Master de Physique Fondamentale et Applications, Sorbonne Université

Laboratoire de Physique Nucléaire et des Hautes Énergies (LPNHE) de Paris

---

Rapport de stage de M1: Reconstruction du boson Z^0 dans l'expérience ATLAS et son utilisation comme chandelle étalon

14 avril 2025 - 04 juin 2025

Élève

Guillem JOSEPH PLANAS
Numéro étudiant: 21105304

Encadrant

Frédéric DERUE

Sommaire

Introduction

Dans le cadre de mon Master 1 de Physique Fondamentale et Applications à Sorbonne Université, j'ai effectué un stage de recherche de deux mois, du 14 avril au 4 juin 2025, au Laboratoire de Physique Nucléaire et de Hautes Énergies (LPNHE), situé sur le campus Pierre et Marie Curie à Jussieu. Ce laboratoire est une unité mixte de recherche (UMR 7585) associée au CNRS/IN2P3, à Sorbonne Université et à l'Université Paris Cité. Il mène des recherches dans plusieurs domaines de la physique fondamentale, notamment la physique des particules, la physique des astroparticules et la cosmologie. Les équipes de recherche du LPNHE participent à plusieurs grandes collaborations internationales, et sont soutenues par des services techniques (informatique, électronique, mécanique), ainsi que par des services d'appui administratif et logistique. J'ai été accueilli au sein du groupe ATLAS, l'une des grandes expériences de physique des particules menées au LHC (Large Hadron Collider) du CERN, conçue pour tester les prédictions du Modèle Standard de la physique des particules.

Ce stage, réalisé entièrement en présentiel, m'a permis de découvrir de l'intérieur le fonctionnement d'un grand laboratoire de recherche, ainsi que de m'initier aux outils et méthodes d'analyse utilisés en physique des particules. Mon encadrement a été assuré par Frédéric Derue, chercheur au LPNHE spécialisé dans l'étude du quark top (la particule élémentaire la plus massive) et dans la mesure précise de sa masse. Le sujet de mon stage portait sur l'analyse de la reconstruction du boson Z^0, une particule du Modèle Standard impliquée dans l'interaction faible. Pour cela, j'ai travaillé à la fois sur des données expérimentales issues du détecteur ATLAS et sur des données simulées. J'ai étudié les performances de l'identification des électrons issus de la désintégration du boson, en comparant une méthode basée sur des coupures simples sur des distributions de variables discriminantes à une approche utilisant des techniques de Machine Learning. Ce travail s'inscrit dans une démarche plus large visant à améliorer la compréhension et la modélisation des performances du détecteur.

La première partie de ce rapport est consacrée au Modèle Standard de la physique des particules ainsi qu'à la présentation du détecteur ATLAS. La seconde partie présente les résultats obtenus sur la reconstruction du boson Z^0 et l'identification des électrons.

1 Contexte scientifique

1.1 - Le Modèle Standard de la physique des particules

Le cadre théorique actuellement utilisé pour décrire les constituants fondamentaux de la matière et leurs interactions est le Modèle Standard de la physique des particules. Il répartit les particules élémentaires en deux catégories: les fermions (spin demi-entier), qui composent la matière, et les bosons (spin entier), qui véhiculent les interactions fondamentales (voir Figure 1). Chaque particule possède une antiparticule de même masse mais de charge opposée. Les fermions se divisent en quarks et leptons, organisés chacun en trois générations. La première, composée des quarks up et down, de l'électron et de son neutrino, constitue la matière ordinaire. Les deux autres regroupent des particules plus lourdes et instables. Les interactions sont assurées par des bosons: le photon (interaction électromagnétique), les gluons (interaction forte), les bosons $W^{\pm}$ et Z^0 (interaction faible). Le boson de Higgs (H^0), découvert en 2012 au LHC [4], explique l'origine de la masse des particules. Bien que très bien validé, le Modèle Standard reste incomplet: il ne prend pas en compte la gravitation, la matière noire ni l'énergie noire, qui dominent pourtant l'univers observable.

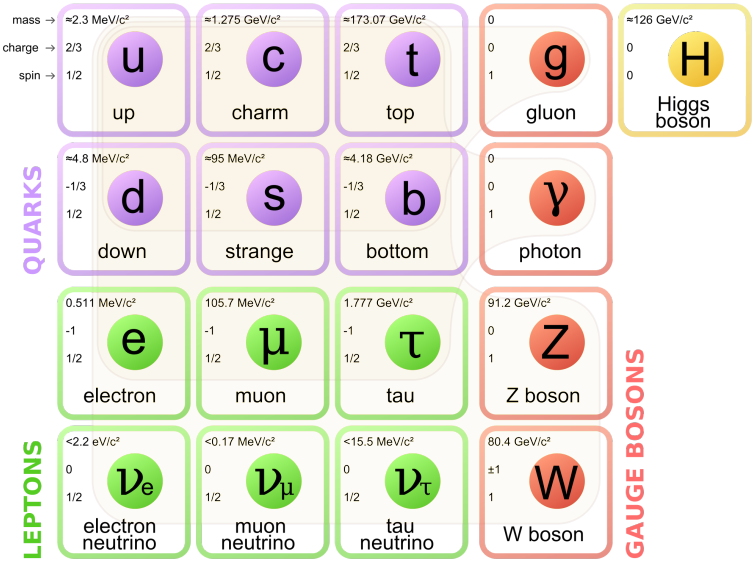


Figure 1: Particules du Modèle Standard [14]

Ce stage porte sur l'étude du boson Z^0, découvert en 1983 au CERN. Il possède une masse d'environ 91.2 GeV/c^2 et une durée de vie extrêmement courte ($\sim 10^{-25}$ s) [8], ce qui empêche sa détection directe. On l'analyse donc via ses produits de désintégration. Le canal étudié est $Z^0 \to e^+e^-$, dont le rapport d'embranchement est de 3.36% [8]. Ce mode est particulièrement intéressant car il permet d'évaluer la capacité du détecteur à reconstruire fidèlement les électrons, une étape cruciale pour l'ensemble des analyses réalisées avec celui-ci.

1.2 - Le Large Hadron Collider et le détecteur ATLAS

L'expérience ATLAS (A Toroidal LHC ApparatuS) est une collaboration internationale regroupant plus de 6000 scientifiques issus de 257 instituts répartis dans 45 pays. Elle repose sur l'exploitation des collisions de protons à très haute énergie produites par le LHC, le plus puissant accélérateur de particules jamais construit, en service depuis 2008 [5]. Cet accélérateur circulaire de 27 km de circonférence permet des collisions jusqu'à une énergie de $\sqrt{s} = 13.6$ TeV. ATLAS est l'un des quatre grands détecteurs installés autour du LHC, en fonctionnement depuis 2009 [3]. Il s'agit d'un détecteur généraliste de 25 m de diamètre pour 46 m de long, pesant plus de 7000 tonnes [7]. Il entoure le point de collision et permet de reconstruire les particules produites.

ATLAS est composé de plusieurs couches successives (voir Figure 2a): le détecteur interne (ou trajectographe), placé autour du point d'interaction, mesure les trajectoires et les impulsions des particules chargées. Le calorimètre électromagnétique, mesure l'énergie des photons et électrons via la formation de gerbes électromagnétiques. Il est entouré du calorimètre hadronique, dédié à la détection des hadrons (comme les protons et neutrons), qui y déposent leur énergie sous forme de gerbes hadroniques. La position de ces gerbes dans les calorimètres est enregistrée et associée aux traces observées dans la trajectographe. Enfin, les chambres à muons, en périphérie du détecteur, sont réservées à la détection des muons, capables de traverser toutes les couches précédentes sans être arrêtés.

Dans le cadre de ce stage, l'analyse a porté sur la détection et l'identification des électrons, principalement à partir des données issues du trajectographe et du calorimètre électromagnétique.

1.3 - Outils utilisés

Chaque collision entre deux protons dans le détecteur ATLAS est appelée un événement. L'objectif est d'identifier les particules produites par cette collision (électrons, muons et jets) afin de reconstruire certains objets physiques. On cherche à sélectionner des paires d'électrons ou de positrons (dans la suite, le terme "électron" désignera indifféremment les deux charges), et à calculer leur masse invariante.

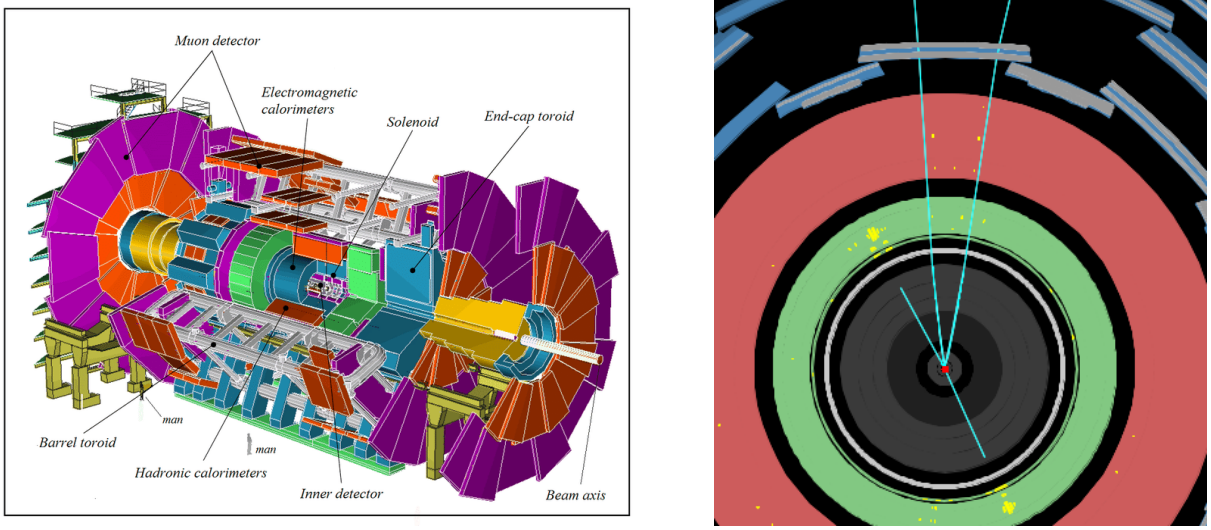
Lorsqu'un boson Z^0 se désintègre, on peut reconstruire sa masse à partir de l'énergie et de l'impulsion de ses produits de désintégration, mesurées par le détecteur. Cela repose sur la relation entre masse, énergie et quantité de mouvement en relativité restreinte (1). Par conservation du quadri-vecteur énergie-impulsion, on obtient la masse invariante du système formé par les deux leptons issus de la désintégration (2).

$$E^2 = (pc)^2 + (mc^2)^2 \quad (1) \qquad m_Z = \sqrt{\left(\frac{E_{e^+} + E_{e^-}}{c^2}\right)^2 - \left(\frac{\vec{p}_{e^+} + \vec{p}_{e^-}}{c}\right)^2} \quad (2)$$

Durant la première semaine du stage, j'ai reconstruit le boson Z^0 à l'aide du logiciel HYPATIA, développé par la collaboration ATLAS. Il permet de visualiser les événements enregistrés

par le détecteur, superposés à une représentation simplifiée de celui-ci (voir Figure 2b). Il est principalement utilisé dans un cadre pédagogique, par exemple lors des Master Classes du CERN [6], et permet d'identifier les particules à partir d'une analyse visuelle. Un lot de 1000 événements enregistrés par ATLAS en 2012 a été utilisé pour cette première approche, où l'on sélectionne les objets reconstruits afin d'accéder à leurs caractéristiques (énergie, impulsion, etc.) et le logiciel reconstruit la masse invariante.

À partir de la deuxième semaine, l'analyse a été étendue à des jeux de données plus volumineux, issus soit de données expérimentales réelles d'ATLAS [2], soit de simulations Monte Carlo. Ces données ont été traitées à l'aide de JupyterHub [9], une plateforme web interactive basée sur des notebooks Python. Ce nouvel environnement a permis de passer à une analyse plus systématique, notamment via l'utilisation d'algorithmes de machine learning de la bibliothèque scikit-learn [11], intégrés pour étudier les performances d'identification des électrons.



(a) **(b)**

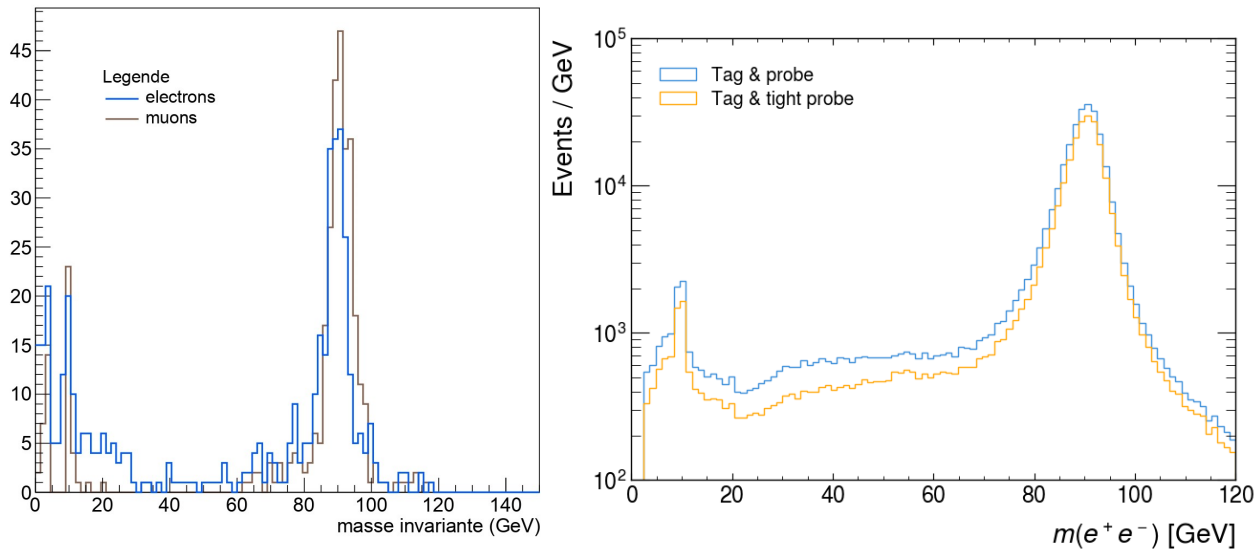
Figure 2: (a) Schéma des différents détecteurs dans ATLAS [13] et (b) Évènement reconstruit dans le logiciel HYPATHIA: $J/\psi + Z^0 \rightarrow \mu^+\mu^- e^+ e^-$, avec en gris le trajectographe, en vert le calorimètre électromagnétique, en rouge le calorimètre hadronique, et en bleu les chambres à muons

2 Travail réalisé et analyses

2.1 - Le boson Z^0 dans des données d'ATLAS

Le lot de données à disposition la première semaine a été pré-filtré manuellement par ATLAS pour qu'il contienne majoritairement des évènements incluant des résonances intéressantes à étudier. Il s'agit du J/ψ [$m_{J/\psi} \approx 3.1$ GeV/c^2], qui est un méson ($c\bar{c}$), de l'Υ [$m_\Upsilon \approx 9.5$ GeV/c^2], un méson ($b\bar{b}$), et le Z^0 [$m_{Z^0} \approx 91.2$ GeV/c^2] [8]. Mon travail a consisté à observer chaque évènement et à classifier visuellement les traces et dépôts d'énergie comme étant des

électrons, des muons ou des photons issus de la désintégration de ces particules, en utilisant leurs interactions caractéristiques dans les détecteurs. Sur la Figure 3a, on voit un histogramme des masses invariantes reconstruites par le logiciel à partir de ma sélection de paires d'électrons et de muons. On observe trois pics aux masses mentionnées plus haut. On peut noter que la distribution pour les électrons a des pics plus larges et que l'histogramme est rempli à des valeurs où il n'y a pas de résonances. Les électrons sont en effet plus difficiles à identifier que les muons, leurs dépôts d'énergie dans le calorimètre étant pas toujours réguliers, et des jets hadroniques de faible énergie pouvant y laisser des traces similaires.



(a) **(b)**

Figure 3: Histogrammes de masse invariante : (a) Des leptons reconstruits par le logiciel HYPATHIA sélectionnés, et (b) Des électrons issus des données d'ATLAS, où seulement un (en bleu) ou les deux (en jaune) ont passé le filtre "tight".

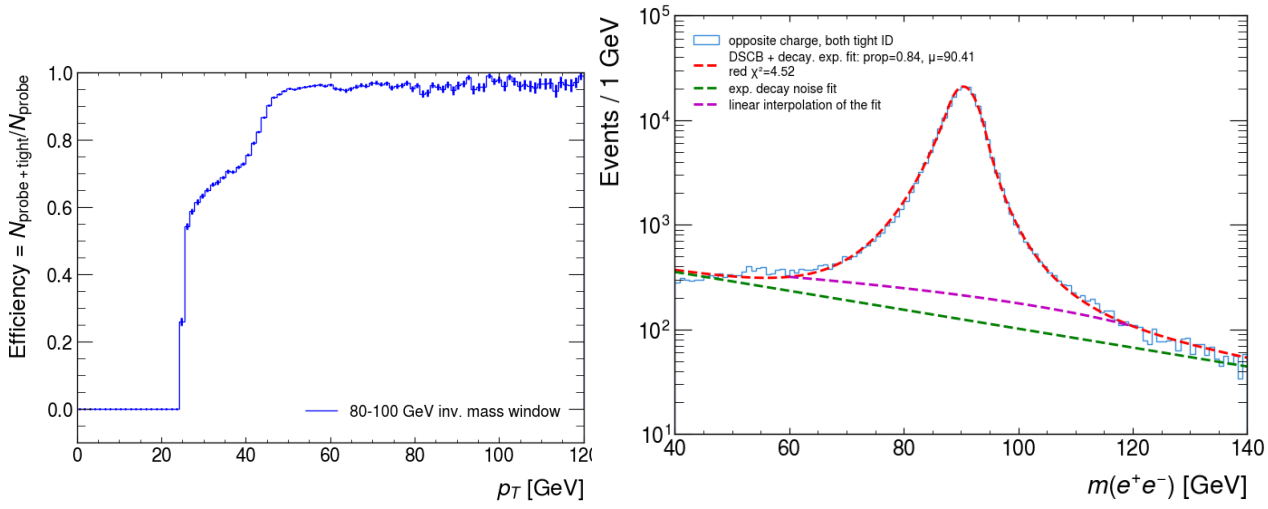
Avec le lot de données de la deuxième semaine, on a accès à des tableaux contenant des milliers d'évènements avec deux leptons ou plus dans l'état final, également filtrés pour inclure majoritairement des résonances, mais avec du bruit de fond aussi présent. Pour chaque évènement, les variables associées étaient l'impulsion, l'énergie, la charge, le type de lepton, ainsi que d'autres variables [2], notamment une variable appelée "is_tight" indiquant si le lepton a été identifié comme tel par les algorithmes de la collaboration. Ceux-ci utilisent des critères stricts tels que la forme étroite du dépôt d'énergie ou la compatibilité des informations de la trace et du dépôt d'énergie dans le calorimètre. Cela permet normalement d'éliminer les objets reconstruits comme électrons mais qui sont en réalité des hadrons légers, et donc de réduire le bruit de fond.

La Figure 3b montre deux histogrammes, construits à partir de l'ensemble d'évènements en sélectionnant des paires d'électrons de charges opposées dont au moins l'un des deux a passé le filtre "tight". On voit nettement le pic autour de la masse du Z^0 et un autre au niveau de l'Υ. Entre ces pics, on voit une population ne correspondant pas à des résonances, que nous avions

déjà vu dans la Figure 3a et qui correspond à du bruit de fond. La figure jaune montre un sous-ensemble de ces évènements, pour lesquels cette fois-ci les deux électrons ont passé le filtre "tight". On remarque que dans la région du pic du Z^0 et de l'Υ, une grande partie des deux candidats électrons passent les critères, alors que dans les régions intermédiaires la proportion est plus faible, car ces objets en majorité ne sont pas des électrons.

2.2 - Méthode de tag & probe, calcul d'efficacité de détection des électrons

À ce stade, la reconstruction du boson Z^0 peut être menée. L'enjeu est de réaliser un ajustement ("fit") sur l'histogramme de la Figure 3 pour retrouver la masse de la particule[1] et le nombre d'événements de signal (paires de vrais électrons) et de bruit de fond (combinaisons aléatoires d'objets autres que des électrons). On utilise ainsi le Z^0 comme **chandelle étalon**. Comme on connaît bien cette particule et son canal de désintégration en di-électrons, on peut l'utiliser pour évaluer comment le détecteur reconstruit les électrons. Nous savons que dans la région du pic de masse invariante, deux électrons ont été émis et devraient être fidèlement reconstruits par le détecteur. Mais il y a aussi du bruit de fond comme nous l'avons vu précédemment.



(a) **(b)**

Figure 4: (a) Efficacité de détection d'électrons en fonction de leur impulsion transverse dans la fenêtre de masse invariante du boson Z^0, et (b) histogramme de masse invariante des paires d'électrons "tight" avec un fit utilisant une Double Sided Crystal Ball pour le signal et un bruit de fond exponentiel ou interpolé linéairement

On a donc effectué un ajustement à l'aide d'une fonction combinant une Double Sided Crystal Ball [15] pour le signal et une exponentielle décroissante pour le bruit de fond. La proportion de signal dans la fenêtre fittée est paramétrée par "prop", visible sur la Figure 4b. Grâce à

[1]Idéalement on pourrait retrouver aussi sa largeur intrinsèque (Γ), mais en raison de processus liés à la QCD, le bruit de fond est trop important. Une reconstruction des caractéristiques du boson Z^0 a été réalisée avec grande précision dès les années 1990 dans des collisionneurs comme le LEP

ce fit, on obtient une estimation de la masse du Z^0, reportée sur le graphique: $m_{Z^0} \approx 90,41$ GeV/c^2, légèrement inférieure à la valeur théorique. D'autres fonctions ont été testées mais donnaient de moins bons résultats.

Par ailleurs, il arrive que certains vrais électrons ne soient pas catégorisés comme "tight" si le détecteur n'est pas bien réglé. Pour déterminer quelle proportion de vrais électrons sont correctement catégorisés, on utilise la méthode dite "tag & probe". Parmi tous les événements dans une fenêtre de masse invariante, on sélectionne ceux contenant au moins un électron "tight", qui sera appelé le "tag". L'autre électron de la paire, du fait qu'elle provient d'une désintégration du Z^0, est également un vrai électron. Celui-ci est notre "probe" (ou "sonde" en français), et on compte combien d'entre eux ont aussi passé le filtre "tight", pour ensuite calculer $\frac{N_{\text{probe+tight}}}{N_{\text{probe}}}$, ce qui définit **l'efficacité** de détection des électrons par notre détecteur.

En moyenne l'efficacité de détection des électrons est de $\epsilon_t = 81.5 \pm 0.1\%$ [10]. Dans la Figure 4a, on a calculé cette efficacité en fonction de l'impulsion transverse du "probe". Les électrons ayant un $p_T \geq 50$ GeV sont identifiés avec une efficacité de 95% en moyenne, quand ceux ayant un p_T inférieur sont identifiés avec une efficacité beaucoup plus faible. On a ensuite essayé une méthode équivalente pour calculer l'efficacité moyenne de détection des électrons: après une intégration du fit réalisé sur l'histogramme des paires "tag & probe" pour obtenir un nombre d'événements, et une sur le fit de l'histogramme des paires "tag & tight probe" (Figure 4b), l'efficacité a été calculée selon la formule définie plus haut, qui donne $\epsilon_t = 81.5 \pm 0.1\%$. Les résultats des deux méthodes sont statistiquement compatibles, ce qui confirme la validité du fit effectué sur le pic.

L'ajustement permet aussi d'estimer le nombre d'évènements de bruit de fond, et donc d'évaluer les efficacités en le soustrayant. La fonction d'ajustement inclut une composante de bruit sous forme exponentielle décroissante, mais comme on peut le voir sur la Figure 4, ce bruit modélisé sous-estime le bruit réel (dans les régions hors du pic du Z^0, on observe encore des événements de signal, alors qu'on ne devrait pas). Une autre manière de modéliser ce bruit est de définir manuellement la fenêtre correspondant au Z^0, et d'utiliser une interpolation linéaire de la distribution comme estimation du bruit. En retirant ce bruit de fond du fit, il ne reste que les événements réellement issus de la désintégration du boson, ce qui devrait permettre, en théorie, d'améliorer l'efficacité. Après soustraction du bruit exponentiel, l'efficacité devient $\epsilon_s = 81.7 \pm 0.1\%$, et après retrait du bruit interpolé, elle est de $\epsilon_s = 81.8 \pm 0.1\%$. En retirant le bruit, on gagne donc, dans le meilleur des cas, 0.5% d'efficacité, sans réelle différence entre les deux méthodes utilisées pour estimer ce bruit.

2.3 - Le boson Z^0 dans des données simulées et étude des variables discriminantes

Pour cette partie, on a utilisé des données simulées représentant à la fois les collisions entre protons et le fonctionnement du détecteur, le tout produit via une méthode de type Monte-Carlo. Ces données proviennent d'un générateur qui simule des collisions proton-proton à

13 TeV ne produisant que des événements de type $Z^0 \to ee$. Une fois ces événements générés, les listes de particules sont injectées dans une simulation complète du détecteur ATLAS, qui construit un modèle virtuel (en tenant compte des sous-détecteurs, des matériaux, de leur disposition, etc.) et simule la réponse de chaque composant au passage des particules. Les événements simulés contiennent une information dite "verité" indiquant entre autres la vraie nature de chaque particule, et cela permet de savoir si celle-ci est un "vrai" électron ou pas. La Figure 5 montre la distribution de masse invariante de toutes les paires de candidats électrons et celle obtenue en utilisant les "vrais" électrons. On observe un pic centré autour de 90 GeV, correspondant à la masse attendue du boson Z^0. Autour de ce pic, on remarque un continuum, qui provient d'événements de bruit de fond. Ce bruit peut être dû à des paires d'électrons qui ne proviennent pas d'une même particule, ou à des objets mal reconstruits, par exemple des hadrons dont la signature a été interprétée à tort comme celle d'un électron.

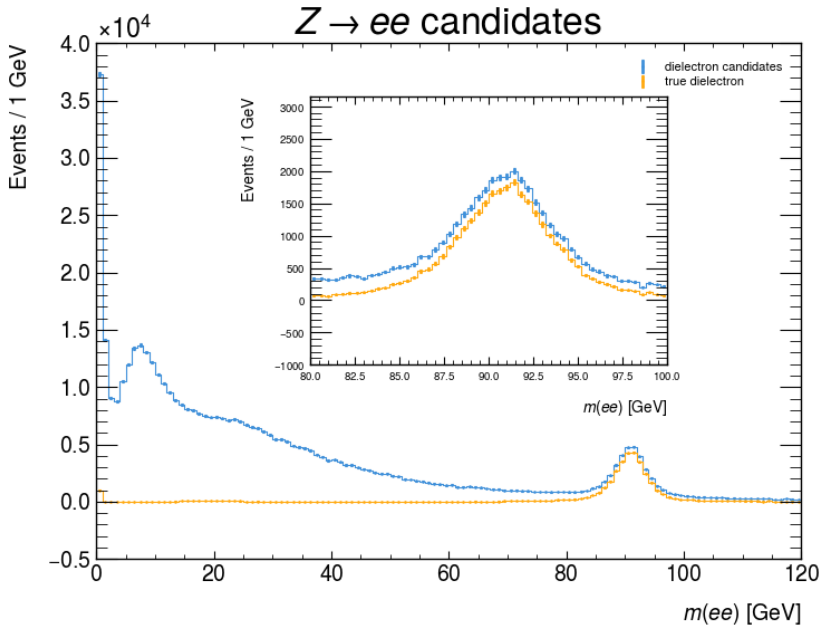
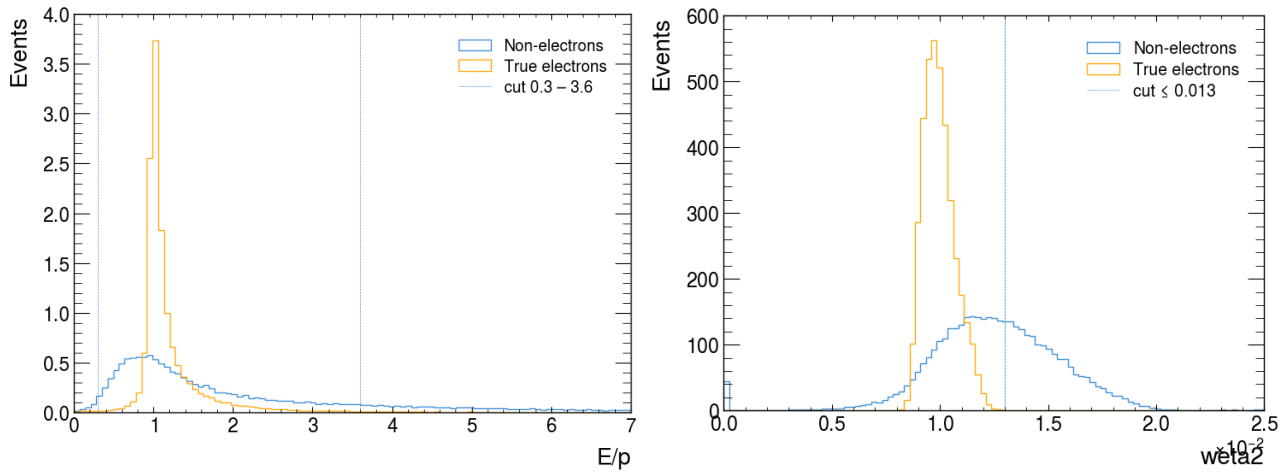


Figure 5: Distribution de la masse invariante de paires d'électrons pour tous les candidats (bleu) et pour les "vrais" électrons (jaune)

Une étape supplémentaire après la reconstruction consiste à identifier les électrons. Un électron ne se manifeste pas seulement par une trace associée à un amas, mais cet amas est généralement fin, concentré, et ne dépasse pas dans le calorimètre hadronique. Les données disponibles contiennent des informations permettant de décrire le développement énergétique de l'amas. L'objectif est alors d'exploiter les caractéristiques des gerbes, dont certaines variables associées sont particulièrement discriminantes, et on peut donc appliquer des coupures dessus.

Les variables sur lesquelles on applique des coupures sont les suivantes: E/p, R_{had}, R_η, $\omega_{\eta 2}$, et ω_{tot1}. L'objectif est de garder un maximum de vrais électrons tout en éliminant un maximum de candidats issus du bruit de fond. E/p est le rapport entre l'énergie déposée dans le calorimètre électromagnétique et l'impulsion mesurée dans le trajectographe. Il est centré autour de 1 pour

les électrons, qui sont généralement ultrarelativistes et de faible masse. R_{had} mesure la fraction d'énergie déposée dans le calorimètre hadronique. Elle doit être très faible pour un vrai électron, qui n'interagit presque pas avec le calorimètre hadronique. R_η est le rapport entre l'énergie dans la zone centrale d'un amas et celle dans une zone plus large. Un électron, qui dépose son énergie de manière concentrée, aura une valeur proche de 1. $\omega_{\eta 2}$ correspond à la largeur latérale de la gerbe dans le second compartiment du calorimètre électromagnétique. Elle doit être faible pour les électrons, dont la gerbe est étroite. ω_{tot1} quantifie l'extension latérale de la gerbe dans le premier compartiment du calorimètre, et doit aussi avoir une petite valeur. Des exemples des coupures décidées visuellement à partir des distributions sont montrées dans la Figure 6, et les valeurs de toutes les coupures sont regroupées dans le Tableau 1.



(a) **(b)**

Figure 6: Distribution normalisée de variables discriminantes pour des électrons (jaune) et du bruit de fond (bleu), avec des coupures. (a) Rapport de l'énergie et de l'impulsion du candidat électron (b) Largeur du dépôt d'énergie dans le second compartiment du calorimètre électromagnétique

Coupure
$p_T \geq 5$ GeV
$E/p \in [0.3, 3.6]$
$R_{had} \leq 0.04$
$R_\eta \geq 0.9$
$\omega_{\eta 2} \leq 0.013$
$\omega_{tot1} \leq 3.2$

Table 1: Coupures appliquées sur les variables de sélection des électrons

$$\epsilon_e = \frac{N_e^{\text{selec}}}{N_e^{\text{tot}}}$$

$$r_b = \frac{1}{\epsilon_b} = \frac{N_b^{\text{tot}}}{N_b^{\text{selec}}}$$

Résultat
$\epsilon_e = 96.8 \pm 0.1\%$
$r_b = 23 \pm 0.5$
$1 - \epsilon_b = 95.6 \pm 0.1\%$

Table 2: Évaluation des sélections

On définit l'efficacité d'identification des électrons (ϵ_e), le facteur de rejet du bruit de fond (r_b), ainsi que la proportion de bruit de fond qu'on élimine ($1 - \epsilon_b$). Pour les coupures selectionnées,

on trouve les valeurs du Tableau 2. Il convient de noter que la coupure sur p_T a été faite au préalable, car il y avait un petit groupe de vrais électrons de très bas p_T qui donnaient des valeurs de E/p non physiques à cause des limites de résolution du détecteur, et donc n'est pas comptée dans le calcul de l'efficacité. L'efficacité de sélection des électrons est de près de 97% pour un facteur de rejet du bruit de fond d'environ 23, éliminant près de 96% de celui-ci.

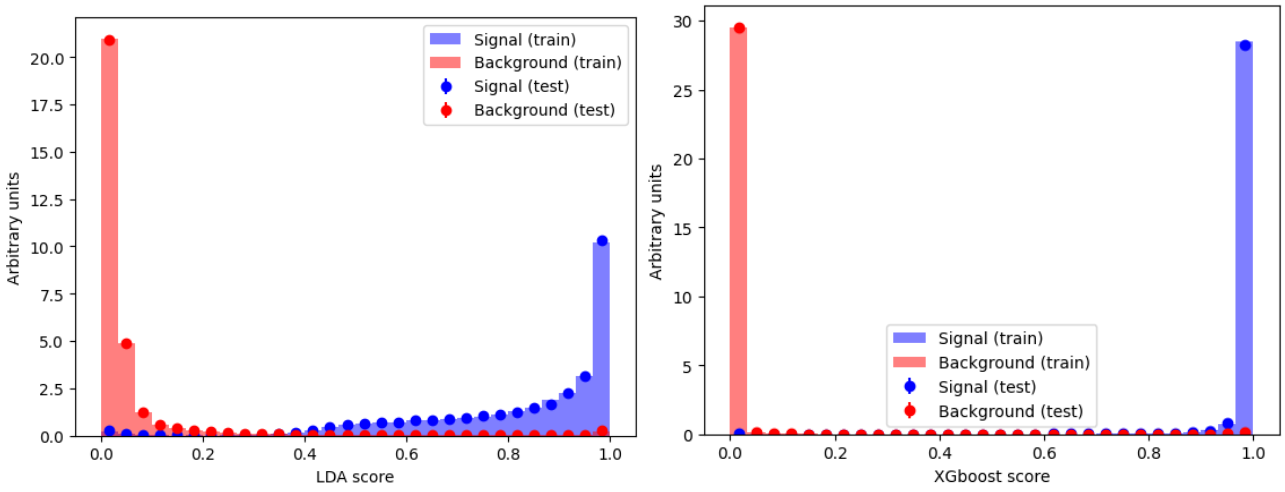
2.4 - Incorporation du Machine Learning

Afin d'évaluer si l'on peut améliorer les performances de l'identification des électrons, nous avons exploré l'utilisation d'algorithmes de Machine Learning, en particulier ceux destinés à des tâches de classification. Ces algorithmes, disponibles dans la bibliothèque scikit-learn [11], ont pour objectif de déterminer automatiquement des critères de sélection optimaux à partir des données. Le principe général est le suivant: le lot de données est divisé en deux sous-ensembles, un échantillon d'entraînement et un échantillon de test. Le modèle apprend à distinguer les vrais électrons du bruit de fond en analysant les relations entre les variables discriminantes associées à chaque événement. Une fois entraîné, le modèle est appliqué à l'échantillon de test pour prédire le statut de chaque électron en utilisant les règles qu'il a apprises, et les performances ϵ_e et r_b sont mesurées [1].

Plusieurs modèles ont été testés au cours du stage. Gaussian Naive Bayes (GNB, ou Likelihood) [1]: Ce modèle suppose que chaque variable suit une distribution gaussienne, indépendante des autres. Il calcule pour chaque candidat la probabilité d'appartenir à la classe "électron", en prenant le produit des probabilités conditionnelles associées à chacune de ses variables. C'est une méthode simple, rapide, mais limitée par ses hypothèses fortes, notamment l'absence de corrélation entre les variables et le fait que les distributions ne sont pas gaussiennes mais ont de grandes queues. Discriminant de Fisher ou LDA [1]: Cette méthode projette les données dans un espace de dimension réduite (au plus $n_{\text{classes}} - 1$), de façon à maximiser la séparation entre les classes. Elle repose sur des combinaisons linéaires des variables d'entrée, ce qui la rend efficace si les distributions sont gaussiennes et les corrélations entre variables sont linéaires. Arbres de décision [1]: Ces modèles fonctionnent par divisions successives du jeu de données selon des coupures sur les variables. À chaque nœud de l'arbre, une condition est testée, et les données sont envoyées dans des sous-ensembles de plus en plus homogènes, jusqu'à obtenir des feuilles contenant essentiellement une seule classe. Pour améliorer les performances des arbres de décision, on peut combiner plusieurs arbres faibles à l'aide de techniques de boosting. Le modèle AdaBoost construit des arbres successifs en accordant plus de poids aux erreurs des arbres précédents. Des variantes plus performantes comme le Gradient Boosting Decision Trees (GBDT) ou XGBoost améliorent la stabilité et la précision, notamment en contrôlant

le sur-apprentissage[2] et en exploitant mieux les corrélations entre variables (corrélations non linéaires).

Les performances des différents algorithmes peuvent être comparées à travers les distributions des scores qu'ils attribuent aux électrons, comme illustré sur la Figure 7. Ces scores représentent la probabilité estimée qu'un candidat soit un vrai électron. Sur les histogrammes, une bonne séparation entre les distributions du signal (vrais électrons) et du bruit de fond indique une capacité de classification efficace. On observe que la méthode XGBoost (Figure 7b) produit des distributions bien distinctes, tandis que le LDA (Figure 7a) donne des distributions plus proches et plus diffuses. Cela montre que XGBoost distingue mieux les deux populations. Par ailleurs, la similarité entre les distributions pour l'échantillon d'entraînement et celui de test traduit une bonne généralisation du modèle et une absence de sur-apprentissage.



(a) **(b)**

Figure 7: Score attribué par les algorithmes de Machine Learning sur des candidats électrons pour (a) Méthode du discriminant de Fisher (b) Arbre de décision boosté (XGBoost)

Une autre façon d'évaluer les performances des modèles est d'utiliser la courbe ROC (Receiver Operating Characteristic), qui montre, pour différentes coupures sur le score, l'efficacité du signal en fonction du taux de rejet du bruit de fond. Un bon modèle produit une courbe proche du coin supérieur droit, avec une aire sous la courbe (AUC) proche de 1. Sur la Figure 8, on observe que GNB et LDA offrent des performances similaires, avec LDA atteignant une meilleure efficacité pour le même taux de rejet. Mon choix de coupures, définies dans la Tableau 2, est proche du "genou" de ces courbes, la où le compromis efficacité/rejet est le meilleur. Il est indiqué par une croix sur la Figure 8. Ces deux algorithmes fonctionnent en fait sur le principe des coupures, donc il est attendu que leurs courbes se ressemblent. Cependant, on voit qu'ils ne sont pas suffisamment performants pour atteindre une efficacité de 100%, du moins sans faire

[2]Il est possible que le modèle essaye de disintiguer les deux classes d'électrons au plus possible sur le lot d'entraînement, provoquant une performance plus faible quan le modèle essaye de faire de même avec les candidats inconnus, donnant un modèle instable

baisser le taux de rejet en dessous d'un seuil acceptable. Leur point de fonctionnement à tout de même servi à définir les critères "tight" dans la collaboration ATLAS jusqu'en 2015 [12]. Un grand pas en avant a pu être fait avec le développement de nouveaux algorithmes, tels que les arbres de décision boostés, bien plus performants. Leurs points de fonctionnement permettent d'atteindre à la fois une efficacité et un taux de rejet supérieurs à 99%. Parmi eux, AdaBoost s'est révélé être le moins performant et le plus lent à entraîner. En revanche, les modèles GBDT et XGBoost donnent des résultats très proches, tout en étant particulièrement bien adaptés à ce type de tâche. Depuis 2015, ce sont d'ailleurs ces algorithmes qui sont utilisés par la collaboration ATLAS pour l'identification des électrons. Différents points de fonctionnement sont choisis en fonction des priorités de l'analyse: par exemple, le critère "loose" vise à conserver un maximum de signal quitte à accepter plus de bruit de fond, tandis que le critère "tight" maximise le rejet du bruit au prix d'une légère perte de signal [12].

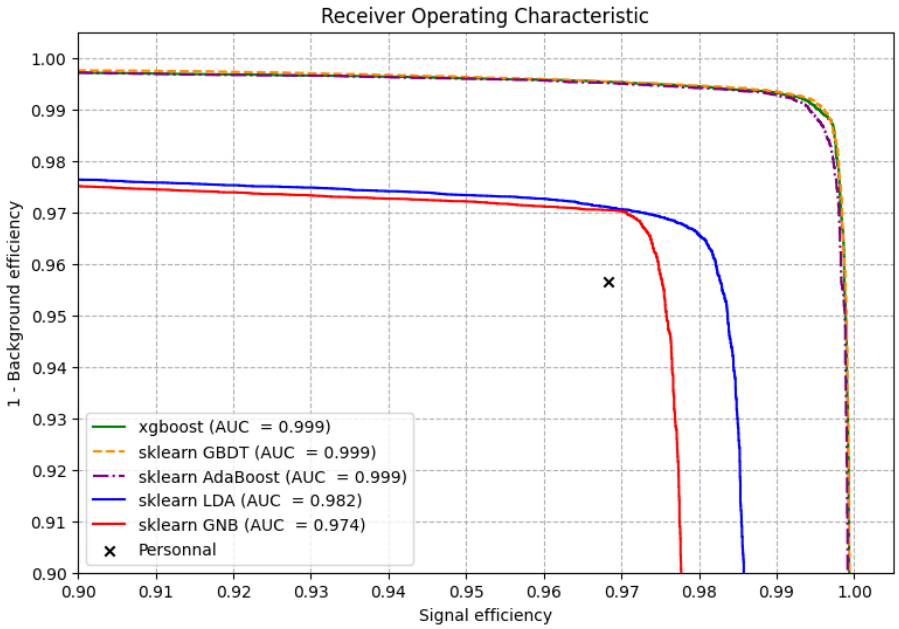


Figure 8: Courbe ROC pour différentes méthodes de Machine Learning, et visualisation des coupures choisies

Finalement, sur toutes les courbes, on remarque qu'un plateau est atteint assez rapidement dès qu'on s'écarte du point de fonctionnement optimal. Cela reflète la taille limitée de nos échantillons: avec trop peu de données, la séparation entre vrais électrons et bruit de fond devient brutale, et cela ne permet pas de trouver les différents points de fonctionnement décrits plus haut. De nombreux seuils de score mènent alors à des performances d'efficacité ou de rejet égales, d'où l'apparition de paliers sur les courbes ROC. On nottera également l'importance de nouveaux algorithmes basées sur des réseaux de neurones, qui n'ont pas pu être testés lors du stage en raison de la taille limitée du lot de données.

Conclusion

Ce stage au LPNHE m'a offert une plongée concrète dans le quotidien d'un travail de recherche en physique des particules. En partant de la problématique de reconstruire le boson Z^0 et identifier des électrons issus de sa désintégration, j'ai progressivement exploré toute une chaîne d'analyse, mêlant physique fondamentale, reconstruction d'événements, statistiques et techniques de Machine Learning.

J'ai d'abord appris à lire les données d'un détecteur complexe comme ATLAS, en observant des événements réels pour y retrouver la signature d'un boson Z^0 à sa désintégration en paire e^+e^-. Cette approche, à la fois intuitive et exigeante, m'a permis de mieux comprendre la logique de sélection des événements, mais aussi les limites de méthodes basées uniquement sur des coupures simples. L'étude de l'efficacité de détection avec la méthode "tag & probe" a ensuite été l'occasion d'aborder une technique classique du domaine.

Le passage à l'étude sur simulation m'a permis de tester plus de variables, d'étudier leurs distributions, et de construire des méthodes de sélection plus efficaces. C'est à cette étape que j'ai commencé à intégrer des outils de Machine Learning pour comparer différentes approches de classification. J'ai ainsi entraîné plusieurs modèles (LDA, GNB, AdaBoost, GBDT, XGBoost), et appris à les évaluer de manière rigoureuse à l'aide de scores, distributions en sortie, et courbes ROC. Cette partie du travail m'a particulièrement intéressé, car elle combine des compétences en physique, en analyse de données et en programmation, et j'ai pu constater de façon très concrète à quel point les outils informatiques modernes (comme les libraires scikit) permettent assez facilement de mettre en oeuvre ces approches qui surpassent en vitesse et en performances celles basées sur des coupures.

Ce stage a aussi été l'occasion de découvrir le fonctionnement d'un grand laboratoire, entre séminaires, réunions d'équipe, discussions informelles et projets en cours. L'environnement au LPNHE est très stimulant, et les échanges que j'ai pu avoir avec les chercheurs du groupe ATLAS m'ont permis d'avoir une vision plus claire de ce que peut être une carrière en recherche. Cela a renforcé ma motivation à continuer dans cette voie, idéalement dans le domaine de la physique des particules.

Je tiens à remercier Frédéric Derue pour son accompagnement et sa bienveillance tout au long de ce stage, ainsi que toute l'équipe du LPNHE pour son accueil chaleureux et les nombreuses discussions qui ont enrichi cette expérience.

Bibliographie

[1] Gareth James et al. *An Introduction to Statistical Learning*. https://www.statlearning.com/. 2023.

[2] ATLAS Collaboration. *Review of the 13 TeV ATLAS Open Data release*. https://cds.cern.ch/record/2707171/files/ANA-OTRC-2019-01-PUB-updated.pdf. 2020.

[3] CERN. *The ATLAS detector*. Consulté le 3 juin 2025. URL: https://greybook.cern.ch/experiment/detail?id=ATLAS.

[4] CERN. *The Higgs Boson*. Consulté le 3 juin 2025. URL: https://home.cern/science/physics/higgs-boson.

[5] CERN. *The LHC*. Consulté le 3 juin 2025. URL: https://home.cern/science/accelerators/large-hadron-collider.

[6] ATLAS Collaboration. *International Physics Masterclasses*. Consulté le 14 avril 2025. URL: https://atlas.physicsmasterclasses.org/fr/index.htm.

[7] Daniel Denegri et al. *L'aventure du grand collisionneur LHC*. EDP sciences, 2014. ISBN: 978-2-7598-0771-0.

[8] Particle Data Group. *2025 Listings and Summary Tables*. Consulté le 3 juin 2025. URL: https://pdglive.lbl.gov/Viewer.action.

[9] Jupyterhub. Consulté le 21 avril 2025. URL: https://jupyter.org/hub.

[10] Louise Heelan. *Calculating Efficiency Uncertainties*. https://indico.cern.ch/event/66256/contributions/2071577/attachments/1017176/1447814/EfficiencyErrors.pdf. 2009.

[11] Machine Learning in Python Scikit-learn. Consulté le 19 mai 2025. URL: https://scikit-learn.org/stable.

[12] Thomas Serre. "Mesure de l'efficacité d'identification des électrons et recherche de SUSY dans le canal 2 leptons avec le détecteur ATLAS". Thèse de doctorat. Centre de Physique des Particules de Marseille, 2014. URL: https://theses.hal.science/tel-01084692.

[13] *The ATLAS detector layout*. Consulté le 3 juin 2025. URL: https://www.researchgate.net/figure/The-ATLAS-detector-layout-40_fig3_289254690.

[14] Standford University. *Elementary particles in Standard Model*. Consulté le 3 juin 2025. URL: https://www.researchgate.net/figure/Fig2-Elementary-particles-in-standard-model-source-Sandford-University-Los-Alamos_fig1_357834237.

[15] z-fit. *Double Sided Crystal Ball definition*. Consulté le 12 mai 2025. URL: https://zfit.readthedocs.io/en/0.6.4/user_api/_generated/pdf/zfit.pdf.DoubleCB.html.