Chapitre X
Reconstruction de la masse d'une résonance grâce au *machine learning*

Sommaire

1 Introduction . . . 2
2 Événements utilisés ou échantillons . . . 3
2.1 Génération avec FastSim . . . 3
2.2 Sélection des événements . . . 4
2.3 Événements obtenus et pondération . . . 5
2.4 Cible et variables d'entrée des modèles . . . 6
3 Concepts communs aux modèles . . . 7
3.1 Fonction de coût . . . 7
3.2 Entraînement, descente de gradient et mini-lots . . . 7
3.3 Sous-entraînement et surentraînement . . . 8
4 Arbres de décision améliorés . . . 9
4.1 Arbres de décision . . . 9
4.2 *Gradient Boosting* et descente de gradient . . . 10
5 Réseaux de neurones profonds . . . 11
5.1 Neurones . . . 11
5.2 Réseaux de neurones . . . 13
5.3 Entraînement . . . 13
6 Optimisation des hyper-paramètres et choix d'un modèle . . . 16
6.1 Variables d'entrée . . . 17
6.2 Type de modèle . . . 19
6.3 Fonction de coût . . . 20
6.4 Algorithme d'optimisation . . . 21
6.5 Autres hyper-paramètres . . . 22
7 Discussions . . . 28
7.1 Effet de l'empilement . . . 28
7.2 Effet de la reconstruction des particules . . . 30
7.3 Effet des faux taus hadroniques . . . 31
7.4 Effet de la séparation des canaux . . . 32
7.5 Effet de la définition de $E_{\mathrm{T}}^{\mathrm{miss}}$ . . . 38
7.6 Effet de l'intervalle de masse . . . 40
7.7 Modèle final . . . 45
8 Utilisation du modèle dans les analyses CMS . . . 45
8.1 Utilisation de m_{ML} comme variable discriminante . . . 45
8.2 Comparaison de m_{ML} à m_{SVFIT} . . . 49
9 Conclusion . . . 54

1 Introduction

L'utilisation de l'intelligence artificielle (IA) s'est grandement développée au cours des dernières années. L'IA regroupe l'ensemble des théories et des techniques développant des programmes informatiques complexes capables de simuler certains traits de l'intelligence humaine tels que le raisonnement et l'apprentissage. L'entreprise Google DeepMind a par exemple développé AlphaGo [1], un programme destiné à jouer au jeu de Go, qui a battu en 2016 le champion du monde de la discipline 4 à 1.

Le *machine learning* (ML) est une branche de l'IA dans laquelle un modèle (algorithme ou programme) s'améliore dans la réalisation d'une tâche par accumulation d'expérience sur des jeux de données d'entraînement, sans pour autant être programmé explicitement pour effectuer cette tâche. Pour y parvenir, les jeux de données d'entraînement comprennent les informations $\{ \vec{x}_i \}$ à donner au modèle ainsi que les « bonnes réponses » $\{ y_{\text{vraie},i} \}$ qu'il doit fournir en sortie. L'objectif est donc d'obtenir un modèle agissant comme une fonction F approximant celle reliant les entrées $\{ \vec{x}_i \}$ aux cibles $\{ y_{\text{vraie},i} \}$. Il peut alors donner une prédiction $y_{\text{préd}}$ sur une nouvelle entrée $\vec{x}$ selon $y_{\text{préd}} = F(\vec{x})$. La tâche du modèle peut être :

une classification lorsque y est discrète, par exemple lorsqu'il s'agit de déterminer si une image représente un chat ou un chien [2] ;

une régression lorsque y est continue, par exemple estimer le prix d'un bien immobilier [3].

Les applications du ML à la physique des particules sont variées et font l'objet de nombreux sujets d'étude [4-7]. Dans les chapitres précédents, le ML est déjà utilisé pour diverses tâches :

— identification des jets issus de quarks b (*b-tagging*) avec DEEPCSV [8-10] ;
— identification des taus hadroniques avec DEEPTAU [11] ;
— catégorisation des événements comme exposé dans le chapitre 4 [12, 13].

Dans les événements $H \to \tau\tau$ présentés au chapitre 1, et plus généralement lors de tout processus physique $X \to \tau\tau$ où une particule X donne une paire de leptons τ, des neutrinos sont émis lors des désintégrations des τ. Or, ils sont invisibles dans les détecteurs tels que CMS ou ATLAS. Il est donc impossible de déterminer la masse invariante totale du système $\tau\tau$ issu de X. Plusieurs méthodes ont été développées afin de reconstruire la masse du système $\tau\tau$ [14-16]. Dans le cadre des analyses $H \to \tau\tau$, la collaboration CMS utilise SVFIT [17], un estimateur de la masse d'une particule se désintégrant en paire de leptons τ par ajustement d'un profil de vraisemblance.

La reconstruction de la masse d'une particule X, ou résonance, se désintégrant en paire de leptons τ grâce au *machine learning* a été étudiée par BÄRTSCHI & coll. [18] dans le cas où X est un boson de Higgs avec une masse comprise entre 80 et 300 GeV. Ils ont obtenu une résolution de 8,4 % sur la masse du Higgs, contre 17 % avec SVFIT. De plus, le temps de calcul nécessaire à l'obtention de la masse est moindre avec le ML. L'utilisation du ML est donc très prometteuse. Cependant, ces travaux utilisent des événements générés avec une simulation grossière du détecteur CMS basée sur DELPHES [19, 20] et sans empilement, notion introduite dans le chapitre 2.

Les travaux présentés dans ce chapitre vont plus loin. La génération des événements, introduite dans la section 2, utilise FASTSIM [21-24] pour modéliser le détecteur CMS. Bien qu'il ne s'agisse pas de la simulation complète basée sur GEANT4 [25-27], FASTSIM est bien plus proche de la réalité que DELPHES. De plus, l'empilement est pris en compte. Les modèles obtenus sont ainsi directement utilisables dans de réelles analyses, telle que celle présentée dans le chapitre 4.

La section 3 introduit des concepts communs aux deux types de modèles étudiés. Les arbres de décision améliorés et les réseaux de neurones profonds sont ensuite introduits dans les sections 4 et 5 respectivement. La comparaison des modèles obtenus et les raisons du choix de l'un d'entre eux sont présentées section 6. Dans la section 7, divers effets sur les performances du modèle sélectionné sont discutés, en particulier la prise en compte de l'empilement. Enfin, l'utilisation en conditions réelles du modèle issu de ces travaux dans des analyses de physique est présentée dans la section 8.

Une note d'analyse [28] est en cours de rédaction. Les scripts utilisés pour la génération des événements peuvent être consultés [29], tout comme ceux permettant d'entraîner les modèles étudiés [30]. Le modèle issu de ces travaux est disponible [31] pour pouvoir être utilisé dans d'autres analyses.

2 Événements utilisés ou échantillons

L'objectif de cette étude est de reconstruire la masse des particules se désintégrant en paire de leptons τ. Il s'agit donc d'une tâche de régression. Dans l'optique d'une utilisation dans les analyses telles que celle présentée dans le chapitre 4, il a été choisi d'utiliser des événements $\mathcal{H} \to \tau\tau$ où $\mathcal{H}$ est le boson de Higgs du modèle standard h dont la masse est modifiée, à l'instar de ce qu'ont fait BÄRTSCHI & coll. [18]. La cible du modèle est donc la masse $m_{\mathcal{H}}$. Un événement est, dans la terminologie du ML, un « échantillon ».

2.1 Génération avec FastSim

Nous avons généré nos propres données simulées [29] afin d'obtenir une distribution continue des valeurs de $m_{\mathcal{H}}$ et suffisamment d'échantillons pour chaque point de masse. Dans le contexte de la collaboration CMS, nous avons utilisé FastSim [21-24]. Cet outil permet de procéder à l'ensemble de la simulation des événements introduite chapitre 2, de la génération du processus initial à la reconstruction des objets physiques par le détecteur.

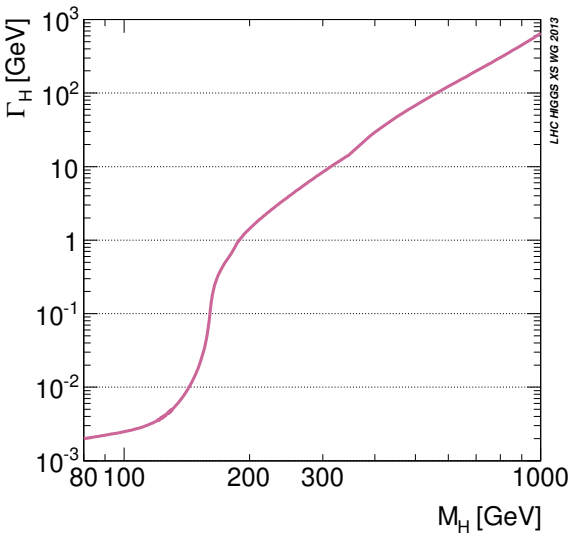
Les données simulées correspondent à des collisions de protons avec une énergie dans le centre de masse de 13 TeV. Les processus physiques sont générés par PYTHIA 8 [32] avec les réglages CUEP8M1 [33, 34]. La production du boson $\mathcal{H}$ se fait par fusion de gluons, il s'agit du mode de production dominant dans le cas du boson de Higgs du modèle standard h. De plus, le rapport de branchement $\mathcal{BR}(\mathcal{H} \to \tau\tau)$ est fixé à 1, c'est-à-dire que $\mathcal{H}$ se désintègre forcément en paires de leptons τ. Tous les événements obtenus sont donc bien du type $\mathcal{H} \to \tau\tau$.

La masse de $\mathcal{H}$ varie de 50 à 800 GeV par pas de 1 GeV. Il est important d'utiliser l'intervalle le plus étendu possible, il correspond à la gamme utile des modèles obtenus. L'effet de l'étendue de cet intervalle est discuté dans la section 7.

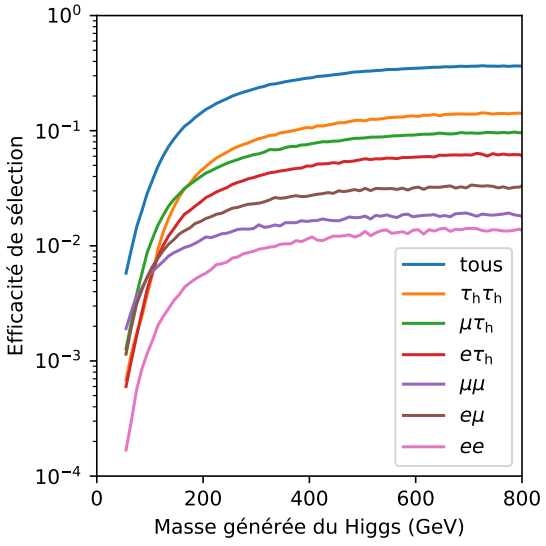
Il n'est pas possible, par la méthode que nous utilisons, de générer des événements avec $m_{\mathcal{H}} \gtrsim 1$ TeV. Cela est dû à la largeur Γ_h de h, représentée figure X.1a en fonction de m_h. La largeur d'une particule est liée à sa durée de vie τ selon

$$\Gamma = \frac{\hbar}{\tau}. \tag{X.1}$$

Ainsi, plus une particule se désintègre rapidement, plus sa largeur est importante. Le principe de Heisenberg mène alors à une incertitude sur la masse, due à la durée de vie τ, égale à Γ. Or, vers



(a) *Largeur du boson de Higgs du modèle standard [35].*



(b) *Efficacité de sélection des événements pour $m_{\mathcal{H}} \in [\![50, 800]\!]$ GeV dans les différents canaux et pour tous les canaux.*

Figure X.1 – *Origine des limites haute (gauche) et basse (droite) de l'intervalle de masse utilisé.*

1 TeV, $\Gamma_{\mathcal{H}} \simeq m_{\mathcal{H}}$. C'est pourquoi la génération de tels événements est compromise avec $\mathcal{H}$ défini comme h avec une masse modifiée. Nous générons donc uniquement des événements en-deçà de 800 GeV.

La sélection des événements est présentée dans la section 2.2. Son efficacité est représentée sur la figure X.1b. Plus de 99 % des événements sont rejetés lorsque $m_{\mathcal{H}} < 50$ GeV. Nous ne considérerons donc pas de masse plus basse.

S'il est possible d'appliquer des poids aux échantillons afin d'équilibrer l'entraînement sur l'ensemble des valeurs de la cible, plus d'événements sont générés à basse masse afin d'obtenir des topologies d'événements variées malgré la faible efficacité de sélection. Ainsi, la quantité d'événements générés pour chaque valeur de $m_{\mathcal{H}}$ est de :

— 60 000 pour 50 GeV $\leqslant m_{\mathcal{H}} <$ 300 GeV ;
— 20 000 pour 300 GeV $\leqslant m_{\mathcal{H}} <$ 500 GeV ;
— 10 000 pour 500 GeV $\leqslant m_{\mathcal{H}} \leqslant$ 800 GeV.

L'empilement est modélisé par superposition du signal $\mathcal{H} \to \tau\tau$ à des événements dits de « biais minimum » [32]. Il s'agit d'événements pouvant contenir des interactions dures, mais n'activant pas de chemin de déclenchement. La quantité d'empilement ajoutée à l'événement $\mathcal{H} \to \tau\tau$ suit le profil d'empilement de l'année 2017. Les conditions des collisions simulées sont ainsi identiques à celles de l'année 2017 au LHC.

2.2 Sélection des événements

2.2.1 Canaux $\tau_h\tau_h$, $\mu\tau_h$, $e\tau_h$ et $e\mu$

Dans le cas des canaux $\tau_h\tau_h$, $\mu\tau_h$, $e\tau_h$ et $e\mu$, la sélection des événements se fait comme exposé dans le chapitre 4 pour l'année 2017. Afin d'obtenir un modèle dont les prédictions auront non seulement un sens dans la région de signal, mais aussi dans les régions de contrôle et de détermination, les coupures sur

— $m_T^{(\mu)}$ dans le canal $\mu\tau_h$;
— $m_T^{(e)}$ dans le canal $e\tau_h$;
— D_ζ dans le canal $e\mu$

ne sont pas appliquées. La construction du dilepton est inchangée. La correspondance des objets du dilepton avec ceux ayant activé le chemin de déclenchement n'est pas vérifiée.

2.2.2 Canal $\mu\mu$

Sélection des muons Tout muon respectant les critères listés ci-après est retenu pour jouer le rôle de L_1 ou L_2 dans le dilepton :

— $p_T{}^\mu > 10$ GeV ;
— $|\eta^\mu| < 2.4$;
— paramètres d'impact $d_z < 0{,}2$ cm et $d_{xy} < 0{,}045$ cm ;
— $I^\mu < 0{,}15\, p_T{}^\mu$;
— passer le point de fonctionnement *medium* du *muonID*.

Sélection du dilepton L'événement est retenu à condition qu'au moins une paire $L_1L_2 = \mu\mu$ puisse être construite avec L_1 et L_2 de charges électriques opposées. Il est de plus requis que L_1 et L_2 soient séparés dans le plan (η, ϕ) tel que $\Delta R > 0{,}3$. Si plus d'une paire possible existe dans l'événement, une seule est retenue selon la logique exposée dans le chapitre 4.

Vetos de leptons supplémentaires Les vetos de leptons supplémentaires doivent être respectés, c'est-à-dire que l'événement ne contient pas :

— de muon supplémentaire tel que $p_T{}^\mu > 10$ GeV, $|\eta^\mu| < 2{,}4$, passant le point de fonctionnement *medium* du *muonID* et d'isolation $I^\mu < 0{,}3\, p_T{}^\mu$;

— d'électron tel que ${p_T}^e > 10\,\text{GeV}$, $|\eta^e| < 2{,}5$, passant le point de fonctionnement à 90 % d'efficacité de l'*electron ID MVA* et d'isolation $I^e < 0{,}3\,{p_T}^e$, l'électron devant passer le veto d'électron de conversion et présenter moins de deux points de passage manquants dans le trajectographe.

2.2.3 Canal *ee*

Sélection des électrons Tout électron respectant les critères listés ci-après est retenu pour jouer le rôle de L_1 ou L_2 dans le dilepton :

— ${p_T}^e > 20\,\text{GeV}$;
— $|\eta^e| < 2.4$;
— paramètres d'impact $d_z < 0{,}2\,\text{cm}$ et $d_{xy} < 0{,}045\,\text{cm}$;
— $I^e < 0{,}1\,{p_T}^e$;
— passer le point de fonctionnement à 90 % d'efficacité de l'*electron ID MVA*.

Sélection du dilepton L'événement est retenu à condition qu'au moins une paire $L_1L_2 = ee$ puisse être construite avec L_1 et L_2 de charges électriques opposées. Il est de plus requis que L_1 et L_2 soient séparés dans le plan (η, ϕ) tel que $\Delta R > 0{,}5$. Si plus d'une paire possible existe dans l'événement, une seule est retenue selon la logique exposée dans le chapitre 4.

Vetos de leptons supplémentaires Les vetos de leptons supplémentaires doivent être respectés, c'est-à-dire que l'événement ne contient pas :

— de muon tel que ${p_T}^\mu > 10\,\text{GeV}$, $|\eta^\mu| < 2{,}4$, passant le point de fonctionnement *medium* du *muonID* et d'isolation $I^\mu < 0{,}3\,{p_T}^\mu$;
— d'électron supplémentaire tel que ${p_T}^e > 10\,\text{GeV}$, $|\eta^e| < 2{,}5$, passant le point de fonctionnement à 90 % d'efficacité de l'*electron ID MVA* et d'isolation $I^e < 0{,}3\,{p_T}^e$, l'électron devant passer le veto d'électron de conversion et présenter moins de deux points de passage manquants dans le trajectographe.

2.2.4 Jets

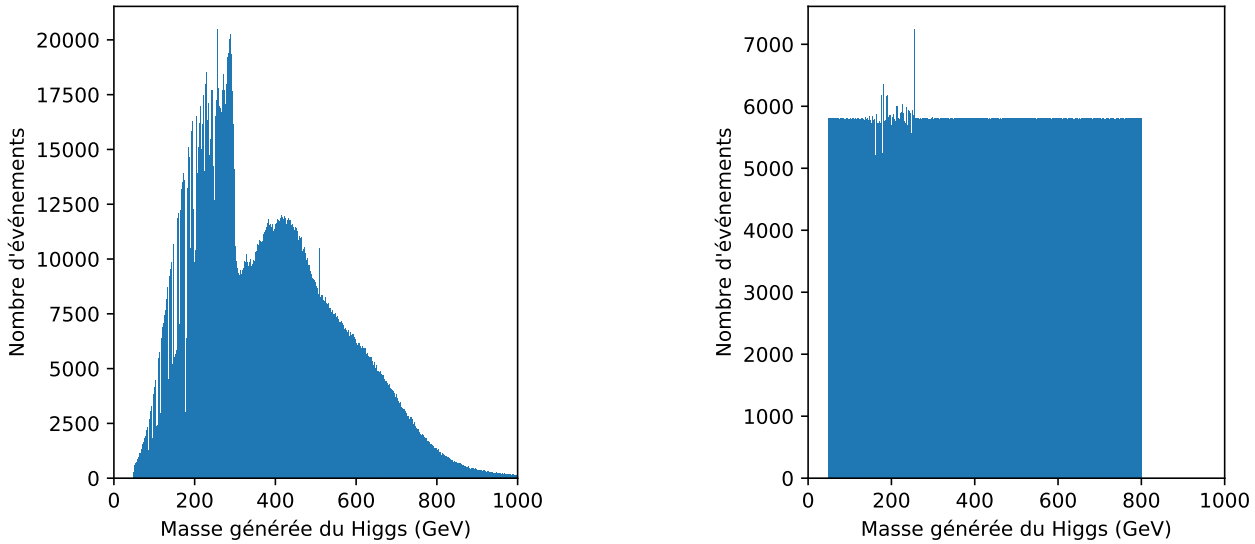
À l'instar de l'analyse présentée au chapitre 4, les jets sont soumis à la procédure de CHS [36] décrite dans le chapitre 3 et reconstruits par l'algorithme anti-k_T [37] avec un paramètre $R = 0{,}4$. Ces jets doivent également passer les critères d'identification présentés dans le chapitre 2. L'identification des jets issus de quarks b (*b-tagging*) est réalisée par l'algorithme DEEPCSV [9, 10]. Les jets tels que $p_T > 20\,\text{GeV}$ et $|\eta| < 2{,}5$ sont considérés comme issus d'un b si leur score est supérieur à 0,3033. Les jets non identifiés comme issus d'un b ne sont retenus que si $p_T > 30\,\text{GeV}$ et $|\eta| < 4{,}7$.

2.3 Événements obtenus et pondération

Environ 3 millions d'événements ont été sélectionnés selon les critères présentés précédemment parmi plus de 22 millions générés. La distribution de $m_{\mathcal{H}}$ dans les événements sélectionnés est représentée sur la figure X.2a. Certains présentent des valeurs de $m_{\mathcal{H}}$ au-delà de 800 GeV, cet effet est dû à la largeur de cette particule, représentée sur la figure X.1a en fonction de sa masse. La largeur à 800 GeV est ainsi d'environ 300 GeV. Le réglage $m_{\mathcal{H}} = 800\,\text{GeV}$ donne donc des événements contenant un boson dont la masse effective se situe entre 500 et 1100 GeV, d'où la queue de la distribution observée à haute masse sur la figure X.2a. À basse masse en revanche, la largeur est inférieure à 100 MeV, cet effet n'est donc pas présent. La cible du modèle est la masse effective du boson. Les événements retenus dans la suite sont ceux où celle-ci se situe bien entre 50 et 800 GeV, d'où la disparition de la queue à haute masse sur la figure X.2b.

Ces événements sont de plus séparés en trois groupes selon les proportions suivantes :

— 70 % pour l'entraînement. Ce sont ces échantillons que les modèles pourront exploiter afin d'apprendre à prédire correctement $m_{\mathcal{H}}$;
— 20 % pour la validation. Ces échantillons permettent de vérifier qu'il n'y a pas de surentraînement, c'est-à-dire que le modèle ne se spécialise pas vis-à-vis du jeu d'entraînement ;



(a) *Distribution brute sur tous les événements.*

(b) *Distribution pondérée pour les événements d'entraînement.*

Figure X.2 – *Distributions de la masse générée de $\mathcal{H}$.*

— 10 % pour les tests. Ces échantillons ne sont pas utilisés lors des entraînements et permettent donc de tester les modèles sur des données inédites. Sauf contre-indication, les figures sont toutes obtenues avec ce groupe d'échantillons.

La répartition des échantillons dans ces trois groupes est faite de manière aléatoire.

Afin de réaliser un entraînement équitable entre les différentes valeurs de $m_{\mathcal{H}}$, un poids est associé à chaque échantillon de manière à ce que la distribution pondérée de $m_{\mathcal{H}}$ soit plate dans chacun des trois groupes précédemment définis. Cette distribution sur les échantillons utilisés pour l'entraînement des modèles est représentée sur la figure X.2b.

2.4 Cible et variables d'entrée des modèles

La cible des modèles est la masse de la particule générée $\mathcal{H}$ se désintégrant en paire de leptons τ. Un tel événement est illustré sur la figure X.3. Les variables d'entrée doivent être des observables accessibles expérimentalement, c'est-à-dire issues de la reconstruction des événements présentée dans le chapitre 2.

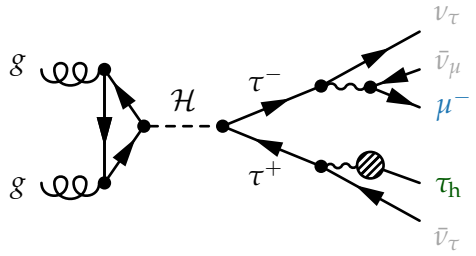


Figure X.3 – *Diagramme de Feynman des événements d'entraînement des modèles dans le cas du canal $\mu\tau_h$.*

Les variables considérées sont :

— les impulsions de L_1 et L_2, les produits de désintégration visibles des τ, c'est-à-dire le muon et le τ_h dans l'exemple de la figure X.3 : ${p_T}^{L_1}$, η^{L_1}, ϕ^{L_1}, ${p_T}^{L_2}$, η^{L_2}, ϕ^{L_2} ;

— l'énergie transverse manquante pour rendre compte de la présence des neutrinos : ${E_T}^{\text{miss}}$, $\phi^{{E_T}^{\text{miss}}}$, obtenus par l'algorithme PUPPI [38] ;

— la matrice M de covariance de ${E_T}^{\text{miss}}$, rendant compte de l'incertitude sur la mesure de ${E_T}^{\text{miss}}$: M_{xx}, M_{xy}, M_{yy} ;

— le nombre attendu de neutrinos lié à l'état final identifié N_ν^{reco}, déterminé à partir du canal obtenu par la sélection des événements, c'est-à-dire sans utilisation des informations générées ;
— les masses transverses $m_{\text{T}}(L_1, {E_{\text{T}}}^{\text{miss}})$, $m_{\text{T}}(L_2, {E_{\text{T}}}^{\text{miss}})$ et $m_{\text{T}}(L_1, L_2)$ définies par

$$m_{\text{T}}(A, B) = \sqrt{2\, {p_{\text{T}}}^A\, {p_{\text{T}}}^B\, (1 - \cos(\phi^A - \phi^B))}\ ; \tag{X.2}$$

— la masse transverse totale $m_{\text{T}}^{\text{tot}}$ définie par

$$m_{\text{T}}^{\text{tot}} = \sqrt{m_{\text{T}}^2(L_1, {E_{\text{T}}}^{\text{miss}}) + m_{\text{T}}^2(L_2, {E_{\text{T}}}^{\text{miss}}) + m_{\text{T}}^2(L_1, L_2)}\ ; \tag{X.3}$$

— les impulsions des deux jets principaux (de plus hauts p_{T}) présents dans l'événement : ${p_{\text{T}}}^{\text{jet 1}}$, $\eta^{\text{jet 1}}$, $\phi^{\text{jet 1}}$, ${p_{\text{T}}}^{\text{jet 2}}$, $\eta^{\text{jet 2}}$, $\phi^{\text{jet 2}}$;
— l'Activité Hadronique Additionnelle (AHA), définie par la somme des impulsions des jets autres que les deux principaux : ${p_{\text{T}}}^{\text{AHA}}$, η^{AHA}, ϕ^{AHA} avec

$$\vec{p}^{\,\text{AHA}} = \sum_{\text{jet } i,\, i>2} \vec{p}^{\,\text{jet } i}\ ; \tag{X.4}$$

— la quantité de jets utilisés pour déterminer $\vec{p}^{\,\text{AHA}}$, $N_{\text{jets}}^{\text{AHA}}$;
— le nombre de vertex principaux d'empilement, N_{PU}.

Des modèles ont été entraînés sur l'ensemble de ces 27 variables ainsi que sur des sous-ensembles de cette liste.

3 Concepts communs aux modèles

3.1 Fonction de coût

Une fonction de coût compare les prédictions d'un modèle aux valeurs vraies de la cible. Elle doit être différentiable et est définie de manière à être minimale lorsque les prédictions sont égales aux valeurs vraies, c'est-à-dire lorsque le modèle est parfait. Les fonctions de coût les plus répandues sont :

MSE *Mean Squared Error* ou erreur quadratique moyenne,

$$\mathsf{L}_{\text{MSE}}(\{\, y_{\text{vraie},i} \,\}, \{\, y_{\text{préd},i} \,\}) = \frac{1}{2N} \sum_{i=1}^{N} \left(y_{\text{préd},i} - y_{\text{vraie},i}\right)^2\ ; \tag{X.5}$$

MAE *Mean Absolute Error* ou erreur absolue moyenne,

$$\mathsf{L}_{\text{MAE}}(\{\, y_{\text{vraie},i} \,\}, \{\, y_{\text{préd},i} \,\}) = \frac{1}{N} \sum_{i=1}^{N} \left|y_{\text{préd},i} - y_{\text{vraie},i}\right|\ ; \tag{X.6}$$

MAPE *Mean Absolute Percentile Error* ou erreur absolue relative moyenne,

$$\mathsf{L}_{\text{MAPE}}(\{\, y_{\text{vraie},i} \,\}, \{\, y_{\text{préd},i} \,\}) = \frac{100}{N} \sum_{i=1}^{N} \left|\frac{y_{\text{préd},i} - y_{\text{vraie},i}}{y_{\text{vraie},i}}\right|\ . \tag{X.7}$$

3.2 Entraînement, descente de gradient et mini-lots

Un modèle peut être vu comme une fonction paramétrique F dont l'application à une entrée $\vec{x}$ donne une prédiction $y_{\text{préd}} = F(\vec{x})$. L'entraînement consiste à régler les paramètres du modèle afin d'obtenir des prédictions fidèles aux valeurs vraies du jeu de données d'entraînement.

La fonction de coût L est minimale lorsque les prédictions du modèle sont parfaites. Il s'agit donc de trouver le minimum de L dans l'espace à D dimensions formé par les $D = N_{\text{params.}}$ paramètres à régler. Cela peut être fait de manière itérative par descente de gradient (GD, *Gradient Descent*) [39].

Il s'agit d'une méthode itérative qui détermine le gradient de L, $\vec{\nabla}(\mathsf{L})$, autour de la « position » du modèle dans l'espace à D dimensions. Chaque paramètre p est alors modifié selon

$$p \to p - \eta \vec{\nabla}\big(\mathsf{L}(\{ y_{\text{vraie},i} \}, \{ y_{\text{préd},i} \})\big) \cdot \vec{e}_p = p - \eta \frac{\partial \mathsf{L}(\{ y_{\text{vraie},i} \}, \{ y_{\text{préd},i} \})}{\partial p} \tag{X.8}$$

avec η le taux d'apprentissage, et $\mathsf{L}(\{ y_{\text{vraie},i} \}, \{ y_{\text{préd},i} \})$ la fonction de coût évaluée sur l'ensemble du jeu de données d'entraînement. Le taux d'apprentissage est généralement pris entre 10^{-5} et 0,5.

Toutefois, l'évaluation du gradient de la fonction de coût sur l'ensemble du jeu de données d'entraînement, contenant éventuellement plusieurs millions d'échantillons comme c'est le cas dans cette thèse, nécessite d'importantes ressources computationnelles (en l'occurrence, de la mémoire vive). Afin de pallier ce problème, la descente de gradient stochastique (SGD, *Stochastic Gradient Descent*) évalue le gradient de L individuellement pour chaque échantillon. Cependant, le SGD amène de fortes fluctuations le long de la descente, ce qui ralenti l'entraînement. De plus, ces mêmes fluctuations une fois proche du minimum mènent à une précision dégradée des prédictions.

Afin d'éviter ce phénomène, le *batch* GD (BGD) [40] évalue le gradient de la fonction de coût par une moyenne sur un « mini-lot » (*mini-batch*), sous-ensemble du jeu de données de taille fixée. Les fluctuations lors de la descente sont moindres que dans le cas de SGD. Cette moyenne introduit un bruit dû à la composition aléatoire des mini-lots qui reste non nul même une fois le minimum de L atteint. Cela permet de s'échapper des minimums locaux, mais dégrade la précision une fois au minimum global.

Une « époque » de l'entraînement correspond à une utilisation de tous les mini-lots, c'est-à-dire de tous les échantillons du jeu de données, pour modifier les paramètres du modèle. Pour ne pas biaiser l'entraînement à cause de l'ordre du jeu de données, il est mélangé aléatoirement à chaque nouvelle époque. La composition des mini-lots est donc également aléatoire. Leur taille est fixée à $2^{11} = 2048$ événements. Une taille de la forme 2^n permet d'optimiser l'utilisation des GPUs (*Graphics Processing Unit*) sur lesquels l'entraînement se fait [41]. Les points de masse générés étant les entiers entre 50 et 800 GeV, soit 750 points de masse, 2048 événements pris au hasard est un compromis entre un petit mini-lot et une bonne probabilité de couvrir une large gamme de masse au sein d'un mini-lot.

3.3 Sous-entraînement et surentraînement

Un modèle doit être suffisamment complexe, c'est-à-dire proposer suffisamment de paramètres réglables à l'entraînement, afin de pouvoir réaliser sa tâche. Dans le cas contraire, ses prédictions ne sont qu'approximatives, voire fausses. Par exemple, en utilisant une droite affine comme modèle, c'est-à-dire avec deux paramètres, il est difficile de prédire correctement une loi polynomiale de degré 2, régie par trois paramètres. Or, le nombre d'itérations d'entraînement nécessaires à l'optimisation d'un modèle augmente avec sa complexité. Avec une quantité limitée d'itérations, le modèle est sous-entraîné. Il est donc nécessaire d'en avoir suffisamment.

Au fur et à mesure des itérations, la valeur de la fonction de coût appliquée au jeu de données d'entraînement diminue. Elle peut donc être un indicateur de l'amélioration des prédictions du modèle d'une itération à une autre. Arrivé à un plateau, le modèle est optimisé et l'entraînement s'arrête.

Cette approche masque toutefois une spécialisation du modèle ou surentraînement. Prédire parfaitement $\{ y_{\text{vraie},i} \}$ sur le jeu de données d'entraînement, ce qui correspond à une fonction de coût nulle, n'est pas équivalent à être optimal sur des données inédites. Or, le but est justement d'utiliser le modèle sur ces dernières. Un autre modèle, moins complexe ou entraîné avec moins d'itérations, peut donc donner de meilleures prédictions sur des données inédites.

Afin d'éviter le surentraînement, il est possible d'utiliser un jeu de données dit de « validation », non utilisé pour régler les paramètres du modèle. L'intérêt du jeu de validation est illustré sur la figure X.4. Un modèle sous-entraîné ou dont l'entraînement est optimal présente des erreurs similaires dans les deux jeux de données. Dans le cas d'un surentraînement, les erreurs continuent à diminuer sur le jeu d'entraînement, mais pas sur le jeu de validation. Une fonction d'évaluation E, éventuellement égale à la fonction de coût L, permet de quantifier ces erreurs et de mettre fin à l'entraînement avant de surentraîner le modèle. Il s'agit de l'arrêt prématuré. La condition activant cet arrêt est

un « hyper-paramètre » du modèle. Les hyper-paramètres, à ne pas confondre avec les paramètres modifiés lors de l'entraînement, sont fixés par l'utilisateur et propres au modèle.

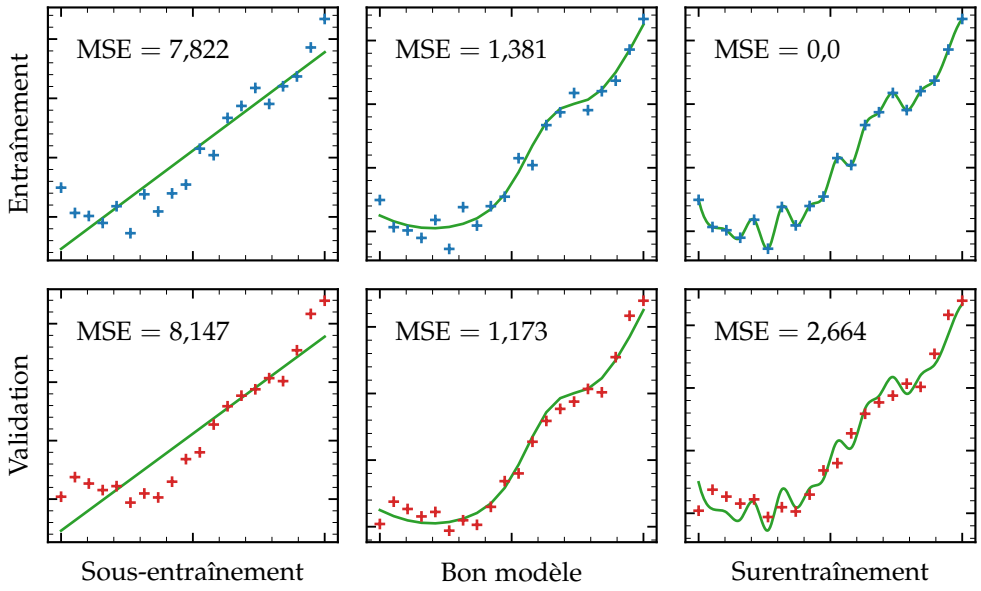


Figure X.4 – *Illustrations du sous-entraînement et du surentraînement. Un même modèle est trop peu (gauche), suffisamment (milieu) ou trop entraîné (droite). Ses prédictions (ordonnées) en fonction de l'entrée (abscisses) sont tracées en vert. Le jeu de données d'entraînement (de validation) est représenté par des croix bleues (rouges) sur la ligne du haut (bas). Les valeurs de* $\mathrm{L}_{\mathrm{MSE}}$ *sont également données afin d'illustrer les variations de l'erreur discutées dans le texte.*

4 Arbres de décision améliorés

La librairie XGBOOST [42] (*eXtreme Gradient Boosting*) permet de construire des arbres de décision améliorés. De nombreuses compétitions Kaggle [43] ont été remportées grâce à eux. Ils présentent l'avantage d'être généralement plus rapides à entraîner que les réseaux de neurones présentés section 5, et peuvent fournir des prédictions même si une des variables d'entrée est manquante, ce qui n'est pas le cas des réseaux de neurones.

4.1 Arbres de décision

Un arbre de décision (non amélioré) est une succession de questions dont les réponses mènent à un résultat final, comme illustré sur la figure X.5. Chaque réponse à une question crée une « branche » menant à une nouvelle question (en bleu) ou à une réponse finale sur une « feuille » (en vert).

De tels arbres peuvent être utilisés avec des variables numériques. Dans ce cas, chaque question consiste en une condition sur l'une des variables, par exemple $p_{\mathrm{T}}{}^{\mu} > 50\,\mathrm{GeV}$. Le choix de la variable ($p_{\mathrm{T}}{}^{\mu}$) et de la coupure correspondante (50 GeV) à utiliser pour former deux nouvelles branches b_1 (condition fausse) et b_2 (condition vraie) se base sur la similarité S. Il s'agit d'une variable quantifiant les écarts d'une liste de valeurs y_i à la moyenne de celles-ci $\langle y \rangle$. Elle est définie comme

$$S = \frac{1}{N}\left(\sum_{i=1}^{N} r_i\right)^2, \quad r_i = y_i - \frac{1}{N}\sum_{j=1}^{N} y_j = y_i - \langle y \rangle \tag{X.9}$$

où N est la taille de la liste de valeurs, r_i le résiduel de y_i et y_i la i^{e} valeur de y.

Le gain G obtenu par la création de deux nouvelles branches b_1 et b_2 s'exprime

$$G = S_{b_1} + S_{b_2} - S_{b_1+b_2} \tag{X.10}$$

avec $S_{b_1+b_2}$ la similarité de la liste non séparée, S_{b_1} (S_{b_2}) la similarité de la liste se retrouvant dans la branche b_1 (b_2). La condition retenue pour former les deux branches est celle présentant le gain

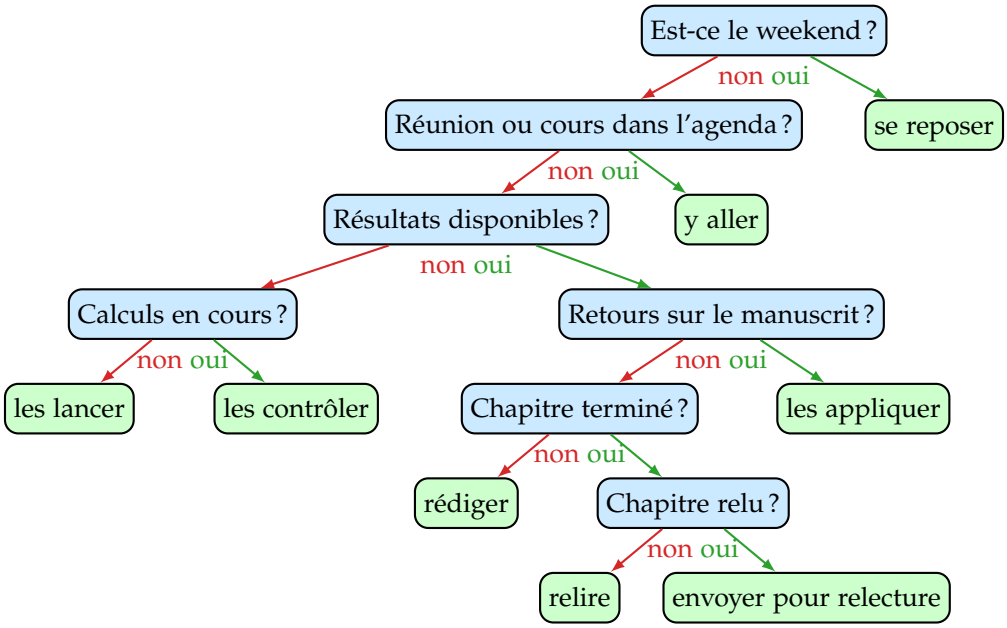


Figure X.5 – *Exemple d'un arbre de décision utilisable par un doctorant.*

le plus élevé. Cela revient à définir deux sous-listes dans lesquelles les valeurs de y sont proches les unes des autres. Ce processus est alors itéré sur chacune des nouvelles branches, jusqu'à ce que :

— le gain soit inférieur à une valeur γ fixée ;

— la profondeur de l'arbre soit égale à une valeur $N_{\text{max}}^{\text{prof.}}$ fixée ;

— la quantité d'échantillons dans une branche soit inférieure à une valeur $N_{\text{min}}^{\text{échant.}}$ fixée.

Les paramètres γ, $N_{\text{max}}^{\text{prof.}}$ et $N_{\text{min}}^{\text{échant.}}$, choisis par l'utilisateur, sont des hyper-paramètres.

4.2 *Gradient Boosting* et descente de gradient

La technique du *Gradient Boosting* est l'utilisation de modèles simples, ici des arbres de décision, pour obtenir un modèle global plus robuste. La construction se fait de manière itérative.

La première étape consiste à créer un arbre de décision, noté M_0, comme exposé dans la section 4.1. La fonction associée à ce modèle est F_0. Puis à chaque étape $k \geqslant 1$, un arbre de décision M_k est construit avec pour objectif de prédire, pour une entrée $\vec{x}_i$,

$$y_{\text{vraie},i} - F_{k-1}(\vec{x}_i) \tag{X.11}$$

avec $y_{\text{vraie},i}$ la valeur que doit prédire le modèle global et F_{k-1} la fonction du modèle issu de l'étape $k-1$. Le modèle M_k corrige donc l'écart résiduel des prédictions $\{ y_{\text{préd},i} \}$ du modèle F_{k-1} à $\{ y_{\text{vraie},i} \}$. Les prédictions F_k du modèle global s'expriment donc

$$y_{\text{préd},i} = F_k(\vec{x}_i) = F_{k-1}(\vec{x}_i) + \eta\, M_k(\vec{x}_i) = M_0(\vec{x}_i) + \eta \sum_{l=1}^{k} M_l(\vec{x}_i) \tag{X.12}$$

avec η le taux d'apprentissage, inférieur à 1, permettant de corriger progressivement l'écart résiduel. L'itération s'arrête lorsque le nombre maximal d'estimateurs $N_{\text{max}}^{\text{estim.}}$ est atteint. Les grandeurs η et $N_{\text{max}}^{\text{estim.}}$ sont également des hyper-paramètres. Le modèle global obtenu est ici un arbre de décision amélioré.

La dérivée partielle de L_{MSE} par rapport à $y_{\text{préd},i}$ s'exprime

$$\frac{\partial \mathsf{L}_{\text{MSE}}(y_{\text{vraie},i}, y_{\text{préd},i})}{\partial y_{\text{préd},i}} = y_{\text{préd},i} - y_{\text{vraie},i}\,. \tag{X.13}$$

Ainsi, la cible de M_k définie précédemment est

$$y_{\text{vraie},i} - F_{k-1}(\vec{x}_i) = -\frac{\partial \mathsf{L}_{\text{MSE}}(y_{\text{vraie},i}, F_{k-1}(\vec{x}_i))}{\partial F_{k-1}(\vec{x}_i)}\,, \tag{X.14}$$

ce qui revient à appliquer la descente de gradient avec $\mathsf{L} = \mathsf{L}_{\text{MSE}}$. À partir de ce constat, il est possible de généraliser le *Gradient Boosting* en considérant que la cible de M_k est

$$-\frac{\partial \mathsf{L}(y_{\text{vraie},i}, F_{k-1}(\vec{x}_i))}{\partial F_{k-1}(\vec{x}_i)} = -\vec{\nabla}_{F_{k-1}(\vec{x}_i)} \left(\mathsf{L}(y_{\text{vraie},i}, F_{k-1}(\vec{x}_i))\right) \tag{X.15}$$

avec L une fonction de coût quelconque.

Une itération de l'entraînement consiste ainsi en l'ajout d'un estimateur au modèle. Un arrêt prématuré est réalisé lorsque l'erreur quadratique moyenne (L_{MSE}) ne diminue pas sur le jeu de validation pendant 5 itérations.

5 Réseaux de neurones profonds

Les réseaux de neurones (NN, *Neural Networks*) sont un autre type de modèle permettant d'approximer la fonction reliant les entrées $\{\vec{x}_i\}$ aux cibles $\{y_{\text{vraie},i}\}$ [41, 44]. Les librairies KERAS [45] et TENSORFLOW [46] sont utilisées afin de construire et d'entraîner ces modèles.

5.1 Neurones

5.1.1 Principe

Un neurone est une entité ayant un certain nombre d'entrées x_j, $j \in \{1, \ldots, n\}$, auxquelles sont associées des poids w_j, un biais b et une fonction f dite d'« activation », discutée section 5.1.2. Les poids w_j et le biais b sont les paramètres du neurone, la fonction d'activation est un hyper-paramètre. La sortie s du neurone s'exprime comme

$$s = f\left(\sum_{j=1}^{n} w_j x_j + b\right). \tag{X.16}$$

Le fonctionnement d'un neurone est résumé sur la figure X.6.

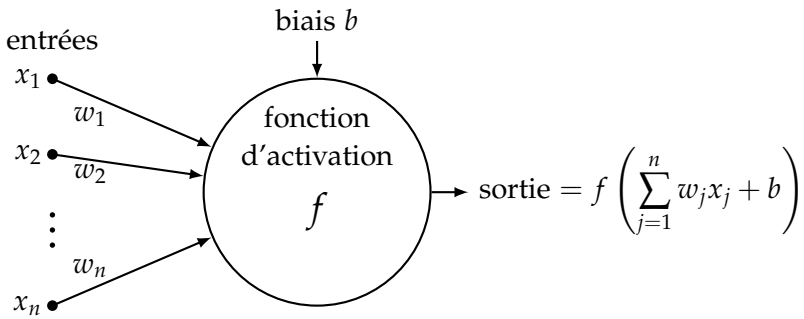


Figure X.6 – *Structure d'un neurone. Une fonction f dite d'« activation » est appliquée à la somme des entrées x_j pondérées par les poids w_j et du biais b afin d'obtenir la valeur de sortie.*

5.1.2 Fonctions d'activation

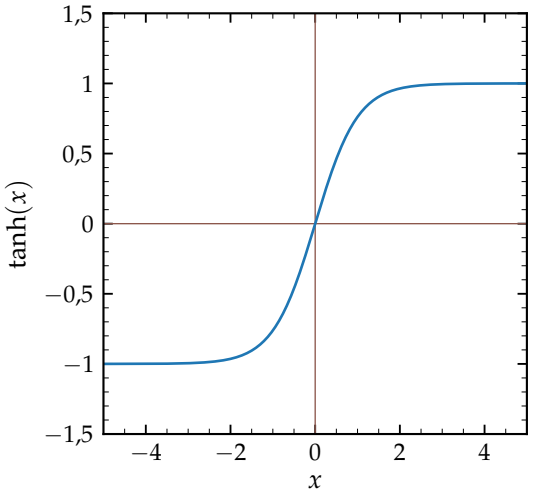
En principe, toute fonction définie sur l'ensemble d'existence de chacune des entrées x_j peut être utilisée comme fonction d'activation. Elles sont ainsi définies sur $\mathbb{R}$. Les plus utilisées sont :

tangente hyperbolique notée tanh, définie par

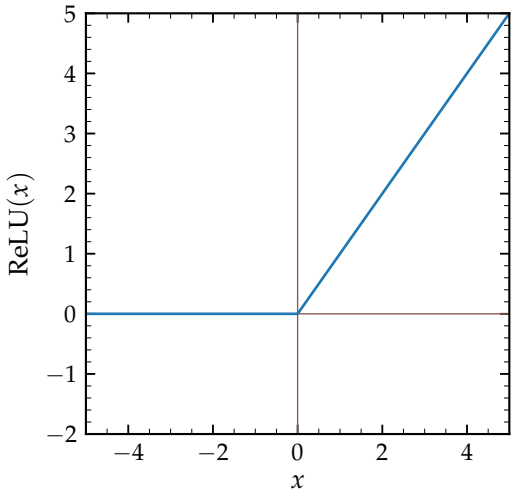
$$\tanh : x \mapsto \frac{e^x - e^{-x}}{e^x + e^{-x}}\ ; \tag{X.17}$$

sigmoïde notée sig, définie par

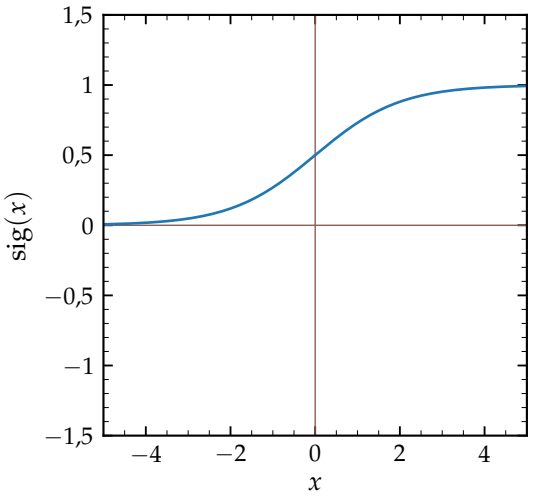
$$\text{sig} : x \mapsto \frac{1}{1 + e^{-x}}\ ; \tag{X.18}$$



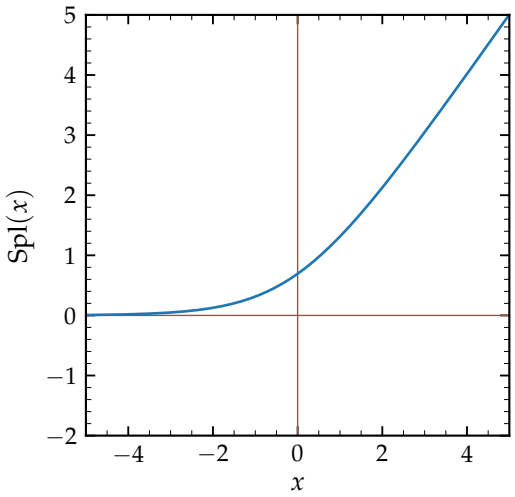
(a) *Tangente hyperbolique.*



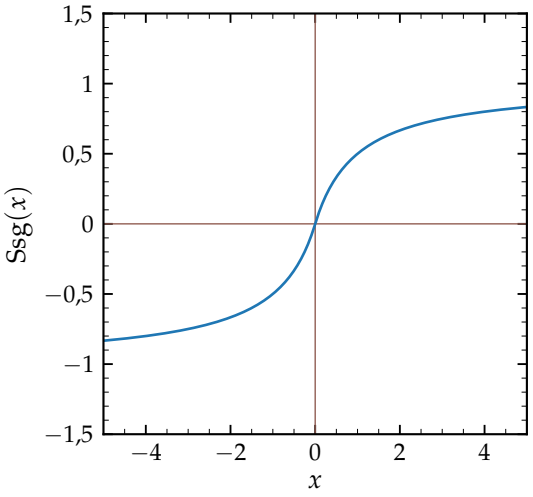
(b) *ReLU.*



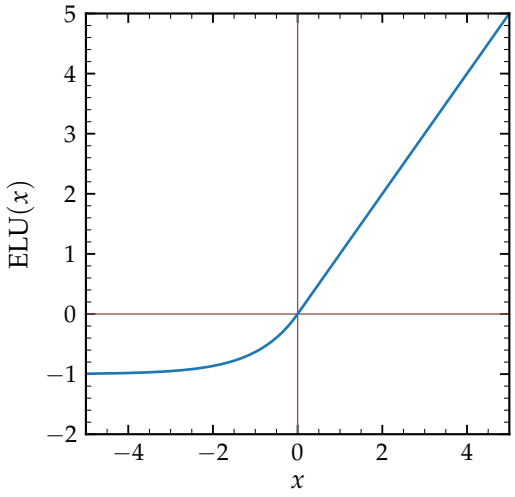
(c) *Sigmoïde.*



(d) *Softplus.*



(e) *Softsign.*



(f) *ELU.*

Figure X.7 – *Exemples de fonctions d'activation. À gauche, des fonctions à valeurs bornées, généralement utilisées en classification. À droite, des fonctions à valeurs non bornées, utilisables pour des tâches de régression.*

Softsign notée Ssg, définie par

$$\mathrm{Ssg} : x \mapsto \frac{x}{1+|x|} \; ; \tag{X.19}$$

ReLU (*Rectified Linear Unit*), définie par

$$\mathrm{ReLU} : x \mapsto \begin{cases} x\,, & x > 0 \\ 0\,, & x \leqslant 0 \end{cases} \; ; \tag{X.20}$$

Softplus notée Spl, définie par

$$\mathrm{Spl} : x \mapsto \ln(1+\mathrm{e}^{x}) \; ; \tag{X.21}$$

ELU (*Exponential Linear Unit*), définie par

$$\mathrm{ELU} : x \mapsto \begin{cases} x\,, & x > 0 \\ \alpha(\mathrm{e}^{x}-1)\,, & x \leqslant 0 \end{cases}, \quad \alpha = 1 \; ; \tag{X.22}$$

SELU (*Scaled Exponential Linear Unit*), similaire à ELU et définie par

$$\mathrm{SELU} : x \mapsto \lambda \times \begin{cases} x\,, & x > 0 \\ \alpha(\mathrm{e}^{x}-1)\,, & x \leqslant 0 \end{cases}, \quad \alpha \simeq 1{,}67\,, \quad \lambda \simeq 1{,}05 \; ; \tag{X.23}$$

ou encore la fonction linéaire identité $\mathbb{1} : x \mapsto x$. Certaines d'entre elles sont représentées sur la figure X.7.

5.2 Réseaux de neurones

Un NN est obtenu par l'interconnexion de plusieurs neurones entre eux. Ces connexions peuvent se faire selon diverses architectures [41, 44]. Nous utilisons ici, comme dans les travaux de BÄRTSCHI & coll. [18], une architecture normale profonde à propagation avant complètement connectée (*normal deep feedforward fully-connected*), représentée sur la figure X.8, c'est-à-dire avec :

— des neurones répartis en couches (normale) ;
— plusieurs couches « cachées », situées entre les couches d'entrée et de sortie (profonde) ;
— toutes les sorties de la couche $k-1$ utilisées comme entrées de chacun des neurones de la couche k (à propagation avant complètement connectée).

Le nombre de neurones par couche cachée est noté $N_{\mathrm{n/c}}$, le nombre de couches cachées N_{cc}. Le NN ayant une structure profonde, il s'agit d'un DNN (*Deep Neural Network*).

La tâche du réseau est une régression vers une seule grandeur, $m_{\mathcal{H}}$, à partir de n variables d'entrée x_j, $j \in \{1, \ldots, n\}$. La couche de sortie est donc composée d'un seul neurone dont la fonction d'activation est l'identité. La couche d'entrée comporte n neurones, chacun se contentant de transmettre la variable d'entrée correspondante. Il s'agit donc d'une couche d'adaptation entre le nombre d'entrées n_{in} et le nombre de neurones dans la couche suivante $N_{\mathrm{n/c}}$. Tous les neurones des couches cachées ont la même fonction d'activation. Plusieurs fonctions d'activation sont testées dans la section 6.

5.3 Entraînement

L'entraînement d'un NN est le réglage des paramètres des neurones du réseau situés sur les couches cachées et la couche de sortie. Il s'agit des poids w_i et du biais b. Pour un DNN avec $n_{\mathrm{in}} = 27$ variables d'entrée, $N_{\mathrm{cc}} = 3$ couches cachées de $N_{\mathrm{n/c}} = 1000$ neurones, le nombre de paramètres est ainsi de

$$\begin{aligned} N_{\mathrm{params.}} &= \underbrace{N_{\mathrm{n/c}} \times (n_{\mathrm{in}}+1)}_{\text{couche cachée 1}} + \underbrace{(N_{\mathrm{cc}}-1) \times N_{\mathrm{n/c}} \times (N_{\mathrm{n/c}}+1)}_{\text{autres couches cachées}} + \underbrace{N_{\mathrm{n/c}}+1}_{\text{couche de sortie}} \\ &= 28\,000 + 2 \times 1\,001\,000 + 1001 = 2\,031\,001\,, \end{aligned} \tag{X.24}$$

soit près de deux millions. Les termes « $+1$ » correspondent aux biais b à ajouter au nombre d'entrées des neurones.

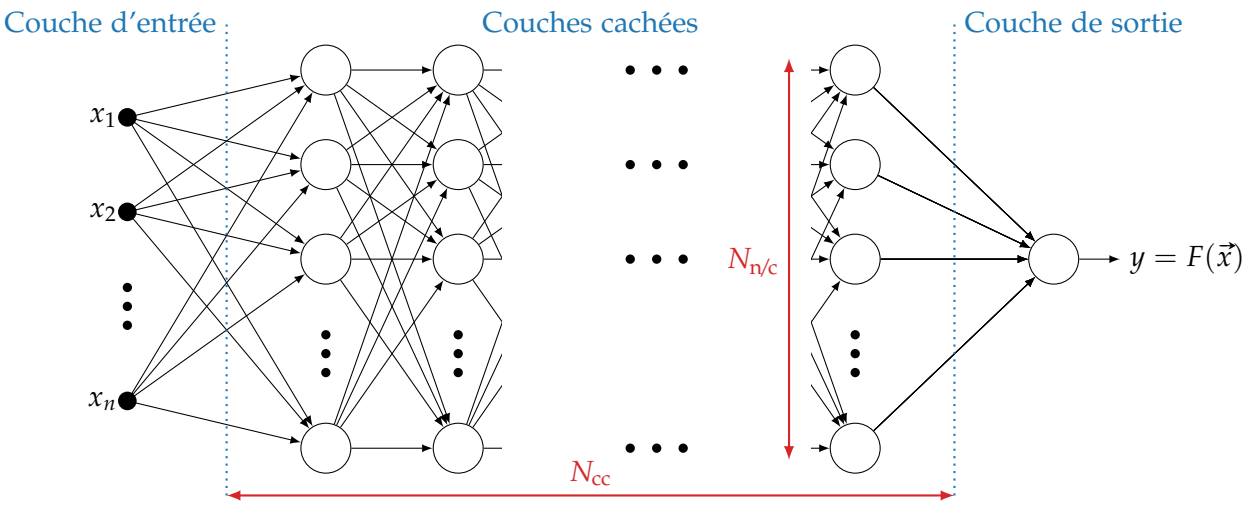


Figure X.8 – *Structure normale profonde à propagation avant complètement connectée d'un réseau de neurones. Une couche d'entrée comporte autant de neurones que de variables x_i. La couche de sortie en comporte autant que de valeurs à donner, c'est-à-dire une. Les fonctions d'activation de ces deux couches sont linéaires. Entre elles se trouvent N_{cc} couches cachées, chacune contenant $N_{n/c}$ neurones. Diverses fonctions d'activation peuvent être utilisées dans les couches cachées.*

5.3.1 Initialisation des paramètres

Les biais b sont initialement fixés à 0, les poids w_i à une valeur constante donnée ou aléatoirement selon une loi de probabilité. Le mode d'initialisation est un hyper-paramètre du modèle. Lors de ces travaux, nous avons testé les lois normale et uniforme. Dans le cas des DNNs, ces modes d'initialisation peuvent être améliorés par la méthode de GLOROT & BENGIO [47] afin de faciliter l'entraînement. Il s'agit alors des lois « Glorot uniforme » et « Glorot normale », également testées.

5.3.2 Optimisation des paramètres

L'optimisation des paramètres est réalisée en utilisant les mini-lots introduits en section 3.2. Un maximum de 500 époques est autorisé, avec un arrêt prématuré au bout de 20 époques sans diminution de l'erreur absolue moyenne (L_{MAE}) sur les données de validation. Plusieurs algorithmes d'optimisation existent [41], présentés de manière non exhaustive ci-après.

***Batch Gradient Descent* (BGD)** [40] L'algorithme BGD, déjà introduit en section 3.2, applique la méthode de descente de gradient sur les mini-lots. Le bruit dû à la composition aléatoire des mini-lots permet de s'échapper des minimums locaux mais dégrade la précision une fois au minimum global. Pour pallier cet effet, le taux d'apprentissage η peut être diminué à chaque époque. La condition sur les taux d'apprentissage η_k avec k l'époque afin de s'assurer de la convergence du modèle optimisé par BGD est [41]

$$\sum_{k=1}^{\infty} \eta_k = \infty\,, \quad \sum_{k=1}^{\infty} \eta_k^2 < \infty\,. \tag{X.25}$$

La mise à jour des paramètres à la fin d'un mini-lot pendant l'époque k est alors réalisée selon

$$p \to p - \eta_k \langle \vec{\nabla}(\mathsf{L}) \rangle_{\text{mini-lot}} \cdot \vec{e}_p = p - \eta_k \left\langle \frac{\partial \mathsf{L}}{\partial p} \right\rangle_{\text{mini-lot}}\,. \tag{X.26}$$

BGD avec moments [41] Les moments sont une « mémoire » des valeurs du gradient de la fonction de coût des époques précédentes. Ce peut être vu comme une inertie du mouvement du modèle dans l'espace des paramètres, prise en compte à travers une vitesse $\vec{v}$ définie initialement par l'utilisateur

et mise à jour à chaque mini-lot selon

$$\vec{v}[t-1] \to \vec{v}[t] = \alpha\vec{v}[t-1] - \eta_k \langle \vec{\nabla}(\mathsf{L})[t] \rangle_{\text{mini-lot}} \tag{X.27}$$

$$\Rightarrow \vec{v}[t] \cdot \vec{e}_p = v_p[t] = \alpha v_p[t-1] - \eta_k \left\langle \frac{\partial \mathsf{L}}{\partial p}[t] \right\rangle_{\text{mini-lot}} \tag{X.28}$$

avec t l'indice d'itération ou indice temporel de l'entraînement, et $0 \leqslant \alpha < 1$ le paramètre des moments. La mise à jour des paramètres lors de l'itération t se fait alors selon

$$p[t-1] \to p[t] = p[t-1] + v_p[t] = p[t-1] + \alpha v_p[t-1] - \eta_k \left\langle \frac{\partial \mathsf{L}}{\partial p}[t] \right\rangle_{\text{mini-lot}} . \tag{X.29}$$

***Adaptive Gradient* (AdaGrad)** [48] L'algorithme AdaGrad adapte le taux d'apprentissage individuellement pour chaque paramètre p à l'aide d'une variable de mémoire $\vec{r}$. Elle est initialement définie à $\vec{0}$ et est modifiée à chaque mini-lot selon

$$\vec{r} \cdot \vec{e}_p = r_p \to r_p + \left\langle \frac{\partial \mathsf{L}}{\partial p} \right\rangle^2_{\text{mini-lot}} . \tag{X.30}$$

La mise à jour des paramètres se fait alors suivant

$$p \to p - \eta \frac{1}{\sqrt{r_p} + \delta} \left\langle \frac{\partial \mathsf{L}}{\partial p} \right\rangle_{\text{mini-lot}} \tag{X.31}$$

où δ est une variable de régularisation évitant les divisions par zéro. Le taux d'apprentissage effectif pour le paramètre p est ainsi η divisé par la somme quadratique des gradients précédents $\sqrt{r_p}$.

Plus un paramètre modifie la valeur de la fonction de coût, plus sa modification est progressive. Dans l'optique de la recherche d'un minimum, cela revient à descendre une pente lentement et à se mouvoir rapidement dans une direction plane. Cependant, l'accumulation depuis le début de l'entraînement des gradients au carré dans r_p peut mener à une diminution excessive du taux d'apprentissage effectif d'un paramètre.

RMSProp [49] L'algorithme RMSProp consiste en une légère modification de AdaGrad. Une décroissance exponentielle de la mémoire des gradients passés est mise en place en remplaçant (X.30) par

$$r_p \to \rho\, r_p + (1-\rho) \left\langle \frac{\partial \mathsf{L}}{\partial p} \right\rangle^2_{\text{mini-lot}} \tag{X.32}$$

où $0 < \rho < 1$ est le taux de diminution de la mémoire. RMSProp est ainsi une version de AdaGrad dont la mémoire est plus adaptée à la situation locale.

***Adaptive Delta* (AdaDelta)** À l'instar de RMSProp, AdaDelta est une modification de AdaGrad visant à améliorer l'effet de mémoire. La variable r_p est mise à jour par (X.30). Cependant, la valeur précédente de r_p est également utilisée lors de la mise à jour de p. Ainsi, lors de l'itération t,

$$p[t-1] \to p[t] = p[t-1] - \frac{\sqrt{r_p[t-1]} + \delta}{\sqrt{r_p[t]} + \delta} \left\langle \frac{\partial \mathsf{L}}{\partial p}[t] \right\rangle_{\text{mini-lot}} . \tag{X.33}$$

Il n'y a donc pas besoin de définir un taux d'apprentissage initial avec AdaDelta.

***Adaptive Moments* (Adam)** [41, 50] L'algorithme Adam est une combinaison de la méthode des moments et de RMSProp. Il adapte donc le taux d'apprentissage pour chaque paramètre à chaque mini-lot. Pour cela sont définis initialement :

— le pas $\epsilon = 0{,}001$;
— les moments d'ordres 1 et 2, $\vec{v} = \vec{0}$ et $\vec{r} = \vec{0}$;
— les taux de diminution de moments d'ordre 1 et 2, $\rho_1 = 0{,}9$ et $\rho_2 = 0{,}999$;

— le paramètre temporel $t = 0$.

Puis, à chaque mini-lot, les moments sont redéfinis selon

$$\vec{v}\cdot\vec{e}_p = v_p \to \rho_1 v_p + (1-\rho_1)\left\langle\frac{\partial \mathsf{L}}{\partial p}\right\rangle_{\text{mini-lot}}, \quad \vec{r}\cdot\vec{e}_p = r_p \to \rho_2 r_p + (1-\rho_2)\left\langle\frac{\partial \mathsf{L}}{\partial p}\right\rangle^2_{\text{mini-lot}}. \tag{X.34}$$

Le biais d'initialisation des moments est corrigé en appliquant

$$t \to t+1, \quad v_p \to \frac{v_p}{1-\rho_1^t}, \quad r_p \to \frac{r_p}{1-\rho_2^t}. \tag{X.35}$$

Les paramètres du modèle sont alors mis à jour selon

$$p \to p - \epsilon\frac{v_p}{\sqrt{r_p}+\delta} \tag{X.36}$$

où $\delta = 10^{-8}$ permet de stabiliser les calculs en évitant une division par zéro.

6 Optimisation des hyper-paramètres et choix d'un modèle

Le choix d'un modèle et de ses hyper-paramètres est l'objet de cette section. Les hyper-paramètres des XGBs, introduits section 4, sont :

— la profondeur maximale des arbres $N_{\text{max}}^{\text{prof.}}$;
— la quantité d'échantillons minimale dans une branche $N_{\text{min}}^{\text{échant.}}$;
— le nombre d'arbres $N_{\text{max}}^{\text{estim.}}$;
— le gain minimal γ ;
— le taux d'apprentissage η ;
— la fonction de coût L ;
— la liste des variables d'entrée.

Les hyper-paramètres des DNNs, introduits section 5, sont :

— le nombre de couches cachées N_{cc} ;
— le nombre de neurones par couche cachée $N_{\text{n/c}}$;
— la fonction d'activation des neurones des couches cachées ;
— l'algorithme d'optimisation ;
— la fonction de coût L ;
— le mode d'initialisation des poids ;
— la liste des variables d'entrée.

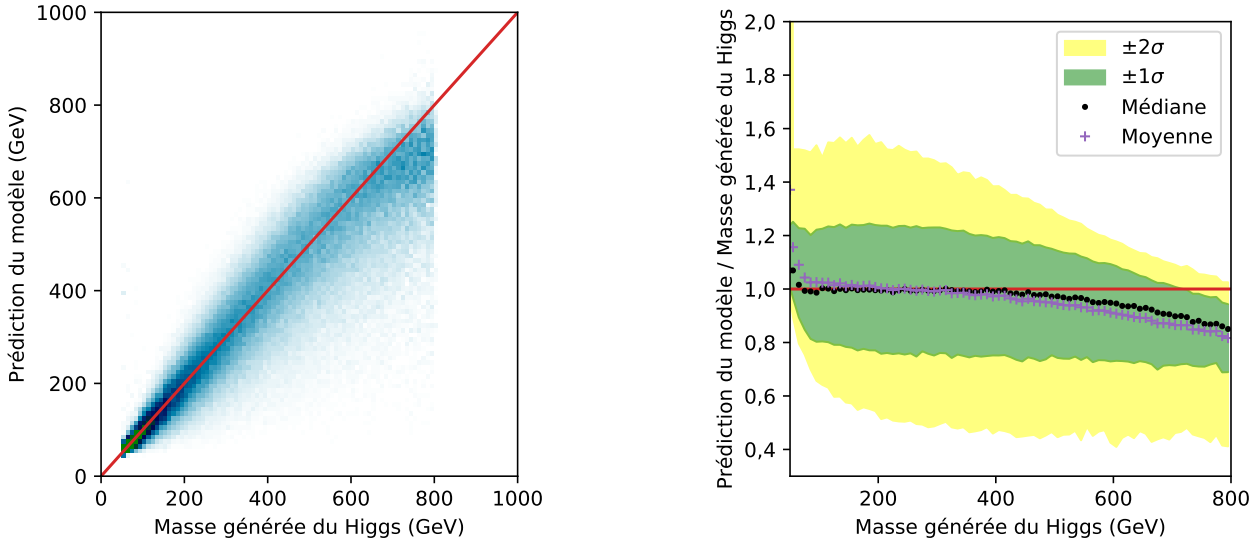
Dans ce chapitre, les modèles ont pour but de prédire la masse générée du boson de Higgs $m_{\mathcal{H}}$. Une représentation graphique possible afin de montrer les performances d'un modèle est de tracer ses prédictions $y_{\text{préd}}$ en fonction de y_{vraie} dans un histogramme à deux dimensions comme sur la figure X.9a. L'objectif des modèles est alors de se rapprocher autant que possible de la première bissectrice, tracée en rouge. Toutefois la large gamme explorée, de 50 à 800 GeV, rend difficile la visualisation des performances à basse masse. Or, cette région est importante car elle contient les bosons Z et h du modèle standard. La réponse R du modèle, définie comme

$$R = \frac{y_{\text{préd}}}{y_{\text{vraie}}} = \frac{F(\vec{x})}{m_{\mathcal{H}}}, \tag{X.37}$$

permet de ramener l'objectif des modèles à 1 sur toute la gamme de masse. La réponse du même modèle est ainsi représentée sur la figure X.9b. Pour chaque intervalle de 10 GeV sur $m_{\mathcal{H}}$, la distribution de r est déterminée. La valeur moyenne et la médiane de cette distribution sont données, ainsi que les largeurs à $\pm1\sigma$ et $\pm2\sigma$, correspondant respectivement aux zones contenant 68 et 95 % des valeurs de R les plus proches de la médiane. Ces zones sont déterminées de manière indépendante sur les valeurs inférieures et supérieures à la médiane, elles peuvent donc être asymétriques.

Il est difficile de définir un seul score quantifiant la qualité d'un modèle. Plusieurs métriques sont considérées afin de l'évaluer :

— les valeurs de L_{MSE}, L_{MAE}, L_{MAPE} ;



(a) *Histogramme à deux dimensions de $y_{\text{préd}}$ en fonction de y_{vraie}.*

(b) *Réponse du modèle $y_{\text{préd}}/y_{\text{vraie}}$ en fonction de y_{vraie}.*

Figure X.9 – *Exemples de graphiques rendant compte des performances des modèles.*

— la largeur de -1σ à $+1\sigma$ de la réponse R du modèle, notée $\Delta_{1\sigma}$, égale à la moyenne sur des intervalles de 10 GeV sur $y_{\text{vraie}} = m_{\mathcal{H}}$ de la distance entre les écarts-types supérieur et inférieur de la distribution de R, c'est-à-dire

$$\Delta_{1\sigma} = \left\langle \left[\sigma_+ \left(\frac{y_{\text{préd}}}{y_{\text{vraie}}} \right) - \sigma_- \left(\frac{y_{\text{préd}}}{y_{\text{vraie}}} \right) \right] \Big|_{y_{\text{vraie}} \in [n, n+1] \times 10\,\text{GeV}} \right\rangle_n . \tag{X.38}$$

Il s'agit donc de la moyenne de la largeur verticale des bandes vertes ($\pm 1\sigma$) sur les graphiques des réponses des modèles comme celui de la figure X.9b, aussi égale à deux fois la résolution relative du modèle.

Pour toutes ces métriques, l'objectif est d'avoir la plus petite valeur possible. De plus, quatre domaines de masse sont définis :

— basse masse : $m_{\mathcal{H}} < 150\,\text{GeV}$, incluant en particulier les bosons Z et h ;
— moyenne masse : $150\,\text{GeV} \leqslant m_{\mathcal{H}} < 500\,\text{GeV}$;
— haute masse : $m_{\mathcal{H}} \geqslant 500\,\text{GeV}$;
— toute masse : aucune restriction sur $m_{\mathcal{H}}$.

Ils permettent de comparer les performances des modèles sur certaines gammes de masse uniquement. Sauf contre-indication, toute la gamme de masse est considérée.

Face à l'immense quantité de combinaisons différentes d'hyper-paramètres, toutes n'ont pas été testées. Nous avons en revanche entraîné suffisamment de modèles afin d'observer les distributions des différentes métriques d'évaluation pour des groupes de modèles ayant une valeur donnée d'un hyper-paramètre. La comparaison des différentes distributions permet dans un premier temps de voir quelles valeurs d'hyper-paramètres donnent des modèles moins performants et ainsi se rapprocher d'une combinaison optimale, comme l'exposent les sections 6.1 à 6.4. Une fois certains hyper-paramètres fixés, la sélection finale d'un seul modèle est réalisée selon la procédure présentée en section 6.5.

6.1 Variables d'entrée

Utiliser le moins de variables d'entrée possible, si cela ne dégrade pas la qualité de nos modèles, pourrait faciliter leur intégration dans les analyses de CMS. En effet, si toutes les variables d'entrée considérées listées section 2.4 sont généralement déjà exploitées dans les analyses en cours, ce n'est toutefois pas toujours le cas, en particulier pour les variables relatives à l'activité hadronique ad-

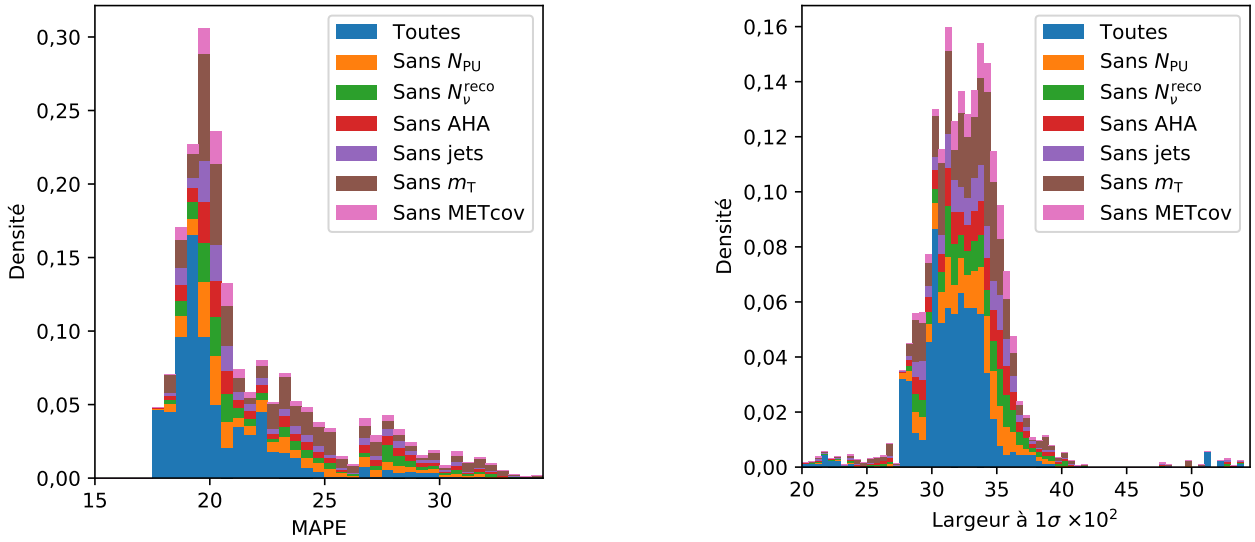
ditionnelle. Il est donc pertinent de déterminer les variables dont les modèles peuvent aisément se passer.

Pour cela, des sous-ensembles des variables d'entrée sont définis par les restrictions suivantes :

— sans N_{PU} : la variable N_{PU} n'est pas utilisée ;
— sans N_ν^{reco} : la variable N_ν^{reco} n'est pas utilisée ;
— sans AHA : les variables d'activité hadronique additionnelle ne sont pas utilisées ;
— sans jets : les variables relatives aux jets (dont AHA) ne sont pas utilisées ;
— sans m_{T} : les masses transverses ne sont pas utilisées ;
— sans METcov : la matrice de covariance de ${E_{\text{T}}}^{\text{miss}}$ n'est pas utilisée.

L'application de plusieurs de ces restrictions simultanément est également testée.

Les performances des modèles entraînés avec les différents ensembles de variables d'entrée sont données figure X.10 pour les XGBs et figure X.11 pour les DNNs. Les modèles concernés par plusieurs restrictions sont comptés de manière pondérée dans chaque groupe correspondant à une restriction unique. Par exemple, un modèle soumis à la restriction « sans N_{PU} » et « sans N_ν^{reco} » a un poids de $\frac{1}{2}$ dans chacun de ces deux groupes. Les histogrammes ainsi créés sont superposés. Il est alors possible de voir les contributions de chacune des restrictions aux valeurs obtenues sur la métrique d'évaluation illustrée.



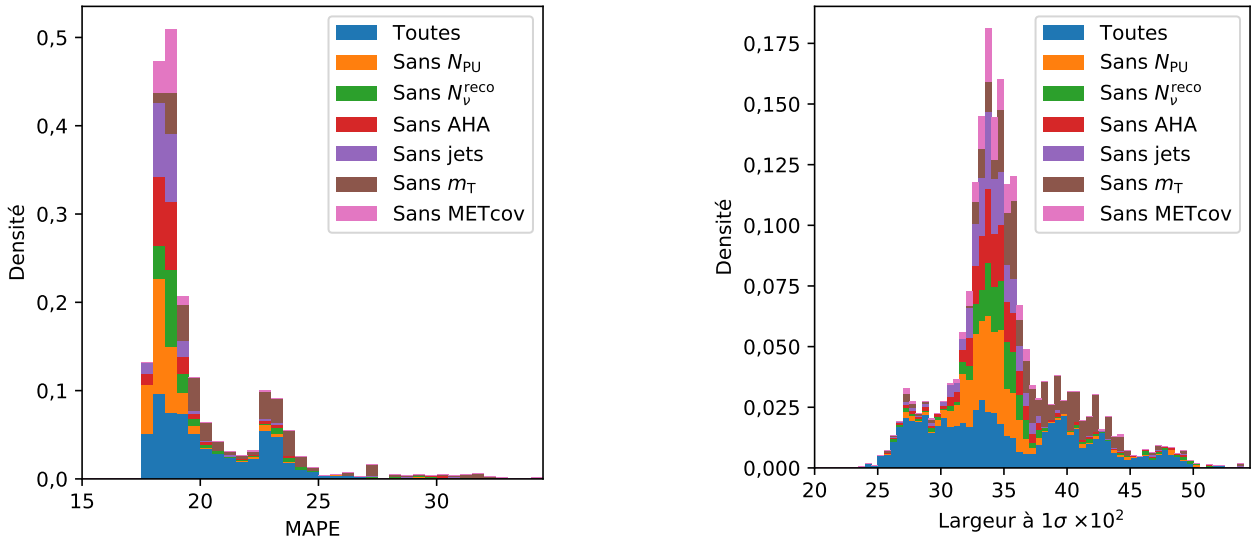
(a) *Évaluation par* L_{MAPE}. **(b)** *Évaluation par* $\Delta_{1\sigma}$.

Figure X.10 – *Évaluations des XGBs regroupés selon les variables d'entrée.*

Dans le cas des XGBs, l'évaluation des modèles par L_{MAPE}, en figure X.10a, donne des valeurs situées entre 17 et 35. Le cœur de la distribution, à $\text{L}_{\text{MAPE}} = 19 \pm 2$, est plutôt constitué de modèles utilisant toutes les entrées dans sa partie gauche ($\text{L}_{\text{MAPE}} < 19$) et de modèles utilisant un sous-ensemble d'entrées dans sa partie droite ($19 < \text{L}_{\text{MAPE}} < 22$). De plus, les basses valeurs de L_{MAPE}, en dessous de 18,5, sont presque exclusivement obtenues avec des modèles utilisant toutes les entrées. À l'inverse, la queue à hautes valeurs de la distribution obtenue ($\text{L}_{\text{MAPE}} > 23$) est largement dominée par les contributions des modèles avec un sous-ensemble d'entrées.

La plupart des XGBs ont une largeur $\Delta_{1\sigma}$, en figure X.10b, située entre 27 et 38. Cependant, les XGBs utilisant toutes les variables d'entrée exhibent une distribution de $\Delta_{1\sigma}$ légèrement décalée vers de plus faibles valeurs.

Dans le cas des DNNs, la distribution de la métrique L_{MAPE}, en figure X.11a, contient des valeurs situées majoritairement entre 17,5 et 25. Les DNNs n'utilisant pas N_ν^{reco} se situent à $\text{L}_{\text{MAPE}} > 18$. Cette variable permet aux modèles de différencier les canaux hadroniques, semi-leptoniques et leptoniques, dont la séparation est discutée dans la section 7. Ceux n'utilisant pas m_{T} présentent également des valeurs de L_{MAPE} uniquement au-delà de 18. L'utilisation de ces variables permet donc d'obtenir de meilleurs modèles. Elles sont de plus facilement obtenues à partir du dilepton, défini chapitre 4, et de ${E_{\text{T}}}^{\text{miss}}$. Les analyses avec deux leptons τ dans l'état final exploitent déjà ces observables, leur uti-



(a) *Évaluation par* L_{MAPE}. **(b)** *Évaluation par* $\Delta_{1\sigma}$.

Figure X.11 – *Évaluations des DNNs regroupés selon les variables d'entrée.*

lisation par nos modèles est donc à la fois pertinente, car les scores de L_{MAPE} obtenus sont meilleurs, et sans incidence sur la facilité d'intégration du modèle à l'analyse. Les DNNs avec $L_{MAPE} \lesssim 18$ exploitent presque tous les variables relatives aux jets, à l'AHA et à la matrice de covariance de E_T^{miss}. Ces entrées sont donc vraisemblablement utiles aux DNNs afin de réaliser la régression. Enfin, la restriction sur N_{PU} ne semble pas dégrader les performances des DNNs selon L_{MAPE}.

La distribution de $\Delta_{1\sigma}$, en figure X.11b, montre que les modèles utilisant toutes les entrées peuvent se répartir en plusieurs groupes, aux alentour des valeurs 0,275, 0,335, 0,395, 0,425 et 0,475. À 0,395 apparaît également un groupe de modèles entraînés sans m_T. À 0,335 se trouvent la majorité des modèles entraînés avec une restriction des entrées. Pour $\Delta_{1\sigma} < 0{,}3$, les modèles sont très majoritairement ceux utilisant l'ensemble des variables proposées.

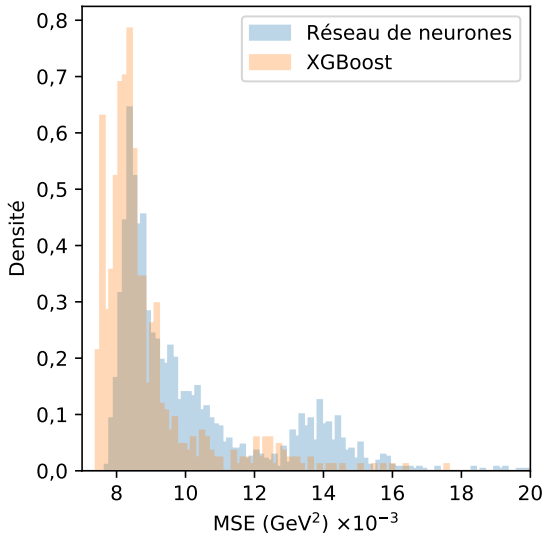
L'utilisation de toutes les variables listées dans la section 2.4 est donc corrélée avec de meilleures performances selon les métriques L_{MAPE} et $\Delta_{1\sigma}$. Par la suite, seuls les modèles utilisant toutes les variables, au nombre de 27, sont considérés.

6.2 Type de modèle

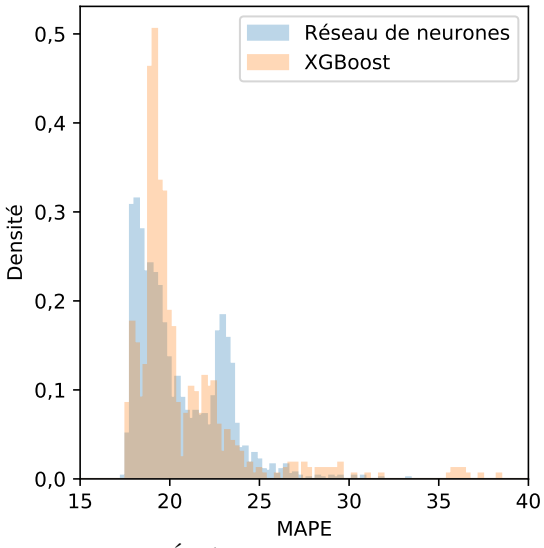
Les figures X.12 et X.13 présentent les distributions des scores de L_{MSE}, L_{MAPE} et $\Delta_{1\sigma}$ pour l'ensemble des DNNs et des XGBs utilisant toutes les variables d'entrée.

L'évaluation par L_{MSE}, en figure X.12a, favorise les XGBs. Le cœur de la distribution de L_{MSE} pour ces modèles est en effet à $8{,}1 \times 10^3\,\text{GeV}^2$ contre $8{,}5 \times 10^3\,\text{GeV}^2$ pour les DNNs. En revanche, l'évaluation par L_{MAPE}, en figure X.12b, favorise les DNNs avec un groupe de DNNs à $L_{MAPE} = 18$ contre 19 pour les XGBs. Un second groupe de DNNs est présent à $L_{MAPE} = 23$. L'existence de ces deux groupes est due à l'utilisation de plusieurs algorithmes d'optimisation, comme discuté dans la section 6.4.

La résolution des modèles est évaluée par $\Delta_{1\sigma}$ en figure X.13a pour toute la gamme de masse et en figure X.13b à basse masse. Sur l'ensemble de la gamme de masse, les XGBs ont un score de $0{,}32 \pm 0{,}04$ et les DNNs se répartissent en plusieurs groupes à environ 0,28, 0,33, 0,40, 0,42 et 0,48. Les XGBs sont ainsi compétitifs d'après cette évaluation. Cependant, les performances des modèles à basse masse, c'est-à-dire pour $m_{\mathcal{H}} < 150\,\text{GeV}$, sont importantes car c'est dans cette gamme de masse que se trouvent les bosons Z et h du modèle standard. En particulier, il faut s'assurer que le signal du Z présente une queue à haute masse, c'est-à-dire dans la région de signal des bosons de Higgs. Dans le cadre des analyses telles que celle présentée au chapitre 4, le Z est en effet un bruit de fond important. Sur la figure X.13b montrant l'évaluation à basse masse par $\Delta_{1\sigma}$ des modèles, les scores des XGBs se situent majoritairement à $0{,}70 \pm 0{,}15$ alors que ceux des DNNs donnent deux groupes, le premier à $0{,}42 \pm 0{,}05$ et le second entre 0,50 et 1,0. Le premier ensemble de DNNs propose les

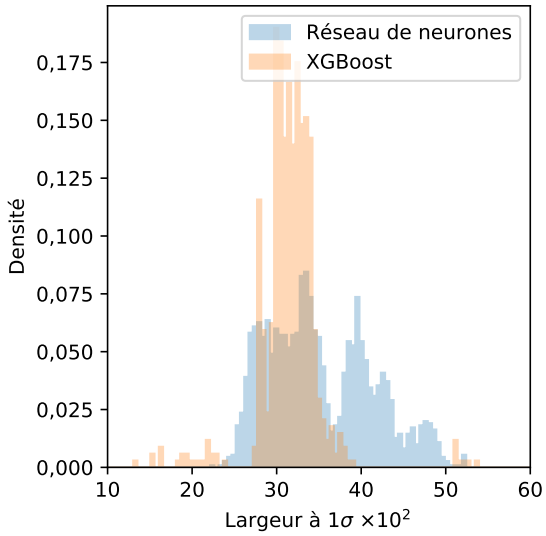


(a) *Évaluation par* $\mathsf{L}_{\mathrm{MSE}}$.

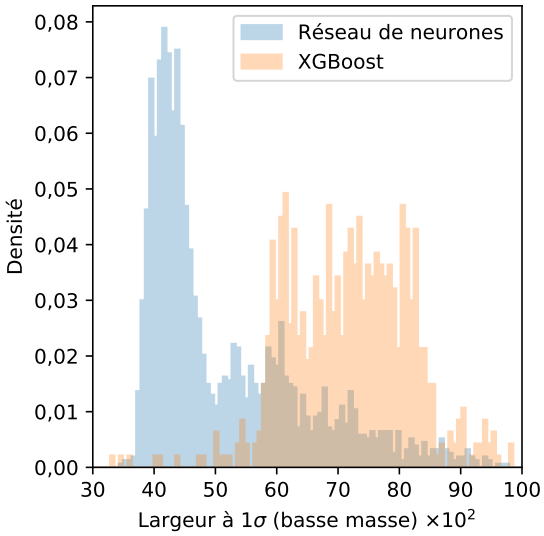


(b) *Évaluation par* $\mathsf{L}_{\mathrm{MAPE}}$.

Figure X.12 – *Évaluations des XGBs et des DNNs par* $\mathsf{L}_{\mathrm{MSE}}$ *et* $\mathsf{L}_{\mathrm{MAPE}}$.



(a) *Sur toute la gamme de masse.*



(b) *À basse masse.*

Figure X.13 – *Évaluations des XGBs et des DNNs par* $\Delta_{1\sigma}$.

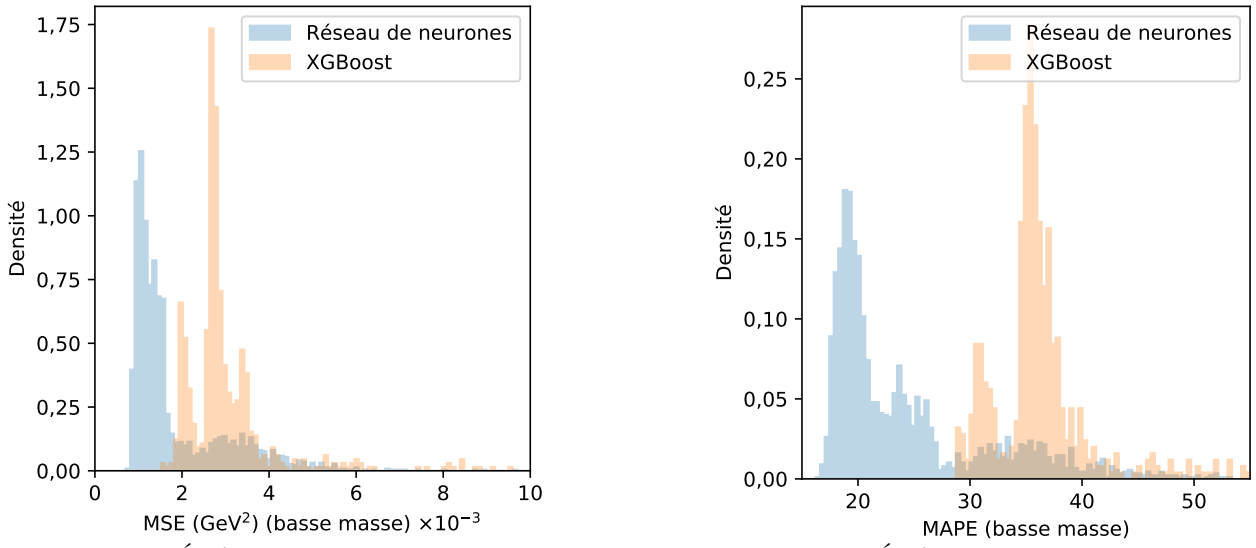
meilleures résolutions sur les masses des particules du modèle standard.

La réévaluation des modèles par $\mathsf{L}_{\mathrm{MSE}}$ et $\mathsf{L}_{\mathrm{MAPE}}$ à basse masse, en figures X.14a et X.14b, confirme l'obtention de meilleures performances avec les DNNs. En effet, les DNNs sont les seuls modèles avec $\mathsf{L}_{\mathrm{MSE}} < 1{,}5\times10^{3}\,\mathrm{GeV}^2$ et $\mathsf{L}_{\mathrm{MAPE}} < 28$ à basse masse. Les XGBs ont des scores de $\mathsf{L}_{\mathrm{MSE}}$ et $\mathsf{L}_{\mathrm{MAPE}}$ généralement compris entre $1{,}5\times10^{3}\,\mathrm{GeV}^2$ et $4{,}0\times10^{3}\,\mathrm{GeV}^2$ et entre 28 et 43, respectivement. Dans la suite, seuls les DNNs seront donc considérés.

6.3 Fonction de coût

Les évaluations des DNNs, regroupés d'après la fonction de coût utilisée lors de leurs entraînements, selon les métriques $\mathsf{L}_{\mathrm{MSE}}$, $\mathsf{L}_{\mathrm{MAPE}}$, $\mathsf{L}_{\mathrm{MAE}}$ sur toute la gamme de masse et $\Delta_{1\sigma}$ à basse masse sont représentées sur la figure X.15.

L'évaluation par $\mathsf{L}_{\mathrm{MSE}}$ est représentée figure X.15a. Les DNNs entraînés avec $\mathsf{L} = \mathsf{L}_{\mathrm{MSE}}$ y présentent un score compris entre $7{,}8\times10^{3}\,\mathrm{GeV}^2$ et $15\times10^{3}\,\mathrm{GeV}^2$, la majorité d'entre eux se trouvant en dessous de $11\times10^{3}\,\mathrm{GeV}^2$ avec un pic de leur distribution à $8{,}7\times10^{3}\,\mathrm{GeV}^2$. Les DNNs entraînés avec



(a) *Évaluation par* $\mathsf{L}_{\mathrm{MSE}}$. **(b)** *Évaluation par* $\mathsf{L}_{\mathrm{MAPE}}$.

Figure X.14 – *Évaluations des XGBs et des DNNs par* $\mathsf{L}_{\mathrm{MSE}}$ *et* $\mathsf{L}_{\mathrm{MAPE}}$ *à basse masse.*

$\mathsf{L} = \mathsf{L}_{\mathrm{MAE}}$ se situent majoritairement entre $7{,}7\times10^3\,\mathrm{GeV}^2$ et $10\times10^3\,\mathrm{GeV}^2$, avec un pic de leur distribution à $8{,}3\times10^3\,\mathrm{GeV}^2$. Les DNNs entraînés avec $\mathsf{L} = \mathsf{L}_{\mathrm{MAPE}}$ se répartissent en deux groupes, le premier entre $7{,}9\times10^3\,\mathrm{GeV}^2$ et $10\times10^3\,\mathrm{GeV}^2$, le second entre $13\times10^3\,\mathrm{GeV}^2$ et $16\times10^3\,\mathrm{GeV}^2$. La fonction de coût $\mathsf{L}_{\mathrm{MAE}}$ semble ainsi préférable à $\mathsf{L}_{\mathrm{MSE}}$ lorsque la comparaison se fait sur $\mathsf{L}_{\mathrm{MSE}}$ elle-même. Il est en revanche plus difficile de conclure quant à $\mathsf{L}_{\mathrm{MAPE}}$.

L'évaluation par $\mathsf{L}_{\mathrm{MAPE}}$, figure X.15b, montre également un avantage de $\mathsf{L}_{\mathrm{MAE}}$ sur $\mathsf{L}_{\mathrm{MSE}}$. En effet, les modèles entraînés avec $\mathsf{L} = \mathsf{L}_{\mathrm{MAE}}$ se situent majoritairement à $\mathsf{L}_{\mathrm{MAPE}} < 21$ alors que ceux entraînés avec $\mathsf{L} = \mathsf{L}_{\mathrm{MSE}}$ sont plutôt dans la région $\mathsf{L}_{\mathrm{MAPE}} > 20$. Les valeurs les plus basses sont obtenues sur les modèles entraînés avec $\mathsf{L}_{\mathrm{MAPE}}$. Or, l'évaluation est basée sur $\mathsf{L}_{\mathrm{MAPE}}$ elle-même, il n'est donc pas équitable de se baser uniquement sur la figure X.15b pour affirmer que $\mathsf{L}_{\mathrm{MAPE}}$ peut être préférable à $\mathsf{L}_{\mathrm{MAE}}$ ou $\mathsf{L}_{\mathrm{MSE}}$.

La figure X.15c représente l'évaluation des DNNs par $\mathsf{L}_{\mathrm{MAE}}$. La distribution obtenue avec les DNNs entraînés avec $\mathsf{L} = \mathsf{L}_{\mathrm{MSE}}$ s'étend de 65 GeV à près de 10 GeV avec un pic à 71 GeV. En revanche, de nombreux modèles entraînés avec $\mathsf{L}_{\mathrm{MAE}}$ ou $\mathsf{L}_{\mathrm{MAPE}}$ se situent à 67 ± 4 GeV.

Enfin, sur la figure X.15d se trouvent les distributions de $\Delta_{1\sigma}$ à basse masse pour ces trois groupes de DNNs. Les modèles utilisant $\mathsf{L}_{\mathrm{MSE}}$ ont tous un score supérieur à 0,5. Ceux entraînés avec $\mathsf{L}_{\mathrm{MAE}}$ se situent entre 0,4 et 0,6. Les modèles basés sur $\mathsf{L}_{\mathrm{MAPE}}$ forment encore deux groupes, le premier entre 0,34 et 0,5, le second entre 0,5 et 0,64. Les fonctions de coût $\mathsf{L}_{\mathrm{MAPE}}$ et $\mathsf{L}_{\mathrm{MAE}}$ permettent donc d'obtenir des modèles avec une meilleure résolution à basse masse que $\mathsf{L}_{\mathrm{MSE}}$.

Les modèles entraînés avec $\mathsf{L} = \mathsf{L}_{\mathrm{MAE}}$ ou $\mathsf{L} = \mathsf{L}_{\mathrm{MAPE}}$ proposent ainsi de meilleurs scores que ceux obtenus avec $\mathsf{L} = \mathsf{L}_{\mathrm{MSE}}$, quelle que soit la métrique d'évaluation utilisée. Lors des évaluations avec $\mathsf{L}_{\mathrm{MSE}}$ ou $\mathsf{L}_{\mathrm{MAE}}$, aucun avantage net n'est visible entre $\mathsf{L} = \mathsf{L}_{\mathrm{MAE}}$ et $\mathsf{L} = \mathsf{L}_{\mathrm{MAPE}}$. En revanche, les métriques $\mathsf{L}_{\mathrm{MAPE}}$ et $\Delta_{1\sigma}$ montrent que certains modèles entraînés avec $\mathsf{L} = \mathsf{L}_{\mathrm{MAPE}}$ donnent de meilleurs résultats. La sélection d'un modèle est donc poursuivie parmi ceux ayant comme fonction de coût $\mathsf{L}_{\mathrm{MAPE}}$.

6.4 Algorithme d'optimisation

Les algorithmes d'optimisation sont présentés dans la section 5.3.2. L'algorithme SGD ne permet pas aux modèles de converger, il est donc exclu de nos investigations. Deux algorithmes sont comparés, AdaDelta et Adam.

Les évaluations des DNNs précédemment sélectionnés, regroupés d'après l'algorithme d'optimisation utilisé lors de leurs entraînements, selon les métriques $\mathsf{L}_{\mathrm{MAPE}}$ sur toute la gamme de masse et $\Delta_{1\sigma}$ à basse masse sont représentées sur la figure X.16. Les deux groupes observés dans les sections

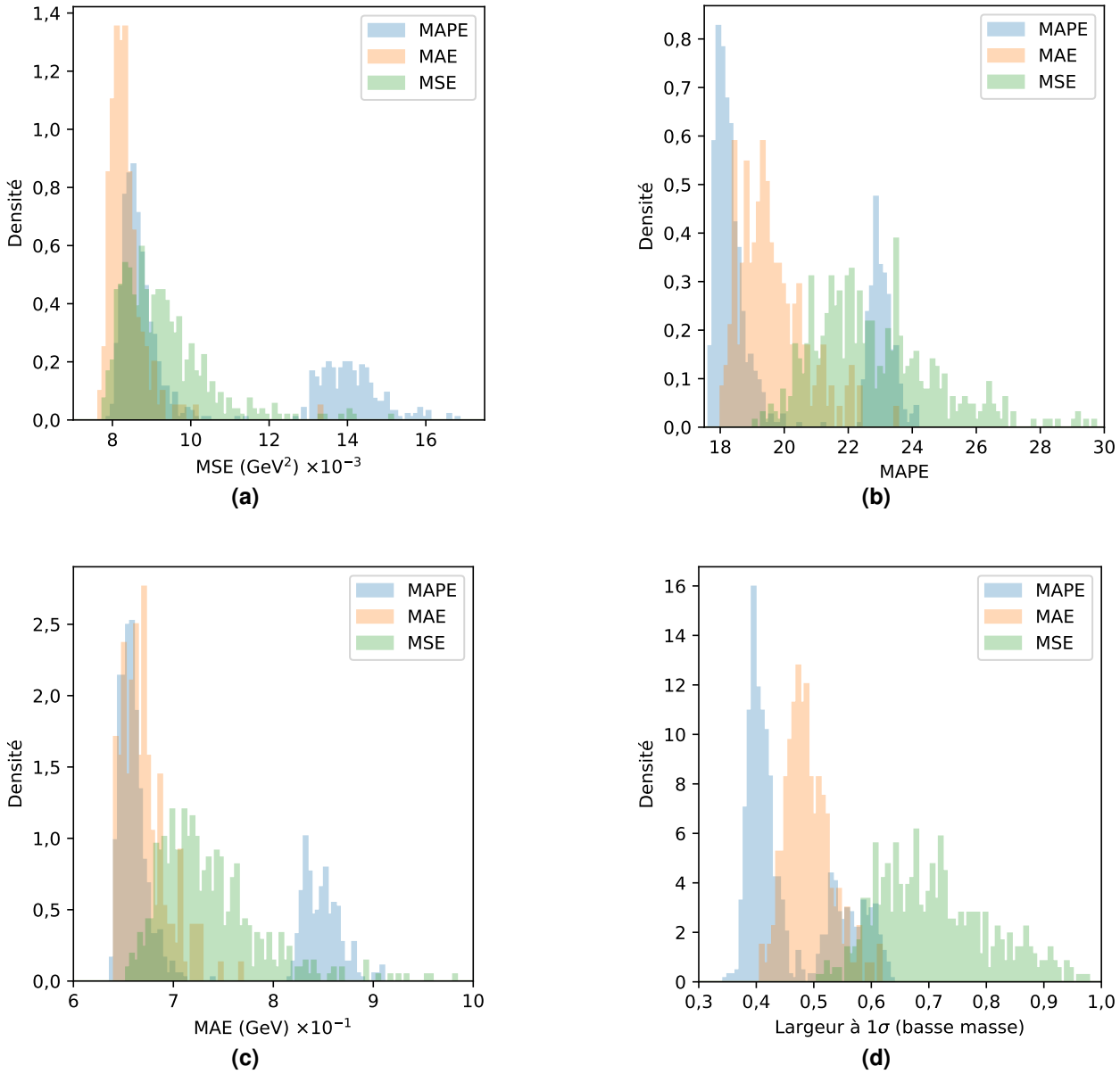


Figure X.15 – *Évaluations des DNNs regroupés selon la fonction de coût par* $\mathrm{L}_{\mathrm{MSE}}$, $\mathrm{L}_{\mathrm{MAPE}}$, $\mathrm{L}_{\mathrm{MAE}}$ *et* $\Delta_{1\sigma}$.

précédentes sont identifiés comme étant les modèles entraînés respectivement par Adam et AdaDelta. Dans le cadre de nos travaux, lors de la recherche de la combinaison optimale d'hyper-paramètres, nous avons initialement utilisé Adam jusqu'à sélectionner le jeu de variables d'entrée (section 6.1) et la fonction de coût (section 6.3) à utiliser. C'est pourquoi ces deux groupes liés à Adam et AdaDelta n'apparaissent que dans certaines sélections de modèles.

Sur la figure X.16a, les modèles optimisés par Adam présentent un score de $\mathrm{L}_{\mathrm{MAPE}}$ entre 17,5 et 20 alors que ceux optimisés par AdaDelta se situent entre 22,2 et 24,3. L'optimisation par Adam semble donc meilleure que celle par AdaDelta. L'évaluation à basse masse par $\Delta_{1\sigma}$ sur la figure X.16b confirme cette observation. Les modèles optimisés par Adam se situent en effet entre 0,35 et 0,48, ceux optimisés par AdaDelta entre 0,49 et 0,64. L'algorithme d'optimisation Adam donne donc de meilleurs modèles qu'AdaDelta avec les hyper-paramètres choisis précédemment.

6.5 Autres hyper-paramètres

Les hyper-paramètres restant à fixer et leurs valeurs testées sont :

— le nombre de couches cachées N_{cc}, 2 à 5 ;

— le nombre de neurones par couche cachée $N_{\mathrm{n/c}}$, 200 à 2000 par pas de 100 ;

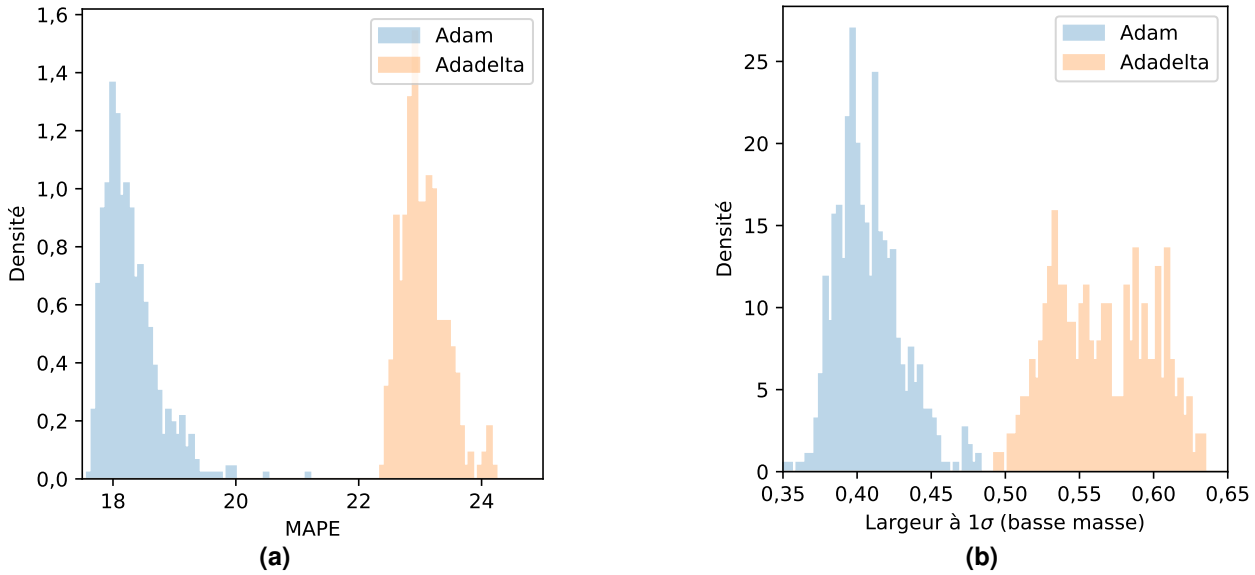


Figure X.16 – *Évaluations des DNNs regroupés selon l'algorithme d'optimisation par* L_{MAPE} *et* $\Delta_{1\sigma}$.

— le mode d'initialisation des poids (WI), uniforme (u), normale (n), Glorot uniforme (gu), Glorot normale (gn) ;
— la fonction d'activation (FA) des neurones des couches cachées, ReLU, SELU, ELU, Softplus.

Les évaluations à basse, moyenne et haute masse des DNNs utilisant les 27 variables d'entrée et entraînés par Adam avec $\mathsf{L} = \mathsf{L}_{\text{MAPE}}$, regroupés par N_{cc}, $N_{\text{n/c}}$, mode d'initialisation des poids et fonction d'activation sont données sur les figures X.17, X.18, X.19 et X.20 respectivement. Les distributions des scores pour les métriques d'évaluation y étant présentées sont semblables pour tous les groupes de modèles formés par une valeur fixée d'un de ces hyper-paramètres. La méthode employée jusqu'ici ne permet donc pas de conclure sur le choix d'une valeur pour un hyper-paramètre.

Nous avons alors choisi de sélectionner un modèle parmi ceux restant à ce stade à l'aide d'une procédure itérative rejetant tout modèle avec un score élevé selon une des métriques d'évaluation jusqu'à ce qu'il n'en reste que 10 au plus. Cette procédure est la suivante :

1. Déterminer la valeur maximale $x_{\text{max}}^{\text{métrique } m}$, sur l'ensemble des modèles sélectionnés, de chacune des métriques d'évaluation m utilisées. La valeur maximale autorisée $x_{\text{OK}}^{\text{métrique } m}$ pour la métrique m est initialement fixée à $x_{\text{max}}^{\text{métrique } m}$;
2. Fixer la valeur maximale autorisée à 99 % de sa valeur actuelle pour chacune des métriques m ;
3. Rejeter tout modèle dont une des métriques donne un score supérieur à $x_{\text{OK}}^{\text{métrique } m}$;
4. Reprendre à l'étape 2 si plus de 10 modèles sont encore sélectionnés.

Les modèles ainsi sélectionnés, au nombre de 7, sont listés dans le tableau X.1 sans ordre particulier. Leurs réponses dont données sur les figures X.21 et X.22.

La même procédure appliquée à tous les modèles entraînés mène à une liste ne contenant que des DNNs, utilisant majoritairement les 27 variables d'entrée, tous entraînés par Adam avec comme fonction de coût L_{MAPE}, ce qui confirme que nos choix d'hyper-paramètres précédents sont pertinents.

Aucun modèle avec $N_{\text{cc}} \in \{2, 5\}$ n'est sélectionné. Pour 4 modèles, $N_{\text{cc}} = 3$. Le nombre de neurones par couche cachée est de 1000 pour 4 modèles sur 7, dont 3 sur les 4 avec $N_{\text{cc}} = 3$. Le WI le plus représenté est Glorot uniforme (5/7). Les FA sont disparates, chacune apparaissant une ou deux fois dans la sélection.

Chacun de ces modèles présente une réponse proche de 1 entre 70 et 400 GeV avec une résolution relative de l'ordre de 22 % à basse masse et 10 % à haute masse. Le modèle F conserve une réponse proche de 1 jusqu'à environ 500 GeV, cependant sa résolution à basse masse est légèrement dégradée par rapport aux autres modèles. Le modèle B présente l'avantage d'avoir des hyper-paramètres « consensus », c'est-à-dire que chacune des valeurs de ses hyper-paramètres correspond à la valeur la plus représentée dans la sélection. C'est à partir de ce modèle que nous avons choisi de continuer

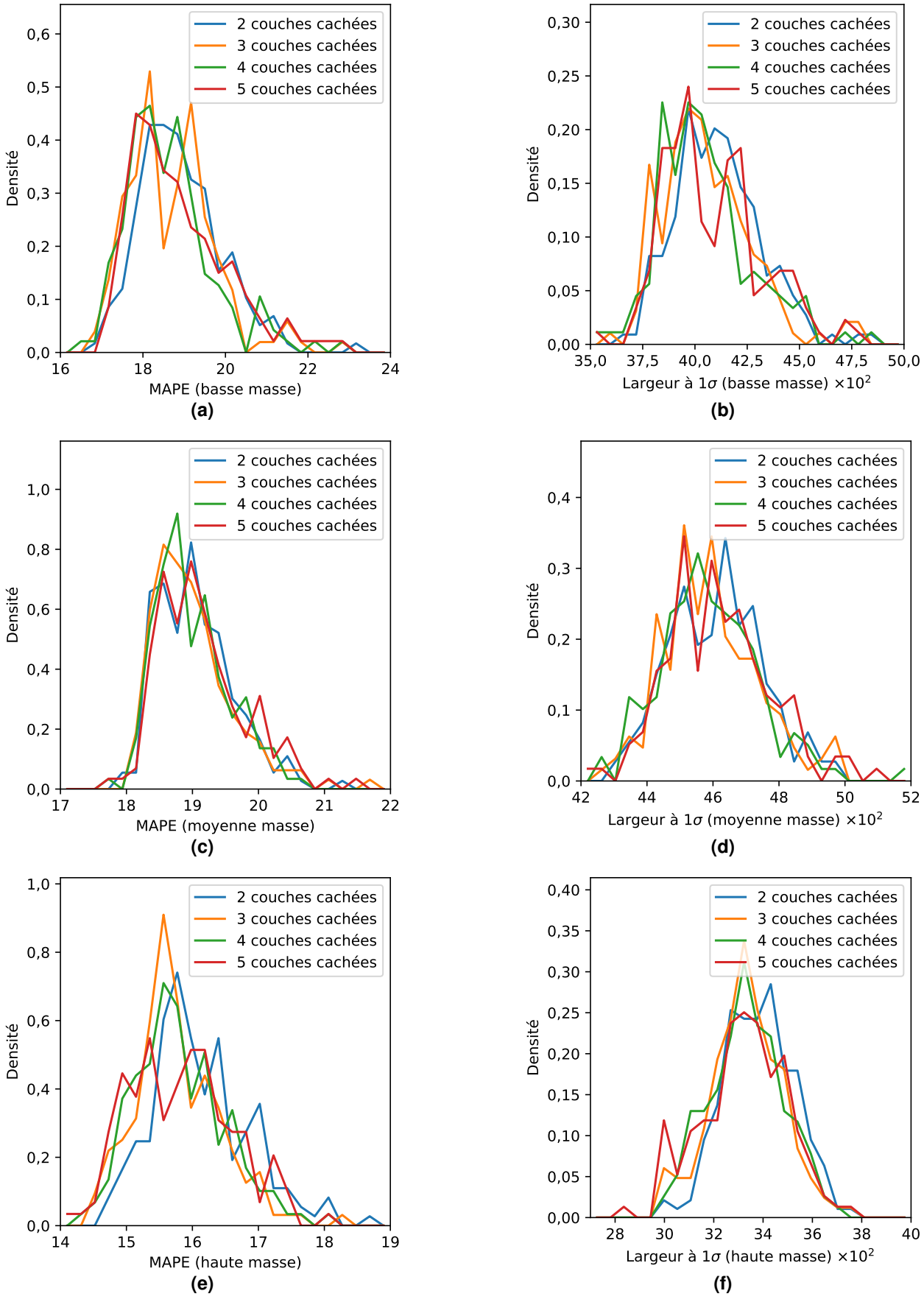


Figure X.17 – *Évaluations des DNNs regroupés selon N_{cc} par* $\mathrm{L}_{\mathrm{MAPE}}$ *et* $\Delta_{1\sigma}$.

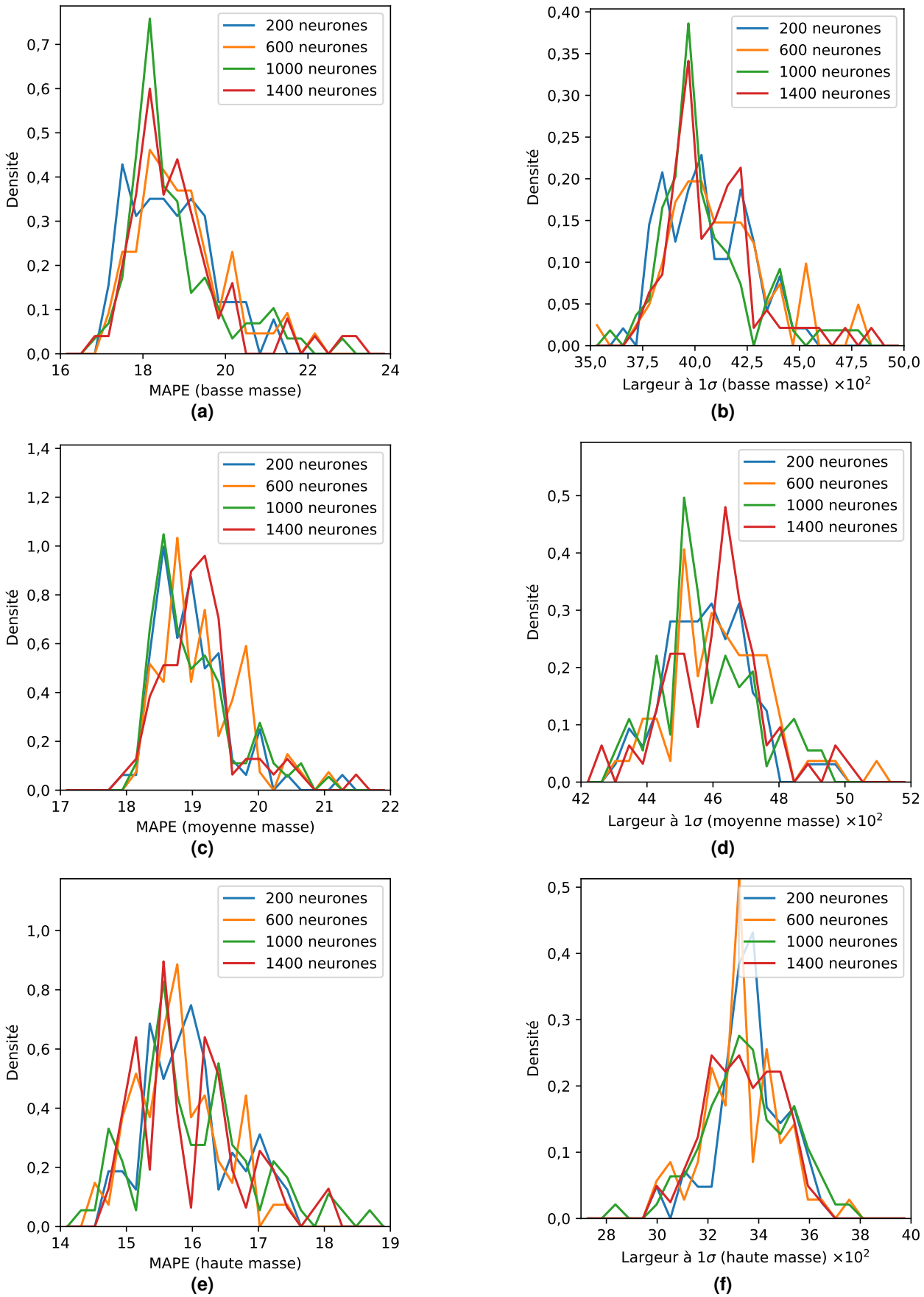


Figure X.18 – *Évaluations des DNNs regroupés selon $N_{n\ell}$ par* $\mathrm{L}_{\mathrm{MAPE}}$ *et* $\Delta_{1\sigma}$.

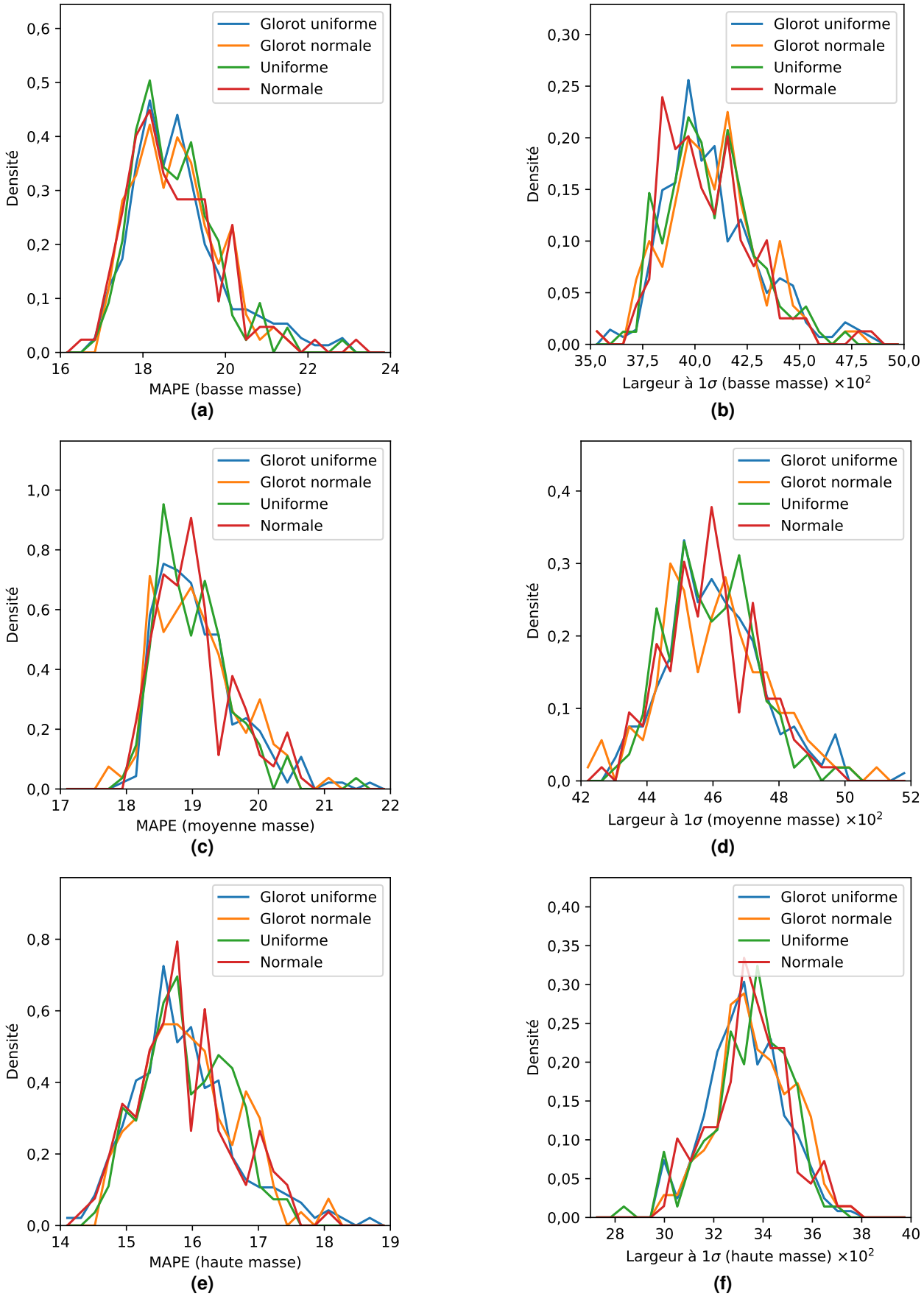


Figure X.19 – *Évaluations des DNNs regroupés selon le mode d'initialisation des poids par* $\mathrm{L}_{\mathrm{MAPE}}$ *et* $\Delta_{1\sigma}$.

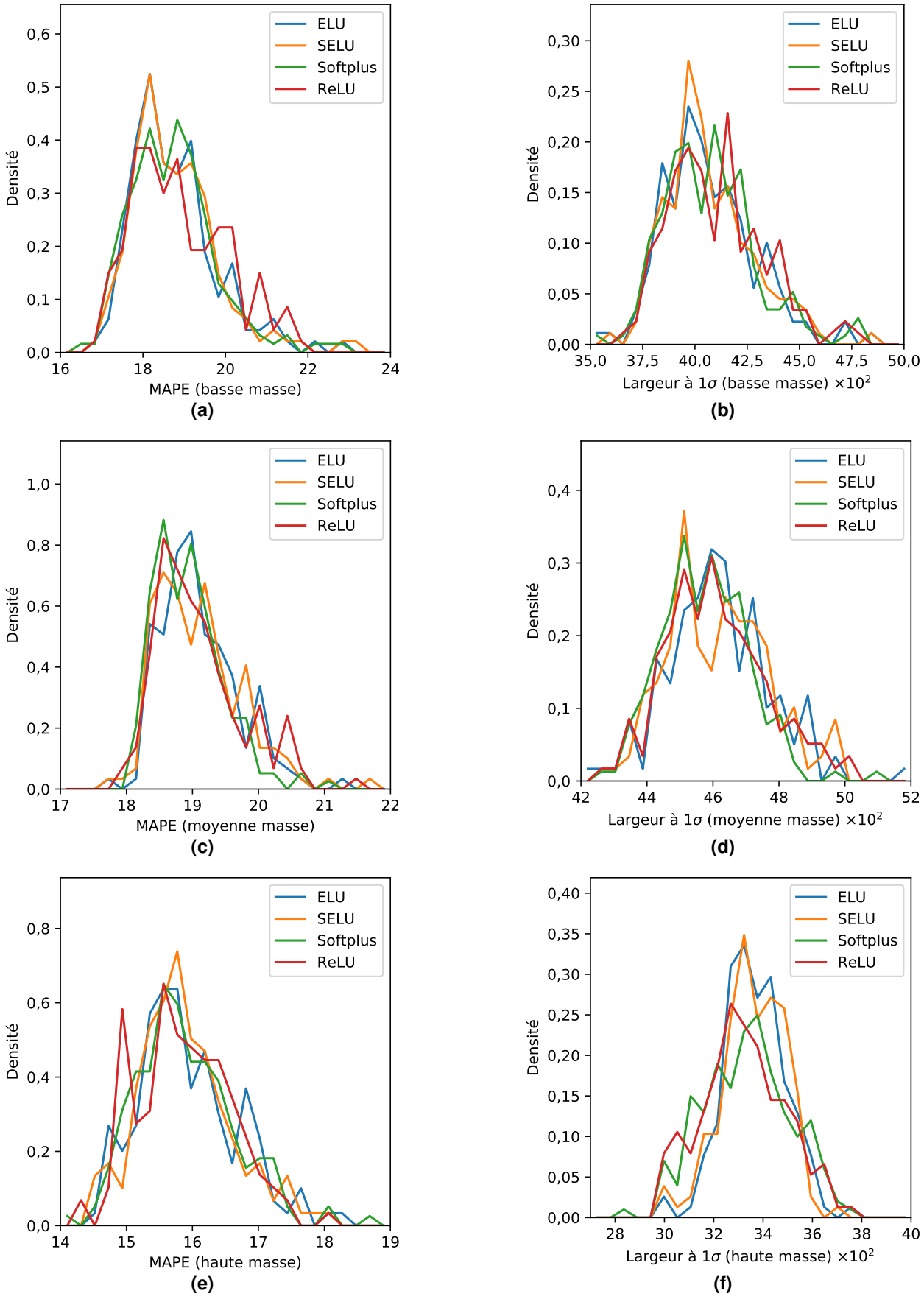


Figure X.20 – *Évaluations des DNNs regroupés selon la fonction d'activation par* $\mathrm{L}_{\mathrm{MAPE}}$ *et* $\Delta_{1\sigma}$.

Modèle	N_{cc}	$N_{n/c}$	WI	FA
A	3	1000	gu	ELU
B	3	1000	gu	Softplus
C	3	1000	n	SELU
D	3	1400	gu	ReLU
E	4	200	gn	ReLU
F	4	1000	gu	ELU
G	4	1400	gu	Softplus

Tableau X.1 – *Liste des 7 modèles sélectionnés.*

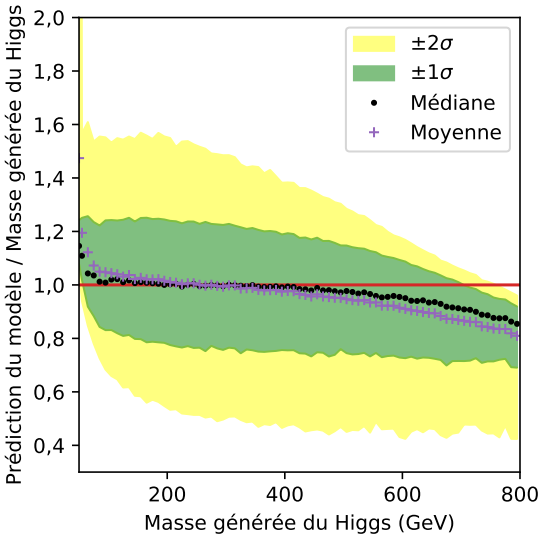


Figure X.21 – *Réponse du modèle A.*

notre étude. Les hyper-paramètres sélectionnés sont donnés dans le tableau X.2 avec une comparaison à ceux utilisés par Bärtschi & coll. [18].

Hyper-paramètre	Notre DNN	DNN de Bärtschi & coll. [18]
Nombre de couches cachées N_{cc}	3	4
Neurones par couche cachée $N_{n/c}$	1000	200
Fonction d'activation	Softplus	ReLU
Algorithme d'optimisation	Adam	Adam
Fonction de coût	L_{MAPE}	L_{MSE}
Initialisation des poids	« Glorot Uniforme » [47]	?
Nombre d'entrées	27 (voir section 2.4)	17 (voir [18])

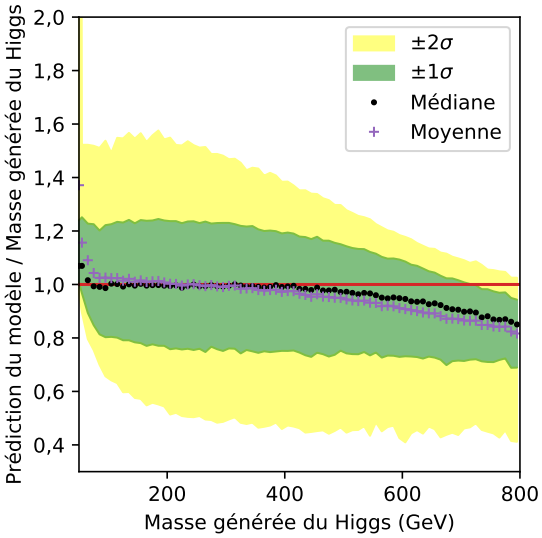
Tableau X.2 – *Comparaison de nos hyper-paramètres à ceux de Bärtschi & coll. Le mode d'initialisation des poids utilisé par Bärtschi & coll. n'est pas donné dans leur article.*

7 Discussions

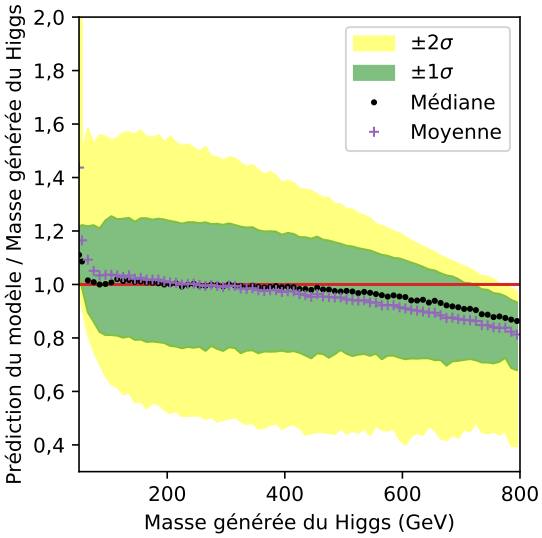
Lors de cette thèse, en parallèle de l'exploration des hyper-paramètres présentée en section 6, les effets de différents facteurs ont été étudiés. À des fins de cohérence dans la comparaison des effets, nous utilisons ici le modèle B sélectionné précédemment comme référence.

7.1 Effet de l'empilement

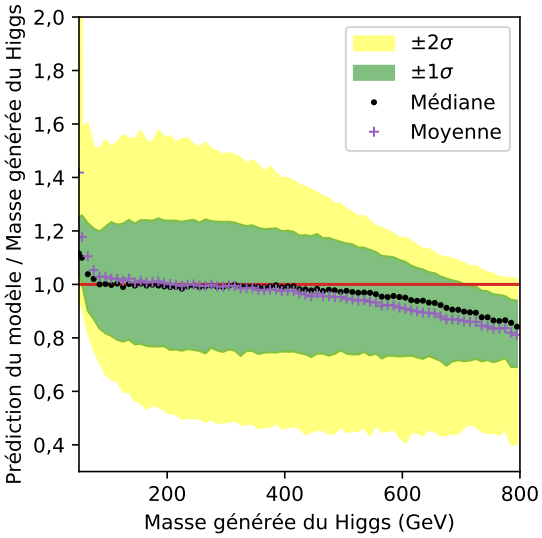
Dans les travaux de Bärtschi & coll. [18], l'empilement (PU, *Pile-Up*) n'est pas considéré. Nous avons donc souhaité déterminer son effet sur les prédictions de notre modèle. Pour cela, les mêmes événements que ceux décrits en section 2 ont été générés sans empilement. Un DNN, noté B^{0PU}, est entraîné sur ces événements sans empilement. Les hyper-paramètres de B^{0PU} sont identiques à ceux de B, à l'exception des variables d'entrée auxquelles N_{PU} est retiré, car $N_{PU} = 0$ pour tous les événements sans empilement. La réponse de B^{0PU} sur ces événements est représentée sur la figure X.23a. Sa moyenne est de $1{,}00 \pm 0{,}05$ pour $m_{\mathcal{H}}$ entre 70 GeV et 600 GeV avec une résolution relative de l'ordre de 22 % à basse masse et 10 % à haute masse. Ces performances sont donc similaires à celles de B sur les événements avec empilement.



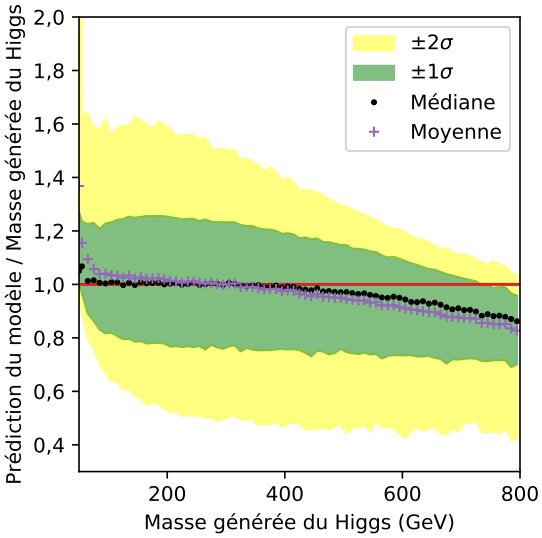
(a) *Réponse du modèle B.*



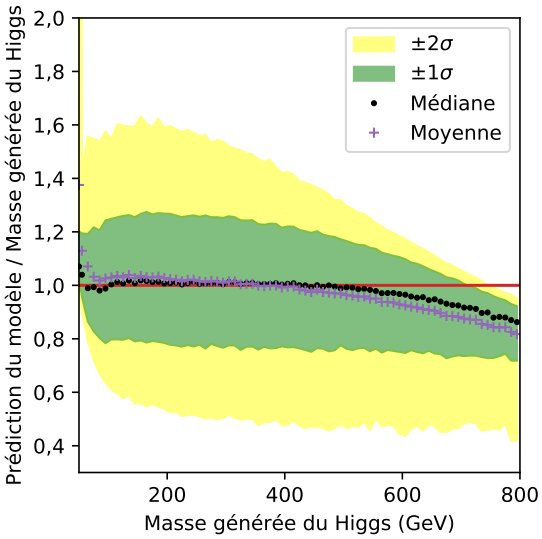
(b) *Réponse du modèle C.*



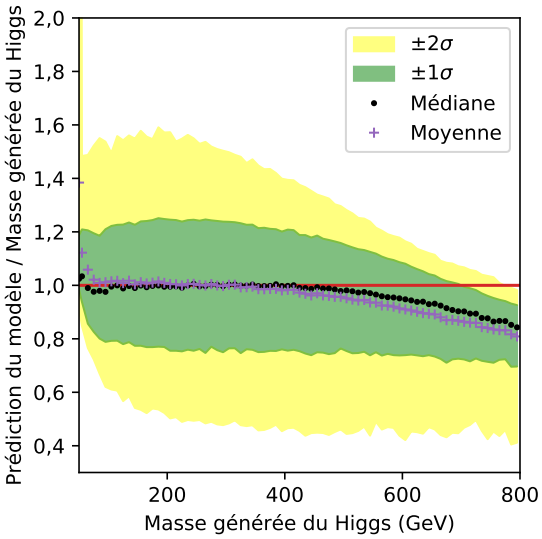
(c) *Réponse du modèle D.*



(d) *Réponse du modèle E.*

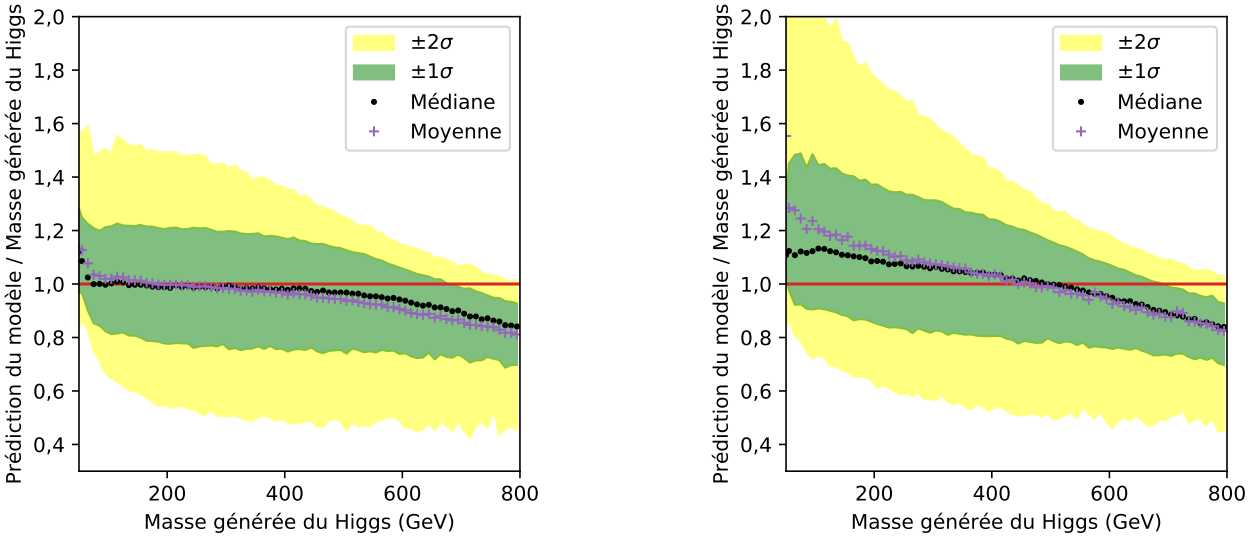


(e) *Réponse du modèle F.*



(f) *Réponse du modèle G.*

Figure X.22 – *Réponse des modèles B à G.*



(a) *Sans empilement.* **(b)** *Avec empilement.*

Figure X.23 – *Réponses du modèle B^{0PU} sur les événements sans et avec empilement.*

Cependant, la réponse de B^{0PU} est dégradée sur des événements contenant de l'empilement, figure X.23b. La réponse médiane se situe en effet à 1,13 à $m_{\mathcal{H}} = 100\,\text{GeV}$ et diminue à 0,83 à $m_{\mathcal{H}} = 800\,\text{GeV}$ avec empilement contre 1,00 et 0,83 sans empilement. La résolution relative à basse masse est de l'ordre de 25 % avec empilement.

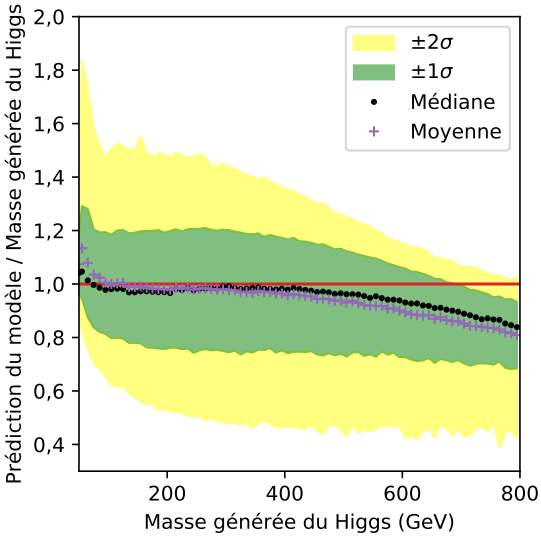
L'empilement peut être considéré comme un bruit blanc, dont l'énergie moyenne est reliée à N_{PU}. L'énergie portée par L_1 et L_2, les éléments visibles de la désintégration de $\mathcal{H}$, est en revanche reliée à $m_{\mathcal{H}}$. À haute masse, elle est grande par rapport à celle de l'empilement. Les prédictions de B^{0PU} ne sont alors pas perturbées, menant à des performances similaires à celles du cas sans empilement. Lorsque $m_{\mathcal{H}}$ diminue, l'énergie disponible pour L_1 et L_2 est moindre et le bruit d'empilement devient significatif. Le modèle B^{0PU} n'est pas entraîné pour traiter ce bruit et ses prédictions sont perturbées. Il est donc primordial d'inclure l'empilement dans l'entraînement dans l'optique d'une utilisation de nos modèles dans les analyses de CMS.

La réponse du modèle B dans le cas d'événements sans empilement, figure X.24a, peut être comparée à celle sur les événements en contenant, figure X.24b (identique à X.22a). Le profil d'empilement utilisé pour générer les événements d'entraînement est celui de l'année 2017. Il apparaît que le modèle B est peu sensible au retrait de l'empilement, les réponses étant similaires sur les figures X.24a et X.24b. L'utilisation de B sur des événements dont le profil d'empilement est légèrement différent de celui de l'année 2017, comme c'est le cas pour les autres années du Run II (2016, 2018), est ainsi directement envisageable.

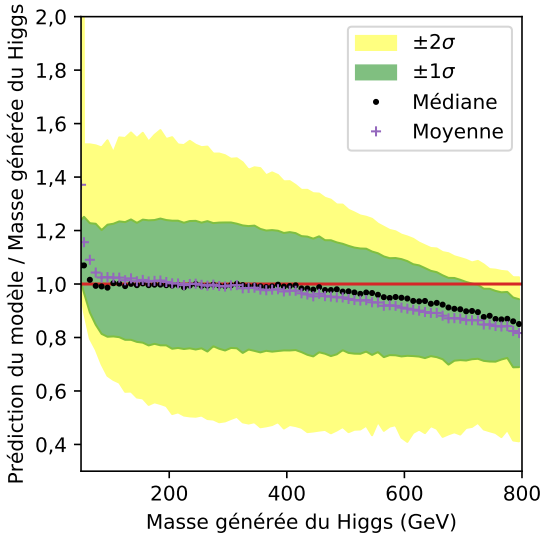
7.2 Effet de la reconstruction des particules

La reconstruction des particules est présentée dans le chapitre 2. Son effet peut être caractérisé par l'étude du modèle B^{gen}, ayant les mêmes hyper-paramètres que B, mais entraîné en utilisant les objets générés au lieu de ceux reconstruits pour L_1, L_2 et ${E_{\text{T}}}^{\text{miss}}$, c'est-à-dire pour les trois objets physiques liés à la désintégration des leptons τ (deux parties visibles L_1 et L_2 et ${E_{\text{T}}}^{\text{miss}}$ pour les neutrinos). En particulier, les valeurs de $\vec{p}_{\text{T}}{}^{L_1}$, $\vec{p}_{\text{T}}{}^{L_2}$ et $\vec{E}_{\text{T}}^{\text{miss}}$ correspondent exactement à la réalité. Il s'agit donc du cas dans lequel les objets physiques issus de $\mathcal{H}$ sont parfaitement reconstruits. Toutes les autres variables, y compris la matrice de covariance de ${E_{\text{T}}}^{\text{miss}}$, restent celles obtenues avec les objets reconstruits.

La figure X.25 montre les réponses du modèle B^{gen} sur les événements avec reconstruction parfaite et réelle. Dans le cas d'une reconstruction parfaite, la réponse médiane de B^{gen} est de l'ordre de $1{,}01 \pm 0{,}02$ de 70 à 800 GeV. La résolution relative est quant à elle de l'ordre de 3 %, soit près de sept fois mieux que B.

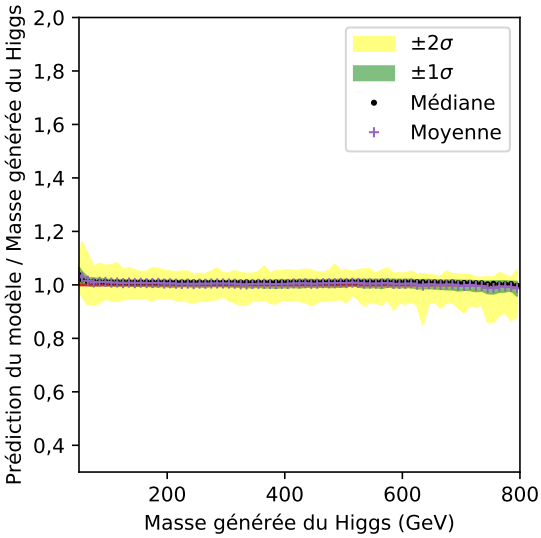


(a) *Sans empilement.*

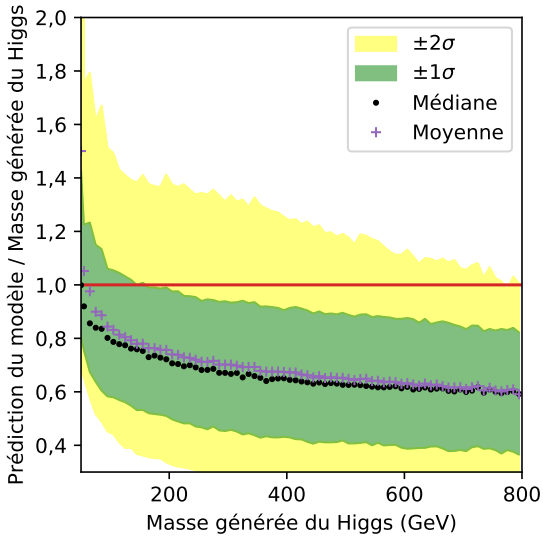


(b) *Avec empilement.*

Figure X.24 – *Réponses du modèle B sur les événements sans et avec empilement.*



(a) *Reconstruction parfaite.*



(b) *Reconstruction réelle.*

Figure X.25 – *Réponses du modèle B^{gen} dans le cas d'une reconstruction des particules parfaite et réelle.*

Les DNNs sont donc en mesure de comprendre la physique des événements $\mathcal{H} \to \tau\tau$ afin d'estimer $m_{\mathcal{H}}$ à partir des objets physiques générés correspondant aux objets effectivement reconstruits par le détecteur. Cependant, comme le montre la figure X.25b, l'utilisation de B^{gen} sur les variables reconstruites, effectivement accessibles expérimentalement, ne permet pas d'obtenir $m_{\mathcal{H}}$. En effet, la réponse moyenne de B^{gen} avec ces variables est inférieure à 1 et de l'ordre de 0,7 à haute masse. De plus, la résolution relative est de l'ordre de 40 %. Une des tâches des DNNs est donc de corriger cet effet de reconstruction.

7.3 Effet des faux taus hadroniques

La phénoménologie des événements contenant une paire de leptons τ est décrite dans le chapitre 1. Ces leptons peuvent se désintégrer hadroniquement en tau hadronique (τ_h) ou leptoniquement en électron (e) ou en muon (μ). Il existe ainsi six canaux différents dans les événements avec une paire de leptons τ, pouvant être répartis en trois groupes :

— complètement hadronique : $\tau_h\tau_h$, avec deux τ_h ;

— semi-leptoniques : $\mu\tau_h$ et $e\tau_h$, ou simplement $\ell\tau_h$, avec un τ_h ;
— leptoniques : $\mu\mu$, $e\mu$ et ee, ou simplement $\ell\ell$, sans τ_h.

Les faux taus hadroniques (faux τ_h) sont des objets physiques tels que des électrons, des muons et surtout des jets identifiés à tort comme des τ_h. Ils représentent près de 70 % des événements dans le canal $\tau_h\tau_h$, 38 % dans le canal $\mu\tau_h$ et 68 % dans le canal $e\tau_h$. Les faux τ_h sont particulièrement difficiles à modéliser dans les simulations [51, 52].

L'identification des τ_h est réalisée dans nos travaux à l'aide de l'algorithme DEEPTAU [11], qui présente un faible taux de mauvaise identification des τ_h, inférieur à 1 %. Cependant, une autre méthode d'identification des τ_h, basée sur un arbre de décision (BDT), peut être utilisée et présente un taux de mauvaise identification de jets en tant que τ_h pouvant atteindre 4 % [53]. Une sélection plus riche en faux τ_h est ainsi obtenue.

Les réponses du modèle B sur chacun des trois groupes de canaux (hadronique, semi-leptoniques et leptoniques) sont représentées figure X.26 pour les deux ensembles de sélection des τ_h. Quel que soit le groupe d'état final, les réponses pour $m_{\mathcal{H}} > 600\,\text{GeV}$ ne sont pas affectées par la sélection des τ_h. En effet, pour de hautes valeurs de $m_{\mathcal{H}}$, les τ_h ont des impulsions suffisamment élevées pour être correctement sélectionnés par la séquence d'analyse. À basse masse en revanche, les impulsions des faux τ_h sont compétitives vis-à-vis de celles des vrais τ_h. La séquence d'analyse forme alors des dileptons contenant des faux τ_h. La réponse du modèle s'en retrouve modifiée, jusqu'à 20 % de plus pour des masses entre 100 GeV et 600 GeV. L'effet le plus important se situe à très basse masse où la résolution est fortement dégradée.

La figure X.27 montre la différence $y_{\text{préd}} - y_{\text{vraie}}$ entre les prédictions du modèle B et la valeur vraie de $m_{\mathcal{H}}$ pour des valeurs de $m_{\mathcal{H}}$ entre 50 et 200 GeV sur chacun des trois groupes de canaux et pour les deux ensembles de sélection des τ_h. Dans les canaux leptoniques ($\ell\ell$), figures X.27e et X.27f, l'effet de la sélection des τ_h est moindre que dans les autres canaux. Il n'y a en effet aucun τ_h dans le dilepton, seule la sélection des événements est modifiée. Un objet physique identifié comme un τ_h par le BDT et non par DEEPTAU peut en effet faire basculer l'événement d'un canal à l'autre, si le τ_h identifié par le BDT permet de construire un dilepton.

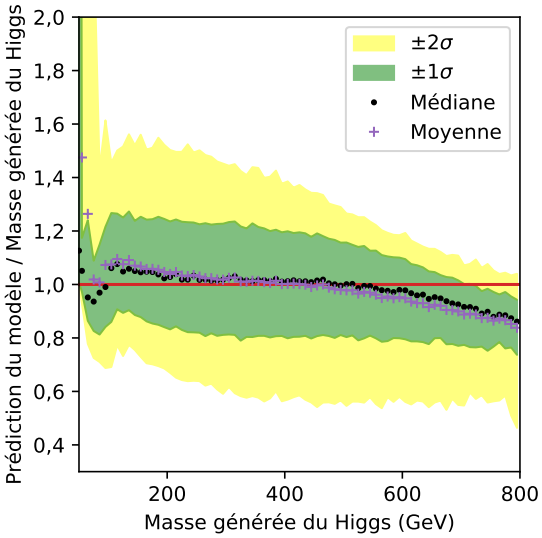
Dans le cas des canaux semi-leptoniques ($\ell\tau_h$), la différence entre $y_{\text{préd}}$ de B et y_{vraie} à basse masse est en moyenne inférieure à 10 GeV pour une sélection des τ_h par DEEPTAU, figure X.27c. La résolution relative est quant à elle inférieure à 25 %. Lorsque les τ_h sont identifiés par le BDT, figure X.27d, le modèle surestime $m_{\mathcal{H}}$ de 25 GeV en moyenne pour $70\,\text{GeV} < m_{\mathcal{H}} < 200\,\text{GeV}$ et de près de 40 GeV à $m_{\mathcal{H}} = 50\,\text{GeV}$. La résolution relative est de l'ordre de 25 % au-dessus de 70 GeV, moins bonne qu'avec DEEPTAU, et augmente drastiquement pour des masses plus basses, ce qui n'est pas le cas avec DEEPTAU. Il s'agit donc de la contribution des faux τ_h.

Dans le canal $\tau_h\tau_h$, figures X.27a et X.27b, un effet similaire existe. La résolution relative est toujours de l'ordre de 22 % au-delà de 100 GeV, mais la présence des faux τ_h mène à une surestimation moyenne de 30 GeV pour $m_{\mathcal{H}} > 110\,\text{GeV}$ et pouvant aller jusqu'à 100 GeV pour $m_{\mathcal{H}} \simeq 50\,\text{GeV}$, soit une erreur de 200 %. La dégradation de la résolution à très basse masse s'étend jusqu'à 100 GeV, au lieu de 70 GeV pour les canaux $\ell\tau_h$. L'effet des faux τ_h est donc plus important que dans les autres canaux, ce qui s'explique facilement par la présence de deux τ_h. Pour $m_{\mathcal{H}} = 50\,\text{GeV}$, la résolution de B sur les événements avec DEEPTAU est également mauvaise. La sélection des τ_h se fait avec $p_T > 40\,\text{GeV}$, ce qui est difficile à obtenir pour $m_{\mathcal{H}} = 50\,\text{GeV}$. Ces événements sont donc peu nombreux et vraisemblablement très contaminés par les faux τ_h.

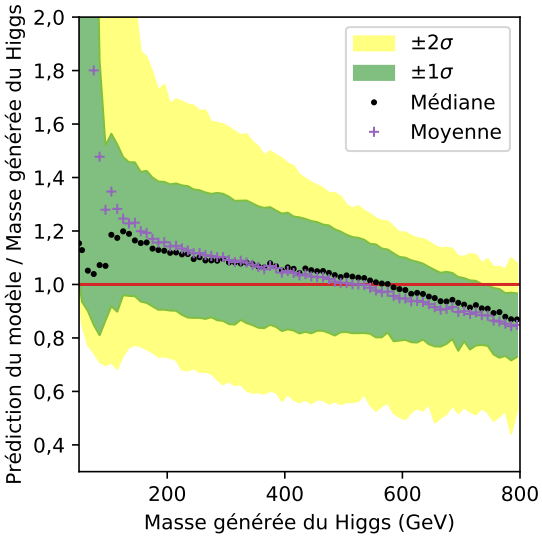
Les faux τ_h introduisent donc un biais important sur une large gamme de masse et en particulier dans la région des bosons Z ($m_Z = 91{,}2\,\text{GeV}$) et h ($m_h = 125{,}1\,\text{GeV}$). L'inclusion des faux τ_h dans l'entraînement est non triviale, car la masse à prédire n'est pas définie, les faux τ_h n'étant pas des objets physiques provenant de $\mathcal{H}$.

7.4 Effet de la séparation des canaux

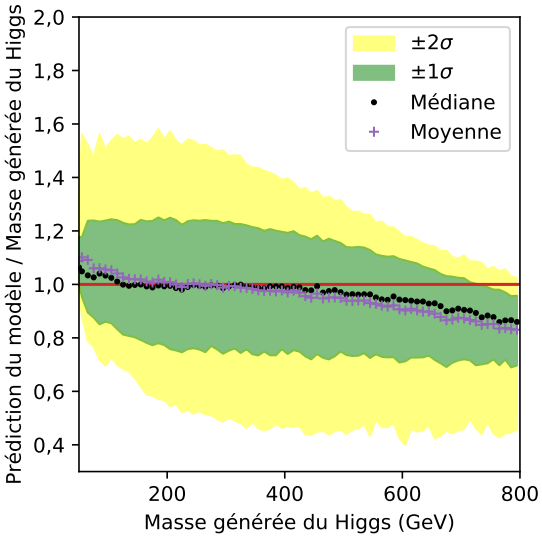
Les modèles construits sont entraînés et testés sur l'ensemble des événements, sans sélection sur le canal. Or, il est possible d'entraîner un DNN par canal afin de le spécialiser à la phénoménologie associée et obtenir, potentiellement, de meilleures estimations de $m_{\mathcal{H}}$.



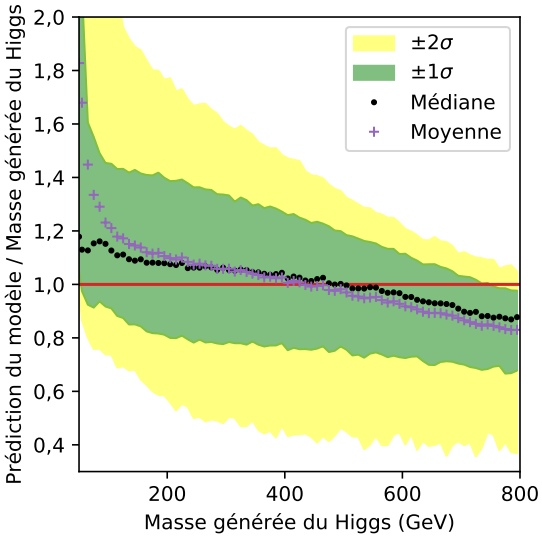
(a) *Canal* $\tau_h\tau_h$, DEEPTAU.



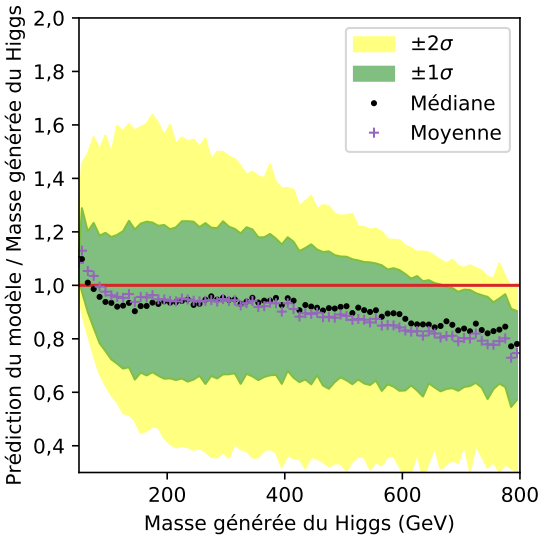
(b) *Canal* $\tau_h\tau_h$, *BDT.*



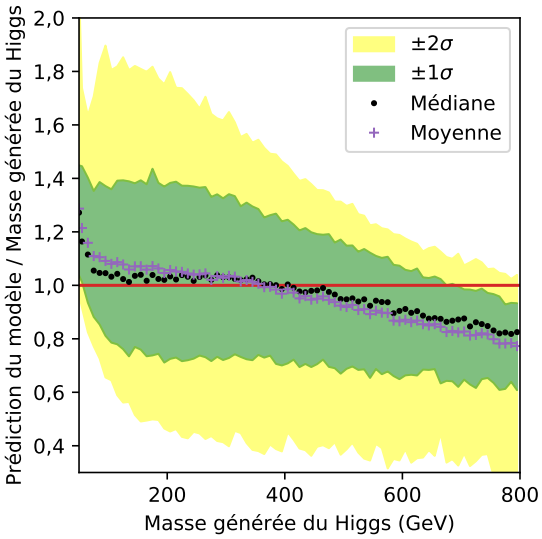
(c) *Canaux* $\ell\tau_h$, DEEPTAU.



(d) *Canaux* $\ell\tau_h$, *BDT.*

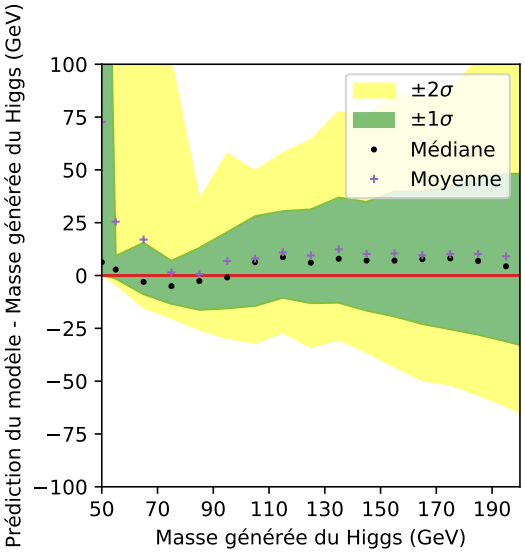


(e) *Canaux* $\ell\ell$, DEEPTAU.

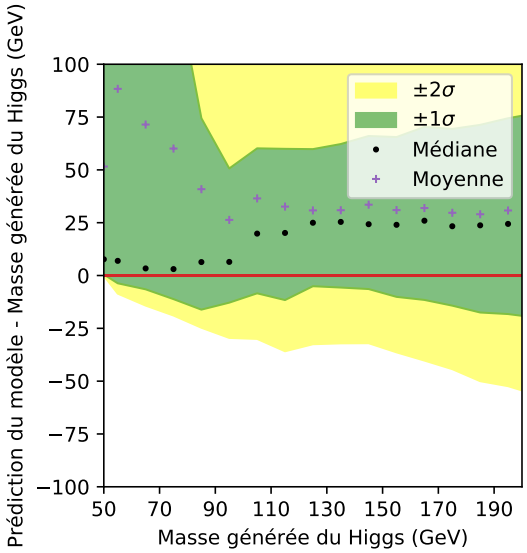


(f) *Canaux* $\ell\ell$, *BDT.*

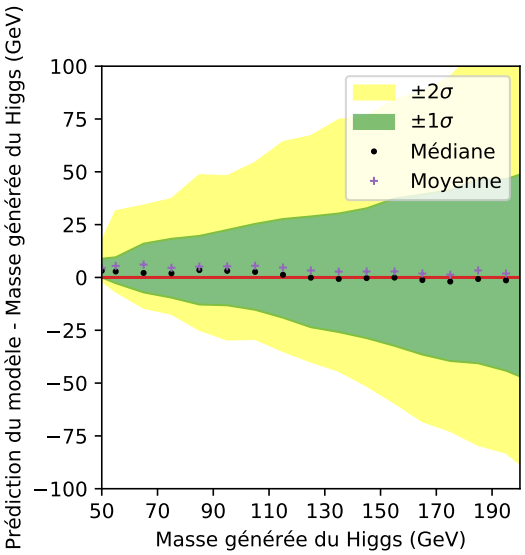
Figure X.26 – *Réponses du modèle B sur les différents types de canaux avec une quantité variable de faux* τ_h.



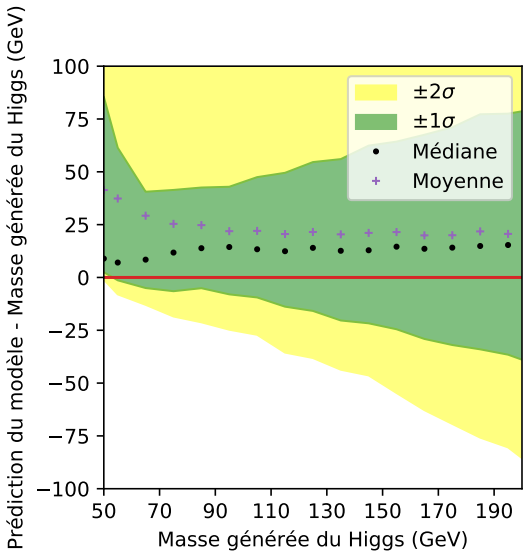
(a) *Canal* $\tau_h\tau_h$, DeepTau.



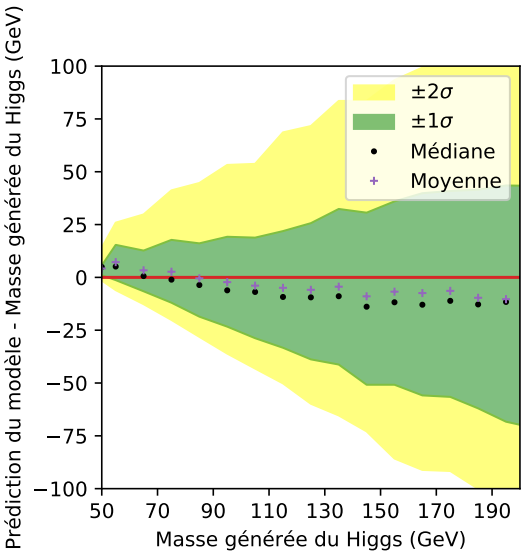
(b) *Canal* $\tau_h\tau_h$, *BDT.*



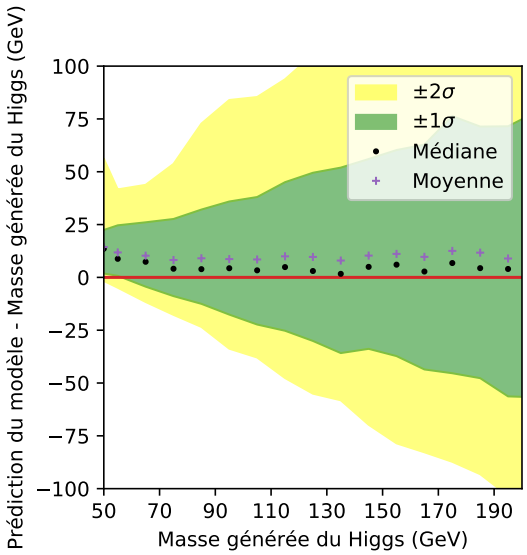
(c) *Canaux* $\ell\tau_h$, DeepTau.



(d) *Canaux* $\ell\tau_h$, *BDT.*



(e) *Canaux* $\ell\ell$, DeepTau.



(f) *Canaux* $\ell\ell$, *BDT.*

Figure X.27 – *Écarts à basse masse du modèle B sur les différents types de canaux avec une quantité variable de faux* τ_h.

Les modèles notés B^{x} possèdent les mêmes hyper-paramètres que B mais sont entraînés uniquement sur les événements du canal x.

7.4.1 Séparation en six canaux

Les figures X.28 et X.29 donnent les réponses des modèles $\mathrm{B}^{\tau_\mathrm{h}\tau_\mathrm{h}}$, $\mathrm{B}^{\mu\tau_\mathrm{h}}$, $\mathrm{B}^{e\tau_\mathrm{h}}$ et $\mathrm{B}^{\mu\mu}$, $\mathrm{B}^{e\mu}$, B^{ee} testés sur leurs canaux respectifs, comparées à celles de B sur les mêmes canaux.

Dans le canal $\tau_\mathrm{h}\tau_\mathrm{h}$, la réponse de $\mathrm{B}^{\tau_\mathrm{h}\tau_\mathrm{h}}$, figure X.28a, est plus proche de 1 que celle de B, figure X.28b, entre 200 et 450 GeV. Au-delà, elle est jusqu'à 3 % plus basse. Elle est en revanche jusqu'à 10 % plus haute à basse masse. Le comportement des modèles est toutefois similaire dans cette région : une baisse locale de la réponse est observable pour $m_{\mathcal{H}} \simeq 80\,\mathrm{GeV}$. La coupure sur l'impulsion transverse des τ_h étant de 40 GeV pour chacun des deux τ_h, il s'agit probablement d'une transition entre les événements avec une majorité de vrais τ_h ($m_{\mathcal{H}} > 80\,\mathrm{GeV}$) et ceux avec une contamination importante par les faux τ_h ($m_{\mathcal{H}} < 80\,\mathrm{GeV}$). Le modèle $\mathrm{B}^{\tau_\mathrm{h}\tau_\mathrm{h}}$ est donc difficilement entraîné dans cette région de masse. Pour $\mathrm{B}^{\tau_\mathrm{h}\tau_\mathrm{h}}$ et B, la résolution relative sur le canal $\tau_\mathrm{h}\tau_\mathrm{h}$ est de 20 %.

Le modèle $\mathrm{B}^{\mu\tau_\mathrm{h}}$, figure X.28c, possède une réponse équivalente à celle de B sur les mêmes événements, figure X.28d, pour des masses inférieures à 400 GeV. À haute masse, la réponse moyenne du modèle B est toutefois plus proche de 1. Le même constat peut être fait dans le cas du canal $e\tau_\mathrm{h}$, figures X.28e et X.28f.

Dans le cas des canaux leptoniques, figure X.29, l'utilisation de B plutôt que $\mathrm{B}^{\mu\mu}$, $\mathrm{B}^{e\mu}$ ou B^{ee} selon le canal permet d'améliorer la résolution relative sur $m_{\mathcal{H}}$ dont les valeurs sont données dans le tableau X.3. Les valeurs des réponses moyennes sont peu modifiées par rapport aux valeurs des résolutions.

Canal x	Modèle B^{x}		Modèle B	
	min	max	min	max
$\mu\mu$	20	50	10	40
$e\mu$	20	40	20	30
ee	20	50	10	30

Tableau X.3 – *Résolutions relatives minimales et maximales sur des intervalles de* 10 GeV *pour* $\mathrm{B}^{\mu\mu}$, $\mathrm{B}^{e\mu}$ *ou* B^{ee} *et B.*

Il semble ainsi préférable d'utiliser un seul modèle global plutôt qu'un modèle par canal. Cet effet peut être dû à la statistique plus faible à disposition lors de l'entraînement des DNNs séparément pour chaque canal. Un compromis peut être obtenu en séparant non pas par canal, mais par groupe de canaux.

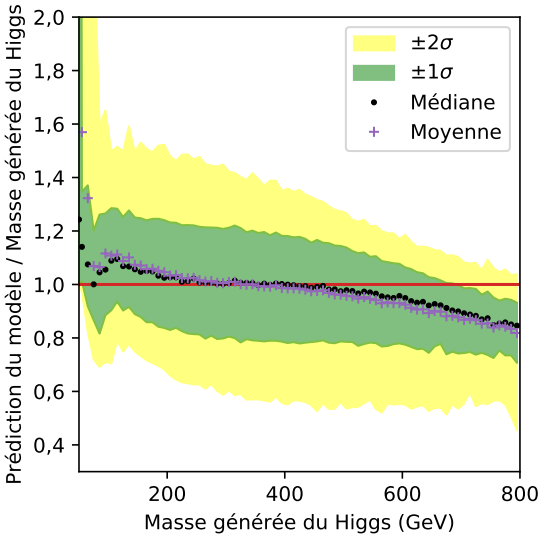
7.4.2 Séparation en trois groupes

En dehors de toute considération de reconstruction des particules, la phénoménologie des canaux d'un même groupe est sensiblement la même. Au lieu de séparer les six canaux ($\tau_\mathrm{h}\tau_\mathrm{h}$, $\mu\tau_\mathrm{h}$, $e\tau_\mathrm{h}$, $\mu\mu$, $e\mu$, ee), il est possible de former trois groupes ($\tau_\mathrm{h}\tau_\mathrm{h}$, $\ell\tau_\mathrm{h}$, $\ell\ell$), dans lesquels les quantités de τ_h et de neutrinos issus des désintégrations des leptons τ sont constantes. Cette nouvelle séparation permet ainsi d'avoir accès à de plus grandes quantités d'événements lors des entraînements, +100 % pour les canaux semi-leptoniques et +100 à +300 % pour les canaux leptoniques.

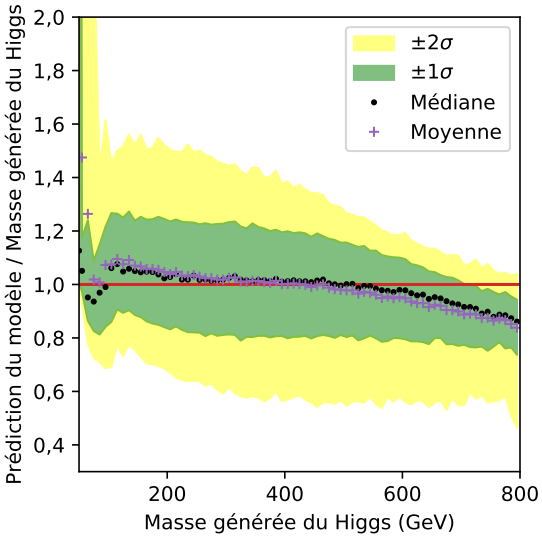
Le canal $\tau_\mathrm{h}\tau_\mathrm{h}$, seul de son groupe, est ainsi déjà traité dans la section précédente.

La figure X.30 compare le modèle $\mathrm{B}^{\ell\tau_\mathrm{h}}$, entraîné sur les canaux semi-leptoniques, à B utilisé sur ces mêmes événements. Pour des masses supérieures à 300 GeV, les deux modèles sont équivalents en termes de réponse et de résolution relative. En revanche, à basse masse, le modèle B a une réponse moyenne de 1,10 contre 1,18 pour $\mathrm{B}^{\ell\tau_\mathrm{h}}$.

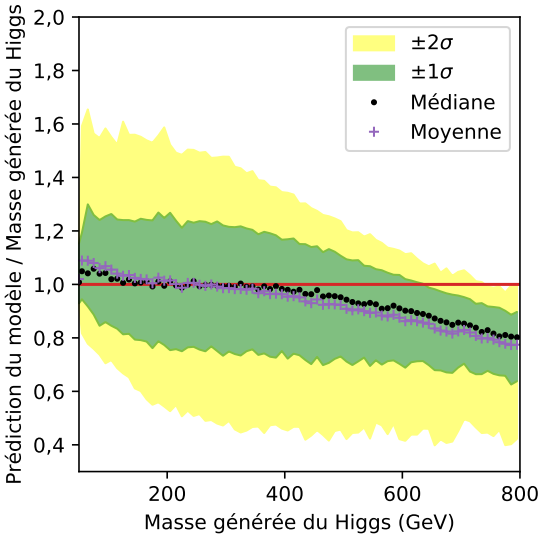
La figure X.31 compare le modèle $\mathrm{B}^{\ell\ell}$, entraîné sur les canaux leptoniques, à B utilisé sur ces mêmes événements. Les réponses de ces deux modèles sont équivalentes. Entre 100 et 200 GeV, la réponse moyenne de $\mathrm{B}^{\ell\ell}$ est légèrement plus proche de 1 que celle de B. Cette différence est toutefois négligeable face à la résolution relative de ces modèles, de l'ordre de 30 % dans cette région.



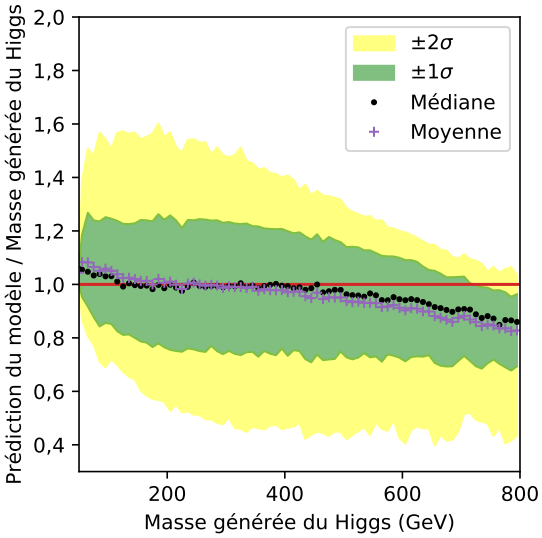
(a) *Modèle $B^{\tau_h\tau_h}$ testé sur $\tau_h\tau_h$.*



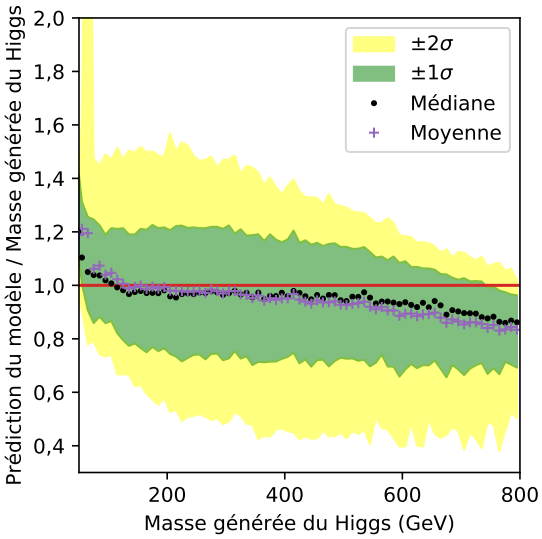
(b) *Modèle B testé sur $\tau_h\tau_h$.*



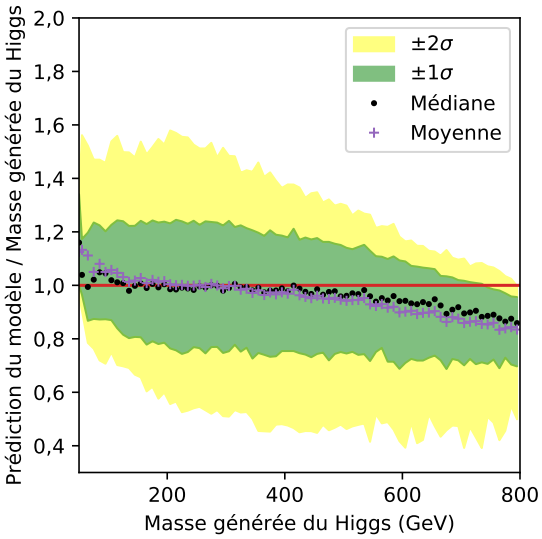
(c) *Modèle $B^{\mu\tau_h}$ testé sur $\mu\tau_h$.*



(d) *Modèle B testé sur $\mu\tau_h$.*

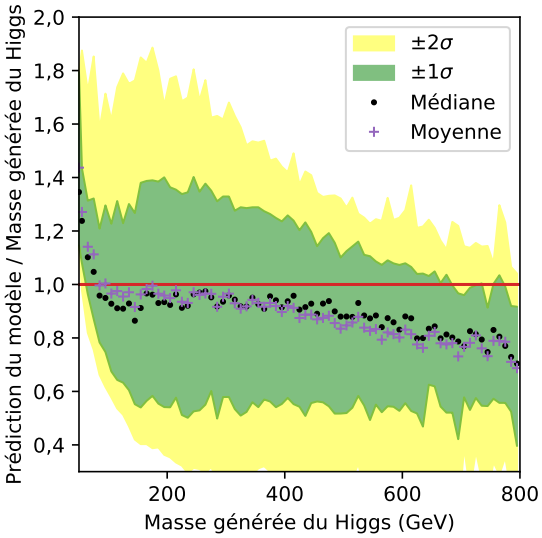


(e) *Modèle $B^{e\tau_h}$ testé sur $e\tau_h$.*

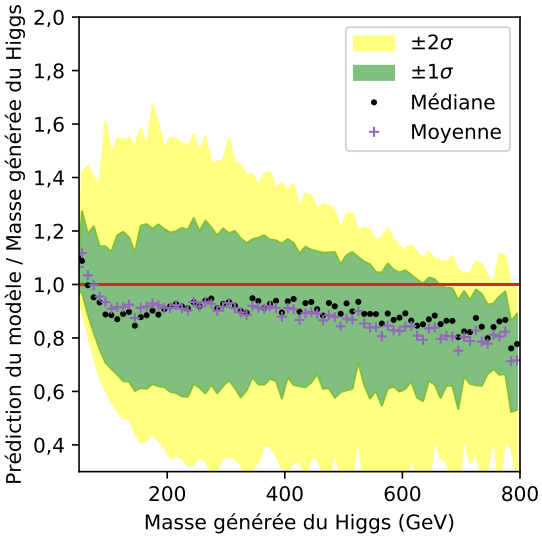


(f) *Modèle B testé sur $e\tau_h$.*

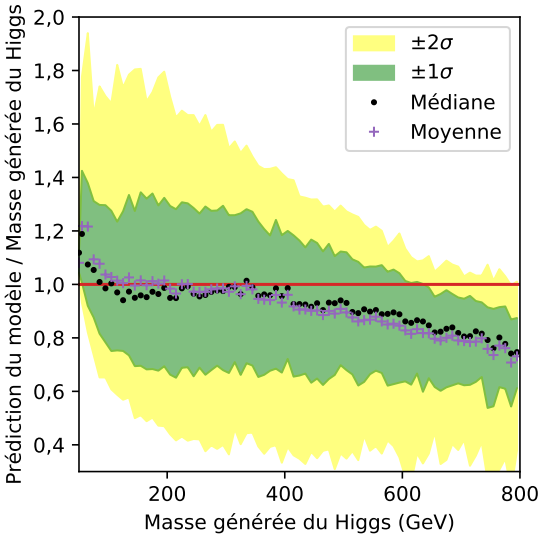
Figure X.28 – *Comparaison des modèles entraînés par canal ($\tau_h\tau_h$, $\mu\tau_h$, $e\tau_h$) au modèle B.*



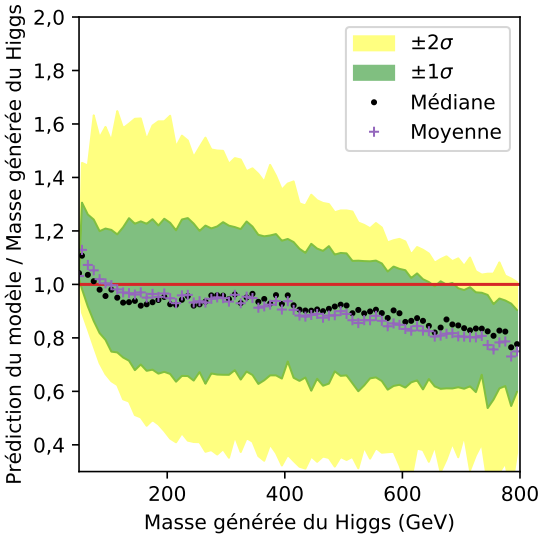
(a) *Modèle $B^{\mu\mu}$ testé sur $\mu\mu$.*



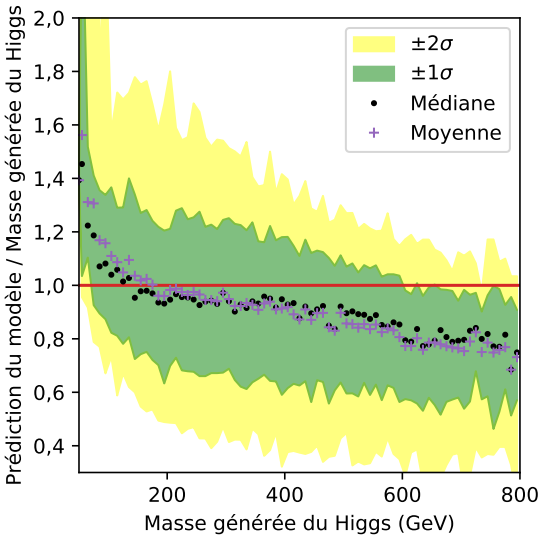
(b) *Modèle B testé sur $\mu\mu$.*



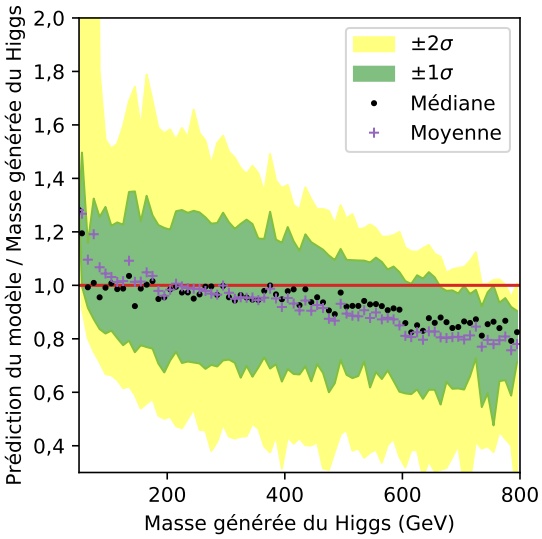
(c) *Modèle $B^{e\mu}$ testé sur $e\mu$.*



(d) *Modèle B testé sur $e\mu$.*

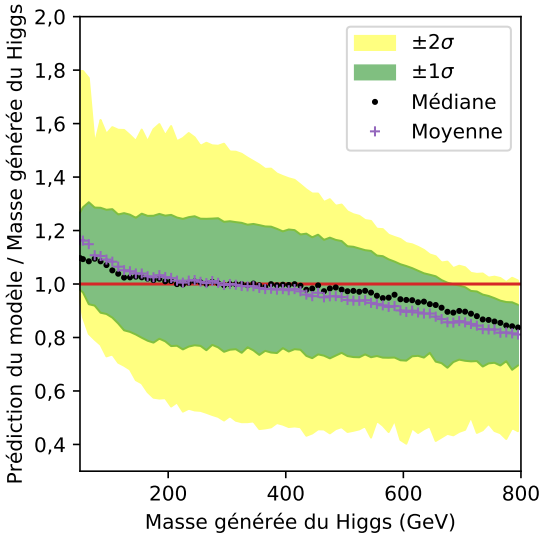


(e) *Modèle B^{ee} testé sur ee.*

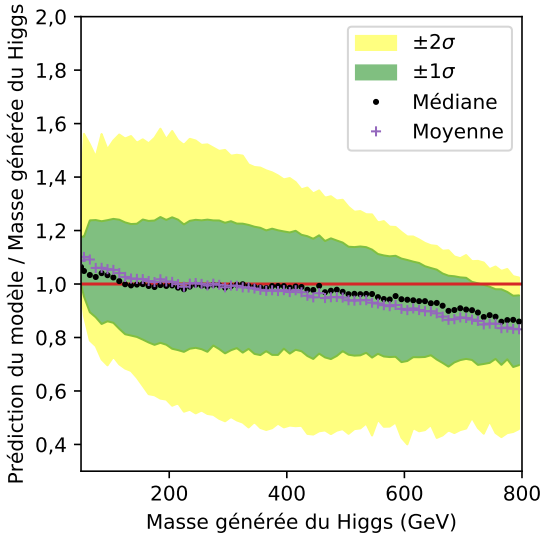


(f) *Modèle B testé sur ee.*

Figure X.29 – *Comparaison des modèles entraînés par canal ($\mu\mu$, $e\mu$, ee) au modèle B.*

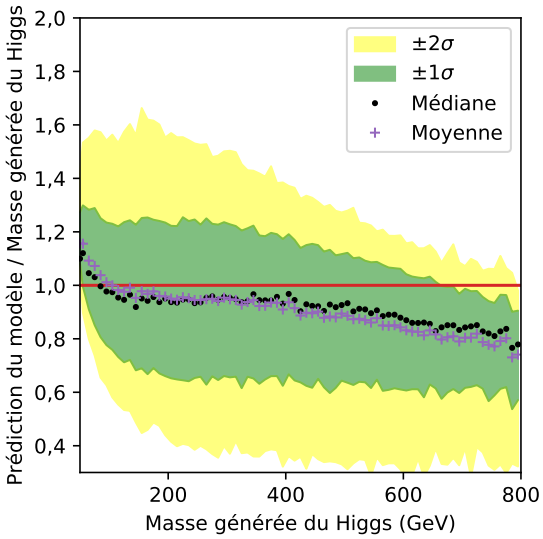


(a) *Modèle $B^{\ell\tau_h}$ testé sur $\ell\tau_h$.*

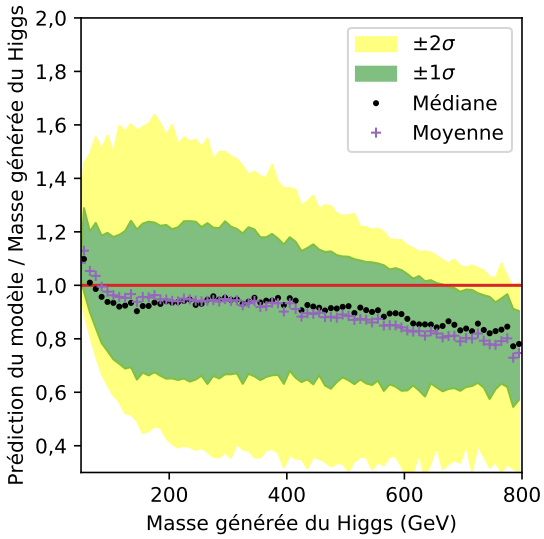


(b) *Modèle B testé sur $\ell\tau_h$.*

Figure X.30 – *Comparaison de $B^{\ell\tau_h}$ à B.*



(a) *Modèle $B^{\ell\ell}$ testé sur $\ell\ell$.*



(b) *Modèle B testé sur $\ell\ell$.*

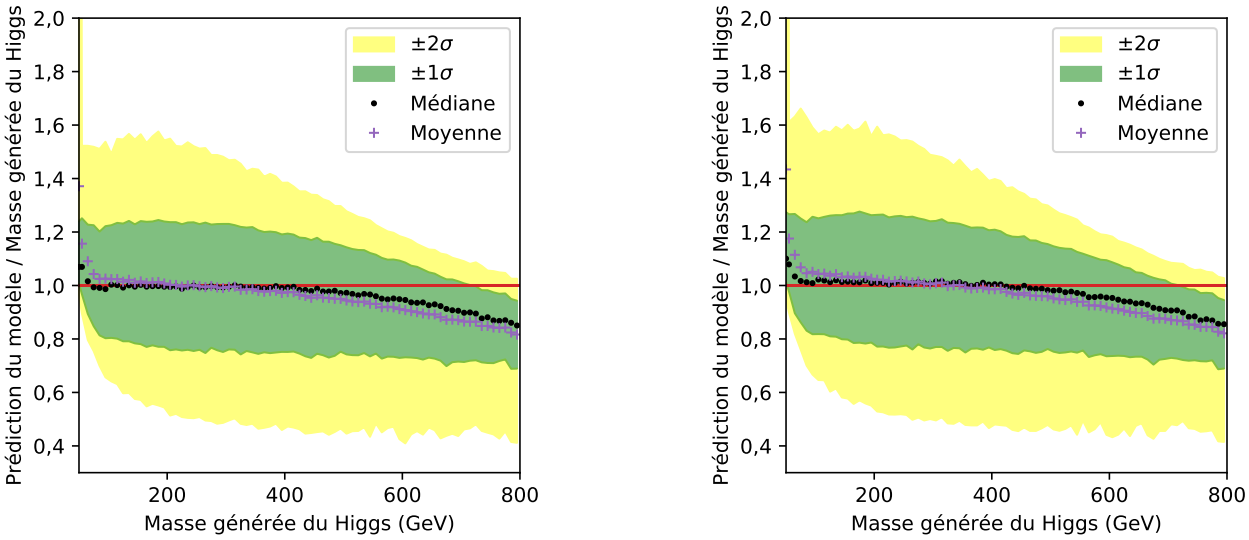
Figure X.31 – *Comparaison de $B^{\ell\ell}$ à B.*

Utiliser un modèle par groupe de canaux ne permet donc pas non plus d'améliorer les estimations obtenues par rapport au modèle de référence entraîné sur l'ensemble des canaux. Ce modèle, B, a en variable d'entrée le nombre attendu de neutrinos dans l'état final, N_ν^{reco}, directement relié au groupe du canal. En effet, $N_\nu^{\text{reco}} = 2$ pour le canal hadronique, 3 pour les semi-leptoniques et 4 pour les leptoniques. Or, comme vu dans la section 6.1, tout modèle privé de cette information a des performances dégradées. Le modèle B identifie donc vraisemblablement correctement le groupe de canal grâce à N_ν^{reco}.

7.5 Effet de la définition de E_T^{miss}

Dans l'analyse présentée au chapitre 4, E_T^{miss} est déterminée par l'algorithme PUPPI [38]. Il s'agit de la « PUPPIMET », avec laquelle les modèles ont donc été entraînés. Toutefois, une autre estimation de E_T^{miss} existe, obtenue directement à partir de l'algorithme de PF, la « PFMET ». Ces deux définitions de E_T^{miss} sont introduites dans le chapitre 2.

Certaines analyses utilisent PFMET plutôt que PUPPIMET. Nous avons donc souhaité vérifier la portabilité du modèle B entraîné avec PUPPIMET à une utilisation avec PFMET. La figure X.32 montre les réponses de B sur les mêmes événements, lorsque les variables reliées à E_T^{miss} sont obtenues à partir de PUPPIMET (figure X.32a) ou PFMET (figure X.32b).

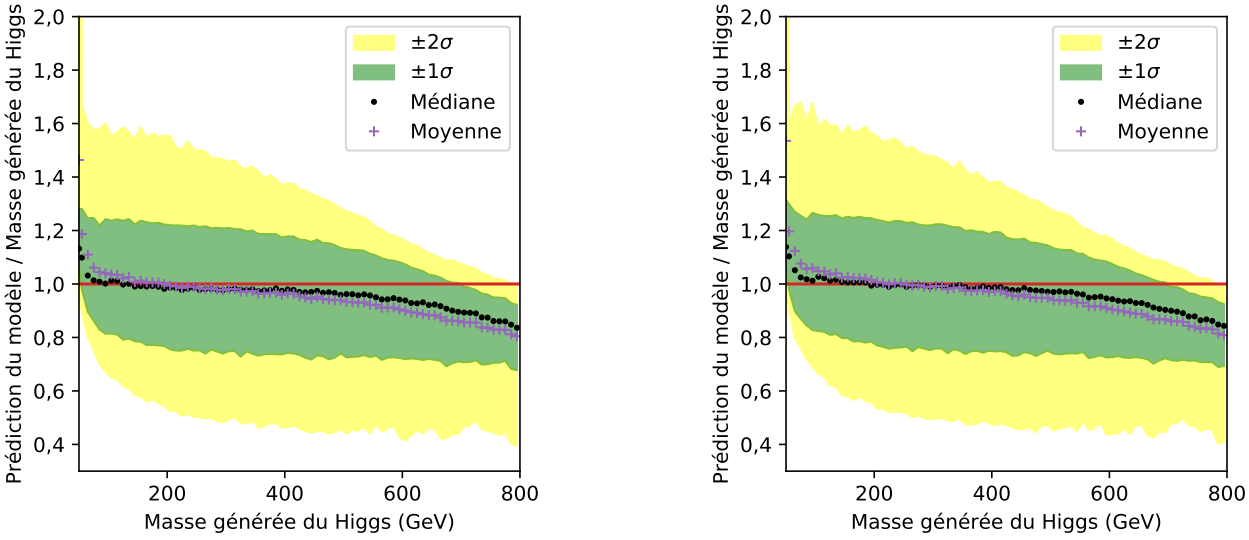


(a) *Modèle B testé avec* PUPPIMET. **(b)** *Modèle B testé avec PFMET.*

Figure X.32 – *Réponses du modèle B avec* PUPPIMET *ou PFMET.*

L'utilisation de PFMET augmente la réponse de B. Cette augmentation est cependant inférieure à 3 %. La résolution relative est inchangée. Il est donc tout à fait possible d'utiliser le modèle B avec PFMET, bien qu'il soit entraîné avec PUPPIMET.

La possibilité d'obtenir de meilleures prédictions en utilisant PFMET à l'aide du modèle B^{PF} entraîné directement avec cette définition de E_T^{miss} a également été étudiée. La figure X.33 présente les réponses obtenues.



(a) *Modèle B^{PF} testé avec* PUPPIMET. **(b)** *Modèle B^{PF} testé avec PFMET.*

Figure X.33 – *Réponses du modèle B^{PF} avec* PUPPIMET *ou PFMET.*

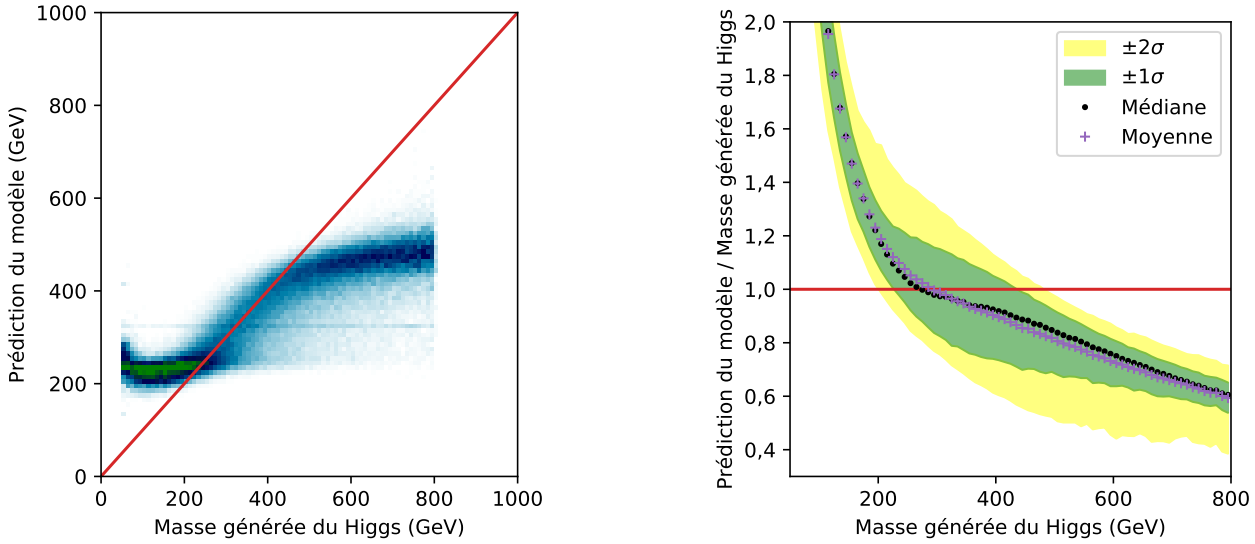
Le même effet de transition entre PFMET et PUPPIMET qu'avec B est observable avec B^{PF}. Cependant, la réponse moyenne de B avec PUPPIMET (figure X.32a) est égale à $1{,}00 \pm 0{,}05$ de 80 GeV à 425 GeV, alors que ce n'est le cas que de 200 GeV à 425 GeV pour B^{PF} avec PFMET (figure X.33b). De

plus, la réponse moyenne de B avec PFMET (figure X.32b) est plus proche de 1 pour $m_{\mathcal{H}} \simeq 100\,\text{GeV}$ que celle de B^{PF} avec PFMET (figure X.33b). Pour les analyses utilisant PFMET, le modèle B basé sur PUPPIMET est donc recommandé plutôt que B^{PF}.

7.6 Effet de l'intervalle de masse

7.6.1 Étendue de l'intervalle

L'intervalle de masse exploré lors de l'entraînement s'étend de 50 à 800 GeV, ce qui correspond aux valeurs de la masse du boson de Higgs modifié $\mathcal{H}$ utilisé lors de la génération des événements présentée dans la section 2. Cet intervalle est également le domaine de validité du modèle. La figure X.34 montre les performances du modèle $\text{B}^{\text{200-500}}$, équivalent au modèle B mais entraîné uniquement entre 200 et 500 GeV.



(a) *Prédictions en fonction des valeurs vraies.* **(b)** *Réponse.*

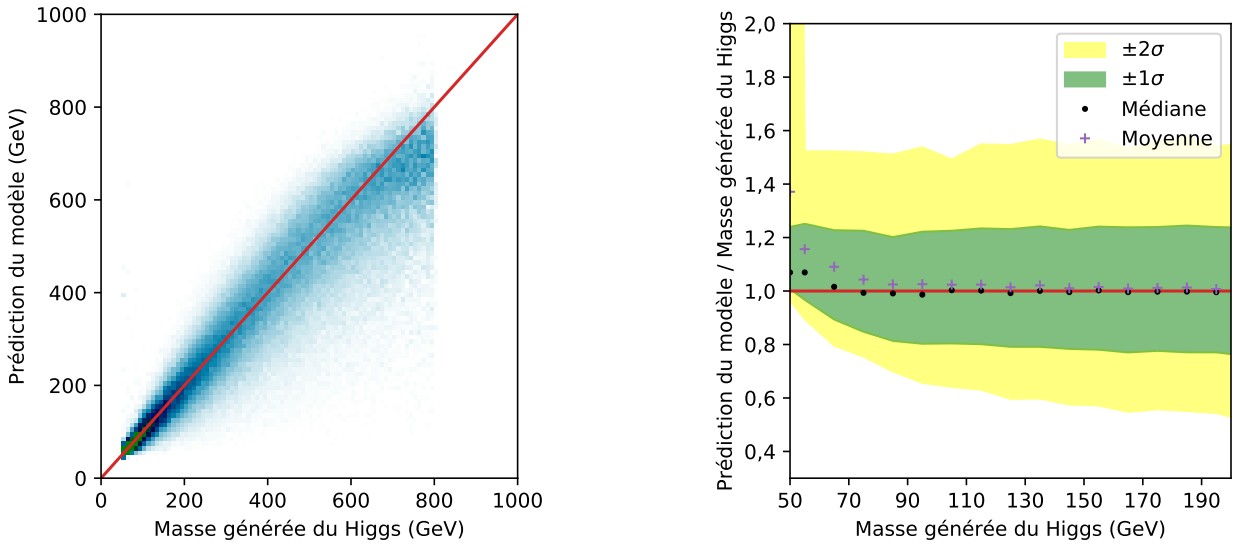
Figure X.34 – *Performances du modèle* $B^{200\text{-}500}$.

L'histogramme à deux dimensions des prédictions de $\text{B}^{\text{200-500}}$ en fonction de la vraie valeur de $m_{\mathcal{H}}$, figure X.34a, montre que les prédictions du modèle sont contenues dans l'intervalle d'entraînement, à quelques exceptions près. Dans l'intervalle d'entraînement, les prédictions sont cohérentes avec $m_{\mathcal{H}}$, la population de l'histogramme est proche de la première bissectrice ($y_{\text{préd}} = y_{\text{vraie}}$) en rouge. Les événements avec $m_{\mathcal{H}} < 200\,\text{GeV}$ sont prédits vers 230 GeV et ceux avec $m_{\mathcal{H}} > 500\,\text{GeV}$ sont prédits vers 480 GeV. Un modèle ne peut donc pas être utilisé afin de prédire des masses en dehors de son intervalle d'entraînement.

Afin d'obtenir un modèle pertinent dans l'optique d'une utilisation dans les analyses de CMS, il est donc important d'utiliser un intervalle contenant la gamme de masse des particules du modèle standard se désintégrant en paire de leptons τ, en particulier des bosons Z et h à 91,2 et 125,1 GeV respectivement. Pour une recherche de bosons de Higgs supplémentaires de haute masse, la limite supérieure de l'intervalle d'entraînement doit être la plus haute possible.

7.6.2 Effet de bord

Origine de l'effet de bord L'intervalle de masse utilisé pour l'entraînement du modèle B s'étend de 50 à 800 GeV. Comme discuté dans la section 2, il ne nous est pas possible de l'étendre avec la méthode utilisée pour générer les événements. Or, il existe un effet de bord sur les prédictions du modèle B lié à cet intervalle. La figure X.35a montre l'histogramme à deux dimensions des prédictions de B en fonction de la vraie valeur de $m_{\mathcal{H}}$. Pour $m_{\mathcal{H}} > 600\,\text{GeV}$, les prédictions de B saturent progressivement en dessous de 800 GeV. De même, la réponse moyenne de B à basse masse, figure X.35b, est de



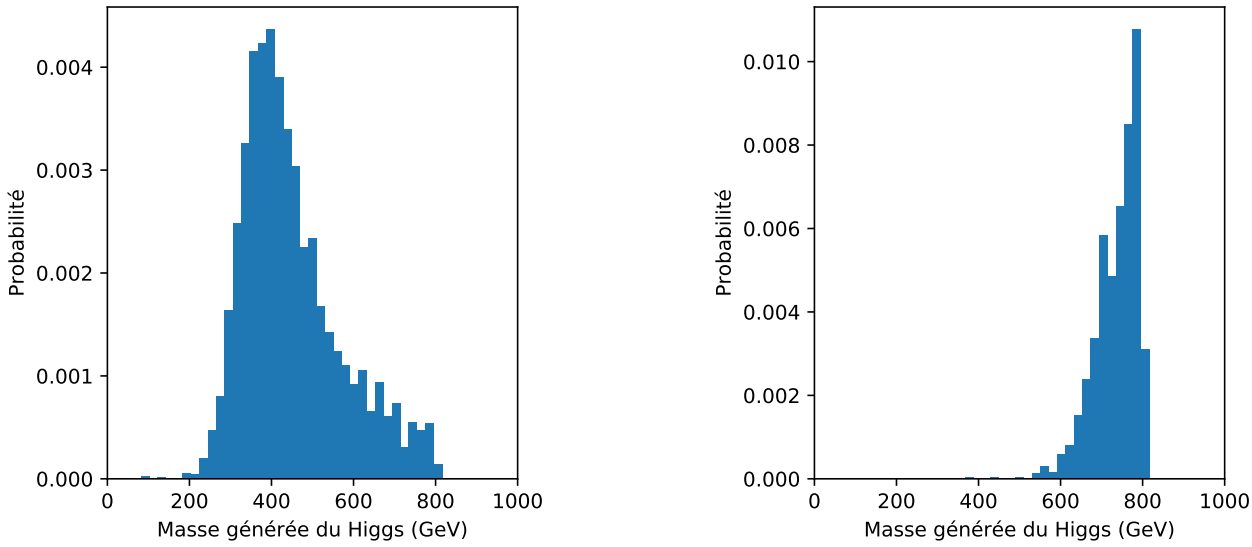
(a) *Prédictions en fonction des valeurs vraies.*

(b) *Réponse du modèle B à basse masse.*

Figure X.35 – *Performances du modèle B.*

$1{,}01 \pm 0{,}01$ pour $80\,\text{GeV} < m_{\mathcal{H}} < 200\,\text{GeV}$. En dessous de 80 GeV, la réponse moyenne augmente et la limite de l'écart-type inférieur (limite basse de la bande verte $\pm 1\sigma$) passe de 0,8 à 1,0.

Ainsi, $m_{\mathcal{H}}$ est surestimée à basse masse et sous-estimée à haute masse. Il s'agit de l'effet de bord de l'intervalle de masse. L'interprétation de cet effet est la suivante. Chaque ensemble d'événements prédits à une valeur de $y_{\text{préd}}$ donnée est une même famille du point de vue du DNN. En termes de classification en un nombre infini de catégories au lieu de régression, cela revient à dire qu'une famille est donc une catégorie identifiée. Sur la figure X.36 sont représentées les distributions de $y_{\text{vraie}} = m_{\mathcal{H}}$ pour des valeurs de la masse prédite par le modèle $y_{\text{préd}}$ de 400 (figure X.36a) et 750 GeV (figure X.36b) à ± 5 GeV. Il s'agit donc de tranches horizontales de l'histogramme de la figure X.35a, ce qui correspond à une famille d'événements selon le DNN.



(a) *Distribution de y_{vraie} pour $y_{\text{préd}} = 400 \pm 5\,\text{GeV}$.*

(b) *Distribution de y_{vraie} pour $y_{\text{préd}} = 750 \pm 5\,\text{GeV}$.*

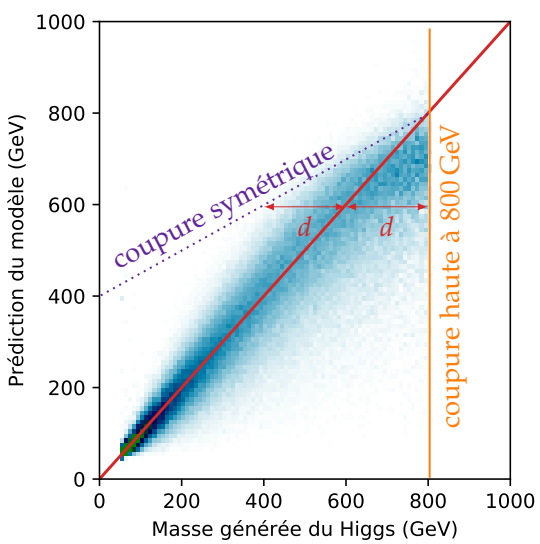
Figure X.36 – *Distributions de y_{vraie} à $y_{\text{préd}}$ fixée.*

Loin des bords de l'intervalle d'entraînement, figure X.36a, la distribution de y_{vraie} pour une famille est complète ; les queues de la distribution sont présentes des deux côtés de la valeur centrale. Lors de l'entraînement, le DNN « apprend » à prédire pour les événements de cette famille la valeur de $y_{\text{préd}}$ minimisant la fonction de coût sur cette distribution. Nous obtenons dans ce cas une valeur

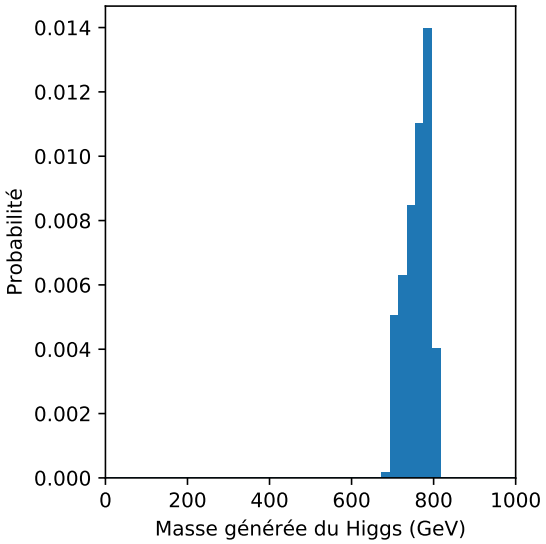
proche de 400 GeV, ce qui est correct.

Au niveau du bord à haute masse, figure X.36b, l'absence d'événements au-delà de 800 GeV donne une distribution tronquée de y_{vraie} pour une famille donnée. Seule l'extrémité à basse masse de la queue de la distribution est présente. Par conséquent, le DNN ne connaît que les basses masses de cette famille, il est ainsi tout à fait cohérent qu'il en sous-estime la masse. La minimisation de la fonction de coût mène donc à des prédictions biaisées. Si les événements au-delà de 800 GeV étaient présents dans cette famille, la distribution serait plus étendue vers les hautes masses, et la minimisation de la fonction de coût mènerait donc le DNN à prédire une masse plus élevée. L'effet est inversé pour le bord à basse masse, d'où la surestimation.

Principe de la correction La minimisation de la fonction de coût est donc réalisée correctement, mais sur des familles d'événements tronquées. Afin de contrer cet effet, l'idée retenue est de reformer des familles équilibrées en les tronquant de manière symétrique par rapport à la valeur de y_{vraie} devant leur correspondre. La figure X.37a illustre le principe de cette coupure symétrique.



(a) *Principe de la coupure symétrique à haute masse.*



(b) *Distribution de y_{vraie} pour $y_{\text{préd}} = 750 \pm 5$ GeV avec coupure symétrique.*

Figure X.37 – *Mise en place de la coupure symétrique.*

La troncature des familles due à l'extrémité de l'intervalle à 800 GeV est symétrisée par rapport au centre de valeur y_{vraie} devant correspondre à chaque famille. Ainsi, tout événement d'une famille (tranche horizontale) situé à une distance d de la valeur centrale à prédire (sur la diagonale rouge) plus grande que la distance de cette valeur centrale à l'extrémité de l'intervalle (trait orange vertical) est rejeté de l'entraînement. La distribution de la famille de la figure X.36b symétrisée par troncature selon cette méthode est présentée sur la figure X.37b. La valeur de la fonction de coût sur cette distribution n'est plus minimale pour la valeur précédemment prédite par le modèle, elle l'est en revanche pour une valeur prédite plus élevée. Le modèle actuel sous-estimant $m_{\mathcal{H}}$ à haute masse, cette correction doit donc permettre d'obtenir un nouveau modèle avec une réponse plus proche de 1.

Modification de la fonction de coût Une coupure symétrique est également mise en place à basse masse en suivant le même principe. Cinq zones peuvent alors être définies dans le plan $(y_{\text{vraie}}, y_{\text{préd}})$ afin d'appliquer les coupures symétriques aux deux extrémités de l'intervalle de masse :

Zone 1 $y_{\text{vraie}} < 50$ GeV : absence d'événements due à la masse minimale de l'entraînement ;

Zone 2 $y_{\text{vraie}} > 800$ GeV : absence d'événements due à la masse maximale de l'entraînement ;

Zone 3 $|y_{\text{préd}} - y_{\text{vraie}}| > |800\,\text{GeV} - y_{\text{préd}}|$: zone d'exclusion à haute masse ;

Zone 4 $|y_{\text{préd}} - y_{\text{vraie}}| > |50\,\text{GeV} - y_{\text{préd}}|$: zone d'exclusion à basse masse ;

Zone 5 zone centrale.

Ces zones sont illustrées sur la figure X.38. Dans la zone centrale, la fonction de coût est utilisée de manière classique, sans changement. Dans les zones 3 et 4 d'exclusion en revanche, pour ne pas prendre en compte les événements afin de symétriser les bords de l'intervalle d'entraînement au sein d'une famille, la fonction de coût est rendue égale à zéro. En d'autres termes, nous procédons au changement $\mathsf{L} = \mathsf{L}_{\text{MAPE}} \to \mathsf{L}'$ avec

$$\mathsf{L}' = \mathsf{L}_{\text{MAPE}} \times \begin{cases} 0 & \text{si } (y_{\text{vraie}}, y_{\text{préd}}) \in \text{zones 3 ou 4} \\ 1 & \text{sinon} \end{cases} . \tag{X.39}$$

Cette fonction de coût ne respecte pas la condition

$$\arg\min_{y_{\text{préd}}} \left(\mathsf{L}(y_{\text{vraie}}, y_{\text{préd}})\right) = \{\, y_{\text{vraie}} \,\} , \tag{X.40}$$

c'est-à-dire que le minimum de L soit atteint lorsque $y_{\text{préd}} = y_{\text{vraie}}$. Des problèmes de convergence lors de l'entraînement peuvent alors survenir. C'est effectivement ce que nous avons pu observer lors de la mise en place de cette fonction de coût avec la condition d'exclusion de la zone 4. Nous avons alors choisi de multiplier la valeur de la fonction de coût par 0,1 dans la zone 4 au lieu de 0. De plus, nous avons observé que la multiplication de L_{MAPE} par la racine de y_{vraie}, conjointement avec les conditions d'exclusion, permettait d'améliorer encore la réponse du modèle. La fonction de coût ainsi utilisée est $\mathsf{L}_{\text{MA}\sqrt{\text{PE}}\times\text{b}}$, définie par

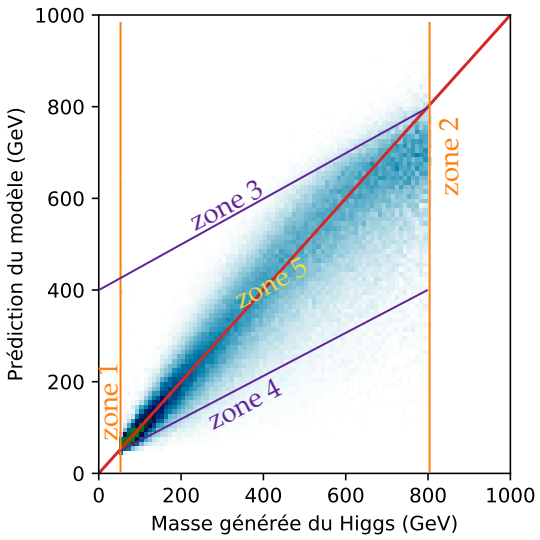


Figure X.38 – *Zones considérées pour l'entraînement.*

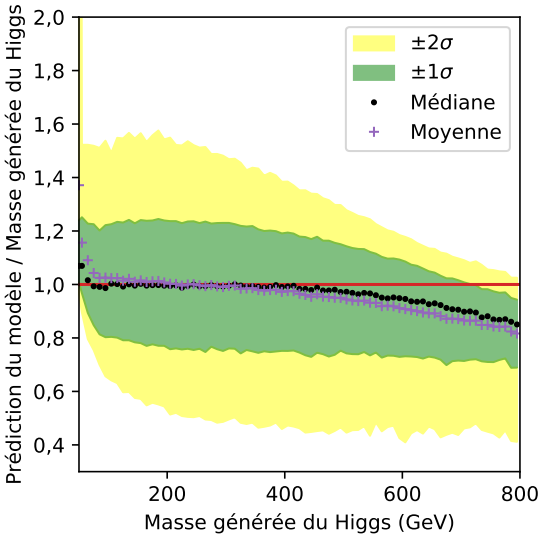
$$\mathsf{L}_{\text{MA}\sqrt{\text{PE}}\times\text{b}}(y_{\text{vraie}}, y_{\text{préd}}) = \mathsf{L}_{\text{MA}\sqrt{\text{PE}}}(y_{\text{vraie}}, y_{\text{préd}}) \times \begin{cases} 0 & \text{si } (y_{\text{vraie}}, y_{\text{préd}}) \in \text{zone 3} \\ 0{,}1 & \text{si } (y_{\text{vraie}}, y_{\text{préd}}) \in \text{zone 4} \\ 1 & \text{sinon} \end{cases} \tag{X.41}$$

avec

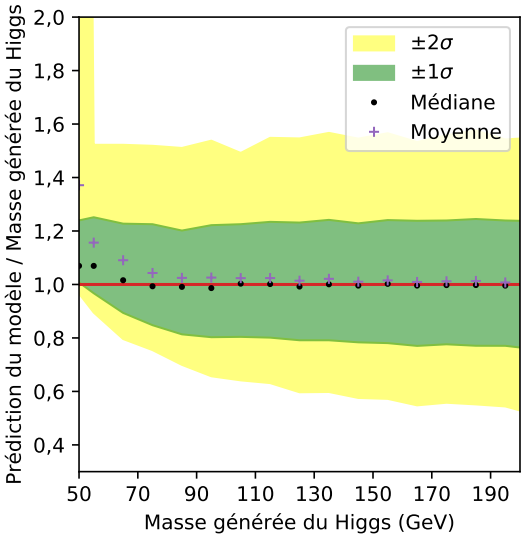
$$\begin{aligned} \mathsf{L}_{\text{MA}\sqrt{\text{PE}}}(y_{\text{vraie}}, y_{\text{préd}}) &= \mathsf{L}_{\text{MAPE}}(y_{\text{vraie}}, y_{\text{préd}}) \times \sqrt{y_{\text{vraie}}} = \left| \frac{y_{\text{préd}} - y_{\text{vraie}}}{y_{\text{vraie}}} \right| \times \sqrt{y_{\text{vraie}}} \\ \Leftrightarrow \mathsf{L}_{\text{MA}\sqrt{\text{PE}}}(y_{\text{vraie}}, y_{\text{préd}}) &= \left| \frac{y_{\text{préd}} - y_{\text{vraie}}}{\sqrt{y_{\text{vraie}}}} \right| . \end{aligned} \tag{X.42}$$

Nouveau modèle obtenu Le nouveau modèle obtenu, noté B', est comparé à B sur la figure X.39. Pour des masses inférieures à 70 GeV, la réponse médiane de B', en figure X.39d, est égale à 1 alors que celle de B, en figure X.39b, est sujette à l'effet de bord. À haute masse, la réponse de B' en figure X.39c est également plus proche de 1 que celle de B, en figure X.39a. L'utilisation de $\mathsf{L}_{\text{MA}\sqrt{\text{PE}}}$ comme fonction de coût permet donc de supprimer l'effet de bord à basse masse et de le réduire à haute masse.

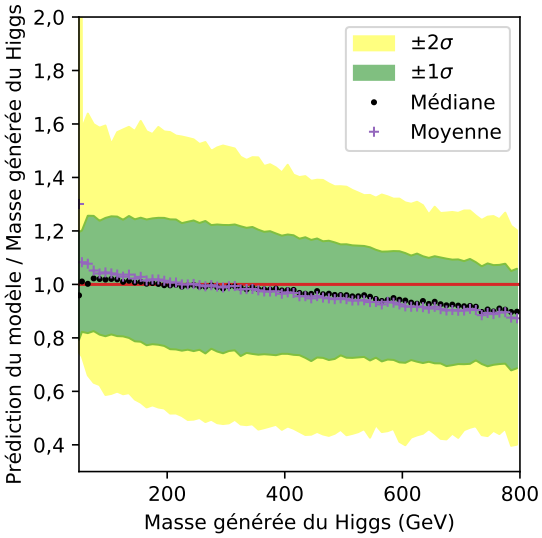
Exploitation de la queue à haute masse des événements générés Lors de la génération des événements à haute masse, la largeur du boson de Higgs permet d'obtenir des événements avec $m_{\mathcal{H}}$ supérieure à 800 GeV, comme discuté en section 2. Jusqu'ici, nous ne considérions que les événements tels que $50\,\text{GeV} \leqslant m_{\mathcal{H}} \leqslant 800\,\text{GeV}$. L'inclusion de la queue à haute masse des événements générés permet d'étendre artificiellement l'intervalle d'entraînement jusqu'à 1 TeV. Le biais éventuel dû à la faible quantité d'événements au-delà de 800 GeV est évité grâce à la pondération présentée dans la section 2. De même, les définitions des cinq zones utilisées pour déterminer la fonction de coût sont adaptées à la nouvelle valeur maximale de $m_{\mathcal{H}}$ fixée à 1 TeV. Le modèle B" ainsi obtenu est comparé à B' sur la figure X.39. La réponse à basse masse de B", figure X.39f, est semblable à celle de B', figure X.39d, ce qui est attendu. En revanche, la réponse moyenne de B" à haute masse, figure X.39e, est de $1{,}00 \pm 0{,}04$ contre $0{,}93 \pm 0{,}07$ pour B', figure X.39c. Enfin, la résolution relative de B' est de



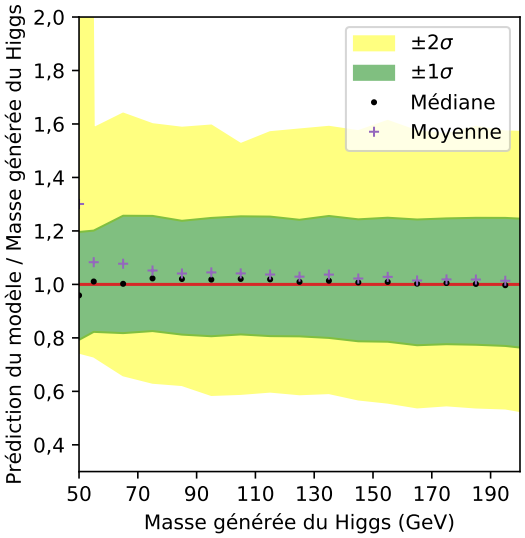
(a) *Réponse du modèle B.*



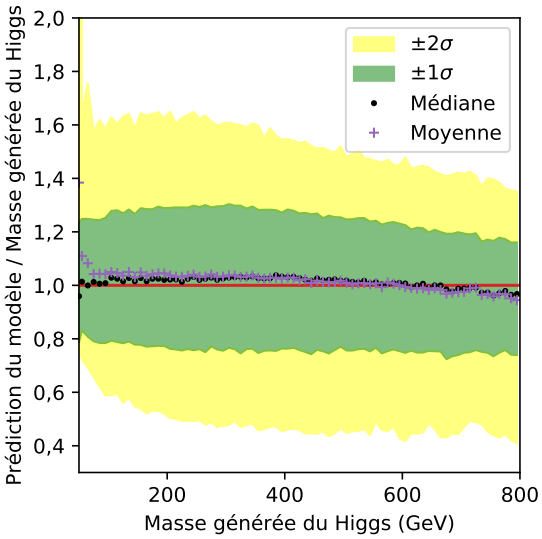
(b) *Réponse du modèle B à basse masse.*



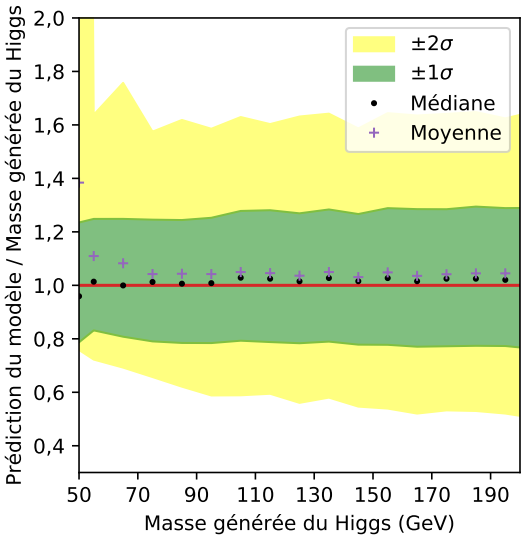
(c) *Réponse du modèle B′.*



(d) *Réponse du modèle B′ à basse masse.*



(e) *Réponse du modèle B″.*



(f) *Réponse du modèle B″ à basse masse.*

Figure X.39 – *Comparaison des modèles B, B′ et B″.*

22 %, celle de B" de 25 %. Malgré une légère dégradation de la résolution, l'utilisation de la fonction de coût modifiée $L_{MA\sqrt{PE}\times b}$ et des événements entre 800 GeV et 1 TeV obtenus grâce à la largeur de $\mathcal{H}$ permet de ramener la réponse moyenne du modèle à des valeurs de $1{,}00 \pm 0{,}05$ pour des valeurs de $m_{\mathcal{H}}$ allant de 80 à 800 GeV.

7.7 Modèle final

Le modèle B" que nous avons construit est donc entraîné sur des événements $\mathcal{H} \to \tau\tau$ où $\mathcal{H}$ est le boson de Higgs du modèle standard h avec une masse modifiée entre 50 et 800 GeV, avec addition d'empilement selon le profil de l'année 2017, dont la sélection est réalisée selon la procédure décrite section 2. En particulier, l'algorithme DEEPTAU [11] est utilisé pour l'identification des taus hadroniques. La largeur de $\mathcal{H}$ permet d'exploiter des événements où sa masse effective est supérieure à 800 GeV, jusqu'à 1 TeV. La gamme d'utilisation de notre modèle est toutefois considérée comme allant de 50 à 800 GeV uniquement. Les hyper-paramètres de B" sont ceux de B à l'exception de la fonction de coût :

— 3 couches cachées ;
— 1000 neurones par couche cachée ;
— fonction d'activation Softplus, $x \mapsto \ln(1 + e^x)$;
— algorithme d'optimisation Adam, présenté en section 5.3.2 ;
— fonction de coût $L_{MA\sqrt{PE}\times b}$, définie équation (X.41) ;
— initialisation des poids selon le mode « Glorot Uniforme » [47] ;
— 27 variables d'entrée données en section 2.4.

L'utilisation de B" dans les analyses de CMS est abordé dans la section suivante.

8 Utilisation du modèle dans les analyses CMS

Dans cette section, l'utilisation du modèle B" introduit dans la section 7 sur les événements de l'analyse présentée dans le chapitre 4 est discutée. Ce modèle, issu du ML, est conçu pour prédire la masse de la particule se désintégrant en paire de leptons τ. Ces prédictions sont notées m_{ML}.

8.1 Utilisation de m_{ML} comme variable discriminante

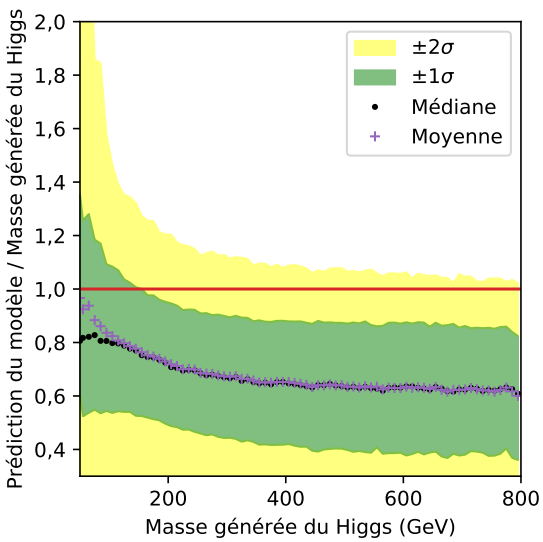
8.1.1 Performances de m_{ML} et de $m_{\text{T}}^{\text{tot}}$ en tant qu'estimateurs de la masse

Dans l'analyse présentée au chapitre 4, la variable discriminante utilisée est $m_{\text{T}}^{\text{tot}}$. Il s'agit d'une approximation de la masse invariante du système dans l'état final, restreinte au plan transverse car l'ensemble des neutrinos est remplacé par $E_{\text{T}}{}^{\text{miss}}$. Nous avons souhaité étudier l'utilisation de m_{ML} comme variable discriminante.

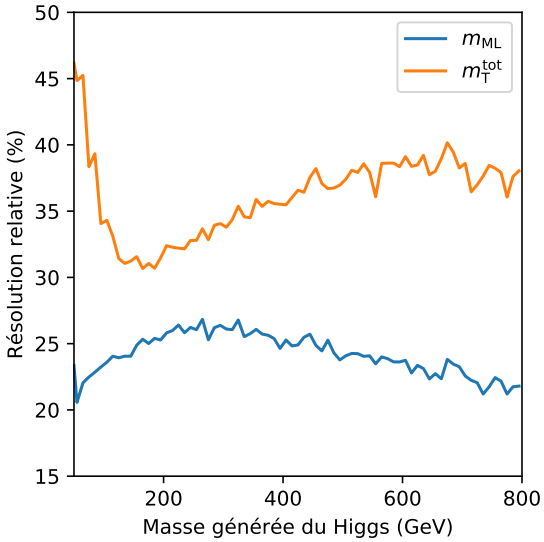
La réponse de $m_{\text{T}}^{\text{tot}}$ est représentée en fonction de $m_{\mathcal{H}}$ sur la figure X.40a. Sa médiane est comprise entre 0,83 et 0,60 sur toute la gamme de masse, les valeurs de $m_{\text{T}}^{\text{tot}}$ ne sont donc absolument pas concordantes avec la masse $m_{\mathcal{H}}$ du boson initial. La réponse médiane du modèle B" en revanche, figure X.39e, est de $1{,}00 \pm 0{,}05$. Les valeurs de m_{ML} sont donc plus fidèles à $m_{\mathcal{H}}$ que celles de $m_{\text{T}}^{\text{tot}}$.

Cependant, pour une variable discriminante, l'enjeu est d'avantage de séparer signal et bruit de fond que de présenter une réponse de 1. Pour cela, la résolution relative est importante. Une mauvaise résolution étale les distributions des différents signaux qui se superposent, rendant difficile leur séparation.

Les résolutions relatives de $m_{\text{T}}^{\text{tot}}$ et m_{ML} sont représentées figure X.40b. Dans le cas de $m_{\text{T}}^{\text{tot}}$, la résolution relative est de 45 % à 50 GeV. Elle diminue à 31 % à 150 GeV, puis augmente jusqu'à 500 GeV où elle est d'environ 37 %. Pour m_{ML} en revanche, elle est de 21 % à basse masse, augmente progressivement jusqu'à 26 % à 270 GeV puis redescend à 22 % à haute masse. La résolution relative de m_{ML} est systématiquement meilleure que celle de $m_{\text{T}}^{\text{tot}}$. L'utilisation de m_{ML} au lieu de $m_{\text{T}}^{\text{tot}}$ en tant que variable discriminante pour l'analyse du chapitre 4 est donc particulièrement intéressante.



(a) *Réponse de* $m_{\mathrm{T}}^{\mathrm{tot}}$.



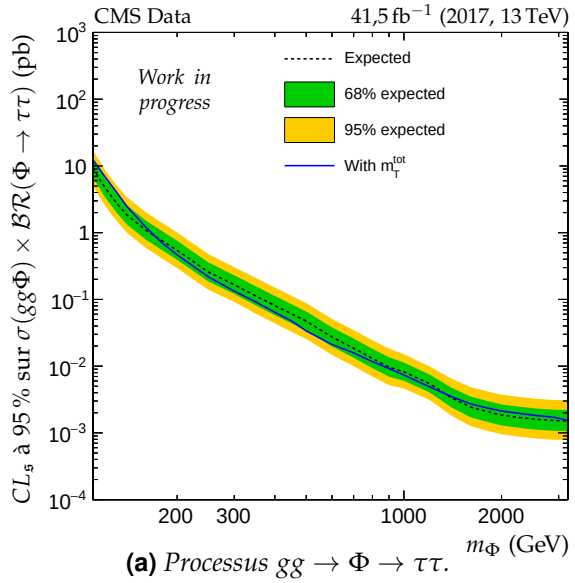
(b) *Résolutions relatives de* $m_{\mathrm{T}}^{\mathrm{tot}}$ *et* m_{ML}.

Figure X.40 – *Réponse de* $m_{\mathrm{T}}^{\mathrm{tot}}$ *et comparaison de sa résolution relative à celle de* m_{ML}.

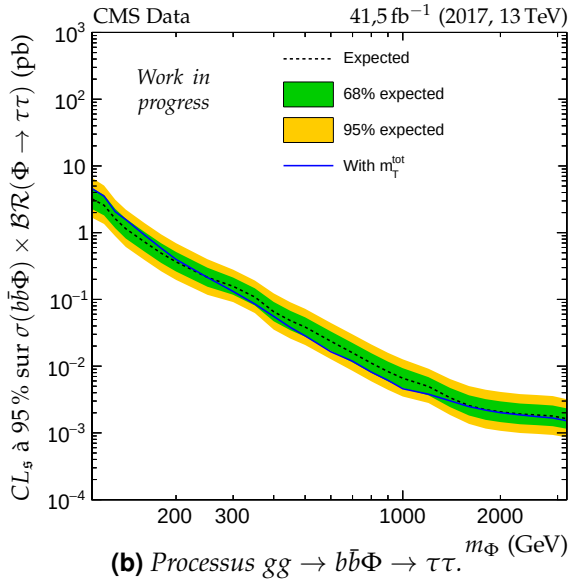
8.1.2 Limites d'exclusion avec m_{ML}

L'analyse indépendante du modèle introduite au chapitre 4 a été réalisée sur l'année 2017 avec $m_{\mathrm{T}}^{\mathrm{tot}}$ et m_{ML} comme variables discriminantes. Les incertitudes systématiques de normalisation et de forme, introduites au chapitre 4, ont été recalculées avec m_{ML} afin de réaliser le traitement statistique des données. Les résultats sont donnés en figure X.41 pour les processus $gg \to \Phi \to \tau\tau$ et $gg \to b\bar{b}\Phi \to \tau\tau$. Il s'agit des limites « attendues », obtenues en considérant que l'estimation du bruit de fond correspond exactement aux observations.

Les limites d'exclusion avec m_{ML} sont équivalentes à celles données par $m_{\mathrm{T}}^{\mathrm{tot}}$ dans la limite de l'incertitude à $\pm 1\sigma$ sur toute la gamme de masse étudiée. À basse masse, $m_\Phi < 180\,\mathrm{GeV}$ pour $gg\Phi$ en figure X.41a et $m_\Phi < 250\,\mathrm{GeV}$ pour $b\bar{b}\Phi$ en figure X.41b, m_{ML} est un peu plus performante que $m_{\mathrm{T}}^{\mathrm{tot}}$. Pour des masses plus élevées et jusqu'à 1 TeV, c'est en revanche $m_{\mathrm{T}}^{\mathrm{tot}}$ qui donne des résultats légèrement meilleurs. Au-delà, les deux variables ont les mêmes performances. Un bon estimateur de la masse d'une résonance n'est donc pas forcément une bonne variable discriminante.



(a) *Processus* $gg \to \Phi \to \tau\tau$.

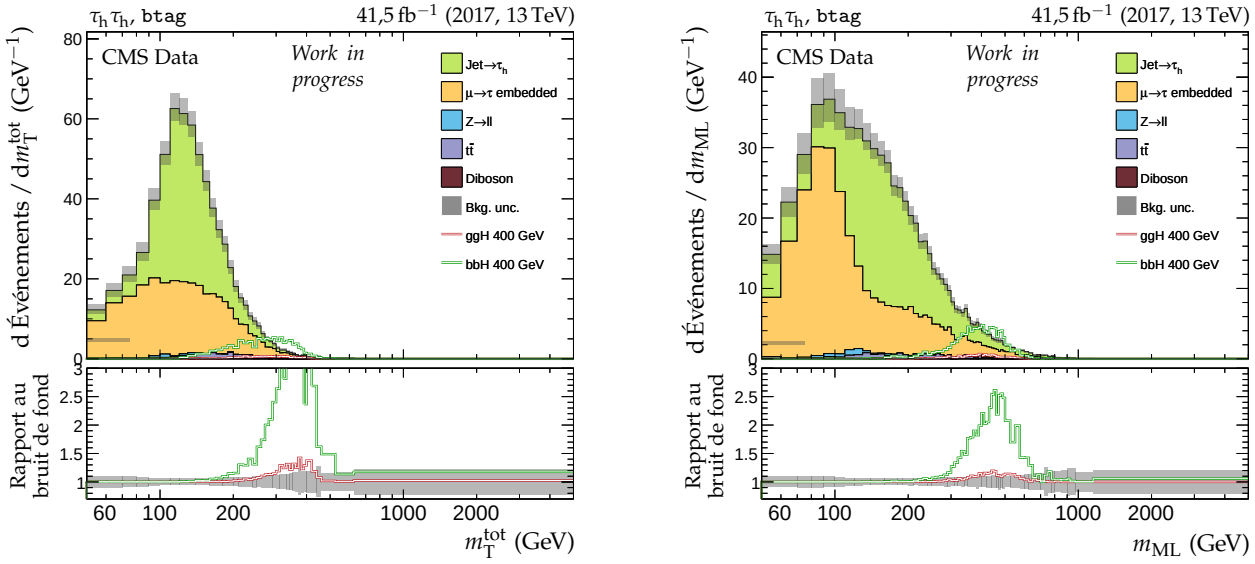


(b) *Processus* $gg \to b\bar{b}\Phi \to \tau\tau$.

Figure X.41 – *Limites d'exclusion attendues avec* m_{ML} *en tant que variable discriminante* (Expected) *obtenues avec l'année 2017. Les limites obtenues avec* $m_{\mathrm{T}}^{\mathrm{tot}}$ *sont également données* (With $m_{\mathrm{T}}^{\mathrm{tot}}$).

8.1.3 Distributions de m_{ML} et de $m_{\mathrm{T}}^{\mathrm{tot}}$

Afin d'interpréter ces résultats, il est possible de se référer aux distributions de $m_{\mathrm{T}}^{\mathrm{tot}}$ et m_{ML}. La figure X.42 les présente pour la catégorie `btag` du canal $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$. D'autres distributions sont disponibles dans l'annexe H.



(a) *Distribution de* $m_{\mathrm{T}}^{\mathrm{tot}}$.

(b) *Distribution de* m_{ML}.

Figure X.42 – *Distributions de* $m_{\mathrm{T}}^{\mathrm{tot}}$ *et* m_{ML} *en 2017 dans la catégorie* `btag` *du canal* $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$. *Un signal correspondant à un boson de Higgs supplémentaire de masse* 400 GeV *est ajouté à titre d'illustration.*

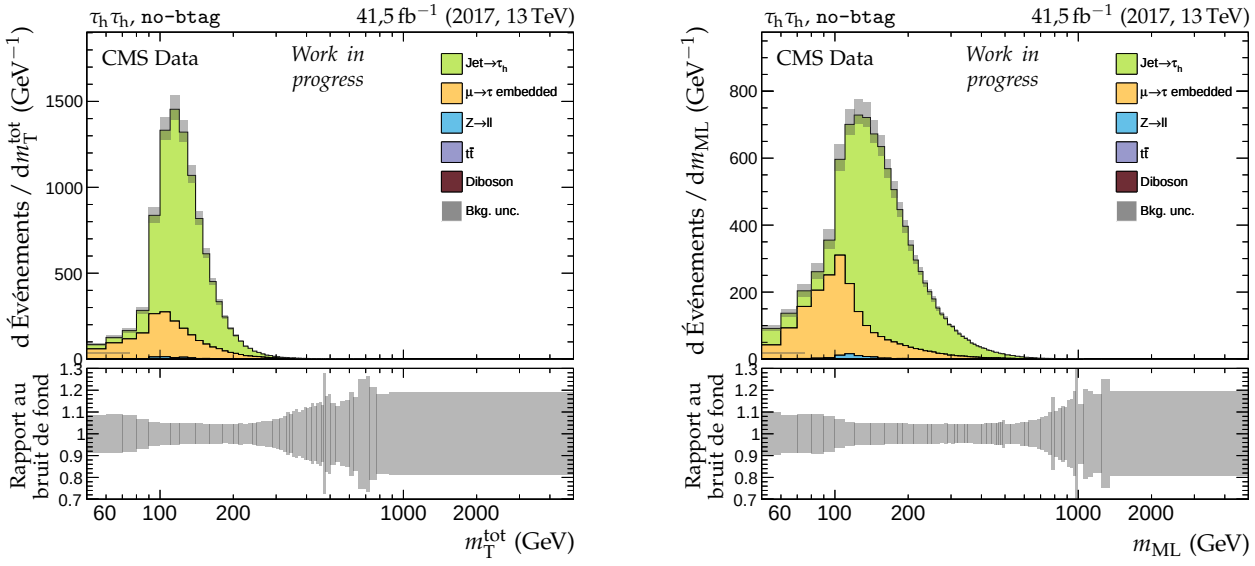
À haute masse, au-delà de 1 TeV, l'effet de la statistique est dominant à cause de la faible quantité d'événements présents. Les deux variables donnent alors des résultats équivalents. Pour les masses intermédiaires en revanche, l'avantage de $m_{\mathrm{T}}^{\mathrm{tot}}$ sur m_{ML} provient de la présence des faux τ_{h}, c'est-à-dire des événements dans lesquels au moins un jet est identifié à tort comme un τ_{h} et utilisé pour former le dilepton. De tels jets ne proviennent pas de la désintégration d'un boson de Higgs en paire de leptons τ. S'il est possible de définir une masse pour un dilepton formé avec un ou deux faux τ_{h}, la valeur obtenue n'a pas de sens physique, ces jets n'étant pas issus de la même désintégration.

Comme cela est visible sur la figure X.42, le bruit de fond dû aux faux τ_{h} (Jet $\rightarrow \tau_{\mathrm{h}}$) forme une queue de distribution à haute masse plus importante avec m_{ML} qu'avec $m_{\mathrm{T}}^{\mathrm{tot}}$. Or, cette queue se retrouve dans la région de signal. Par exemple, selon $m_{\mathrm{T}}^{\mathrm{tot}}$, le signal d'un boson de Higgs supplémentaire de masse 400 GeV se trouve entre 120 et 500 GeV avec un maximum de 5 GeV^{-1} à 300 GeV. À cet endroit, le nombre d'événements de bruit de fond pour $m_{\mathrm{T}}^{\mathrm{tot}}$ est d'environ 3 GeV^{-1}. Selon m_{ML} en revanche, le même signal se trouve entre 200 et 650 GeV avec un maximum de 6 GeV^{-1} à 400 GeV. À cet endroit, le nombre d'événements de bruit de fond pour m_{ML} est d'environ 5 GeV^{-1}.

Bien que les valeurs données par m_{ML} soient plus proches de la valeur vraie de la masse du boson et que le signal présente une valeur maximale plus élevée qu'avec $m_{\mathrm{T}}^{\mathrm{tot}}$, la queue à haute masse due aux faux τ_{h} dégrade le rapport du signal au bruit de fond. Ce dernier est en effet plus faible avec m_{ML} qu'avec $m_{\mathrm{T}}^{\mathrm{tot}}$.

Un effet intéressant est par contre la meilleure séparation par m_{ML} que par $m_{\mathrm{T}}^{\mathrm{tot}}$ des événements des données encapsulées ($\mu \rightarrow \tau$ embedded), introduites dans le chapitre 4, vis-à-vis des faux τ_{h}. Les données encapsulées décrivent principalement le boson Z et une fraction d'événements $t\bar{t}$. Dans le cas de $m_{\mathrm{T}}^{\mathrm{tot}}$, leur distribution est largement étalée jusqu'à 300 GeV, avec un maximum d'événements de 20 GeV^{-1} à 95 GeV. La distribution des valeurs de m_{ML} sur ces mêmes événements exhibe deux composantes. La première forme un pic avec un maximum d'événements de 30 GeV^{-1} à 90 GeV, il s'agit du signal du boson Z. La seconde se situe à plus haute masse, jusqu'à 600 GeV, avec un maximum d'événements de 7 GeV^{-1} à environ 200 GeV. Ce groupe peut correspondre à la composante $t\bar{t}$ des données encapsulées. Pour le vérifier, la figure X.43 montre les distributions de $m_{\mathrm{T}}^{\mathrm{tot}}$ et m_{ML} pour

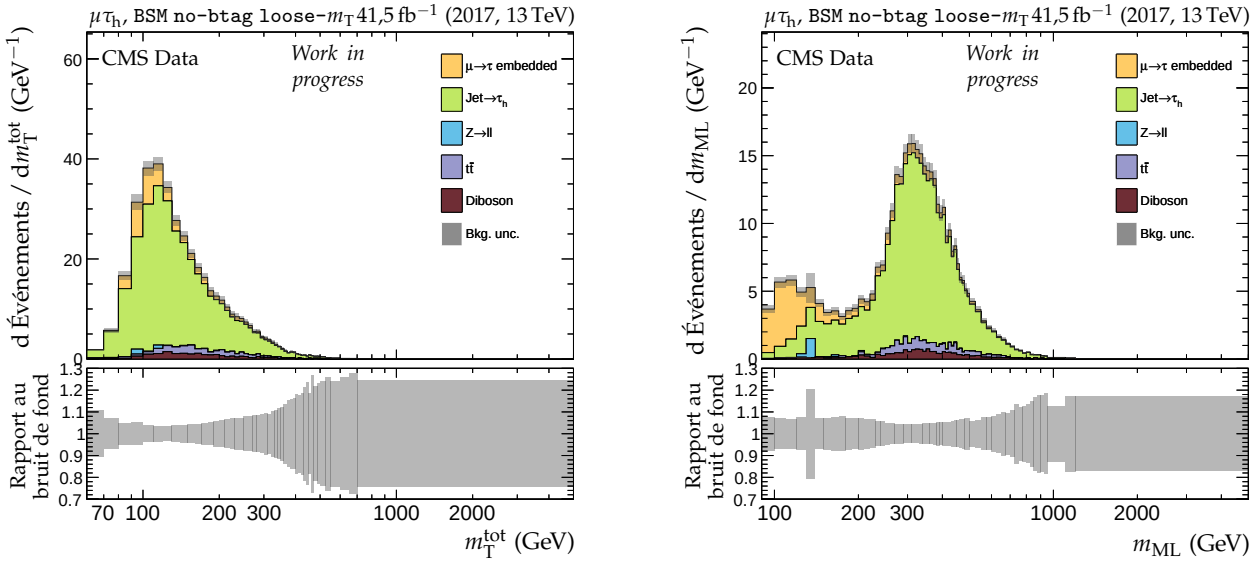
la catégorie `no-btag` du canal $\tau_h\tau_h$, où l'exclusion des événements $t\bar{t}$ est réalisée par le rejet des jets issus de quarks b. La distribution de m_{ML} dans les données encapsulées ne montre pas deux parties comme dans la catégories `btag`, ce qui confirme cette hypothèse. Notre modèle est donc en mesure de séparer les composantes du boson Z et $t\bar{t}$ des données encapsulées du canal $\tau_h\tau_h$, bien qu'il n'ait pas été entraîné dans ce but, alors que m_T^{tot} ne le permet pas.



(a) *Distribution de* m_T^{tot}. **(b)** *Distribution de* m_{ML}.

Figure X.43 – *Distributions de* m_T^{tot} *et* m_{ML} *en 2017 dans la catégorie* `no-btag` *du canal* $\tau_h\tau_h$.

Notre modèle permet de plus de mieux séparer les événements du boson Z des faux τ_h. Par exemple, les distributions de m_T^{tot} et m_{ML} sont données dans la catégorie `BSM no-btag loose-`m_T pour l'année 2017 en figure X.44. La composante du boson Z apparaît nettement avec m_{ML} aux alentours de 100 GeV, alors qu'elle est confondue avec les faux τ_h d'après m_T^{tot}. Or, la sélection des événements pour les catégories `BSM`, introduite au chapitre 4, contient en particulier la coupure $m_{SVFIT} \geqslant 250$ GeV, avec m_{SVFIT} la masse prédite par l'algorithme SVFIT [17]. Notre modèle identifie donc des événements correspondant au boson Z, de masse 91,2 GeV, alors que SVFIT les estime au-delà de 250 GeV.



(a) *Distribution de* m_T^{tot}. **(b)** *Distribution de* m_{ML}.

Figure X.44 – *Distributions de* m_T^{tot} *et* m_{ML} *en 2017 dans la catégorie* `BSM no-btag loose-`m_T *du canal* $\mu\tau_h$.

8.2 Comparaison de m_{ML} à m_{SVFIT}

Au sein de la collaboration CMS, l'algorithme SVFIT [17] a le même objectif que notre modèle B", estimer la masse d'une particule se désintégrant en paire de leptons τ, bien que la méthode employée soit différente. Cependant, nous avons observé précédemment que certains événements estimés au-delà de 250 GeV par SVFIT semblent correspondre au boson Z de masse 91,2 GeV selon notre modèle. Nous avons donc souhaité comparer les valeurs de m_{ML} prédites par B" à celles de m_{SVFIT} fournies par SVFIT.

8.2.1 Distributions inclusives

La figure X.45 montre les distributions inclusives, c'est-à-dire sur les événements de toutes les catégories confondues, de m_{SVFIT} et m_{ML} en 2017 dans les canaux $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$, $e\tau_{\mathrm{h}}$ et $e\mu$. Ces distributions sont deux à deux très similaires, ce qui est attendu car m_{SVFIT} comme m_{ML} sont deux estimateurs de la même grandeur physique. Les faux τ_{h} sont toutefois prédits à des masses légèrement plus élevées par B" que par SVFIT. Il en résulte des queues de distribution à haute masse plus importantes avec m_{ML}, effet déjà observé lors de la comparaison de m_{ML} à $m_{\mathrm{T}}^{\mathrm{tot}}$. Ces comparaisons sur l'ensemble des topologies d'événements d'un canal ne permet pas de comparer B" et SVFIT plus en détail. C'est pourquoi des comparaisons sur des sous-ensembles d'événements sont réalisées dans les sections suivantes.

8.2.2 Distributions des événements de haute masse

D'après la figure X.44, notre modèle identifie des événements correspondant au boson Z parmi ceux que SVFIT estime au-delà de 250 GeV. Il y a donc un désaccord entre les deux estimateurs. Toutefois, SVFIT pourrait aussi trouver un signal à basse masse parmi les événements estimés à haute masse par B". Pour le vérifier, les distributions de m_{ML} sur les événements tels que $m_{\mathrm{SVFIT}} > 250$ GeV sont représentées en figure X.46 pour les canaux $\mu\tau_{\mathrm{h}}$, $e\tau_{\mathrm{h}}$ et $e\mu$ en 2017, en regard des distributions de m_{SVFIT} sur les événements tels que $m_{\mathrm{ML}} > 250$ GeV.

Les distributions de m_{ML} confirment que B" parvient à trouver des événements correspondant au boson Z alors que SVFIT les estime au-delà de 250 GeV. Les distributions de m_{SVFIT} en revanche ne font apparaître aucun signal en particulier parmi les événements estimés au-delà de 250 GeV par B". Des investigations plus poussées peuvent être menées en sélectionnant des topologies particulières d'événements.

8.2.3 Distributions par topologies

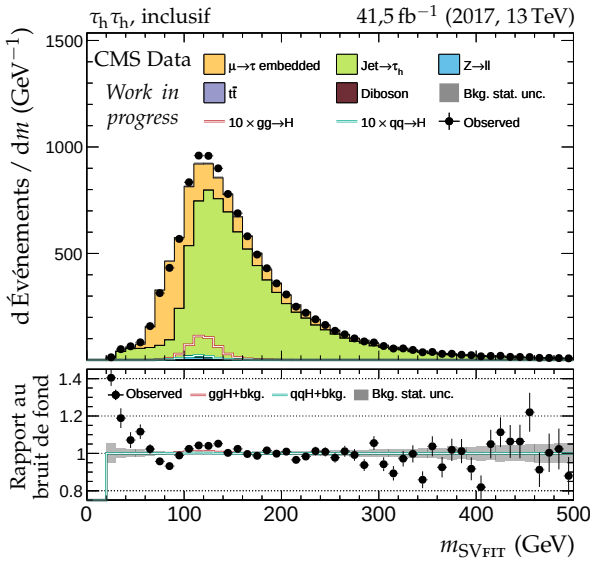
Les distributions de m_{SVFIT} et m_{ML} sont comparées dans le cas de topologies spécifiques d'événements. Ces topologies correspondent à certaines catégories de l'analyse du boson de Higgs du modèle standard dans sa désintégration en paire de leptons τ n'utilisant pas de réseau de neurones pour obtenir une variable discriminante [54, 55].

La figure X.47 correspond aux événements du canal $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ sans jets et tels que la distance ΔR entre les deux éléments du dilepton dans le plan (η, ϕ) soit inférieure à 3,2. Avec m_{SVFIT}, le signal du boson Z est situé au-dessus de 100 GeV et est complètement mélangé avec les faux τ_{h}. En revanche, le passage à m_{ML} a pour effet de décaler les faux τ_{h} d'une dizaine de GeV vers les hautes valeurs alors que le signal du Z est en partie estimé à plus basse masse.

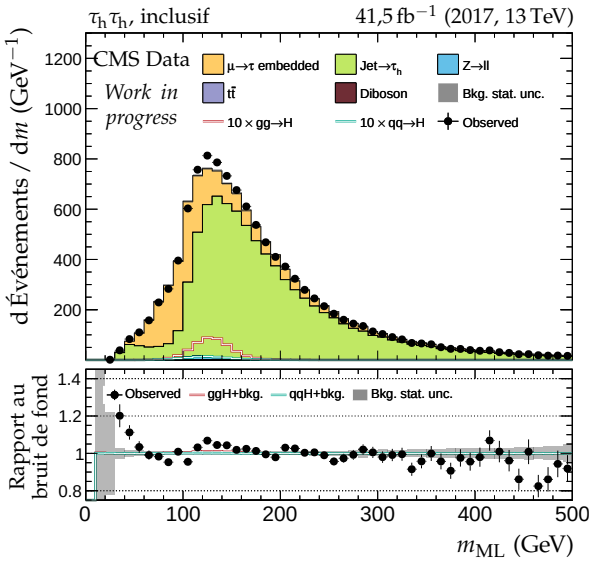
Cet effet est encore plus visible dans le cas des événements vérifiant l'une des deux conditions suivantes :

— moins de deux jets et tels que $\Delta R \geqslant 3{,}2$;
— au moins deux jets et tels que $\Delta R \geqslant 2{,}5$;

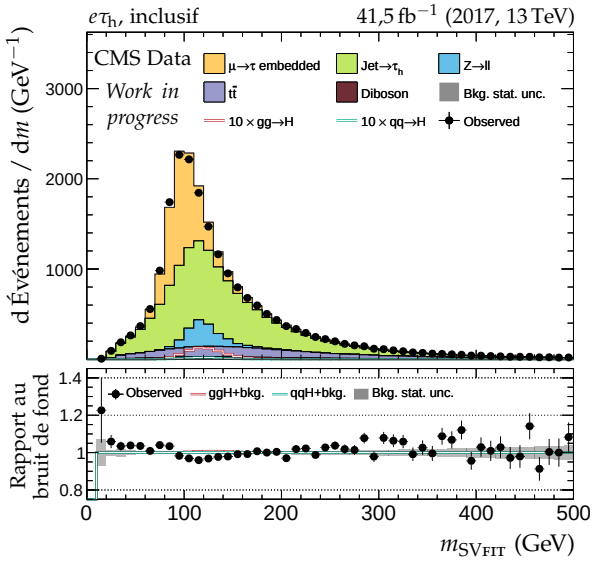
dont les distributions de m_{SVFIT} et m_{ML} sont données en figure X.48. Dans ce cas, la distribution de m_{SVFIT} sur les données encapsulées est complètement confondue avec celle des faux τ_{h}, tandis que les valeurs de m_{ML} font apparaître un signal aux alentours de 100 GeV correspondant au boson Z, sa masse étant donc surestimée de 10 GeV environ.



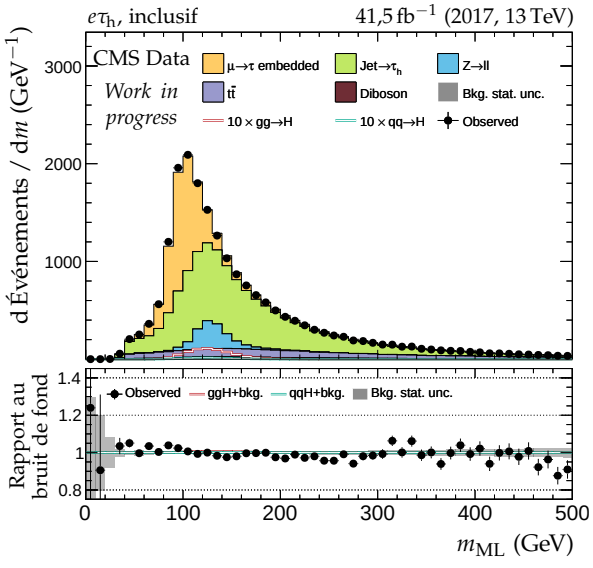
(a) m_{SVFIT}, *canal* $\tau_h\tau_h$.



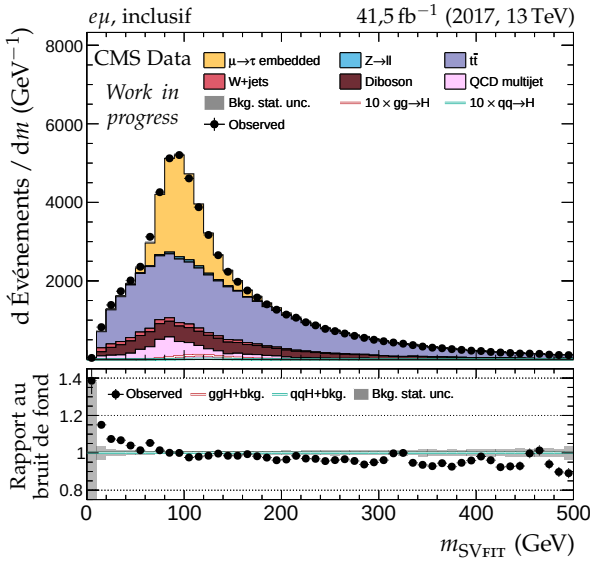
(b) m_{ML}, *canal* $\tau_h\tau_h$.



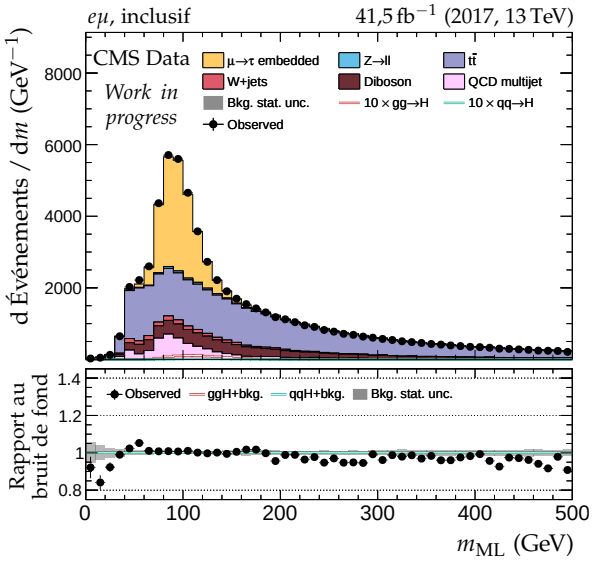
(c) m_{SVFIT}, *canal* $e\tau_h$.



(d) m_{ML}, *canal* $e\tau_h$.

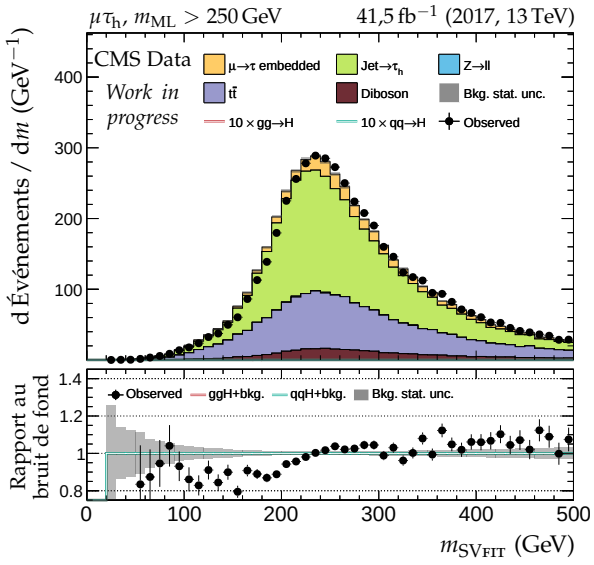


(e) m_{SVFIT}, *canal* $e\mu$.

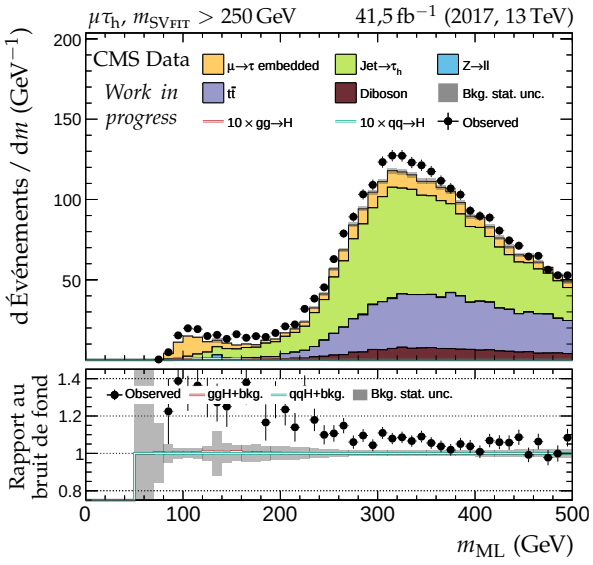


(f) m_{ML}, *canal* $e\mu$.

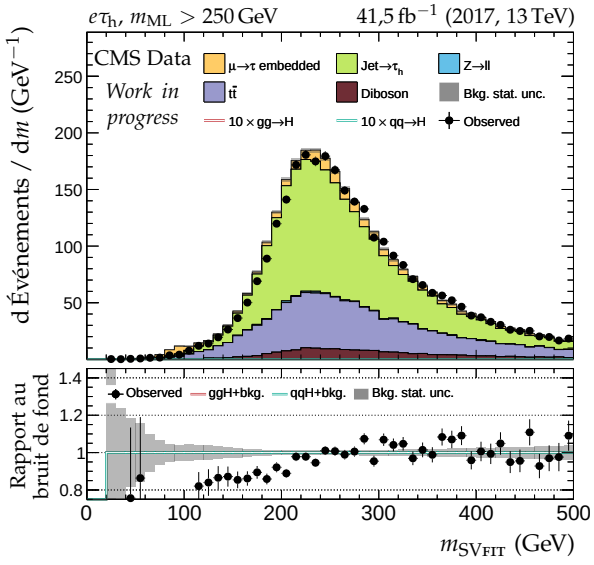
Figure X.45 – *Distributions de* m_{SVFIT} *et* m_{ML} *en 2017 dans les canaux* $\tau_h\tau_h$, $e\tau_h$ *et* $e\mu$.



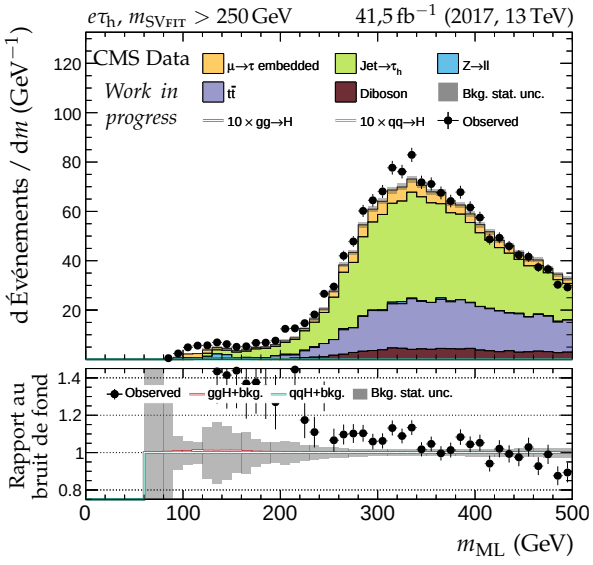
(a) m_{SVFIT} *pour* $m_{\text{ML}} > 250\,\text{GeV}$, *canal* $\mu\tau_{\text{h}}$.



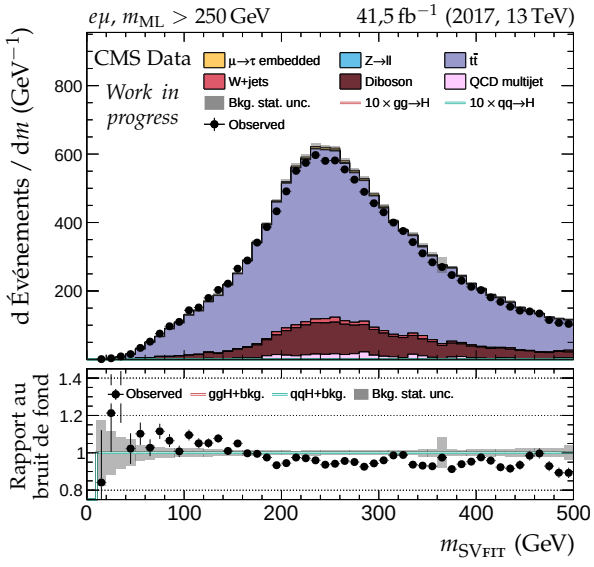
(b) m_{ML} *pour* $m_{\text{SVFIT}} > 250\,\text{GeV}$, *canal* $\mu\tau_{\text{h}}$.



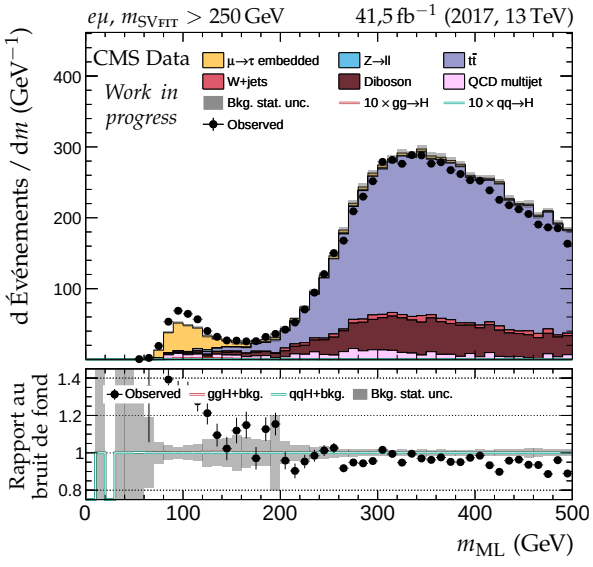
(c) m_{SVFIT} *pour* $m_{\text{ML}} > 250\,\text{GeV}$, *canal* $e\tau_{\text{h}}$.



(d) m_{ML} *pour* $m_{\text{SVFIT}} > 250\,\text{GeV}$, *canal* $e\tau_{\text{h}}$.

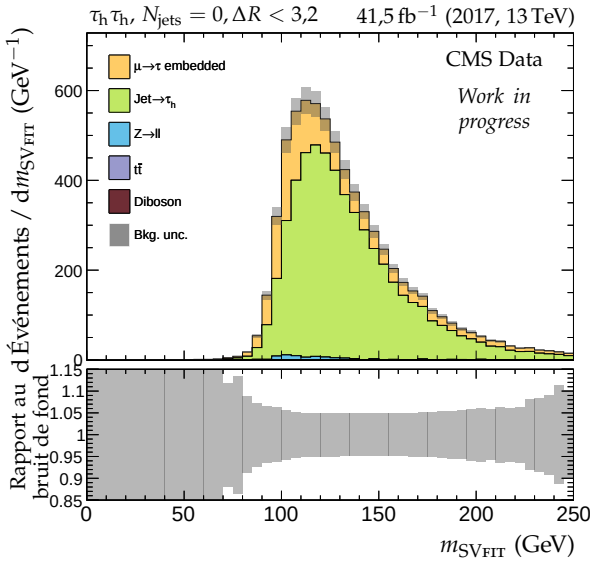


(e) m_{SVFIT} *pour* $m_{\text{ML}} > 250\,\text{GeV}$, *canal* $e\mu$.

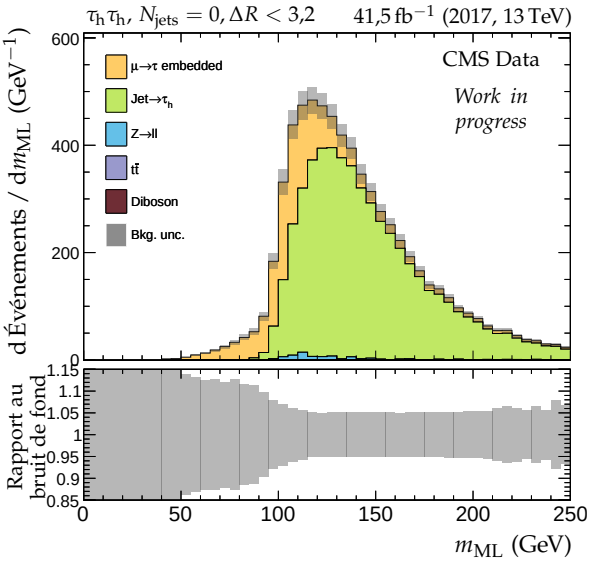


(f) m_{ML} *pour* $m_{\text{SVFIT}} > 250\,\text{GeV}$, *canal* $e\mu$.

Figure X.46 – *Distributions de* m_{SVFIT} *et* m_{ML} *en 2017 sur les événements de haute masse des canaux* $\mu\tau_{\text{h}}$, $e\tau_{\text{h}}$ *et* $e\mu$.

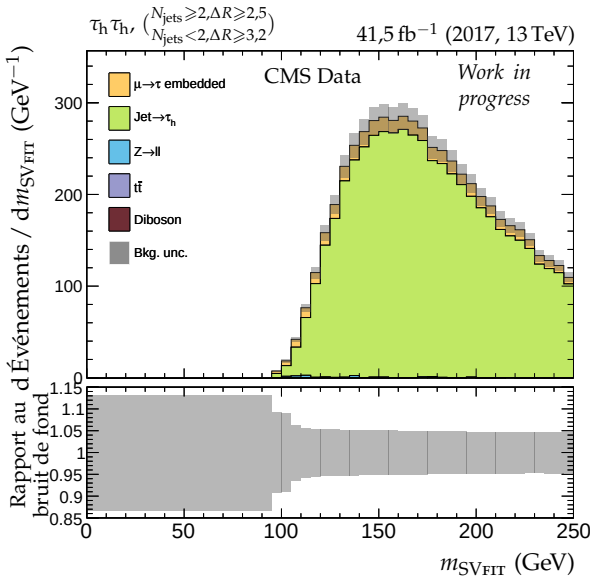


(a) m_{SVFIT}, *canal* $\tau_h\tau_h$.

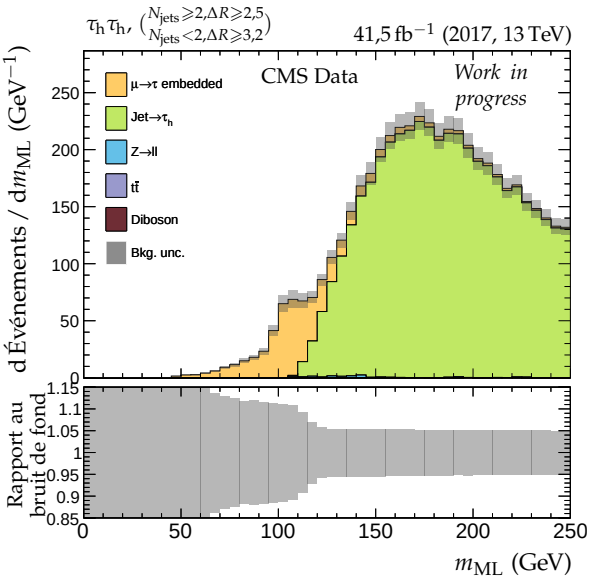


(b) m_{ML}, *canal* $\tau_h\tau_h$.

Figure X.47 – *Distributions de* m_{SVFIT} *et* m_{ML} *en 2017 dans le canal* $\tau_h\tau_h$ *avec les événements sans jets et tels que la distance* ΔR *entre les deux éléments du dilepton dans le plan* (η, ϕ) *soit inférieure à* 3,2.



(a) m_{SVFIT}, *canal* $\tau_h\tau_h$.



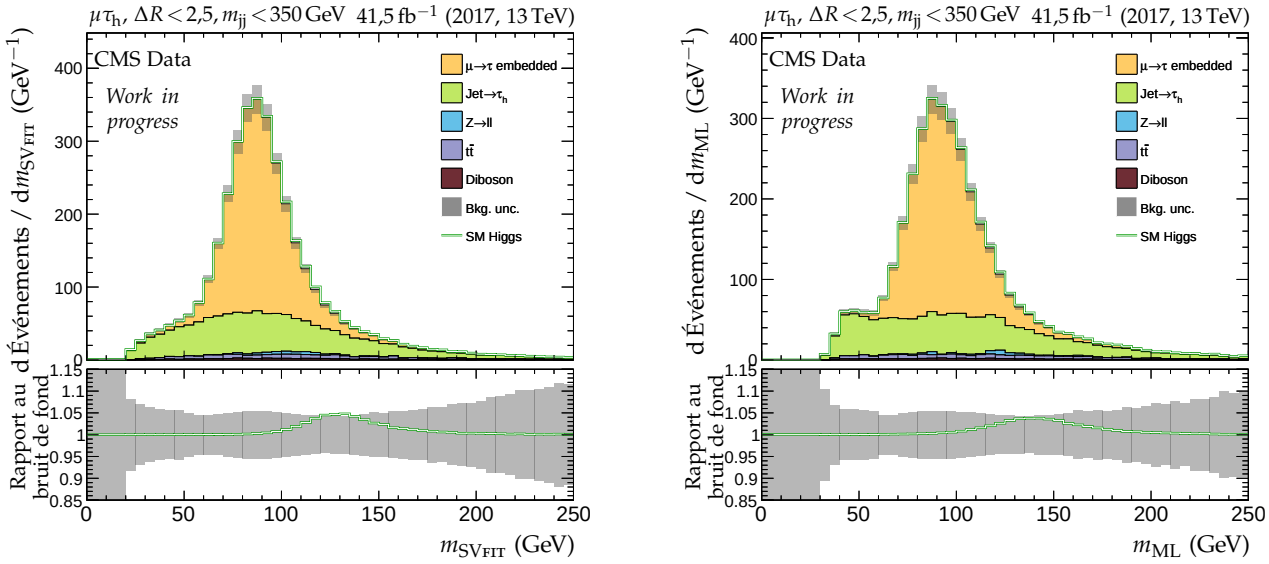
(b) m_{ML}, *canal* $\tau_h\tau_h$.

Figure X.48 – *Distributions de* m_{SVFIT} *et* m_{ML} *en 2017 dans le canal* $\tau_h\tau_h$ *avec les événements contenant moins de deux jets et tels que la distance* ΔR *entre les deux éléments du dilepton dans le plan* (η, ϕ) *soit supérieure à* 3,2 *ou contenant au moins deux jets et* $\Delta R \geqslant 2{,}5$.

La sensibilité de B" au boson de Higgs du modèle standard h est comparable avec celle de SVFIT. Par exemple, avec les événements du canal $\mu\tau_h$ contenant au moins deux jets avec la masse invariante des deux jets principaux m_{jj} inférieure à 350 GeV et tels que la distance ΔR entre les deux éléments du dilepton dans le plan (η, ϕ) soit inférieure à 2,5, les distributions de m_{SVFIT} et m_{ML} en figure X.49 montrent un rapport au bruit de fond similaire pour le signal associé à h.

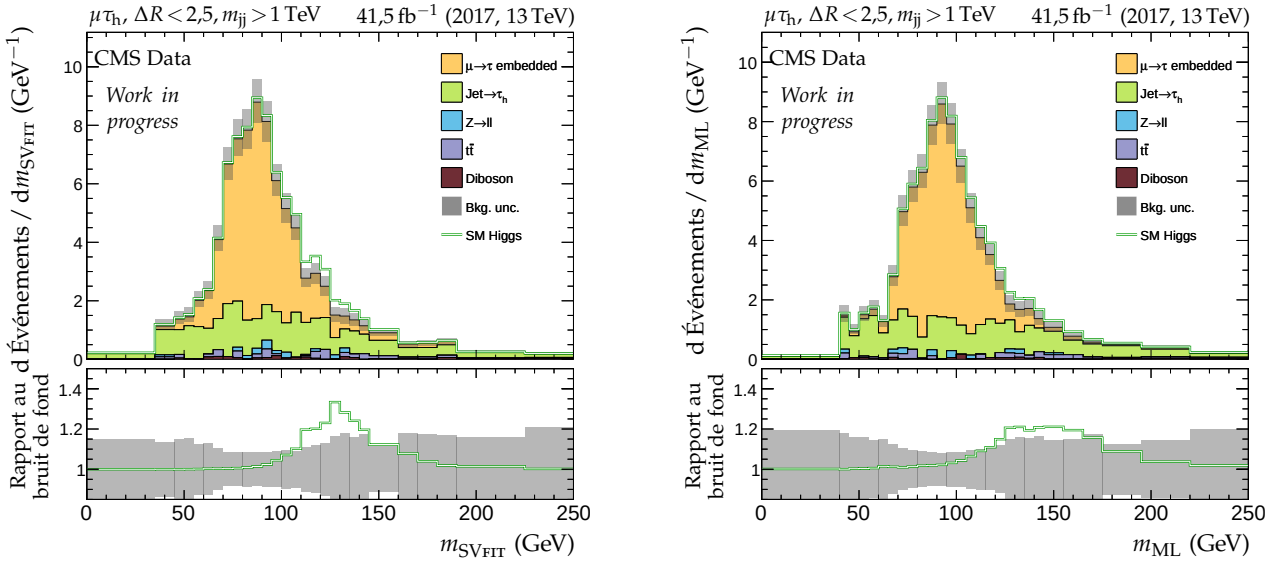
Lorsque m_{jj} est supérieure à 1 TeV en revanche, en figure X.50, la distribution du signal de h est plus étalée avec m_{ML} qu'avec m_{SVFIT}. Il en résulte une sensibilité diminuée.

Notre modèle propose donc des estimations de la masse des particules se désintégrant en paires de leptons τ comparables à SVFIT, actuellement utilisé au sein de la collaboration CMS. Selon les topologies des événements considérés, la description du boson de Higgs h est similaire. Celle du boson Z est parfois meilleure avec notre modèle.



(a) m_{SVFIT}, *canal* $\mu\tau_h$. **(b)** m_{ML}, *canal* $\mu\tau_h$.

Figure X.49 – *Distributions de* m_{SVFIT} *et* m_{ML} *en 2017 dans le canal* $\mu\tau_h$ *avec les événements contenant au moins deux jets avec la masse invariante des deux jets principaux* m_{jj} *inférieure à* 350 GeV *et tels que la distance* ΔR *entre les deux éléments du dilepton dans le plan* (η, ϕ) *soit inférieure à* 2,5.



(a) m_{SVFIT}, *canal* $\mu\tau_h$. **(b)** m_{ML}, *canal* $\mu\tau_h$.

Figure X.50 – *Distributions de* m_{SVFIT} *et* m_{ML} *en 2017 dans le canal* $\mu\tau_h$ *avec les événements contenant au moins deux jets avec la masse invariante des deux jets principaux* m_{jj} *supérieure à* 1 TeV *et tels que la distance* ΔR *entre les deux éléments du dilepton dans le plan* (η, ϕ) *soit inférieure à* 2,5.

8.2.4 Vitesses d'exécution

L'algorithme de SVFIT est connu pour être relativement long en termes de temps d'exécution. Cela provient de sa démarche qui consiste à réaliser un ajustement de la valeur de m_{SVFIT} aux observables de chaque événement [17]. Dans le cas d'un DNN, l'ajustement en fonction des variables d'entrées est réalisé une fois pour toutes lors de l'entraînement du modèle. Le DNN B" a été entraîné en 15 minutes environ. Puis, l'obtention des prédictions m_{ML} de B" consiste à appliquer la fonction F de B" aux variables d'entrée. La figure X.51 compare les temps d'exécution moyens par événement de deux scripts, l'un en C++ pour obtenir les valeurs de m_{SVFIT}, l'autre en Python pour obtenir celles de m_{ML}.

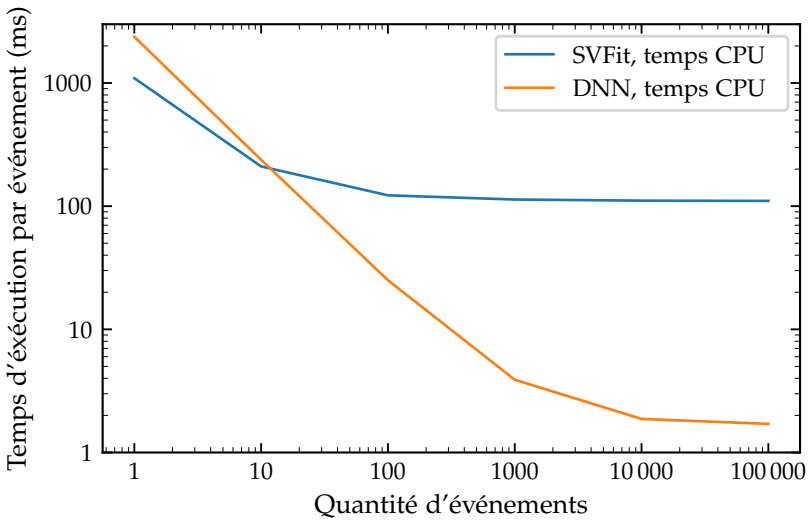


Figure X.51 – *Temps nécessaire à l'obtention des prédictions de masse avec* SVFIT *et avec le DNN B".*

Pour 10 événements ou moins à traiter, le script de SVFIT est plus rapide que celui du DNN. Ceci est dû au langage utilisé ainsi qu'au temps de chargement du DNN. Cependant, une quantité si faible d'événements à traiter est rarement rencontrée dans les analyses du Run II à CMS dans lesquelles les événements se comptent au moins en millions. Sur la figure X.51, le temps moyen par événement pour chacun des deux scripts stagne à partir de 100 événements avec SVFIT et 10 000 avec le DNN. Le script de SVFIT nécessite alors 110 ms par événement contre seulement 1,71 ms avec le DNN.

Les prédictions de notre modèle sont donc obtenues 60 fois plus rapidement que celles de SVFIT. Dans le contexte des futurs *runs* du LHC et surtout du HL-LHC, l'augmentation de la luminosité et donc du nombre d'événements à traiter rend l'utilisation d'un DNN à la place de SVFIT pertinente dans l'optique de minimiser les temps d'exécution des analyses.

9 Conclusion

Le *machine learning* (ML) est une branche de l'intelligence artificielle permettant d'obtenir des modèles pouvant réaliser des classifications ou des régressions. De tels modèles sont déjà exploités en physique des particules afin de réaliser diverses tâches, comme l'identification des jets issus de quarks b par exemple.

Nous avons étudié la possibilité de prédire la masse d'une résonance se désintégrant en paire de leptons τ grâce au ML. En effet, la phénoménologie de ces événements ne permet pas d'obtenir la masse invariante totale du système dans l'état final. Les travaux réalisés par BÄRTSCHI & coll. sur ce sujet donnent des résultats prometteurs, mais l'empilement n'y est pas pris en compte et la modélisation du détecteur CMS approximée. À partir d'événements que nous avons nous-mêmes simulés à l'aide de FASTSIM, l'empilement a été pris en compte et le détecteur modélisé plus précisément.

Nous avons construit et entraîné des arbres de décision améliorés à l'aide de la librairie XGBOOST et des réseaux de neurones profonds à l'aide des librairies KERAS et TENSORFLOW. Le principe et l'entraînement de ces types de modèle ont été présentés. De nombreuses combinaisons d'hyper-paramètres, propriétés des modèles régissant par exemple leur structure interne, ont été étudiées et comparées. Il en ressort que certaines variables d'entrée sont des informations pertinentes afin d'estimer plus fidèlement la masse de la résonance. Les réseaux de neurones proposent de meilleures performances que les arbres de décision améliorés d'après les métriques d'évaluation que nous avons utilisées, en particulier pour les valeurs de masse de la résonance correspondant aux bosons Z et h du modèle standard. Les performances des réseaux de neurones dépendent également fortement de l'algorithme d'optimisation utilisé lors de l'entraînement.

Nous avons alors déterminé une combinaison performante d'hyper-paramètres correspondant au modèle B. Divers effets sur ses prédictions ont été étudiés. Ainsi, l'empilement doit être pris en

compte lors de l'entraînement afin d'obtenir un modèle pertinent pour les analyses.

Un effet majeur sur la précision des prédictions est lié à la reconstruction des particules. Dans le cas d'une reconstruction parfaite des particules, c'est-à-dire en utilisant les objets générés correspondants au lieu des objets reconstruits, la résolution relative sur la masse de la résonance est de 3 % contre 20 à 25 % sinon. De plus, les faux τ_h perturbent les prédictions des modèles à basse masse, en particulier dans la région des bosons Z et h. Cependant, l'entraînement de modèles spécifiques aux différents canaux ou aux différentes phénoménologies de canaux n'apporte pas de gain en termes de précision des prédictions. L'utilisation de la PFMET au lieu de la PUPPIMET a un effet négligeable sur les prédictions du modèle face à sa résolution.

La gamme de masse explorée lors de l'entraînement définit la zone utile du modèle, ses prédictions étant en bonne approximation restreintes à cet intervalle. Cependant, il n'est pas possible d'étendre cet intervalle à l'infini et des effets de bord apparaissent sur les prédictions du modèle. Nous avons modifié la fonction de coût afin de rejeter dynamiquement certains événements de l'entraînement pour réduire cet effet de bord avec succès. L'exploitation de la queue à hautes valeurs de la distribution de la masse de la résonance, objectif des prédictions des modèles, a permis d'améliorer encore les prédictions moyennes obtenues avec le modèle B". Ce modèle permet de reconstruire avec succès la masse d'une résonance se désintégrant en paire de leptons τ entre 50 GeV et 800 GeV avec une précision de 20 à 25 %.

La prédiction du modèle B", m_{ML}, a été utilisée en tant que variable discriminante à la place de $m_{\mathrm{T}}^{\mathrm{tot}}$ pour obtenir les limites d'exclusion indépendantes du modèle de l'analyse présentée dans le chapitre 4 sur l'année 2017. Malgré des valeurs plus proches de la vraie masse de la résonance ainsi qu'une meilleure résolution que $m_{\mathrm{T}}^{\mathrm{tot}}$, m_{ML} ne permet pas de repousser les limites d'exclusion obtenues. Ceci est dû aux processus physiques tels que les faux τ_h ne correspondant pas à une résonance se désintégrant en paire de leptons τ mais passant les critères de sélection des événements appliqués. L'utilisation de m_{ML} en tant que variable discriminante pour la recherche de bosons de Higgs supplémentaires de haute masse n'est donc pas pertinente. D'autres analyses peuvent en revanche bénéficier de ce projet. Son utilisation dans un autre but que d'obtenir une variable discriminante est également envisageable, par exemple pour la sélection des événements.

Notre modèle permet en effet de mieux séparer les événements $Z \to \tau\tau$ des faux τ_h que $m_{\mathrm{T}}^{\mathrm{tot}}$. De plus, une comparaison des valeurs de m_{ML} à celles de m_{SVFIT}, obtenues par l'algorithme SVFIT déjà utilisé par la collaboration CMS, a permis de mettre en lumière une meilleure description du boson Z par notre modèle. En effet, certains événements $Z \to \tau\tau$ sont prédits au-delà de 250 GeV par SVFIT, alors que la masse du Z est de 91,2 GeV. Parmi ces événements, notre modèle en prédit aux alentours de 100 GeV, ce qui est plus proche de la valeur vraie. L'effet inverse, c'est-à-dire des événements tels que $m_{\mathrm{ML}} > 250$ GeV alors que $m_{\mathrm{SVFIT}} \simeq 100$ GeV pour le boson Z, n'est pas observé. La sensibilité au boson de Higgs du modèle standard h est similaire avec m_{SVFIT} ou m_{ML} dans certaines topologies d'événements. Pour d'autres, le signal de h est plus étendu avec m_{ML} qu'avec m_{SVFIT}, donnant une sensibilité moindre. L'utilisation de processus physiques plus variés pour entraîner les modèles pourrait améliorer leurs prédictions sur de telles topologies. Une résolution similaire voire meilleure que celle de SVFIT est donc envisageable avec des réseaux de neurones.

Le temps nécessaire pour obtenir les prédictions de masse est 60 fois plus court avec notre modèle qu'avec SVFIT. Les futures analyses de la collaboration CMS seront basées sur de plus grandes quantité d'événements, l'utilisation des réseaux de neurones au lieu de SVFIT présente donc un intérêt certain afin de minimiser leur coût computationnel.

Le modèle B" développé au cours de ma thèse [30] peut être récupéré [31] et utilisé dans d'autres analyses. Le groupe en charge de l'analyse des événements avec une paire de leptons τ dans le cadre du NMSSM (*Next to MSSM*), modèle contenant sept bosons de Higgs contre cinq dans le MSSM introduit au chapitre 1, a déjà manifesté un intérêt pour notre modèle. De même, l'utilisation de B" pour l'analyse des événements $hh \to b\bar{b}\tau\tau$, c'est-à-dire avec deux bosons de Higgs dont l'un se désintègre en paire de quarks b et l'autre en paire de leptons τ, est déjà étudiée. La topologie de ces événements, différente de ceux contenant uniquement $h \to \tau\tau$, permet de tester notre modèle dans des situations inédites. Les résultats obtenus à cette occasion pourront permettre d'améliorer le modèle et d'étendre son domaine d'application.

Références

[1] DeepMind. *AlphaGo*. URL : https://www.deepmind.com/research/case-studies/alphago-the-story-so-far.

[2] C. Bernet. *The Data Frog – Image Recognition : Dogs vs Cats !* URL : https://thedatafrog.com/en/articles/dogs-vs-cats/.

[3] M. Mir. *House Prices Prediction Using Deep Learning*. URL : https://towardsdatascience.com/house-prices-prediction-using-deep-learning-dea265cc3154.

[4] G. Touquet. « Search for an additional neutral MSSM Higgs boson decaying to tau leptons with the CMS experiment ». Thèse de doct. Université Claude Bernard Lyon 1, oct. 2019. URL : https://hal.archives-ouvertes.fr/tel-02526393.

[5] M. Scham. « Standard Model $H \to \tau\tau$ Analysis with a Neural Network Trained on a Mix of Simulation and Data Samples ». Mém. de mast. Fakultät für Physik des Karlsruher Instituts für Technologie (KIT), juin 2020. URL : https://publish.etp.kit.edu/record/21993.

[6] T. Kopf. « Recoil Calibration as a Neural Network Task ». Mém. de mast. Fakultät für Physik des Karlsruher Instituts für Technologie (KIT), fév. 2019. URL : https://publish.etp.kit.edu/record/21500.

[7] P. Baldi, P. Sadowski & D. Whiteson. « Enhanced Higgs Boson to $\tau^+\tau^-$ Search with Deep Learning ». *Physical Review Letters* **114**.11 (mar. 2015). DOI : 10.1103/physrevlett.114.111801.

[8] D. Guest & coll. « Jet flavor classification in high-energy physics with deep neural networks ». *Physical Review* **D94**.11 (déc. 2016). DOI : 10.1103/physrevd.94.112002.

[9] The CMS Collaboration. « Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV ». *Journal of Instrumentation* **13**.05 (mai 2018). DOI : 10.1088/1748-0221/13/05/p05011.

[10] The CMS Collaboration. *DeepJet : deep learning based on physics objects for jet reconstruction*. URL : https://twiki.cern.ch/twiki/bin/viewauth/CMS/DeepFlavour.

[11] The CMS Collaboration. « Performance of the DeepTau algorithm for the discrimination of taus against jets, electron, and muons » (oct. 2019). URL : https://cds.cern.ch/record/2694158.

[12] J. Andrejkovic & coll. « Measurement of Higgs(125) boson properties in decays to a pair of tau leptons with full Run II data using Machine-Learning techniques ». *CMS analysis Note* (sept. 2020). URL : https://cms.cern.ch/iCMS/jsp/db_notes/noteInfo.jsp?cmsnoteid=CMS%5C%20AN-2019/177.

[13] J. Andrejkovic & coll. « Multi-class neural network architecture and training for measurements of Higgs(125) boson decays to two tau leptons on full Run II data ». *CMS analysis Note* (mai 2020). URL : https://cms.cern.ch/iCMS/jsp/db_notes/noteInfo.jsp?cmsnoteid=CMS%5C%20AN-2019/178.

[14] A. Elagin & coll. « A new mass reconstruction technique for resonances decaying to $\tau\tau$ ». *Nuclear Instruments and Methods in Physics Research* **A654**.1 (2011), p. 481-489. DOI : 10.1016/j.nima.2011.07.009.

[15] A. J. Barr & coll. « Speedy Higgs boson discovery in decays to tau lepton pairs : $h \to \tau\tau$ ». *Journal of High Energy Physics* **2011**.10 (oct. 2011). DOI : 10.1007/JHEP10(2011)080.

[16] B. Gripaios & coll. « Reconstruction of Higgs bosons in the di-tau channel via 3-prong decay ». *Journal of High Energy Physics* **2013**.3 (mar. 2013). DOI : 10.1007/JHEP03(2013)106.

[17] L. Bianchini & coll. « Reconstruction of the Higgs mass in $H \to \tau\tau$ Events by Dynamical Likelihood techniques ». *Journal of Physics : Conference Series* **513**.2 (juin 2014). DOI : 10.1088/1742-6596/513/2/022035.

[18] P. Bärtschi & coll. « Reconstruction of τ lepton pair invariant mass using an artificial neural network ». *Nuclear Instruments and Methods in Physics Research* **A929** (2019), p. 29-33. DOI : 10.1016/j.nima.2019.03.029.

[19] J. de Favereau & coll. « Delphes 3 : a modular framework for fast simulation of a generic collider experiment ». *Journal of High Energy Physics* **2** (fév. 2014). doi : 10.1007/jhep02(2014)057.

[20] A. Mertens. « New features in Delphes 3 ». *Journal of Physics : Conference Series* **608**.1 (2015). Sous la dir. de L. Fiala, M. Lokajicek & N. Tumova. doi : 10.1088/1742-6596/608/1/012045.

[21] S. Abdullin & coll. « The Fast Simulation of the CMS Detector at LHC ». *Journal of Physics : Conference Series* **331**.3 (déc. 2011). doi : 10.1088/1742-6596/331/3/032049.

[22] A. Giammanco. « The Fast Simulation of the CMS Experiment ». *Journal of Physics : Conference Series* **513**.2 (juin 2014). doi : 10.1088/1742-6596/513/2/022012.

[23] M. Komm. « Fast emulation of track reconstruction in the CMS simulation ». *Journal of Physics : Conference Series* **898** (oct. 2017). doi : 10.1088/1742-6596/898/4/042034.

[24] S. Sekmen. *Recent Developments in CMS Fast Simulation*. 2017. arXiv : 1701.03850.

[25] S. Agostinelli & coll. « Geant4 – A simulation toolkit ». *Nuclear Instruments and Methods in Physics Research* **A506**.3 (2003), p. 250-303. doi : 10.1016/S0168-9002(03)01368-8.

[26] J. Allison & coll. « Geant4 developments and applications ». *IEEE Transactions on Nuclear Science* **53**.1 (fév. 2006), p. 270-278. doi : 10.1109/tns.2006.869826.

[27] J. Allison & coll. « Recent developments in Geant4 ». *Nuclear Instruments and Methods in Physics Research* **A835** (2016), p. 186-225. doi : 10.1016/j.nima.2016.06.125.

[28] E. Aşılar, L. Torterotot & C. Bernet. « Reconstruction of di-tau mass using deep neural networks ». *CMS analysis Note* (2021). url : https://cms.cern.ch/iCMS/jsp/db_notes/noteInfo.jsp?cmsnoteid=CMS%5C%20AN-2021/054.

[29] E. Aşılar. *How to produce nanoAOD events of $h \to \tau\tau$ where Higgs has a* 130 GeV *mass*. url : https://github.com/easilar/cmssw/blob/from-CMSSW_10_2_22/README.

[30] L. Torterotot, E. Aşılar & C. Bernet. *Reconstruction of di-tau mass using Machine Learning*. url : https://github.com/lucastorterotot/DL_for_HTT_mass.

[31] L. Torterotot. *DiTau_ML_mass – Estimations of di-tau mass using Machine Learning*. url : https://github.com/lucastorterotot/DiTau_ML_mass.

[32] T. Sjöstrand & coll. « An Introduction to pythia 8.2 ». *Computer Physics Communications* **191** (2015), p. 159-177. doi : 10.1016/j.cpc.2015.01.024. arXiv : 1410.3012 [hep-ph].

[33] The CMS Collaboration. « Event generator tunes obtained from underlying event and multiparton scattering measurements ». *European Physical Journal* **C76**.3 (2016). doi : 10.1140/epjc/s10052-016-3988-x. arXiv : 1512.00815 [hep-ex].

[34] The CMS Collaboration. « Extraction and validation of a new set of CMS pythia 8 tunes from underlying-event measurements ». *European Physical Journal* **C80** (mar. 2019). url : https://cds.cern.ch/record/2669320.

[35] LHC Higgs Cross Section Working Group. « Higgs Properties ». *Handbook of LHC Higgs Cross Sections*. **3**. CERN Yellow Reports : Monographs. Geneva : CERN, 2013. url : https://cds.cern.ch/record/1559921.

[36] The CMS Collaboration. *Pileup Removal Algorithms*. Rapp. tech. CMS-PAS-JME-14-001. Geneva : CERN, 2014. url : https://cds.cern.ch/record/1751454.

[37] M. Cacciari, G. P. Salam & G. Soyez. « The Anti-k_T jet clustering algorithm ». *Journal of High Energy Physics* **04** (avr. 2008). doi : 10.1088/1126-6708/2008/04/063. arXiv : 0802.1189 [hep-ph].

[38] D. Bertolini & coll. « Pileup per particle identification ». *Journal of High Energy Physics* **10** (oct. 2014). doi : 10.1007/jhep10(2014)059.

[39] A. Cauchy. « Méthode générale pour la résolution des systèmes d'équations simultannées ». *Comptes rendus hebdomadaires des séances de l'Académie des Sciences*. **25**. 1847, p. 536-538. url : https://gallica.bnf.fr/ark:/12148/bpt6k2982c/f3.item.

[40] L. Bottou. « Online Algorithms and Stochastic Approximations ». *Online Learning and Neural Networks*. Sous la dir. de D. Saad. Cambridge, UK : Cambridge University Press, 1998. URL : http://leon.bottou.org/papers/bottou-98x.

[41] I. Goodfellow, Y. Bengio & A. Courville. *Deep Learning*. MIT Press, 2016. URL : http://www.deeplearningbook.org.

[42] T. Chen & C. Guestrin. « XGBoost : A Scalable Tree Boosting System ». *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (août 2016). DOI : 10.1145/2939672.2939785.

[43] *Kaggle Competitions*. URL : https://www.kaggle.com/competitions.

[44] W. Sarle. « Neural Networks and Statistical Models ». 1994. URL : https://people.orie.cornell.edu/davidr/or474/nn_sas.pdf.

[45] F. Chollet & coll. Keras. 2015. URL : https://keras.io.

[46] M. Abadi & coll. TensorFlow *: Large-scale machine learning on heterogeneous distributed systems*. Software available from tensorflow.org. 2015. URL : https://www.tensorflow.org/.

[47] X. Glorot & Y. Bengio. « Understanding the difficulty of training deep feedforward neural networks ». *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Sous la dir. d'Y. W. Teh & M. Titterington. **9**. Proceedings of Machine Learning Research. PMLR, mai 2010, p. 249-256. URL : http://proceedings.mlr.press/v9/glorot10a.html.

[48] J. Duchi, E. Hazan & Y. Singer. « Adaptive Subgradient Methods for Online Learning and Stochastic Optimization ». *Journal of Machine Learning Research* **12**.61 (2011), p. 2121-2159. URL : http://jmlr.org/papers/v12/duchi11a.html.

[49] G. Hinton. *Neural Networks for Machine Learning*. Coursera Video Lectures, Academic Torrents. 2012. URL : https://archive.org/search.php?query=creator%3A%22Geoffrey+Hinton%22&and[]=year%3A%222012%22.

[50] D. P. Kingma & J. Ba. « Adam : A Method for Stochastic Optimization » (2017). arXiv : 1412.6980 [cs.LG].

[51] J. Andrejkovic & coll. « Data-driven background estimation of fake-tau backgrounds in di-tau final states with 2016 and 2017 data ». *CMS analysis Note* (oct. 2018). URL : https://cms.cern.ch/iCMS/jsp/db_notes/noteInfo.jsp?cmsnoteid=CMS%5C%20AN-2018/257.

[52] J. Andrejkovic & J. Bechtel. « Data-driven background estimation of fake-tau backgrounds in di-tau final states with the full Run-II dataset ». *CMS analysis Note* (juin 2020). URL : https://cms.cern.ch/iCMS/jsp/db_notes/noteInfo.jsp?cmsnoteid=CMS%5C%20AN-2019/170.

[53] The CMS Collaboration. « Reconstruction and identification of tau lepton decays to hadrons and tau neutrino at CMS ». *Journal of Instrumentation* **11**.1 (2016). DOI : 10.1088/1748-0221/11/01/P01019. arXiv : 1510.07488 [physics.ins-det].

[54] A. Gottmann. « Global Interpretation of $\tau\tau$ Events in the Context of the Standard Model and Beyond ». Thèse de doct. Fakultät für Physik des Karlsruher Instituts für Technologie (KIT), juin 2020. URL : https://publish.etp.kit.edu/record/22014.

[55] The CMS Collaboration. « Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector ». *Physics Letters* **B779** (avr. 2018), p. 283-316. DOI : 10.1016/j.physletb.2018.02.004.