment*, JHEP **02** (2014) 057, arXiv:1307.6346 [hep-ex].

[102] ATLAS Collaboration, *Evidence for the higgs-boson yukawa coupling to tau leptons with the atlas detector*, Journal of High Energy Physics **2015** (Apr., 2015) . http://dx.doi.org/10.1007/JHEP04(2015)117.

[103] D. Vohra, *Apache Parquet*, pp. 325–335. Apress, Berkeley, CA, 2016. https://doi.org/10.1007/978-1-4842-2199-0_8.

[104] "Nersc: Perlmutter." https://www.nersc.gov/systems/perlmutter/, 2022.

[105] B. Nachman and T. Rudelius, *Evidence for conservatism in LHC SUSY searches*, Eur. Phys. J. Plus **127** (2012) 157, arXiv:1209.3522 [stat.AP].

[106] B. Nachman and T. Rudelius, *A Meta-analysis of the 8 TeV ATLAS and CMS SUSY Searches*, JHEP **02** (2015) 004, arXiv:1410.2270 [hep-ph].

[107] T. Chen and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM, New York, NY, USA, 2016. http://doi.acm.org/10.1145/2939672.2939785.

[108] ROOT, K. Cranmer, G. Lewis, L. Moneta, A. Shibata, and W. Verkerke, *HistFactory: A tool for creating statistical models for use with RooFit and RooStats*, tech. rep., New York U., New York, 2012. https://cds.cern.ch/record/1456844.

[109] L. Benato, C. Giordano, C. Krause, A. Li, R. Schöfbeck, D. Schwarz, M. Shooshtari, and D. Wang, *Unbinned inclusive cross-section measurements with machine-learned systematic uncertainties*, Phys. Rev. D **112** (Sep, 2025) 052006. https://link.aps.org/doi/10.1103/zwzt-1rrw.

[110] I. Elsharkawy and Y. Kahn, *Contrastive Normalizing Flows for Uncertainty-Aware Parameter Estimation*, arXiv:2505.08709 [physics.data-an].

[111] S. Rizvi, M. Pettee, and B. Nachman, *Learning likelihood ratios with neural network classifiers*, Journal of High Energy Physics **2024** (Feb., 2024) . http://dx.doi.org/10.1007/JHEP02(2024)136.

[112] R. Schöfbeck, *Refinable modeling for unbinned SMEFT analyses*, Mach. Learn. Sci. Tech. **6** (2025) 015007, arXiv:2406.19076 [hep-ph].

[113] IRIS-HEP Collaboration, "Nsbi lhc toolkit." Github, 2026. https://github.com/iris-hep/nsbi-lhc-toolkit. GitHub repository, accessed 2026-06-17.

[114] ATLAS Collaboration, *Jet reconstruction and performance using particle flow with the atlas detector*, The European Physical Journal C **77** (July, 2017) . http://dx.doi.org/10.1140/epjc/s10052-017-5031-2.

[115] M. Cacciari, G. P. Salam, and G. Soyez, *The anti-k_tjet clustering algorithm*, Journal of High Energy Physics **2008** (Apr., 2008) 063–063. http://dx.doi.org/10.1088/1126-6708/2008/04/063.

[116] M. Cacciari, G. P. Salam, and G. Soyez, *Fastjet user manual: (for version 3.0.2)*, The European Physical Journal C **72** (Mar., 2012) . http://dx.doi.org/10.1140/epjc/s10052-012-1896-2.

[117] ATLAS Collaboration, *Jet energy scale measurements and their systematic uncertainties in proton-proton collisions at $\sqrt{s} = 13$ tev with the atlas detector*, Physical Review D **96** (Oct., 2017) . http://dx.doi.org/10.1103/PhysRevD.96.072002.

[118] ATLAS Collaboration, *Forward jet vertex tagging using the particle flow algorithm*, tech. rep., CERN, Geneva, 2019. https://cds.cern.ch/record/2683100. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2019-026.

[119] ATLAS Collaboration, *Forward Jet Vertex Tagging: A new technique for the identification and rejection of forward pileup jets*, tech. rep., CERN, Geneva, 2015. https://cds.cern.ch/record/2042098. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2015-034.

[120] ATLAS Collaboration, *Tagging and suppression of pileup jets with the ATLAS detector*, tech. rep., CERN, Geneva, 2014. https://cds.cern.ch/record/1700870. All figures including auxiliary figures are available at

https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2014-018.

[121] ATLAS Collaboration, *Transforming jet flavour tagging at ATLAS*, Nature Commun. **17** (2026) 541, arXiv:2505.19689 [hep-ex].

[122] ATLAS Collaboration, *ATLAS b-jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s} = 13$ TeV*, Eur. Phys. J. C **79** (2019) 970, arXiv:1907.05120 [hep-ex].

[123] ATLAS Collaboration, *Calibration of the light-flavour jet mistagging efficiency of the b-tagging algorithms with Z+jets events using 139 fb^{-1} of ATLAS proton–proton collision data at $\sqrt{s} = 13$ TeV*, Eur. Phys. J. C **83** (2023) 728, arXiv:2301.06319 [hep-ex].

[124] ATLAS Collaboration, *Measurement of the c-jet mistagging efficiency in $t\bar{t}$ events using pp collision data at $\sqrt{s} = 13$ TeV collected with the ATLAS detector*, Eur. Phys. J. C **82** (2022) 95, arXiv:2109.10627 [hep-ex].

[125] ATLAS Collaboration, *Reconstruction, Identification, and Calibration of hadronically decaying tau leptons with the ATLAS detector for the LHC Run 3 and reprocessed Run 2 data*, ATL-PHYS-PUB-2022-044, 2022. https://cds.cern.ch/record/2827111.

[126] ATLAS Collaboration, *Identification of hadronic tau lepton decays using neural networks in the ATLAS experiment*, ATL-PHYS-PUB-2019-033, 2019. https://cds.cern.ch/record/2688062.

[127] V. Nair and G. E. Hinton, *Rectified linear units improve restricted Boltzmann machines*, in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, J. Fürnkranz and T. Joachims, eds. Omnipress, Haifa, Israel, June, 2010. http://www.icml2010.org/papers/432.pdf.

[128] ATLAS Collaboration, *Preliminary analysis of the luminosity calibration of the ATLAS* 13.6 *TeV data recorded in 2022*, ATL-DAPR-PUB-2023-001, 2023. https://cds.cern.ch/record/2853525.

[129] ATLAS Collaboration, *Preliminary analysis of the luminosity calibration for the ATLAS* 13.6 *TeV data recorded in 2023*, ATL-DAPR-PUB-2024-001, 2024. https://cds.cern.ch/record/2900949.

[130] H. Läuter, *Silverman, b. w.: Density estimation for statistics and data analysis. chapman & hall, london – new york 1986, 175 pp., £12.—*, Biometrical Journal **30** (1988) 876, https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.4710300745. https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.4710300745.

[131] A. V. Dorogush, V. Ershov, and A. Gulin, *Catboost: gradient boosting with categorical features support*, CoRR **abs/1810.11363** (2018) , 1810.11363. http://arxiv.org/abs/1810.11363.

[132] W. Bhimji, P. Calafiura, R. Chakkappai, P.-W. Chang, Y.-T. Chou, S. Diefenbacher, J. Dudley, S. Farrell, A. Ghosh, I. Guyon, C. Harris, S.-C. Hsu, E. E. Khoda, B. Nachman, P. Nugent, D. Rousseau, B. Thorne, I. Ullah, and Y. Zhang, "Fair universe generation." https://github.com/FAIR-Universe/genHEPdata, 2025.

[133] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyon, B. Kegl, and D. Rousseau, *Learning to discover: the Higgs boson machine learning challenge - Documentation*, . http://opendata.cern.ch/record/329.

[134] P. Baldi, P. Sadowski, and D. Whiteson, *Searching for Exotic Particles in High-Energy Physics with Deep Learning*, Nature Commun. **5** (2014) 4308, arXiv:1402.4735 [hep-ph].

[135] P. Baldi, P. Sadowski, and D. Whiteson, *Enhanced Higgs Boson to tau+ tau- Search with Deep Learning*, Physical Review Letters **114** (Mar., 2015) . http://dx.doi.org/10.1103/PhysRevLett.114.111801.

[136] J. Ma, V. Thomas, R. Hosseinzadeh, A. Labach, J. C. Cresswell, K. Golestan, G. Yu, A. L. Caterini, and M. Volkovs, *TabDPT: Scaling tabular foundation models on real data*, in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 2026. https://openreview.net/forum?id=pIZxEOZCId.

[137] Y. Gorishniy, A. Kotelnikov, and A. Babenko, *Tabm: Advancing tabular deep learning with parameter-efficient ensembling*, in *The Thirteenth International Conference on Learning Representations*. 2025. https://openreview.net/forum?id=Sd4wYYOhmY.

[138] J. QU, D. Holzmüller, G. Varoquaux, and M. L. Morvan, *TabICL: A tabular foundation model for in-context learning on large data*, in *Forty-second International Conference on Machine Learning*. 2025. https://openreview.net/forum?id=0VvD1PmNzM.

[139] N. Hollmann, S. Müller, L. Purucker, A. Krishnakumar, M. Körfer, S. B. Hoo, R. T. Schirrmeister, and F. Hutter, *Accurate predictions on small data with a tabular foundation model*, Nature **637** (Jan., 2025) 319–326. http://dx.doi.org/10.1038/s41586-024-08328-6.

Appendix A
FAIR Universe Dataset

This chapter gives a detailed explanation of the FAIR Universe HiggsML Uncertainty challenge. The Public dataset is published in Zenodo[12] and the details described here were part of the dataset and benchmark paper[4]

A.1 Crossection tables for FAIR Universe Dataset

The following are the cross-sections used to produce the FAIR Universe dataset. The files and production codes are available in GitHub[132].

Table A.1: Crossection tables for FAIR Universe Dataset

ID	Process	X Section (pb)
	Cross sections for Higgs production	
901	$f\bar{f} \to H(SM)$	57.65
902	$gg \to H(SM)$	1,634.0
903	$\gamma\gamma \to H(SM)$	0.0493
904	$f\bar{f} \to HZ^0(SM)$	38.17
905	$f\bar{f} \to HW^{\pm}(SM)$	69.79
906	$ff' \to Hff'(Z^0Z^0 fusion)(SM)$	60.54
907	$f_1f_2 \to Hf_3f_4(W+W-fusion)(SM)$	162.7
908	$gg \to Ht\bar{t}(SM)$	16.09
909	$q\bar{q} \to Ht\bar{t}(SM)$	6.577
	Cross sections for Z + jets production	
221	$f\bar{f} \to \gamma^*/Z^0$	6,563,000.0
231	$f\bar{f} \to \gamma^*/Z^0\gamma^*/Z^0$	7,192.0
243	$f\bar{f} \to \gamma^*/Z^0\gamma$	348,200.0
244	$f\gamma \to \gamma^*/Z^0 f$	11,430.0
	Cross sections for $t\bar{t}$ production	
601	$gg \to t\bar{t}$	470,400.0
602	$q\bar{q} \to t\bar{t}$	68,040.0
603	$qq \to tq(t-channel W^{\pm})$	151,100.0
604	$f\bar{f} \to t\bar{t}(s-channel \gamma^*/Z^0)$	363.0
605	$f\bar{f} \to t\bar{q}(s-channel W^{\pm})$	7,160.0
606	$\gamma\gamma \to t\bar{t}$	2.784

A.2 List of features for FAIR Universe Dataset

In this section, we present the list of features that describe the events, as provided in Zenodo[12].

Prefix-less variables `Weight`, `Label`,`DetailedLabel`, have a special role and should not be used as regular features for the model[1]:

`Weight` The event weight w_i. Not to be used as a feature. Not available in the test sample.

`Label` The event label (integer) y_i 1 for signal, 0 for background . Not to be used as a feature. Not available in the test sample.

`DetailedLabel` The event detailed label (string) "htautau" for signal (when Label==1), "ztautau", "ttbar" and "diboson" for the three background categories (when Label==0). Not to be used as a feature. Not available in the test sample. This feature is used to implement some systematic biases; see section 4.3. It could be used to train a multi-category classifier.

The variables prefixed with `PRI` (for PRImitives are "raw" quantities about the bunch collision as measured by the detector, essentially parameters of the momenta of particles.

In addition:

- Features are float unless specified otherwise.
- All azimuthal ϕ angles are in radian in the $]-\pi, +\pi]$ range.
- Energy, mass, and momentum are all in GeV
- All other features are unitless.
- Features are indicated as "undefined" when it can happen that they are meaningless or cannot be computed; in this case, their value is -25, which is outside the normal range of all variables.
- The mass of particles has not been provided, as it can safely be neglected for the challenge.

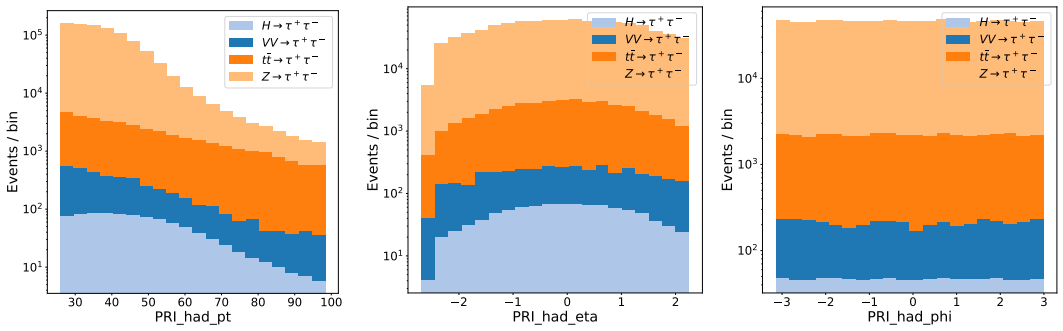
`PRI_had_pt` The transverse momentum $\sqrt{p_x^2 + p_y^2}$ of the hadronic tau.

`PRI_had_eta` The pseudorapidity η of the hadronic tau.

`PRI_had_phi` The azimuth angle ϕ of the hadronic tau.

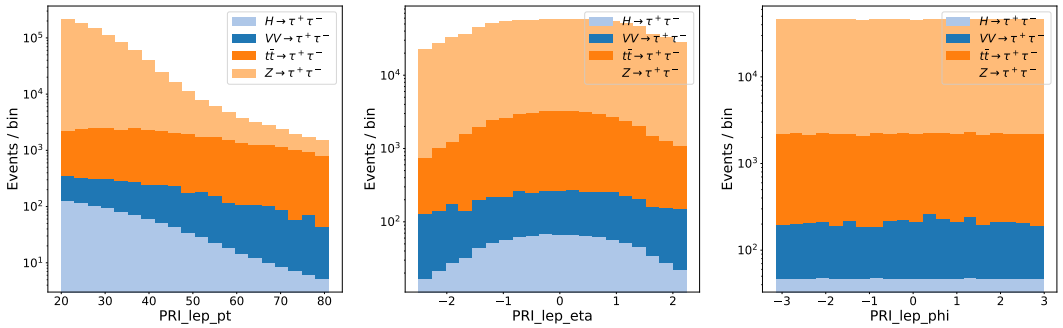
`PRI_lep_pt` The transverse momentum $\sqrt{p_x^2 + p_y^2}$ of the lepton (electron or muon).

`PRI_lep_eta` The pseudorapidity η of the lepton.



(a) (b) (c)

Figure A.1: Distributions of the primary hadronic τ kinematic variables as stacked histogram, (a) transverse momentum $p_{\mathrm{T}}^{\tau_{\mathrm{had}}}$, (b) pseudorapidity $\eta^{\tau_{\mathrm{had}}}$, and (c) azimuthal angle $\phi^{\tau_{\mathrm{had}}}$.



(a) (b) (c)

Figure A.2: Distributions of the primary lepton kinematic variables as stacked histograms: (a) transverse momentum, p_{T}^{ℓ}; (b) pseudorapidity, η^{ℓ}; and (c) azimuthal angle, ϕ^{ℓ}.

`PRI_lep_phi` The azimuth angle ϕ of the lepton.

`PRI_met` The missing transverse energy $\vec{E}_{\mathrm{T}}^{\mathrm{miss}}$.

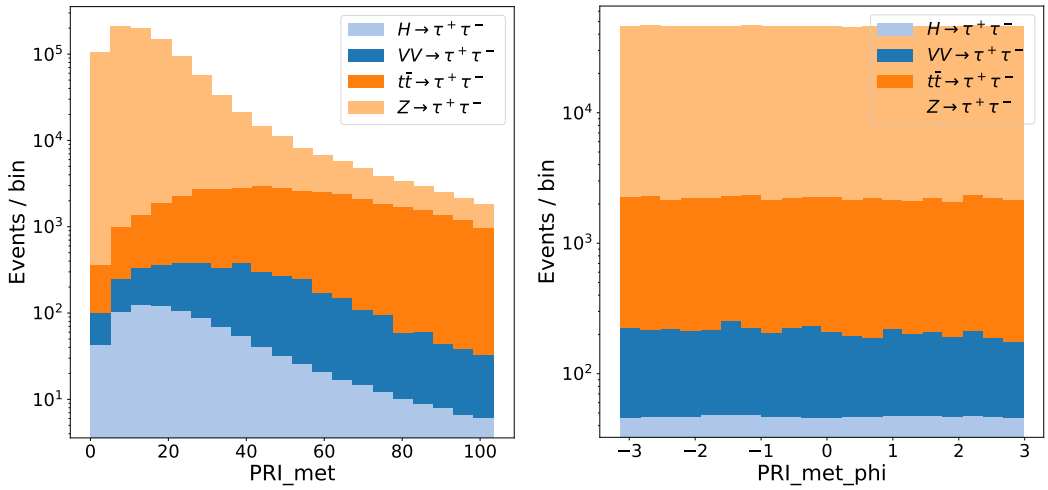
`PRI_met_phi` The azimuth angle ϕ of the missing transverse energy vector.

`PRI_jet_all_pt` The scalar sum of the transverse momentum of all the jets of the events (not limited to the first 2).

`PRI_n_jets` The number of jets.

`PRI_jet_leading_pt` The transverse momentum $\sqrt{p_x^2 + p_y^2}$ of the leading jet, that is the jet with the largest transverse momentum (undefined if `PRI_jet_num` $= 0$).

[1] In the starting kit, they are split away in separate numpy arrays while the regular features are stored in a Dataframe



(a) (b)

Figure A.3: Distributions of the missing transverse momentum variables as stacked histograms: (a) missing transverse momentum, $p_{\mathrm{T}}^{\mathrm{miss}}$; and (b) the azimuthal angle of the missing transverse momentum, ϕ^{miss}.

`PRI_jet_leading_eta` The pseudorapidity η of the leading jet (undefined if `PRI_jet_num` = 0).

`PRI_jet_leading_phi` The azimuth angle ϕ of the leading jet (undefined if `PRI_jet_num` = 0).

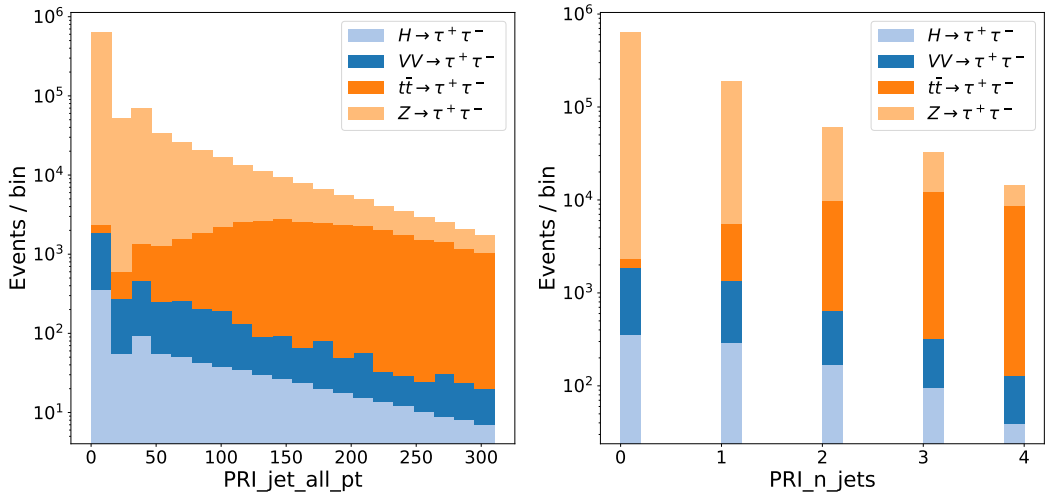
`PRI_jet_subleading_pt` The transverse momentum $\sqrt{p_x^2 + p_y^2}$ of the sub leading jet, that is, the jet with the second largest transverse momentum (undefined if `PRI_jet_num` ≤ 1).

`PRI_jet_subleading_eta` The pseudorapidity η of the subleading jet (undefined if `PRI_jet_num` $\leq$ 1).

`PRI_jet_subleading_phi` The azimuth angle ϕ of the subleading jet (undefined if `PRI_jet_num` $\leq$ 1).

Variables prefixed with `DER` (for Derived) are quantities computed from the primitive features on the fly from PRImary features (including possible systematics shifts)[2]. These quantities were selected by the physicists of ATLAS in the reference document [102] either to select regions of interest or as features for the Boosted Decision Trees used in this analysis in order to enhance the separation of signal Higgs boson events from background events. Derived features were already present in the

[2]The code to compute DERived features from PRImitive features can be seen at https://github.com/FAIR-Universe/FAIR_Universe_dataset/blob/main/hep_challenge/derived_quantities.py



(a) (b)

Figure A.4: Distributions of the primary jet-related variables as stacked histograms: (a) the scalar sum of the transverse momenta of all reconstructed jets, Scalar Sum($p_{\mathrm{T}}^{\mathrm{jet}}$); and (b) the number of reconstructed jets, N_{jets}.

HiggsML dataset [133][3]). The derived features correspond to feature engineering; an ideal model trained on infinite statistics should not need them. This distinction between primary and derived features (or "low-level" and "high-level," or "raw variables" and "human-assisted variables") is rather standard in the AI for HEP literature; see, for example, [134, 135]. There is no guarantee that all derived features are useful for this challenge (they could even be detrimental in the context of systematics). The challenge participant is free to keep these derived features, remove them altogether, keep a few, or do more feature engineering.

The invariant mass of two particles is the invariant mass of their 4-momentum sum, that is (still neglecting the mass of the two particles),

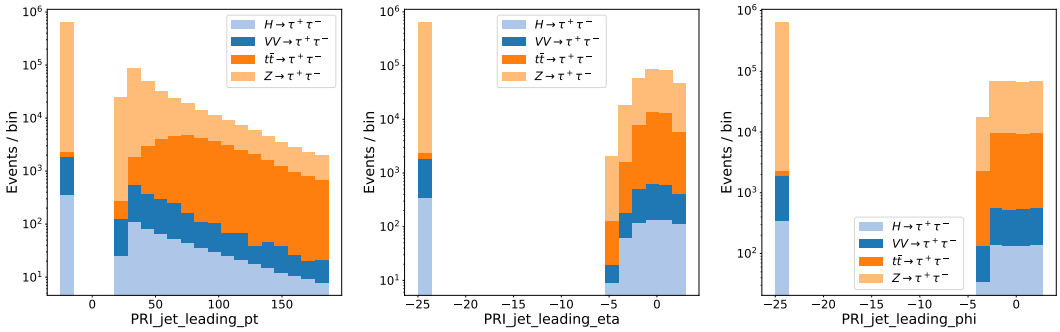
$$m_{\mathrm{inv}}(\vec{a},\vec{b}) = \sqrt{\left(\sqrt{a_x^2+a_y^2+a_z^2}+\sqrt{b_x^2+b_y^2+b_z^2}\right)^2-(a_x+b_x)^2-(a_y+b_y)^2-(a_z+b_z)^2}. \tag{A.1}$$

`DER_mass_vis` The invariant mass (Equation A.1) of the hadronic tau and the lepton.

`DER_mass_jet_jet` The invariant mass (Equation A.1) of the two jets (undefined if `PRI_jet_num` ≤ 1).

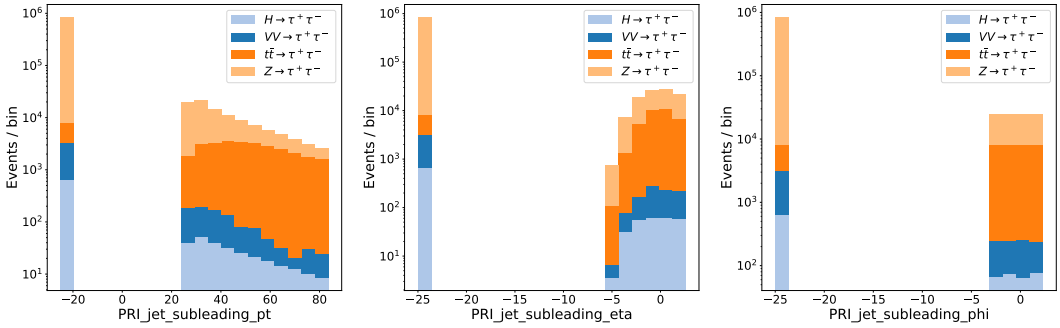
The transverse mass of two particles is the invariant mass of the vector sum, but this time The third component is set to zero, indicating that only the projection onto the transverse plane is considered. That is (still neglecting

[3]The notable exception of `DER_mass_MMC`, which was in the HiggsML dataset but is deliberately absent from the Fair-Universe dataset because it was the result of a complex and lengthy Monte-Carlo Markov Chain integration, which is not practical to rerun.



(a) (b) (c)

Figure A.5: Distributions of the leading jet kinematic variables as stacked histograms: (a) transverse momentum, $p_{\mathrm{T}}^{\mathrm{jet_leading}}$; (b) pseudorapidity, $\eta^{\mathrm{jet_leading}}$; and (c) azimuthal angle, $\phi^{\mathrm{jet_leading}}$.



(a) (b) (c)

Figure A.6: Distributions of the leading jet kinematic variables as stacked histograms: (a) transverse momentum, $p_{\mathrm{T}}^{\mathrm{jet_subleading}}$; (b) pseudorapidity, $\eta^{\mathrm{jet_subleading}}$; and (c) azimuthal angle, $\phi^{\mathrm{jet_subleading}}$.

the mass of the two particles),

$$m_{\mathrm{tr}}(\vec{a},\vec{b}) = \sqrt{\left(\sqrt{a_x^2+a_y^2}+\sqrt{b_x^2+b_y^2}\right)^2-(a_x+b_x)^2-(a_y+b_y)^2}. \qquad \text{(A.2)}$$

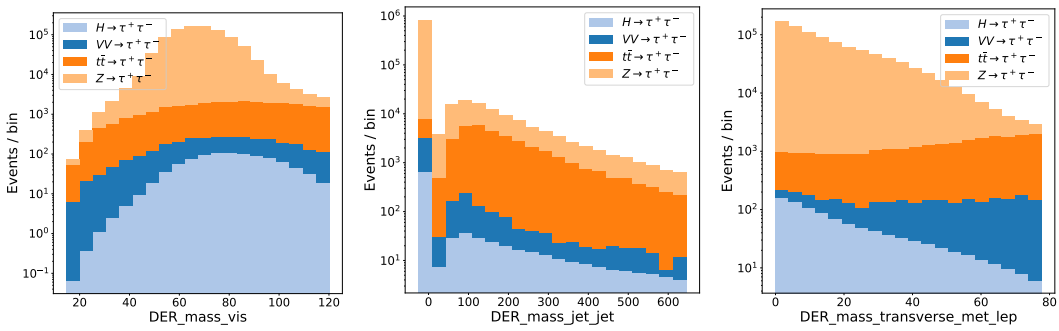
DER_mass_transverse_met_lep The transverse mass (Equation A.2) between the missing transverse energy and the lepton.

DER_deltaeta_jet_jet The absolute value of the pseudorapidity separation $|\eta_{j1}-\eta_{j2}|$. between the two jets (undefined if PRI_jet_num ≤ 1).

DER_prodeta_jet_jet The product of the pseudorapidities of the two jets (undefined if PRI_jet_num ≤ 1).

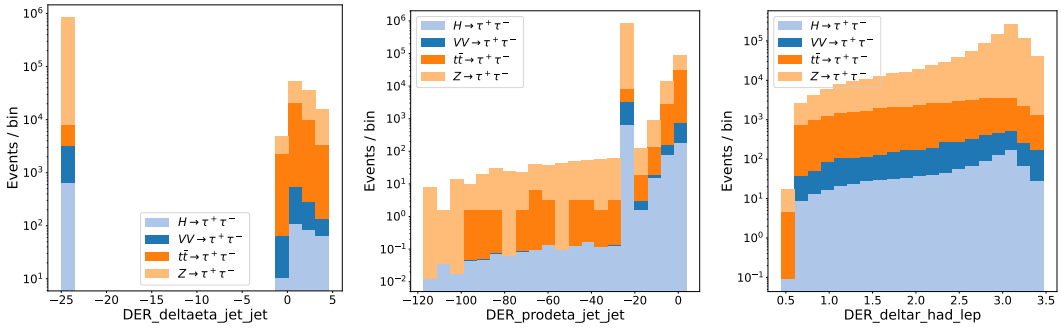
DER_deltar_had_lep The R separation between the hadronic tau and the lepton.

$$\Delta R(\tau_{\mathrm{had}},\ell) = \sqrt{(\eta_{\tau_{\mathrm{had}}}-\eta_\ell)^2+(\phi_{\tau_{\mathrm{had}}}-\phi_\ell)^2}, \qquad \text{(A.3)}$$



(a) (b) (c)

Figure A.7: Distributions of derived mass-related variables as stacked histograms: (a) visible invariant mass of the hadronic τ and lepton system, m_{vis}; (b) invariant mass of the two leading jets, $m^{\text{jet-jet}}$; and (c) transverse mass of the lepton and missing transverse momentum system, $m_{\text{T}}(\ell, p_{\text{T}}^{\text{miss}})$.



(a) (b) (c)

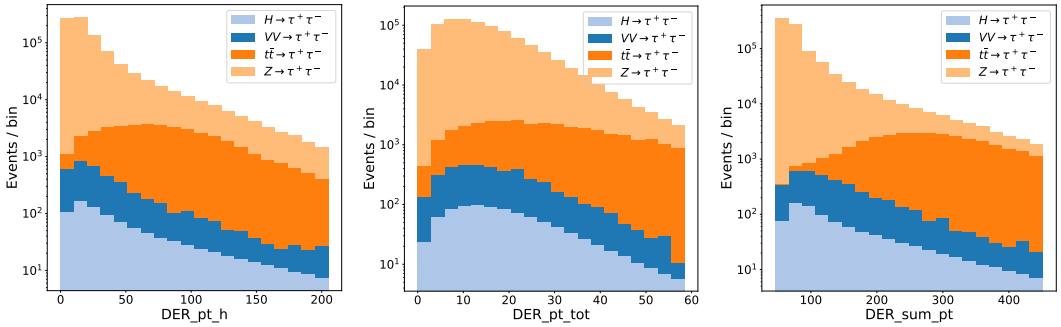
Figure A.8: Distributions of derived angular separation variables as stacked histograms: (a) the pseudorapidity separation between the two leading jets, $\Delta\eta^{\text{jet-jet}}$; (b) the product of the pseudorapidities of the two leading jets, $\text{Prod}(\eta^{\text{jet-jet}})$; and (c) the angular separation between the hadronically decaying τ lepton and the lepton, $\Delta R(\tau_{\text{had}}, \ell)$.

$$p_{\text{T}} \times \cosh \eta, \tag{A.4}$$

`DER_pt_h` The modulus (Equation A.4) of the vector sum of the transverse momentum of the hadronic tau, the lepton, and the missing transverse energy vector.

`DER_pt_tot` The modulus (Equation A.4) of the vector sum of the missing transverse momenta and the transverse momenta of the hadronic tau, the lepton, the leading jet (if `PRI_jet_num` ≥ 1) and the subleading jet (if `PRI_jet_num` $= 2$) (but not of any additional jets).

DER_sum_pt The sum of the moduli (Equation A.4) of the transverse momenta of the hadronic tau, the lepton, the leading jet (if PRI_jet_num ≥ 1), the subleading jet (if PRI_jet_num $= 2$) and the other jets (if PRI_jet_num >= 3).



(a) (b) (c)

Figure A.9: Distributions of derived transverse momentum variables as stacked histograms: (a) the transverse momentum of the reconstructed Higgs candidate, $p_{\mathrm{T}}^{\mathrm{H}}$, obtained from the vector sum of the hadronic τ, lepton, and missing transverse momentum; (b) the total transverse momentum, $p_{\mathrm{T}}^{\mathrm{tot}}$, defined as the magnitude of the vector sum of the transverse momenta of the reconstructed objects; and (c) the scalar sum of the transverse momenta of all reconstructed objects, Scalar Sum(p_{T}).

DER_pt_ratio_lep_tau The ratio of the transverse momenta of the lepton and the hadronic tau.

DER_met_phi_centrality The centrality of the azimuthal angle of the missing transverse energy vector w.r.t. the hadronic tau and the lepton

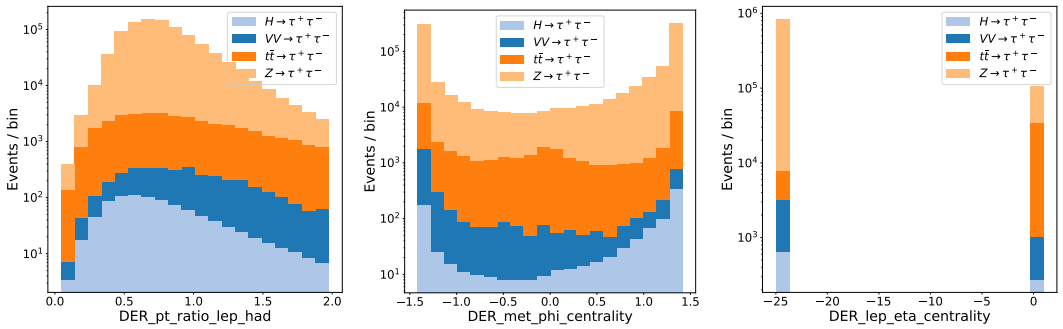
$$C = \frac{A + B}{\sqrt{A^2 + B^2}},$$

where $A = \sin(\phi_{\mathrm{met}} - \phi_{\mathrm{lep}}) * \mathrm{sign}(\sin(\phi_{\mathrm{had}} - \phi_{\mathrm{lep}}))$, $B = \sin(\phi_{\mathrm{had}} - \phi_{\mathrm{met}}) * \mathrm{sign}(\sin(\phi_{\mathrm{had}} - \phi_{\mathrm{lep}}))$, and ϕ_{met}, ϕ_{lep}, and ϕ_{had} are the azimuthal angles of the missing transverse energy vector, the lepton, and the hadronic tau, respectively. The centrality is $\sqrt{2}$ if the missing transverse energy vector $\vec{E}_{\mathrm{T}}^{\mathrm{miss}}$ is on the bisector of the transverse momenta of the lepton and the hadronic tau. It decreases to 1 if $\vec{E}_{\mathrm{T}}^{\mathrm{miss}}$ is collinear with one of these vectors, and it decreases further to $-\sqrt{2}$ when $\vec{E}_{\mathrm{T}}^{\mathrm{miss}}$ is exactly opposite to the bisector. The logic behind this feature is that if the neutrinos are colinear to the lepton and the hadronic tau (which is a good approximation), then the missing transverse energy vector should be between the lepton and the hadronic tau.

DER_lep_eta_centrality The centrality of the pseudorapidity of the lepton w.r.t. the two jets (undefined if PRI_jet_num ≤ 1)

$$\exp\left[\frac{-4}{(\eta_1 - \eta_2)^2}\left(\eta_{\mathrm{lep}} - \frac{\eta_1 + \eta_2}{2}\right)^2\right],$$

where η_{lep} is the pseudorapidity of the lepton and η_1 and η_2 are the pseudorapidities of the two jets. The centrality is 1 when the lepton is on the bisector of the two jets, decreases to $1/e$ when it is collinear to one of the jets, and decreases further to zero at infinity. The logic behind this feature is that if the two jets are emitted together with the Higgs boson, then the Higgs decay product should be on average between the two jets.



(a) (b) (c)

Figure A.10: Distributions of derived kinematic and angular variables as stacked histograms: (a) the ratio of the lepton and hadronic τ transverse momenta, $p_{\text{T}}^{\ell}/p_{\text{T}}^{\tau_{\text{had}}}$; (b) the azimuthal centrality of the missing transverse momentum, $C^{\text{miss}} * \phi$; and (c) the pseudorapidity centrality of the lepton, $C^{\ell} * \eta$.

The feature list and event sample are primarily inspired by [102]. One crucial difference is that the dataset was produced using a simpler (leading-order) event generator (Pythia), and the detector effects were simulated with a simpler detector simulation (Delphes rather than Geant4 ATLAS Geant4 Simulation). The simplifications allowed us to provide participants with a large sample, allowing the development of sophisticated models while preserving the complexity of the original problem.

Appendix B
Study of Tabular Foundational Models for HEP Analysis

In collider-based particle physics experiments, independent events are commonly represented as tabular datasets of high-level variables, a representation widely used in multivariate and machine-learning analyses. There are many ML classifiers available that could be applied to tabular datasets. The baseline example model for the FAIR Universe HiggsML Uncertainty Challenge utilised XGB Classifiers as the discriminators. In the last decade, more complex algorithms, such as Dense Neural Networks and Graph Neural Networks, have been extensively utilised in analysis across the LHC. But their analysis often requires large montecarlo training samples, which may not be available for a very rare process. This problem becomes even more prevalent in the context of simulation-based inference. The new HiggsML datasets provide an opportunity to explore these questions in a stable, easier-to-diagnose stage before applying them to more complex analyses, such as Di-Higgs.

Inspired by the success of foundation models in language and vision, recent developments have introduced Tabular Foundational Models (TFM) such as TabDPT[136], TabM[137], TabICL[138], and TabPFN[139]. Tabular foundation models are used across many industries that rely on structured data. For example, they support disease prediction in healthcare, fraud detection and credit scoring in finance, customer analytics in retail, predictive maintenance in manufacturing, and demand forecasting in energy. Their ability to transfer knowledge across different tabular datasets makes them effective for a wide range of real-world applications. This chapter explores the use of such TFMs in data-limited environments in high-energy physics. The foundational models are benchmarked against boosted decision trees and neural networks across classification, density-ratio estimation, and histogram-based analysis. They are assessed for their potential role in HEP analysis workflows.

B.1 Tabular Foundational Models

Tabular Foundational Models (TFM) are a class of models trained to work with a wide variety of Tabular Datasets without explicit retraining. It works similarly to an LLM, but instead of tokenising words in a sentence, it tokenises table elements, unlike traditional ML algorithms like BDTs or NNs, which treat each row of elements as the input. TFMs compute on these tokens. This enables training the model on large collections of unrelated datasets of varying sizes and shapes. Our study focuses on two TFMs. **Tabular Prior-Fitter Networks (TabPFN)** and **Tabular**

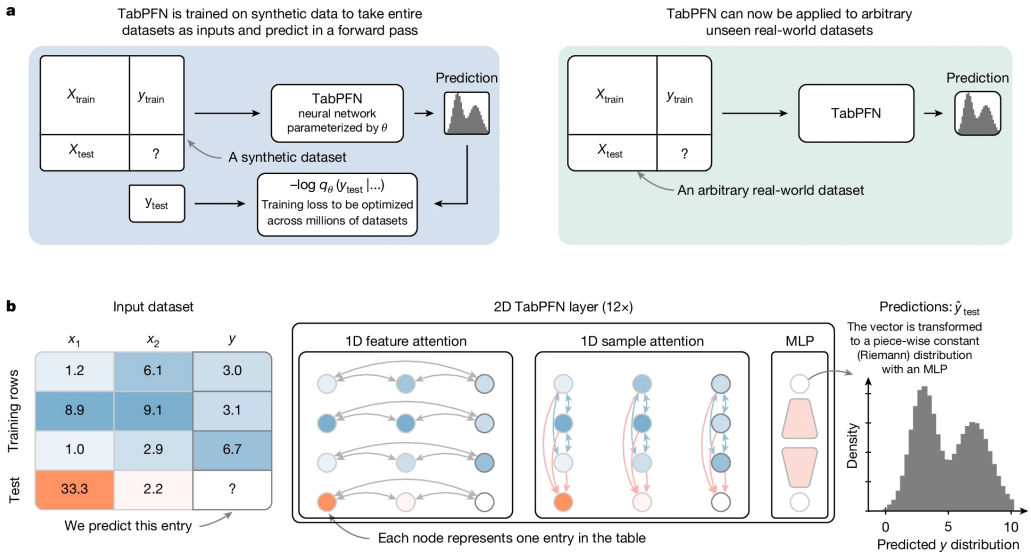


Figure B.1: Basic architecture of TabPFN[139]

In-Context Learning (TabICL). TabPFN and TabICL share many architectural similarities. They both use Row-wise attention and Feature-Attention to learn relationships between different rows or events in the HEP context and between different features. This is similar to how LLMs use self-attention to learn the importance of a token relative to the tokens before it. The difference between the 2 architectures lies in the step after the attention maps.Figure B.1 gives the architecture of TabPFN. The last step in the process is an MLP, which makes the final prediction of the label (y_{test}).

Figure B.2 gives the architecture of TabICL. The last step in the chain is the In-Context Transformer, which learns the relationship between datasets.

The core challenge when using TFMs for HEP analysis is that, so far, they don't handle sample weights. But weights are critical for HEP analysis, as Monte Carlo simulators often use them to indicate the importance of an event's physics. TFMs are pretrained, so there is no straightforward way to change the loss function or architecture to help them learn the weights. This is critical. In this chapter, we will study and compare TFMs' performance in the absence of sample events.

B.2 Analysis Methodology

For the analysis, we used a 1M events subsample of the FAIR Universe HiggsML dataset. The dataset was then further divided into 800k events for training and validation, and 200K for testing. For comparison, we primarily used 4 models: an

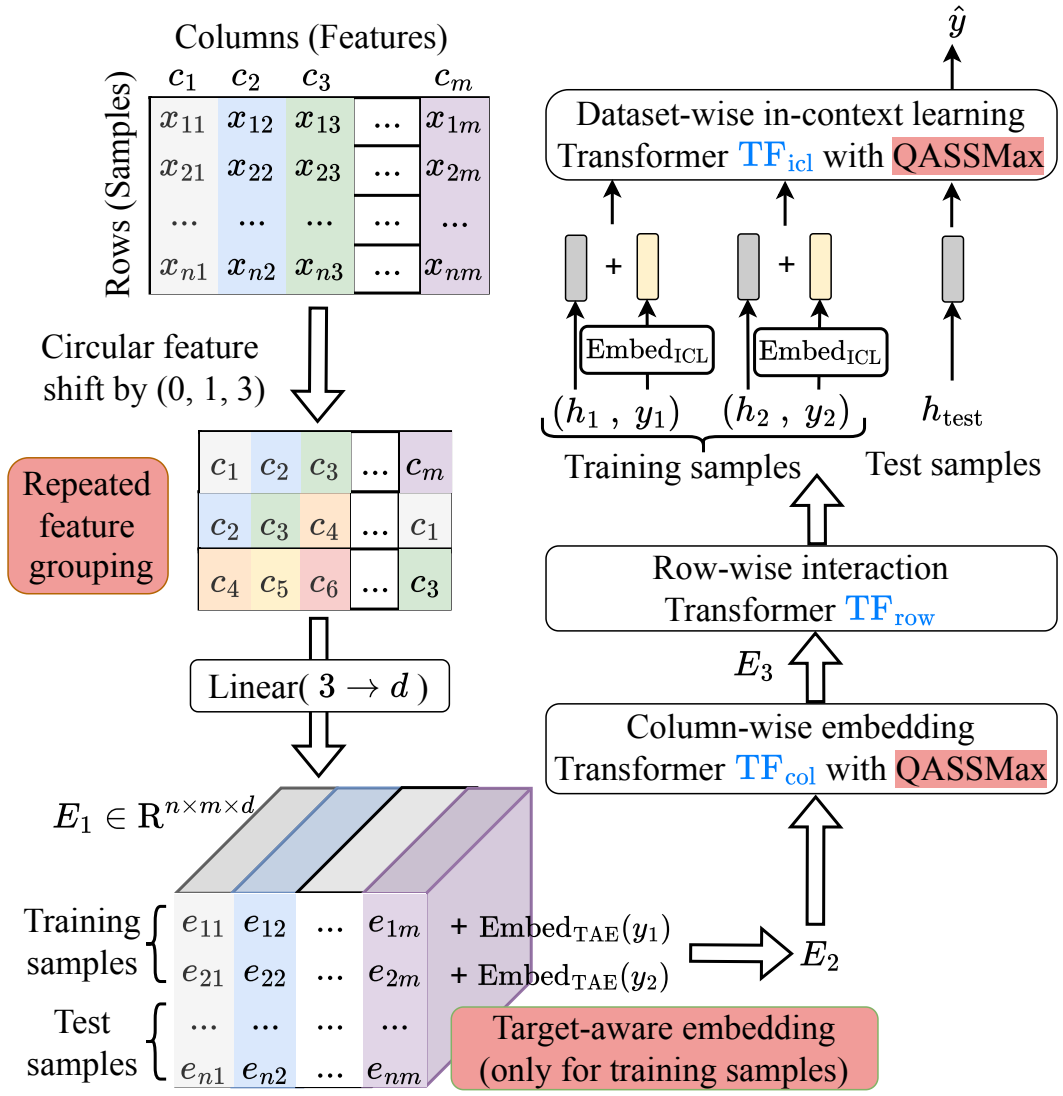


Figure B.2: Basic architecture of TabICL[138]

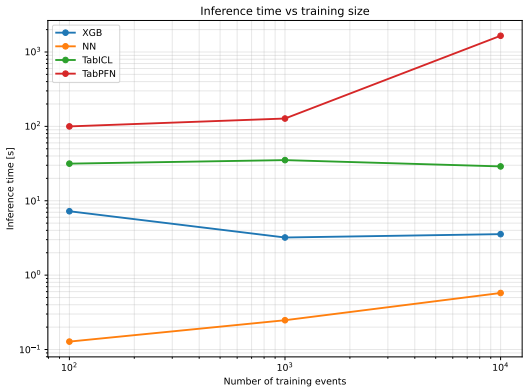


Figure B.3: Inference time vs size of training dataset for 4 different ML models, BDT in Blue, DNN in orange, TABPFN in red and TABICL in green.

XGBoost-based BDT, a Torch-based neural network, and the 2 TFMs mentioned earlier. Since TFMs cannot accommodate sample weights, to ensure a fair comparison, we removed sample weights from the training data and trained all classifiers without them. For each model type, an ensemble of 10 models was trained with different training sizes. We used the ensemble to find the uncertainty in our metrics. Unlike BDTs or DNNs, TFMs' inference time can depend on the training dataset size, since the attention map for the training events is used for the inference. Hence, we also studied the inference time of various models and how they change with respect to the training dataset size. Figure B.3 shows inference time versus training dataset size for 4 different ML models. We can clearly see that TFMs are slower than BDTs and DNNs, and among the TFMs, TabICL is almost 10 times faster on average. This can be attributed to the in-context learning that ICL performs.

Before applying physics-based metrics, we evaluated classification performance using the Receiver Operating Characteristic (ROC) curve and its Area under the curve (AUC). Figure B.4 shows the evolution of the ROC curve as the number of training events increases from 100 to 10,000.

For very low training events like the 100 or 200, we see that the TFMs (TabICL and TabPFN) are clearly better, but as the number of training events increases, the difference between the models blurs. This is all the more clear in Figure B.5, which shows the AUC of each model as a function of training sample size (N).

B.3 Physics Analysis with Tabular Fundamental Models

Although the Area Under the Receiver Operating Characteristic Curve (AUC) is an effective measure of a classifier's ability to rank signal events above background events, it is a threshold-independent metric that does not consider the experimental context or the relative consequences of signal and background selections. In contrast, particle physics analyses require selecting a specific operating point that maximises the probability of discovering a rare process while controlling the number of background events. For this reason, we use Approximate Median Significance (AMS) as

our primary evaluation metric and later perform a Binned likelihood fit to obtain a more accurate estimate.

B.3.1 Approximate Median Significance

Approximate Median Significance(AMS) estimates the expected statistical significance of a potential Higgs boson discovery by combining the numbers of selected signal (s) and background (b) events into a significance measure. The expression for AMS is given by:

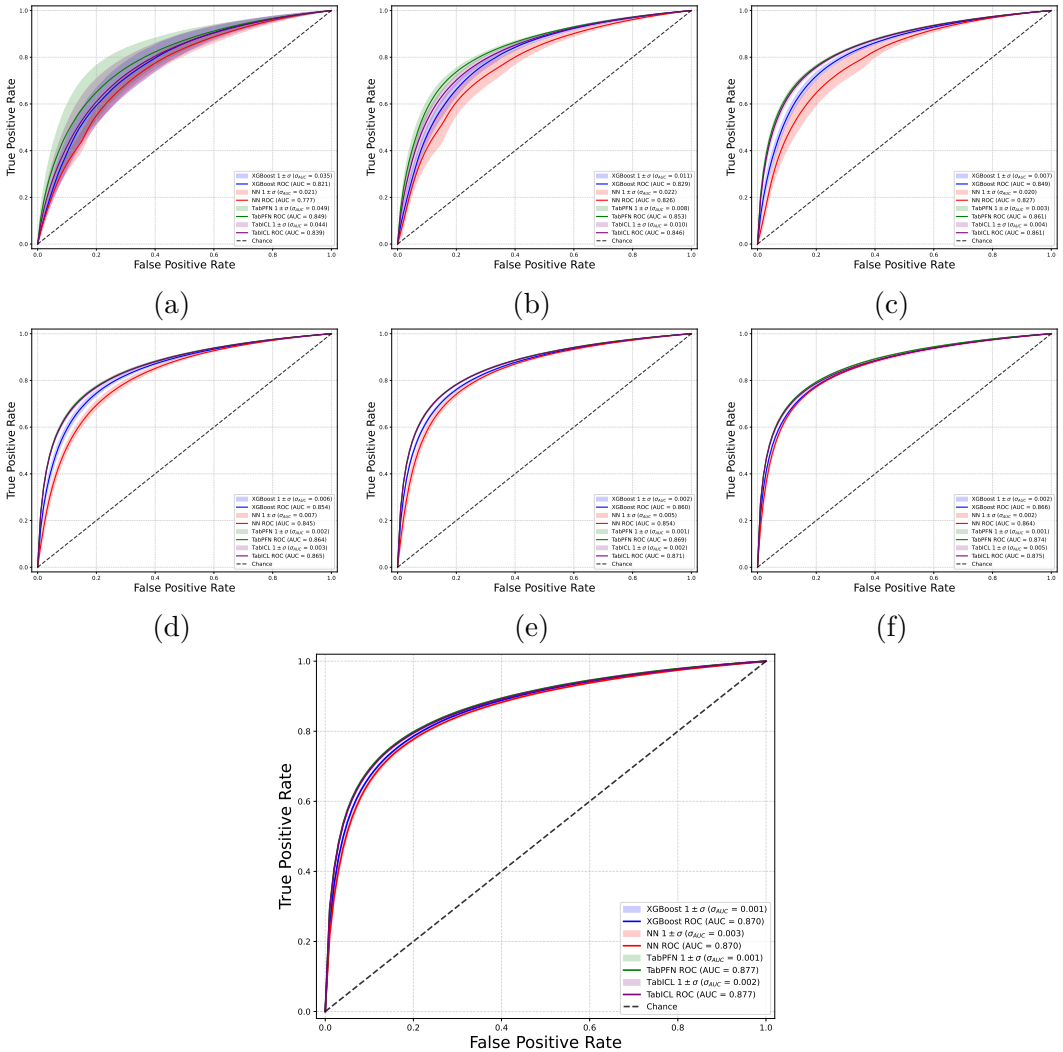


Figure B.4: The evolution of the ROC curve as the number of training events. (a) Trained with 100 events, (b)Trained with 200 events, (c) Trained with 500 events, (d) Trained with 1000 events, (e) Trained with 2000 events. (f) Trained with 5000 events and (g) Trained with 10000 events

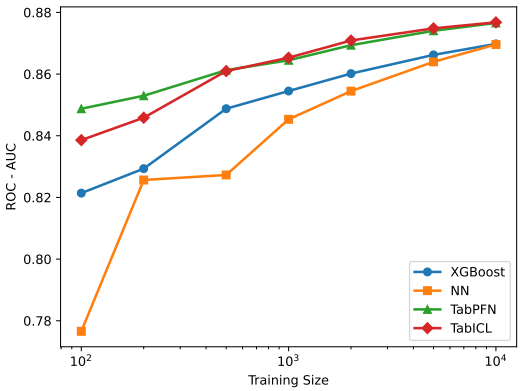
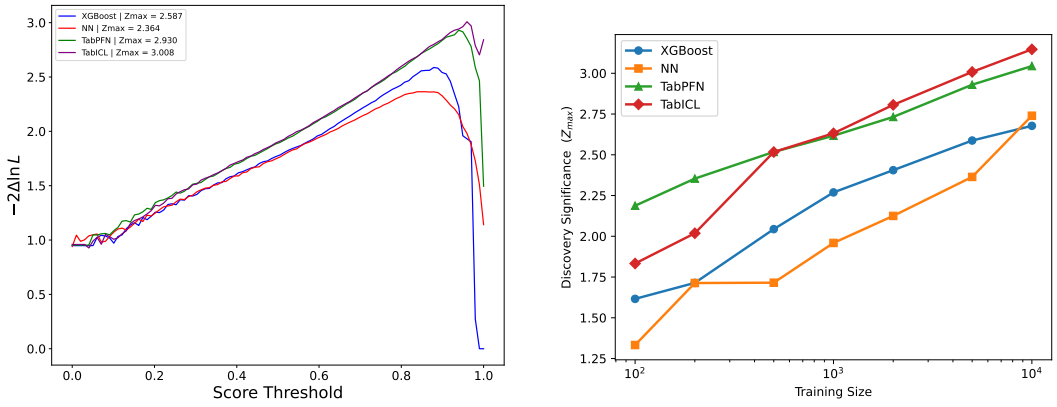


Figure B.5: The AUC of each model as a function of training sample size. ($N \in \{100, 200, 500, 1000, 2000, 5000, 10000\}$).

$$\text{AMS} = \sqrt{2\left[(s+b)\,\ln\left(1+\frac{s}{b}\right)-s\right]}, \tag{B.1}$$

Figure B.6a shows the evolution of discovery significance as a function of decision threshold, where all the models were trained with 5000 events. The ideal threshold is at the maximum of this curve; Z max. Figure B.6b gives the evolution of the Zmax for different models as a function of Training sample size.



(a) (b)

Figure B.6: Comparison of the tabulated results (a) Z vs threshold for Training sample size of 5000 and (b) Evolution of $Z_{\max}$ as a function of Training sample size.

From Figure B.6b we can clearly see the advantage of TFMs over tractional methods. Additionally, we see that for 100 and 200 events, TabPFN has a clear advantage over TabICL.

B.3.2 Summary Histogram with Machine Learning Discriminants

In section 4.5, we discussed how to extract more information from the events. Performing a binned analysis of the classifier's score can yield greater discovery

significance. So, perform a binned analysis of the scores for the various models we discussed, using the same set of 25 uniform bins from the minimum to the maximum score. Figure B.7(a - g) shows the evolution of the NLLs with increasing number of training samples.

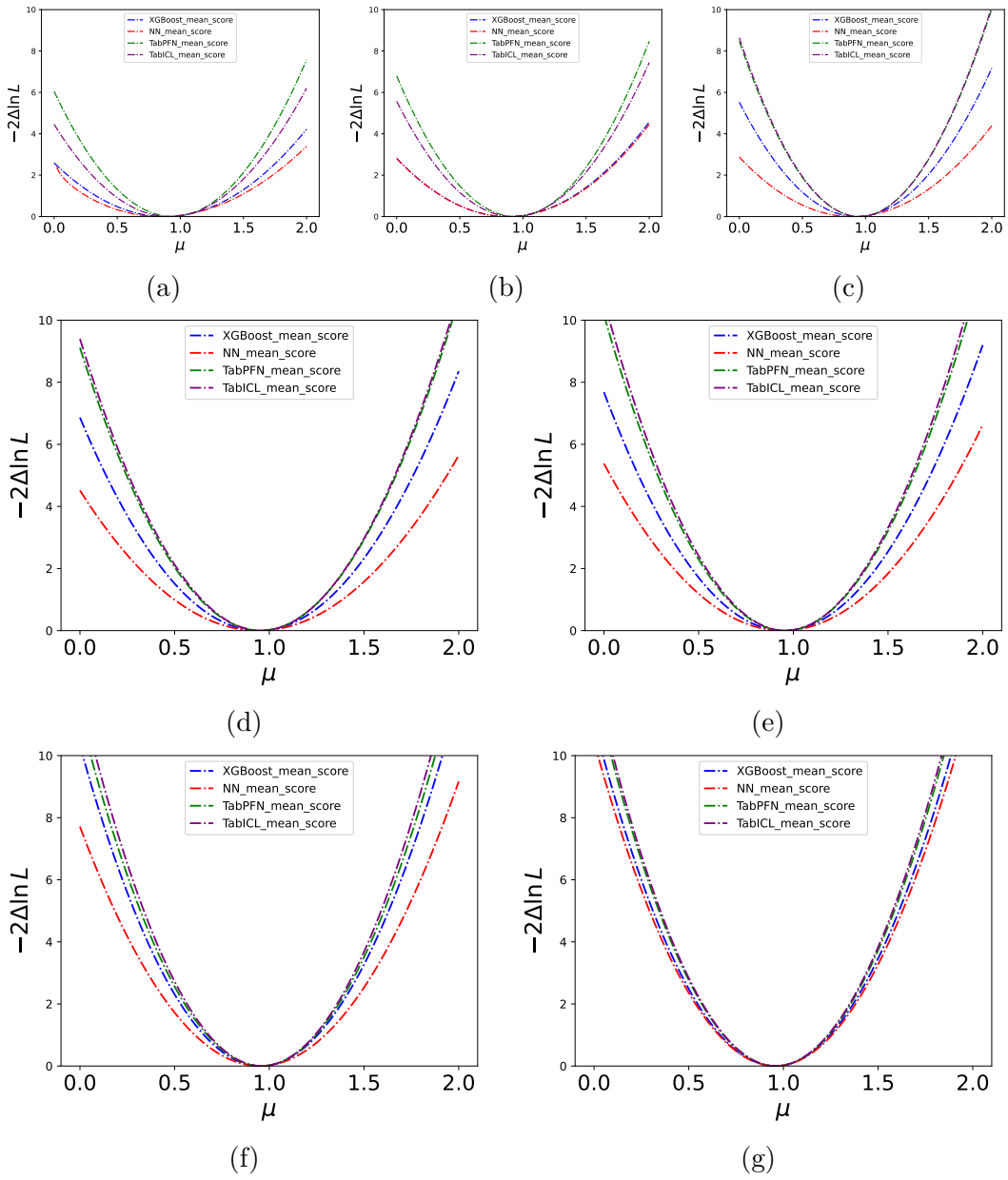


Figure B.7: The evolution of the likelihood curves as the number of training events. (a) Trained with 100 events, (b)Trained with 200 events, (c) Trained with 500 events, (d) Trained with 1000 events, (e) Trained with 2000 events. (f) Trained with 5000 events and (g) Trained with 10000 events

As mentioned in section 4.5, the narrower the NLL, the better the confidence interval. The square root of the point at which the NLL touch the y-axis ($\sqrt{t_{\mu=0}}$)

is the statistical significance for a null hypothesis rejection or discovery significance. Figure B.8 shows the evolution of $\sqrt{t_{\mu=0}}$ and its discovery significance as a function of the number of training events, for various models.

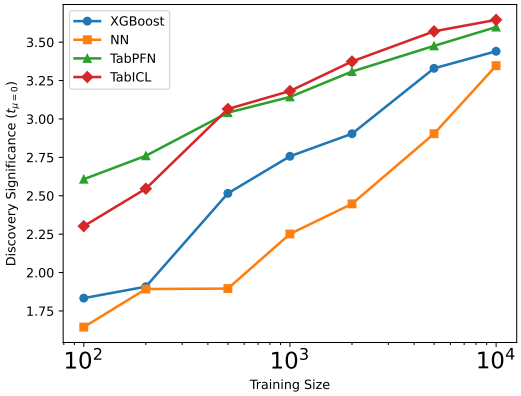


Figure B.8: The discovery significance($\sqrt{t_{\mu=0}}$) of each model as a function of training sample size,($N \in \{100, 200, 500, 1000, 2000, 5000, 10000\}$).

As one would expect, the trends we see in $\sqrt{t_{\mu=0}}$ are similar to those in AMS.TFMs perform better than the BDTs and NN, but the gap between them is decreasing as the number of events increases.

B.3.3 Simulation-Based Inference with Tabular Foundational Models

Traditionally, SBIs have mostly been done with Neural Networks. We have so far discussed in detail the advantages of using Tabular Foundation over traditional ML algorithms, such as DNNs or BDTs, when we have a limited number of training events. We can expand that study to evaluate the viability of using Tabular Foundation models for SBI by computing calibration curves for 2 Tabular Foundation models, TabICL and TabPFN, and comparing them with those of the BDT and DNN when trained on the same 2000 events and tested on 800k events. Figure B.9 compares the calibration curves for $N \in \{100, 200, 500, 1000\}$ and Figure B.10 for $N \in \{2000, 5000, 10000\}$. As the models themselves are trained without weights, the calibration curves are also built without weights.

Even though, for $N \in \{100, 200, 500, 1000\}$, the calibration curve is subpar, we still clearly see that the foundation model has better calibration than BDT and DNN. From $N = 2000$ onwards, the TFMs have well-calibrated curves, while the BDTs and DNN only reach close to that $N = 10,000$. This motivates us to use TFMs for SBI analysis. But a crucial problem is that we don't yet have a way to add sample weights, and it gets worse when we consider adding negative weights.

B.4 Discussion of Tabular Foundational Models for HEP

The study showed that we have strong potential to use the Tabular foundation model in regimes with few events. Between the 2 TFM models, we can clearly see that the first 2 TFMs outperform the BDT and DNNs across all metrics when we

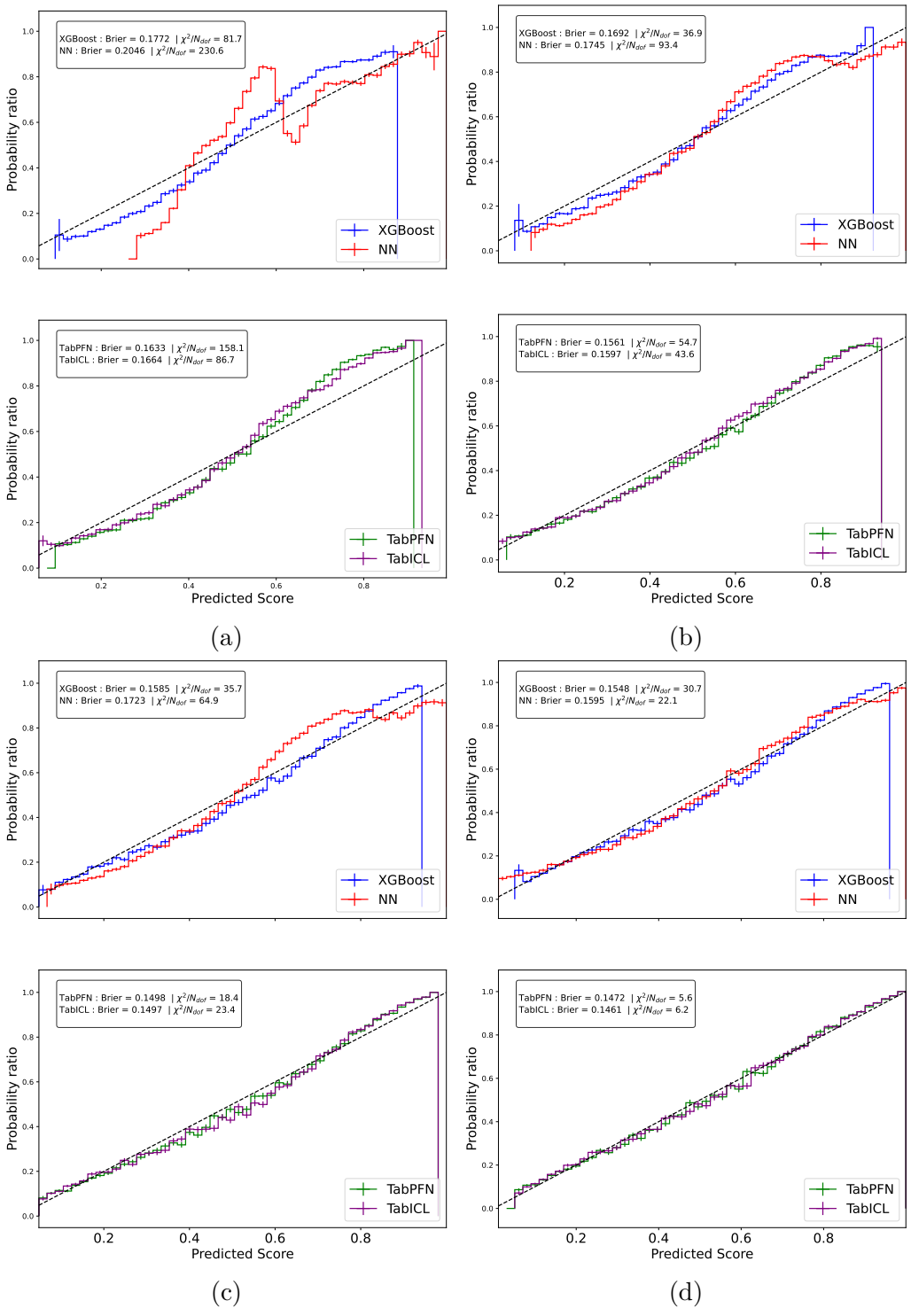


Figure B.9: Comparing calibration curves using different Binary Classifiers (BDT in blue, DNN in red, TAbPFN in green and TabICL in purple) for various sizes of training events. (a) Trained with 100 events, (b)Trained with 200 events, (c) Trained with 500 events, (d) Trained with 1000 events

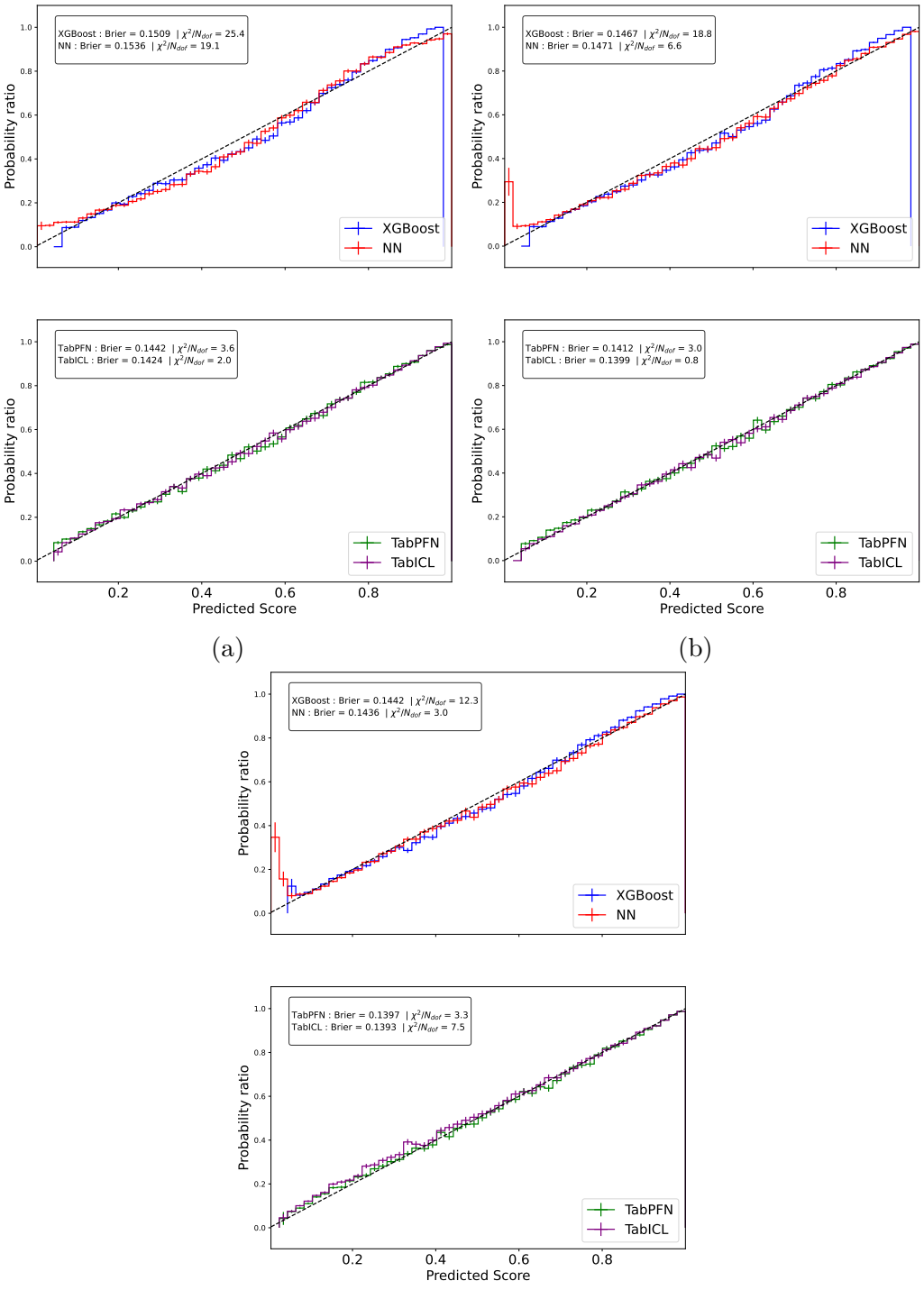


Figure B.10: Comparing calibration curves using different Binary Classifiers (BDT in blue, DNN in red, TAbPFN in green and TabICL in purple) for various sizes of training events. (a) Trained with 2000 events, (b) Trained with 5000 events, (c) Trained with 10,000 events

have fewer than 5000 training events. Among the TFMs, the most noticeable aspect is that for 100 and 200 training events, TabICL performs worse than TabPFM. This was because TabICL weren't trained with such a small dataset in mind. On the flip side, the timing plot (Figure B.3) suggests that TAbICL makes up for it with improved inference speed and can hence scale up to 1M events. On that note, TabPFN2.5 had a hard limit of 50k events, which motivated us to limit our studies to 10k events. The test with the calibration curve indeed showed that the TFMs are learning probabilities correctly.

But there are a few critical limitations that need to be addressed before we can use it in realist analysis. First is the case for weights. As we noted, there are currently no methods for either of the TFMs to incorporate sample weights, which critically impacts usability in HEP. Solving the weight issue for Tabular Foundational Models for HEP, including in Simulation-Based Inference.

Appendix C
Summary Statistics for Di-Higgs Analysis samples

This chapter summarises various processes, their yields, average weight, the number of MC samples, and effectiveness. In Table 7.1, we discussed this for each category. Here is a further breakdown in terms of processes in each category.

Table C.1: Summary Statistics for **Diboson** samples

Sample	**Yield**	**Number of Samples**	**Average Weight**	**Effectiveness**
ZZ	27.464 ± 0.180	113823	0.0002	0.205
WZ	10.069 ± 0.294	5534	0.002	0.212
WW	2.803 ± 0.589	47	0.060	0.482
VVEWK	2.388 ± 0.142	1812	0.001	0.155
Total	42.725 ± 0.697	121216	0.016	0.031

Table C.2: Summary Statistics for **Single H** samples

Sample	**Yield**	**Number of Samples**	**Average Weight**	**Effectiveness**
(ggF) $H \to \tau^+\tau^-$	12.727 ± 0.329	1816	0.007	0.824
(ggF) $H \to \tau^+\tau^-$	1.230 ± 0.017	6392	0.0002	0.775
(ggF) $ZH \to \tau^+\tau^-$ ggZHtautau	2.009 ± 0.051	1761	0.001	0.874
$Z(gg)H(b\bar{b})$	3.317 ± 0.025	20511	0.0002	0.878
$Z(qq)H(b\bar{b})$	7.395 ± 0.046	67898	0.0001	0.374
$Z(qq)H(\tau^+\tau^-)$	7.273 ± 0.064	37719	0.0002	0.345
$t\bar{t}H(\tau^+\tau^-)$	26.073 ± 0.111	74110	0.0004	0.741
$t\bar{t}H$	7.948 ± 0.218	3001	0.0026	0.443
$WH(b\bar{b})$	0.379 ± 0.019	1051	0.0004	0.365
$WH(\tau^+\tau^-)$	0.522 ± 0.029	418	0.001	0.761
Total	68.874 ± 0.423	214677	0.001	0.123

Table C.3: Summary Statistics for **Rest** samples

Sample	**Yield**	**Number of Samples**	**Average Weight**	**Effectiveness**
$t\bar{t}Z$	9.124 ± 0.469	2692	0.003	0.141
$t\bar{t}W$	16.535 ± 0.265	9983	0.002	0.389
stopst	61.768 ± 3.926	939	0.066	0.264
tWincl	246.689 ± 7.299	1369	0.180	0.834
tWdil	151.260 ± 1.838	7930	0.019	0.854
$Z + \ell\ell$	5.097 ± 4.108	47	0.108	0.033
$Z + bb$	16.983 ± 1.680	1436	0.012	0.071
$Z + cc$	0.188 ± 0.368	47	0.004	0.006
$Z + cl$	1.501 ± 0.514	39	0.038	0.218
$Z + bc$	1.424 ± 0.295	118	0.012	0.198
$Z + bl$	2.493 ± 0.672	158	0.016	0.087
Zttcl	41.179 ± 3.209	2578	0.016	0.064
Zttbl	97.478 ± 4.168	7205	0.014	0.076
Zttbb	794.945 ± 10.387	73997	0.011	0.079
Zttcc	43.757 ± 2.968	2893	0.015	0.075
Zttbc	68.982 ± 3.007	7138	0.010	0.074
Zttll	38.290 ± 14.134	2064	0.019	0.004
Total	1600.185 ± 21.134	120791	0.031	0.047

Appendix D
Analysis plots Di-Higgs SBI

This chapter gives the detailed analysis plots for various classifiers. The calibration curves and Reweighting plots we discussed in subsection 7.3.5, and the NLL comparison from section 7.4, are further explored here.

D.1 Analysis plots of Classifier for SBI

To properly validate the models, we need to build calibration curves and Reweighting plots for each classifier and each category. Further, for reweighting plots, we have also presented m_{bb} and $m_{\tau\tau}^{\mathrm{MMC}}$, in addition to the m_{HH} we discussed in subsection 7.3.5.

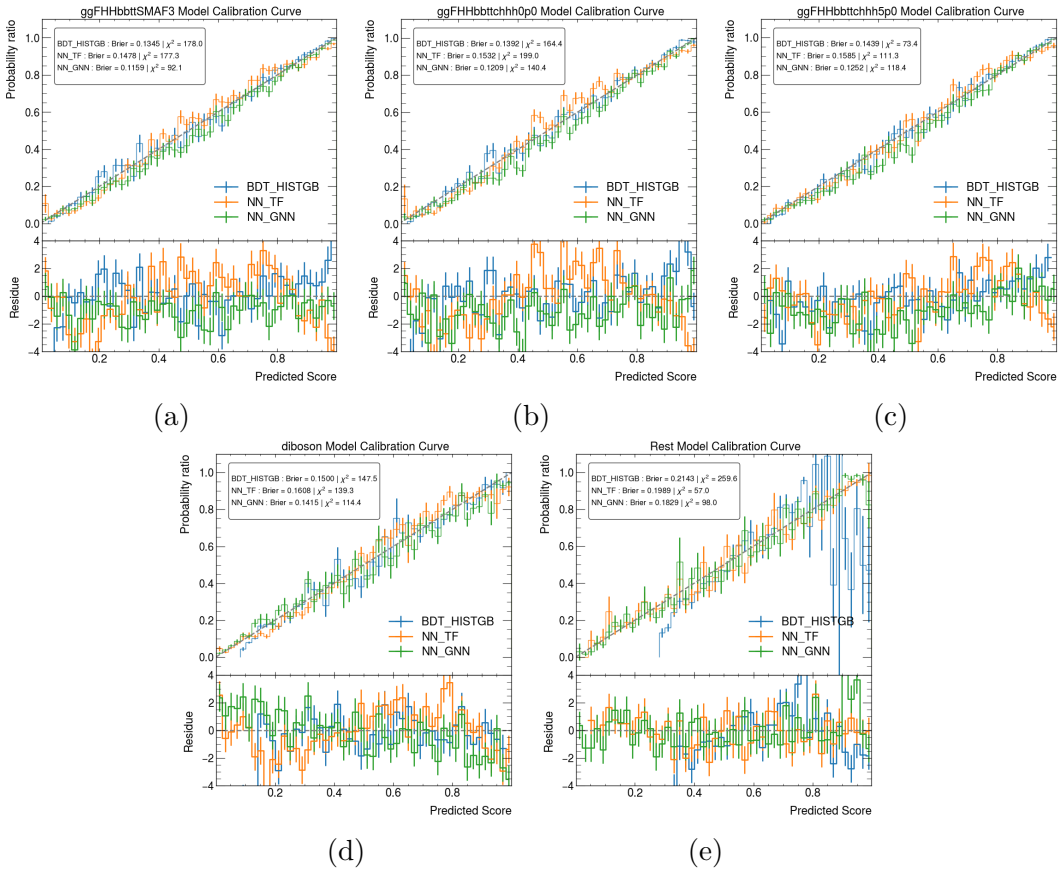


Figure D.1: Receiver operating characteristic Curve using different Binary Classifiers Graph Attention Transformers (GAT) in green, Dense Neural Networks (DNN) in orange and Boosted Decision Tree (BDT) in blue. For different samples (a)SM $HH \to b\bar{b}\tau^+\tau^-$, on different samples (a)SM $HH \to b\bar{b}\tau^+\tau^-$, (b)$\kappa_\lambda = 0$, $HH \to b\bar{b}\tau^+\tau^-$, (c) $\kappa_\lambda = 5$, $HH \to b\bar{b}\tau^+\tau^-$, (d) Di-Boson and (e) Rest ($Z+$ jets and tops)

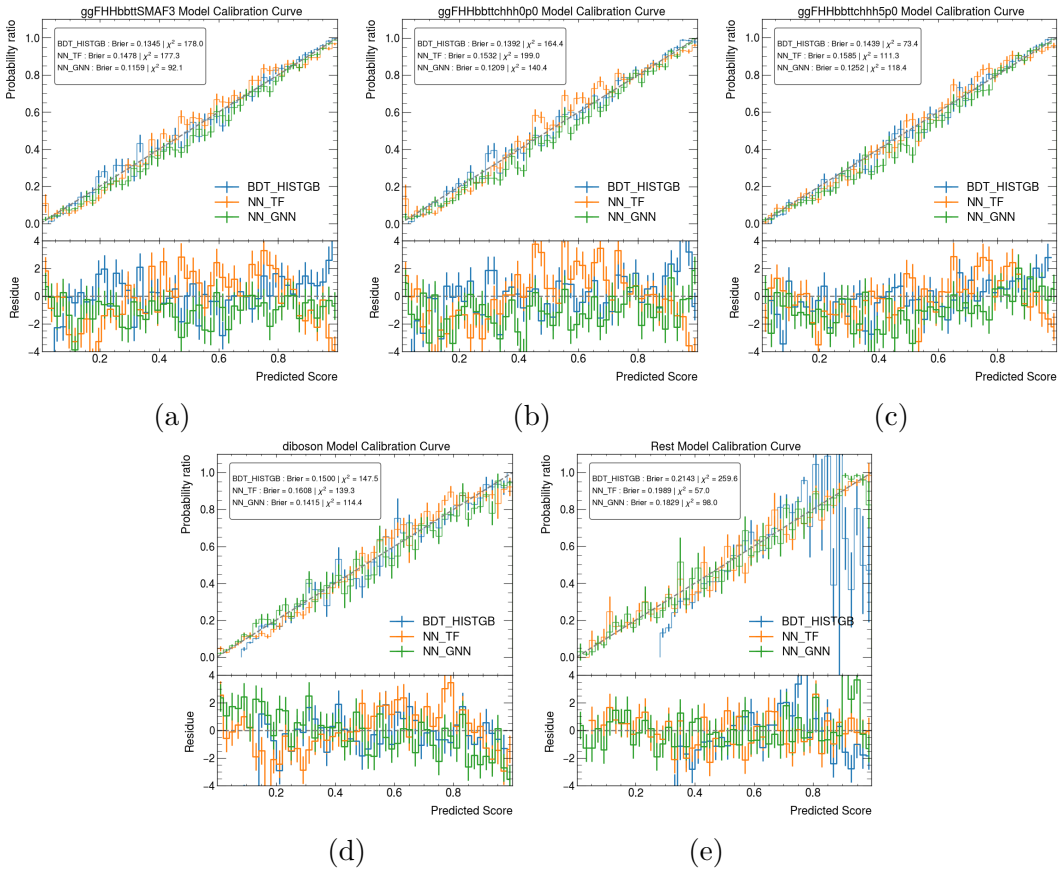


Figure D.2: Comparing Calibration Curves using different Binary Classifiers Graph Attention Tranformers (GAT) in green, Dense Neural Networks (DNN) in orange and Boosted Decision Tree (BDT) in blue, on different samples (a)SM $HH \to b\bar{b}\tau^+\tau^-$, (b)$\kappa_\lambda = 0$, $HH \to b\bar{b}\tau^+\tau^-$, (c) $\kappa_\lambda = 5$, $HH \to b\bar{b}\tau^+\tau^-$, (d) Di-Boson and (e) Rest ($Z+$ jets and tops)

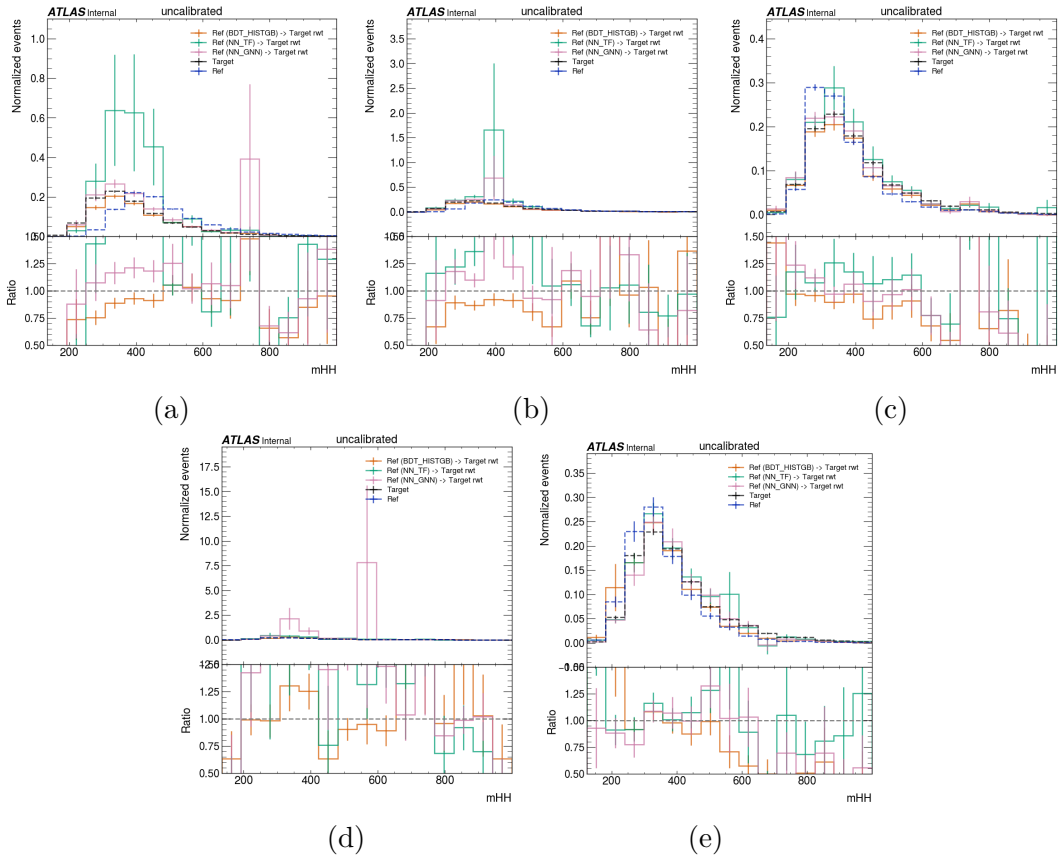


Figure D.3: Comparing reweighting plots of Di-Higgs Invariant mass (m_{HH}) using different Binary Classifiers; BDT (orange), DNN (green) and GAT (pink), reference distribution (dashed blue) and target distribution in the dashed black line. On different samples (a)SM $HH \to b\bar{b}\tau^+\tau^-$, (b)$\kappa_\lambda = 0$, $HH \to b\bar{b}\tau^+\tau^-$, (c) $\kappa_\lambda = 5$, $HH \to b\bar{b}\tau^+\tau^-$, (d) Di-Boson and (e) Rest (Z+ jets and tops)

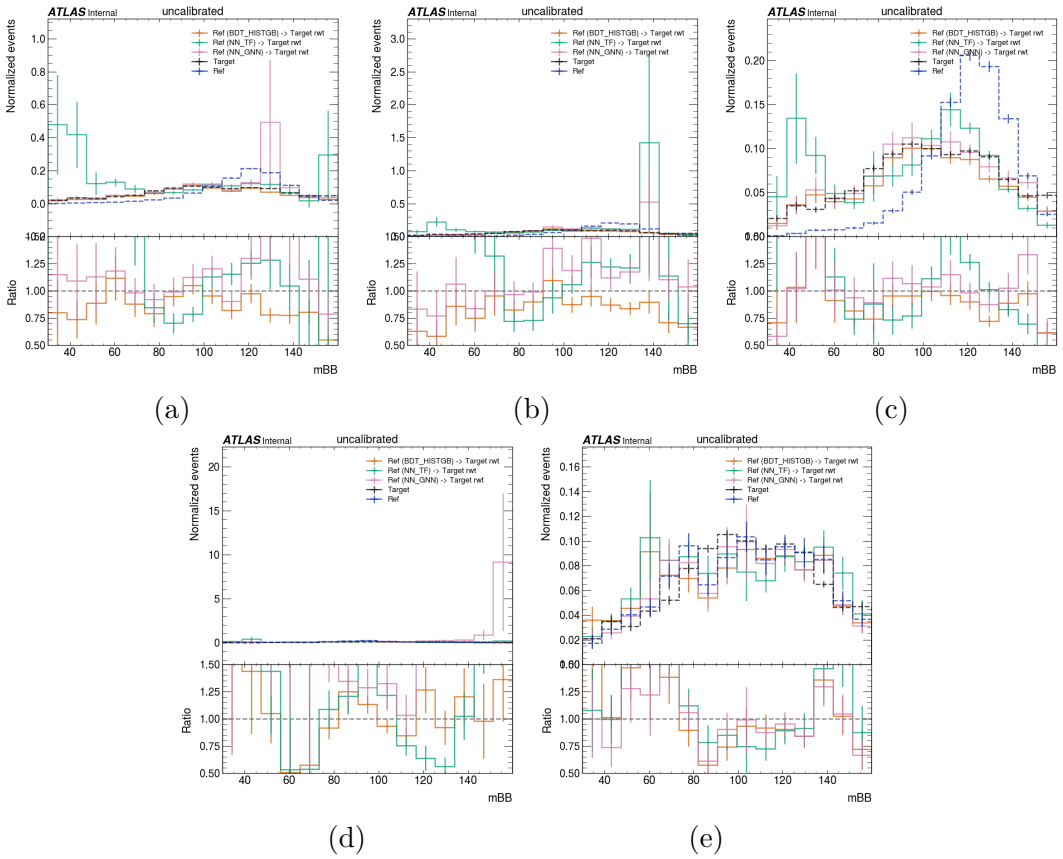


Figure D.4: Comparing reweighting plots of invariant mass of the 2B system (m_{bb}) using different Binary Classifiers: BDT (orange), DNN (green) and GAT (pink), reference distribution (dashed blue) and target distribution in the dashed black line. On different samples (a)SM $HH \to b\bar{b}\tau^+\tau^-$, (b)$\kappa_\lambda = 0$, $HH \to b\bar{b}\tau^+\tau^-$, (c) $\kappa_\lambda = 5$, $HH \to b\bar{b}\tau^+\tau^-$, (d) Di-Boson and (e) Rest (Z+ jets and tops)

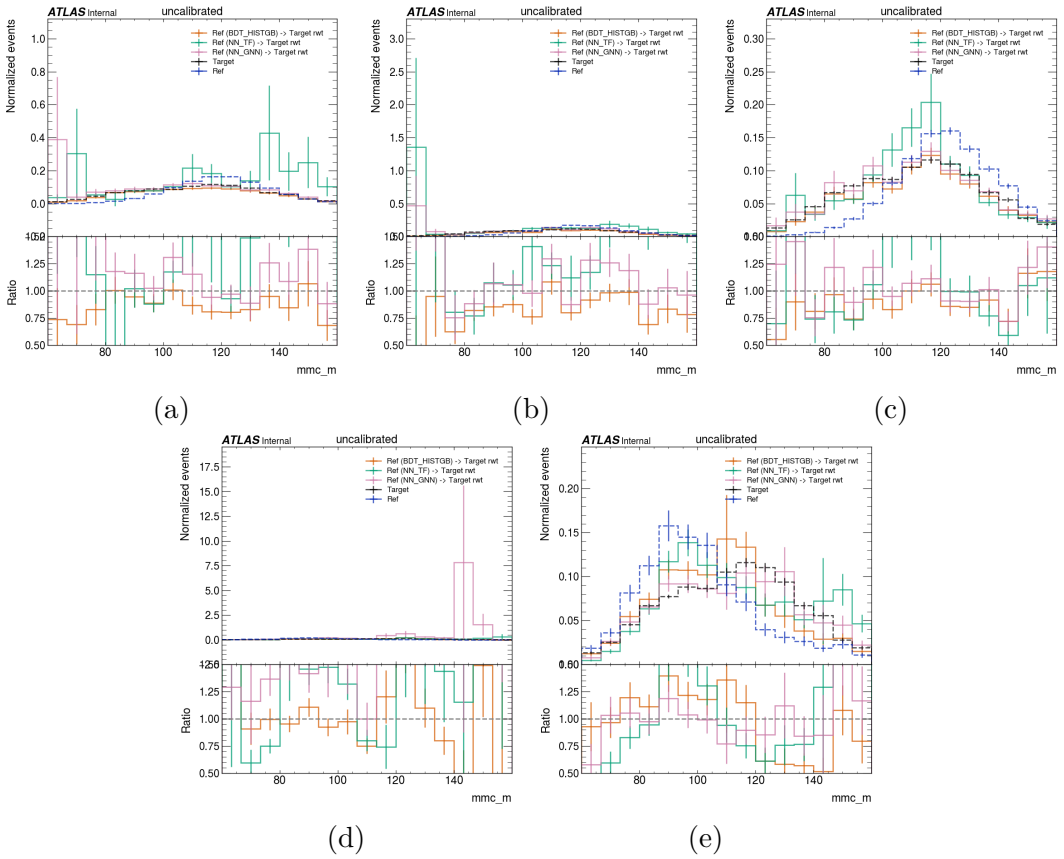
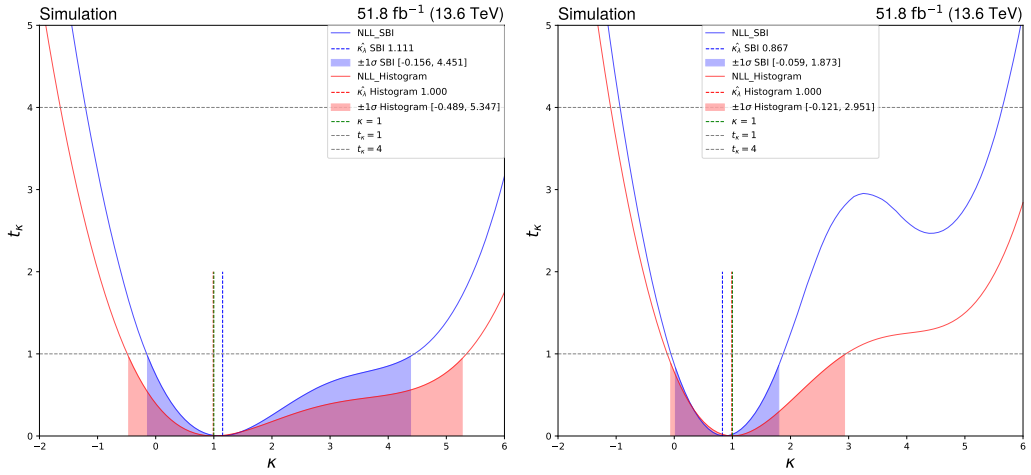


Figure D.5: Comparing reweighting plots of MMC mass of the Di-tau system ($m_{\tau\tau}^{\mathrm{MMC}}$) using different Binary Classifiers: BDT (orange), DNN (green) and GAT (pink), reference distribution (dashed blue) and target distribution in the dashed black line. On different samples (a)SM $HH \to b\bar{b}\tau^+\tau^-$, (b)$\kappa_\lambda = 0$, $HH \to b\bar{b}\tau^+\tau^-$, (c) $\kappa_\lambda = 5$, $HH \to b\bar{b}\tau^+\tau^-$, (d) Di-Boson and (e) Rest ($Z+$ jets and tops)

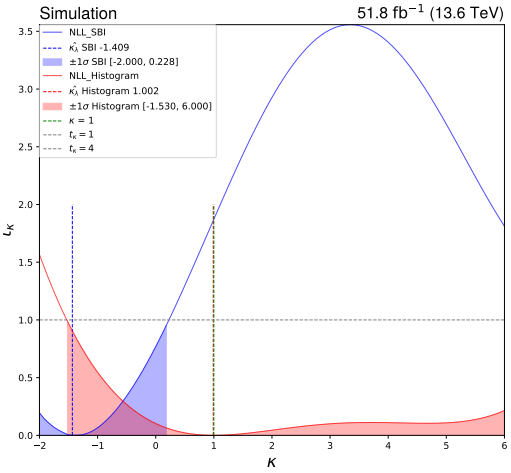
D.2 SBI pseudo-univertse Anlysis

In section 7.4 we discussed the case where only single Higgs and Di-boson backgrounds were present, and here we present the other background and various combinations of them.

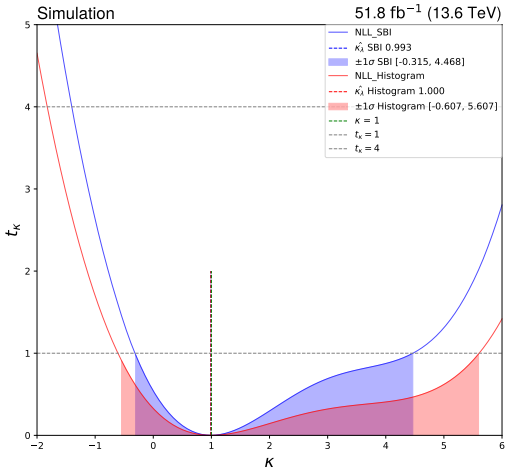


(a)

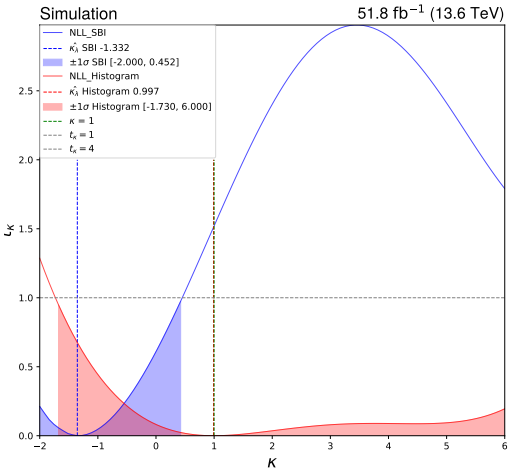
(b)



(c)



(d)



(e)

Figure D.6: NLL scan over Higgs self coupling κ_λ for stat-only results with $\kappa_{\lambda true} = 1$, with different combination if backgrounds considered starting with no background (a), only single Higgs, (b) only di-boson (c), only Rest (top and $Z+$jets) (d) only single Higgs and Di-boson (e) and all background (single Higgs, Di-boson, top and $Z+$jets)(f)