

# **Rapport de Stage**

Amélioration de la reconstruction des événements dans le  
détecteur Hyper-Kamiokande

**Dubrulle Thibaud**

Stage de deuxième année du CPES  
du 1er juillet au 1er août 2025

**Superviseur : M. Erwan Le Blevec**

Laboratoire Leprince-Ringuet (LLR)  
École Polytechnique

# Table des matières

1. Introduction .....	4
2. Présentation de la structure d'accueil .....	5
3. Contexte scientifique et technique .....	6
3.1. Les neutrinos .....	6
3.1.a. Oscillation et états de masse .....	7
3.2. Le Super-Kamiokande détecteur .....	8
3.2.a. Description du détecteur .....	8
3.2.b. Principe du détection .....	8
3.2.c. Reconstruction d'événements .....	9
3.3. Le détecteur Hyper-Kamiokande .....	10
4. Nouvelles améliorations de l'algorithme de traitement de données .....	11
4.1. Contexte technique .....	11
4.2. Séparation des ensembles de données .....	11
4.3. Génération d'un fichier d'extremas et d'une structure de clés .....	12
5. Première estimation approximative de l'Énergie .....	14
5.1. Exploration et Préparation du Jeu de Données .....	14
5.1.a. Description du jeu de données .....	14
5.1.b. Prétraitement des données .....	14
5.1.c. Analyse exploratoire initiale .....	14
5.2. Régression Linéaire Simple .....	16
5.2.a. Théorie .....	16
5.2.b. Modélisation avec la charge totale .....	16
5.2.c. Modélisation avec le nombre de hits .....	17
5.2.d. Impact de la variable <code>towall</code> .....	18
5.2.e. Bilan Régression Linéaire Simple .....	20
5.3. Régression Linéaire Multiple .....	22
5.3.a. Théorie .....	22
5.3.b. Choix des variables explicatives .....	22
5.3.c. Analyse des résidus .....	23
5.3.c.i. Théorie .....	23
5.3.c.ii. Distribution des résidus .....	24
5.3.c.iii. Événements leviers ayant un résidu élevé .....	25
5.3.d. Résultats de performance .....	28
5.3.d.i. Résultats sur l'ensemble des données .....	28
5.3.d.ii. Résultats sur les événements ayant un <code>towall</code> > 500 .....	28
5.3.d.iii. Résultats sur une segmentation de <code>towall</code> .....	29
5.4. Modèle cyclique (Pipeline) .....	30
5.4.a. Corrélation entre <code>nhits</code> et <code>towall</code> .....	30
5.4.b. Architecture du pipeline .....	30
5.4.c. Résultats .....	31
5.5. Résumé de résultats par modèles .....	33

5.6. Régression Ridge .....	34
5.7. Correction des charges captées .....	34
5.7.a. Théorie sur l'absorption et dépendance angulaire .....	34
5.7.a.i. Absorption des photons .....	34
5.7.a.ii. Dépendance angulaire .....	35
5.7.b. Méthodes .....	35
5.7.b.i. Coefficient d'absorption de photons .....	35
5.7.b.ii. Correction de la charge en fonction de la distance au vertex .....	36
5.7.b.iii. Calcul de l'angle Incident .....	38
5.7.b.iv. Correction de la charge en fonction de l'angle .....	39
5.7.c. Résultats .....	39
5.8. Aspects non aboutis .....	41
6. Conclusion .....	42

# 1. Introduction

Ce rapport présente les travaux que j'ai réalisés dans le cadre de mon stage de deuxième année du cycle CPES (Cycle Pluridisciplinaire d'Études Supérieures), effectué du 1er juillet au 1er août 2025 au sein du Groupe Neutrinos du Laboratoire Le Prince Ringuet(LLR), sous la supervision de Erwan Le Blevec.

L'objectif principal de ce stage était de contribuer à l'amélioration de la reconstruction des événements dans le détecteur Hyper-Kamiokande, un détecteur de neutrinos de nouvelle génération situé au Japon.

Mon travail s'est articulé en deux parties :

## **1. Mise à jour et amélioration d'un logiciel de traitement de données.**

J'ai modifié et optimisé un outil existant permettant la conversion des fichiers bruts au format « .root » vers un format exploitable (« .pt »), afin de faciliter leur utilisation dans des environnements d'analyse basés sur le machine learning.

## **2. Développement d'un nouvel algorithme d'estimation de l'énergie.**

J'ai conçu et implémenté un modèle supervisé de régression, s'appuyant sur des techniques de machine learning, afin de réaliser une première estimation de l'énergie des événements enregistrés par le détecteur Hyper-Kamiokande.

Ce travail s'inscrit dans un contexte de recherche fondamentale en physique des particules, dont l'un des enjeux majeurs est la compréhension du comportement des neutrinos. Les neutrinos sont des particules élémentaires encore largement mystérieuses malgré leur abondance dans l'univers. Le projet s'appuie sur les avancées techniques et méthodologiques du détecteur Super-Kamiokande, prédecesseur d'Hyper-Kamiokande, et vise à optimiser les outils de traitement et d'interprétation des événements détectés.

## 2. Présentation de la structure d'accueil

Mon stage s'est déroulé au Laboratoire Leprince-Ringuet (LLR), une unité de recherche conjointe de l'École Polytechnique et du CNRS-IN2P3 (Institut National de Physique Nucléaire et de Physique des Particules). Situé sur le campus de l'École Polytechnique à Palaiseau, le LLR est un acteur majeur de la recherche en physique des particules, physique des astroparticules et instrumentation.

J'ai été intégré à l'équipe Neutrinos, qui participe à plusieurs projets internationaux visant à étudier ces particules fondamentales. Cette équipe est notamment impliquée dans les collaborations Super-Kamiokande et Hyper-Kamiokande, deux détecteurs japonais conçus pour observer les interactions rares des neutrinos avec la matière.

### 3. Contexte scientifique et technique

#### 3.1. Les neutrinos

Les neutrinos sont les particules au centre de ce rapport, le détecteur Super-Kamiokande (SK) ayant été construit pour pouvoir les détecter.

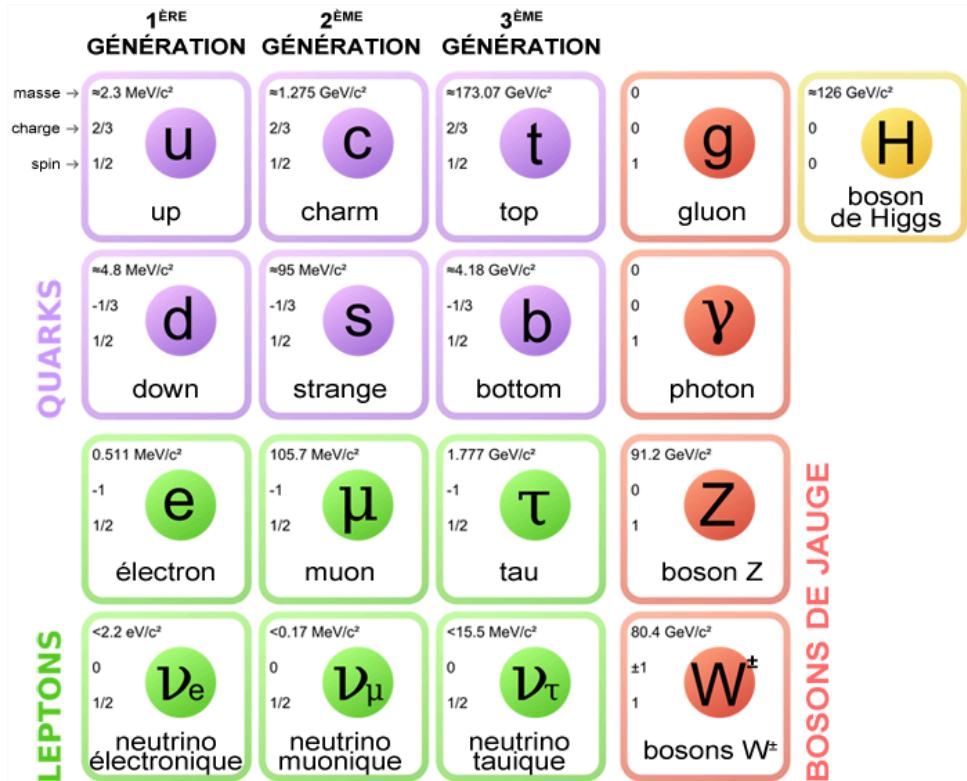


Fig. 1. – Particules du Modèle Standard

Ils font partie des particules élémentaires du modèle standard de la physique des particules. Découvertes en 1956 pour la première fois par Frederick Reines et Clyde Cowan, ces particules sont très difficiles à détecter car elles interagissent seulement par interaction faible et par gravité (les neutrinos ont une masse très faible et une charge neutre). Après le photon, le neutrino est la particule la plus abondante dans l'univers.

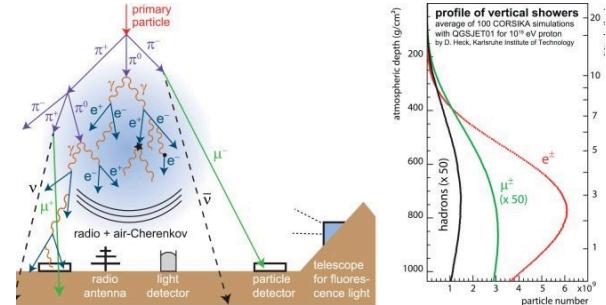
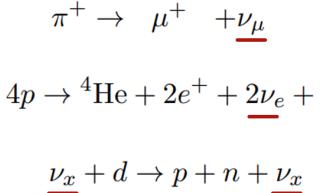


Fig. 2. – Différentes réactions dans lesquelles le neutrino intervient.

Fig. 3. – Rayon cosmique arrivant sur la Terre.

Les neutrinos sont produits dans divers processus physiques, notamment dans les réactions nucléaires au cœur des étoiles comme notre Soleil, dans les interactions des rayons cosmiques avec l'atmosphère terrestre, et dans des faisceaux de neutrinos (neutrino beam) produits artificiellement.

### 3.1.a. Oscillation et états de masse

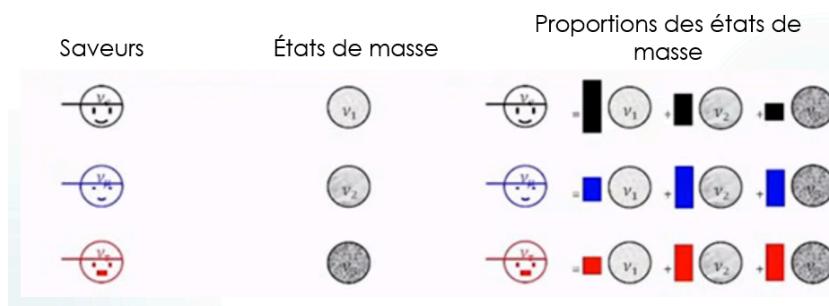


Fig. 4. – Schéma de l'oscillation des neutrinos

Un des phénomènes les plus fascinants concernant les neutrinos est leur oscillation. Il existe trois « saveurs » de neutrinos (électronique, muonique et tauique), et un neutrino peut changer de saveur au cours du temps. Ce phénomène d'oscillation a été mis en évidence pour la première fois en 1998 dans le détecteur Super Kamiokande. Lors d'une expérience supposant qu'il n'y avait pas d'oscillation on a observé un large déficit de neutrinos atmosphérique muoniques provenant d'en dessous du détecteur (ayant traversé toute la Terre avant d'avoir été détecté) alors que le nombre de neutrinos muoniques provenant du dessus du détecteur est cohérent avec cette supposition (faible de distance parcourue avant d'être détecté). Les neutrinos ayant parcourus une plus longue distance ont donc eu le temps d'osciller ce qui expliquerai ce déficit.

Ces oscillations signifient aussi qu'un neutrino est observé avec une certaine saveur à un instant t, qui n'est pas nécessairement celle de sa création.

## 3.2. Le Super-Kamiokande détecteur

### 3.2.a. Description du détecteur

Super-Kamiokande est un détecteur de neutrinos de type Tcherenkov à eau, situé à 1000 mètres sous terre dans la mine de Kamioka au Japon. C'est une structure cylindrique contenant 50 000 tonnes d'eau ultrapure. Il est équipé de plus de 11 000 tubes photomultiplicateurs (PMT) qui détectent la lumière émise lors d'un événement (une interaction de neutrino avec l'eau).

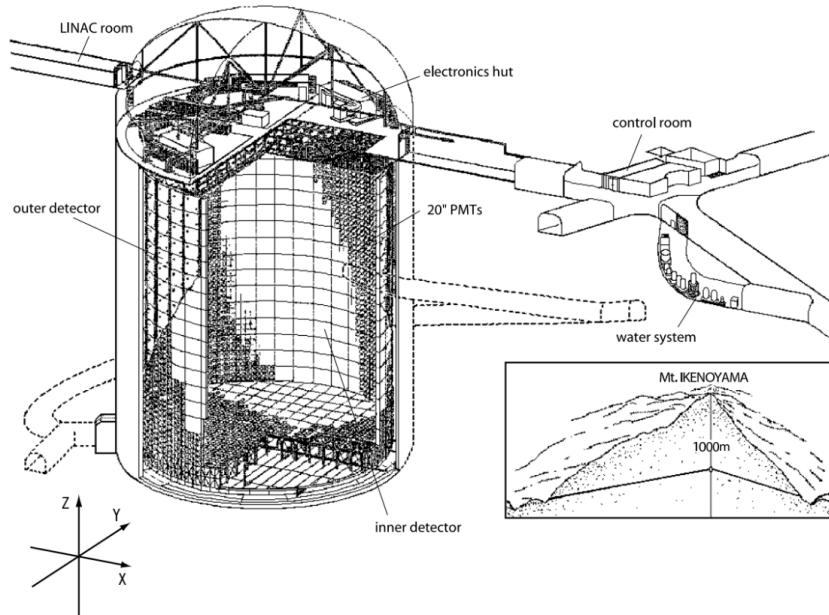


FIGURE 2.1: Super-Kamiokande detector and its location in Mountain Ikenoyama.

Fig. 5. – Détecteur Super-Kamiokande et sa localisation dans la montagne Ikenoyama

### 3.2.b. Principe du détection

Lorsqu'une particule chargée traverse l'eau à une vitesse supérieure à celle de la lumière dans ce milieu, elle émet un cône de lumière appelé rayonnement Tcherenkov. C'est ce rayonnement qui est détecté par les PMTs.

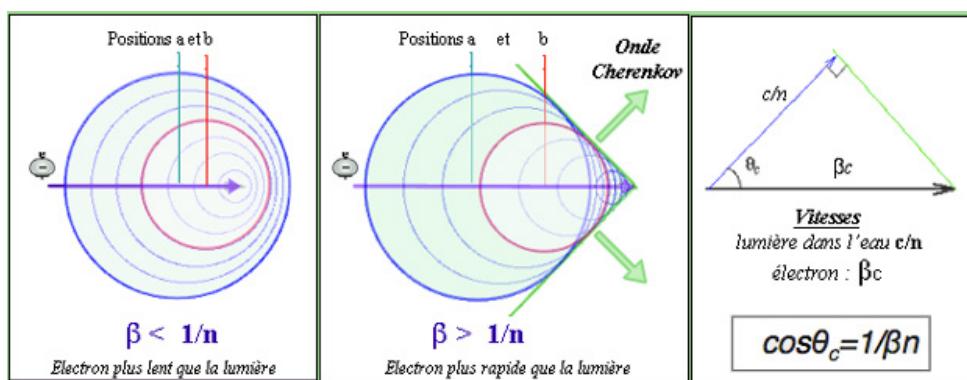


Fig. 6. – Le front d'onde de lumière Tcherenkov. Les ondes sphériques se rattrapent, générant un front d'onde de forme conique qui suit la particule.

Les PMTs sont des capteurs de photons extrêmement sensibles. La surface interne du détecteur en est entièrement recouverte.

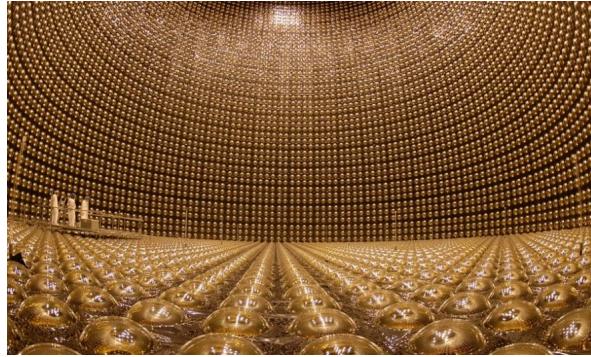


Fig. 7. – SK, vue de l'intérieur

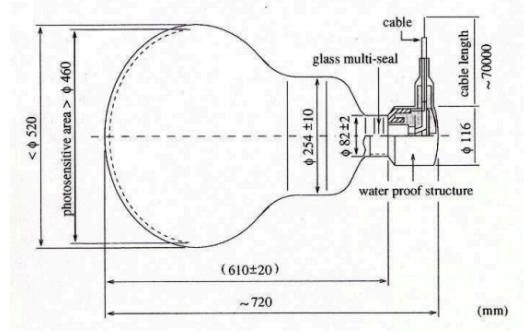


Fig. 8. – Un tube photomultiplicateur (PMT)

### 3.2.c. Reconstruction d'événements

Lors d'un événement, on récupère toutes les données captées par les PMTs, ce qui nous aide à reconstruire l'événement. C'est à dire qu'à partir des données brutes détectée par l'ensemble des PMTs on cherche à reconstituer l'interaction ayant eu lieu. Pour cela On cherche à déterminer les caractéristiques physiques de l'événement : son énergie, le vertex (point d'interaction), la direction et le type de neutrino ( $e$ ,  $\mu$ ,  $\tau$ ).

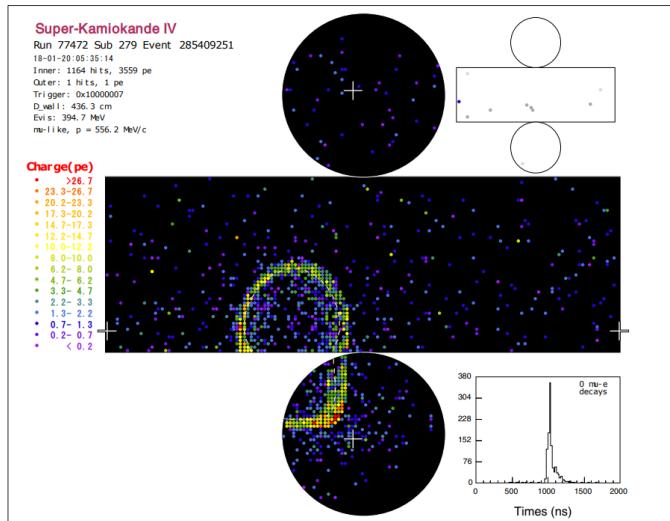


FIGURE 2.2: A example of neutrino event observed by Super-K detector. Each point denotes a PMT hit, whose charge is represented by the color. The Cherenkov ring pattern can be seen clearly.

Fig. 9. – Exemple d'un événement neutrino observé par le détecteur Super-Kamiokande. Les couleurs indiquent le temps d'arrivée de la lumière.

### 3.3. Le détecteur Hyper-Kamiokande

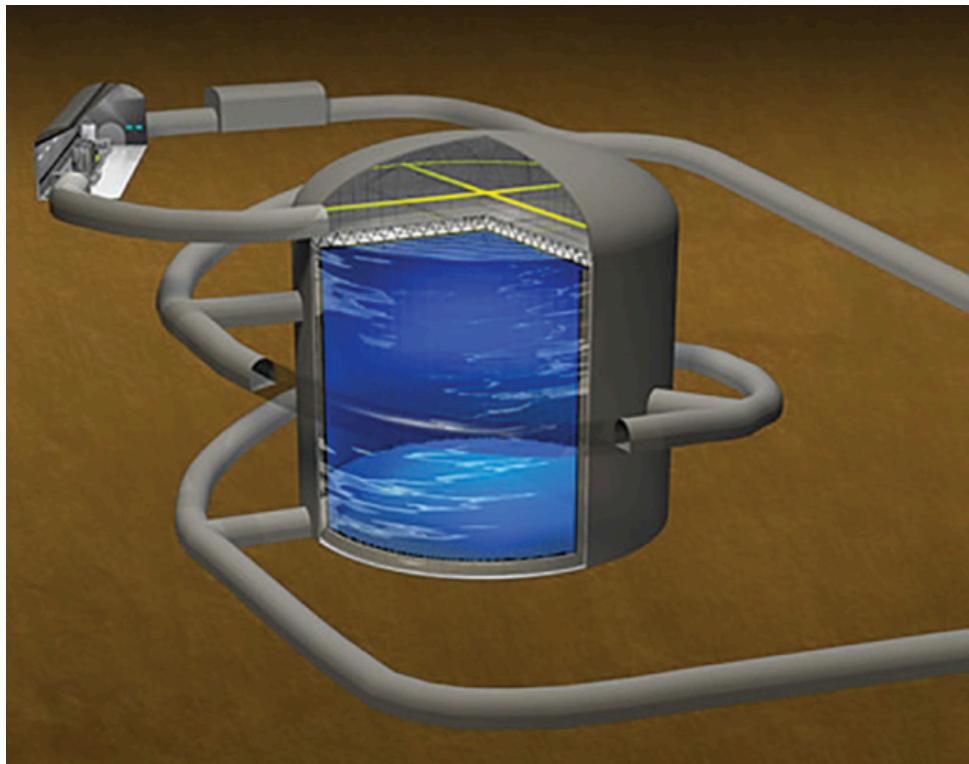


Fig. 10. – Hyper Kamiokande en modélisation 3D

# todo intro T2K

Hyper-Kamiokande (HK) est la prochaine génération du détecteur de neutrinos japonais, conçu comme la suite directe du Super-Kamiokande (SK). Situé à Kamioka, dans la préfecture de Gifu (Japon), à environ 650 mètres sous terre, HK en reprend le principe de détection basé sur l'eau ultra-pure et la lumière Cherenkov, mais avec un volume environ 8 fois plus grand et des photomultiplicateurs de nouvelle génération. La construction a débuté en 2020, avec un démarrage prévu pour 2027, et jouera un rôle central dans le programme T2K. Hyper-K réunit environ 300 chercheurs de 15 pays et s'appuie sur l'expertise acquise avec Super-K. L'objectif du HK est d'augmenter significativement la sensibilité aux phénomènes rares comme la désintégration du proton, les oscillations de neutrinos et les neutrinos cosmiques, poursuivant ainsi le travail initié par SK avec une précision inégalée.

## 4. Nouvelles améliorations de l'algorithme de traitement de données

La première mission de mon stage a consisté à modifier un algorithme appelé `RootToGraph`, utilisé pour convertir des fichiers de données au format `.root` en fichiers compatibles avec PyTorch (`.pt`), contenant des graphes utilisables pour l'entraînement de réseaux de neurones.

### 4.1. Contexte technique

Les fichiers `.root` issus des simulations contiennent de nombreuses variables par événement. Ces événements doivent être transformés en **graphes** pour permettre leur exploitation par des modèles d'apprentissage automatique.

Chaque graphe encode :

- Des **nœuds** représentant les PMTs activés,
- Des **features** pour chaque nœud (ex : charge, temps, position),
- Et des **arêtes** reliant les nœuds.

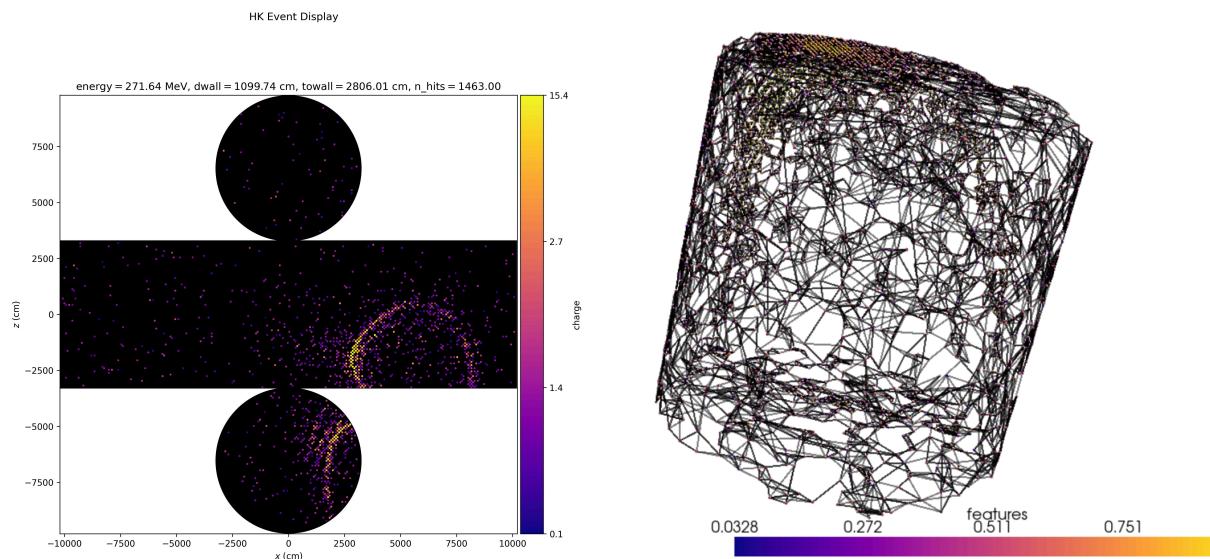


Fig. 11. – Visualisation de l'événement

Fig. 12. – Graphe produit correspondant à cet événement (vue en 3D).

Ici sur cet exemple de ce qui se trouve sur le fichier Pytorch à la fin chaque point du graphe est un PMT qui a reçu une certaine quantité de photons sur une fenêtre de temps. Ici le graphe a été formé grâce à l'algorithme KNN ( $k$  plus proches voisins).

### 4.2. Séparation des ensembles de données

J'ai intégré une gestion automatique de la séparation des données en trois sous-ensembles : **entraînement**, **validation** et **test**, à partir d'un fichier de splitting externe.

Cette étape est essentielle car la pipeline actuelle charge l'intégralité du fichier .pt en mémoire. Si les graphes de test y sont inclus, cela implique une utilisation de la mémoire inutile pendant les phases train/val. Ma modification permet donc un chargement plus ciblé et plus efficace, avec un fichier .pt distinct pour chaque sous-ensemble.

📄	data_meta_test.pt	10/07/2025 15:10
📄	data_meta_train_val.pt	10/07/2025 15:10
📄	data_pos_test.pt	10/07/2025 15:10
📄	data_pos_train_val.pt	10/07/2025 15:10
📄	data_test.pt	10/07/2025 15:10
📄	data_train_val.pt	10/07/2025 15:10
📄	extrema_key_order.npz	10/07/2025 16:24
📄	pre_filter.pt	10/07/2025 15:10
📄	pre_transform.pt	10/07/2025 15:10

Fig. 13. – Dossier /processed en résultat

### 4.3. Génération d'un fichier d'extremas et d'une structure de clés

J'ai également ajouté un module générant automatiquement dans un même fichier :

- Les **extremas** pour chaque variable (minima et maxima),
- L'**ordre des clés** (features) présentes dans les tenseurs.

L'objectif de ce second fichier est pratique : un tenseur brut n'indique pas explicitement quelles colonnes correspondent à quelles variables. Sans cette information, il devient compliqué de savoir comment interpréter chaque dimension du tenseur. Le fichier que j'ai généré aide donc à l'interprétation des données.

```

charge.npy
[0.10028630495071411, 738.3221435546875
dwall.npy
[29.287353515625, 2268.763671875]
energy.npy
[105.69336700439453, 981.109375]
eventType.npy
[11, 11]
hitx.npy
[-3242.76611328125, 3242.76611328125]
hity.npy
[-3242.76611328125, 3242.76611328125]

```

Fig. 14. – Exemple du fichier contenant les extremas (sous forme de tableau numpy) dans le fichier npz

```
train_keys_order.npy  
['charge', 'time', 'hitx', 'hity', 'hitz']  
label_keys_order.npy  
['eventType', 'energy', 'vertex_x', 'vertex_y', 'vertex_z', 'particleDir_x', 'particleDir_y', 'particleDir_z']  
edge_keys_order.npy  
['time', 'hitx', 'hity', 'hitz', 'charge']
```

Fig. 15. – Exemple du fichier contenant l'ordre des variable (sous forme de tableau numpy)  
dans le même fichier npz

# 5. Première estimation approximative de l'Énergie

## 5.1. Exploration et Préparation du Jeu de Données

### 5.1.a. Description du jeu de données

Le jeu de données utilisé est un dataset d'environ 50 000 événements simulés de neutrinos électroniques, avec une énergie cinétique allant de 100 MeV à 1000 MeV. Pour chaque événement, nous avons accès aux informations suivantes :

- **Informations sur la particule simulée :**
  - `energy`: Énergie cinétique de l'électron
  - `vertex`: Position initiale de la particule
  - `particleDir`: Direction de la particule
  - `dwall`: Distance au mur le plus proche
  - `towall`: Distance au mur le long de la direction de la particule
- **Données des PMTs :**
  - `n_hits`: Nombre de PMTs ayant détecté de la lumière
  - `charge`: Charge détectée par chaque PMT
  - `time`: Temps associé à chaque PMT
  - `hitx`, `hity`, `hitz`: Coordonnées de chaque PMT activé

### 5.1.b. Prétraitement des données

À partir des données brutes, j'ai créé trois nouvelles variables agrégées par événement : la charge totale (`charge_totale`), la charge maximale (`max_charge`) et la charge minimale (`min_charge`). De plus, les variables vectorielles (comme `vertex`) ont été décomposées en leurs composantes scalaires (par exemple `vertex_x`, `vertex_y`, `vertex_z`) pour pouvoir être utilisées dans les modèles de régression.

### 5.1.c. Analyse exploratoire initiale

Une matrice de corrélation a été calculée pour visualiser les relations linéaires entre les variables.

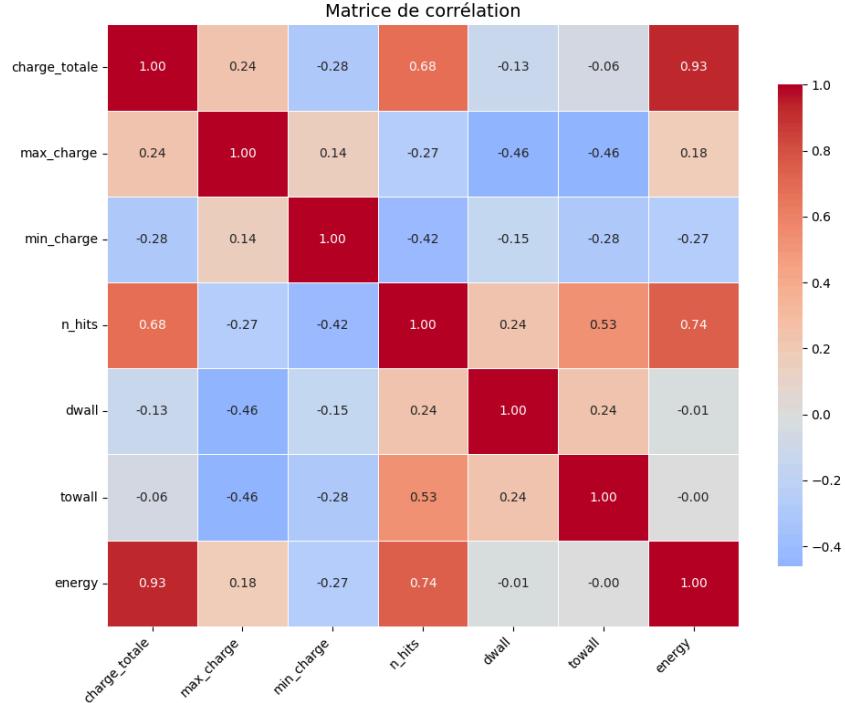


Fig. 16. – Matrice de corrélation entre les principales variables

Comme le montre la matrice de corrélation, l'énergie de l'événement est fortement reliée à la charge totale (`charge_totale`, coefficient de 0.93) et au nombre de PMTs touchés (`n_hits`, 0.74). Ces deux variables traduisent directement l'intensité globale du signal détecté, et constituent donc les meilleures candidates pour une première modélisation.

À l'inverse, des variables comme `towall` ou `dwall` contiennent surtout une information spatiale (position de l'événement), qui n'a pas d'influence directe sur l'énergie (du moins dans un cadre global). De même, `max_charge` et `min_charge` reflètent uniquement la mesure d'un seul PMT et apportent donc peu d'information globale, si ce n'est dans le cas particulier où un neutrino interagit très proche d'un PMT par exemple, produisant un signal exceptionnellement élevé (limité à 1000 dans le traitement).

Enfin, on pourrait aussi se demander pourquoi la corrélation de la charge totale est plus importante que celle du nombre de hits. On pourrait expliquer cette différence de pouvoir explicatif par l'information que chacune d'entre elles contient. En effet, le nombre de hits indique une quantité de PMTs qui ont capté une charge cependant elle n'indique pas du tout l'intensité de ces charges captées. Contrairement au nombre de hits, la charge totale nous donne une mesure (corrigable) de la quantité de photon qui a été émise lors du trajet de l'électron. Ainsi, la charge totale paraît être une variable pouvant mieux expliquer l'énergie d'un événement.

## 5.2. Régression Linéaire Simple

### 5.2.a. Théorie

Pour analyser la relation entre l'énergie de l'événement et les signaux mesurés, j'ai d'abord employé la méthode de la régression linéaire par les moindres carrés. Le principe des moindres carrés consiste à déterminer la droite qui « s'ajuste » le mieux au nuage de points des données en minimisant la somme des carrés des écarts verticaux (résidus) entre chaque point et la droite.

Le modèle s'écrit sous la forme :

$$y = aX + b$$

où  $X$  représente la variable explicative (par exemple `n_hits` ou `charge_totale`),  $y$  l'énergie réelle,  $a$  la pente de la droite et  $b$  l'ordonnée à l'origine. Les coefficients  $a$  et  $b$  sont calculés en minimisant la somme des carrés des résidus, ce qui conduit aux formules classiques de la régression par les moindres carrés :

Pour évaluer la qualité des prédictions, nous utilisons un indicateur appelé **Résolution**, qui mesure la dispersion relative entre les valeurs prédites et les valeurs réelles. Elle est définie comme l'écart-type du résidu relatif :

$$\text{Résolution} = \sigma \left( 100 \times \frac{\hat{y} - y}{y} \right)$$

Où  $\hat{y}$  est la valeur prédite et  $y$  la valeur réelle. Une faible résolution (en %) indique une bonne précision. Nous distinguons deux types de résolution :

- La **Résolution globale**, calculée sur l'ensemble du jeu de test.
- La **Résolution par bin d'énergie**, calculée sur des intervalles d'énergie distincts, ce qui permet d'évaluer la performance du modèle de manière locale.

### 5.2.b. Modélisation avec la charge totale

Un premier modèle a été construit pour prédire l'énergie en fonction de `charge_totale`.

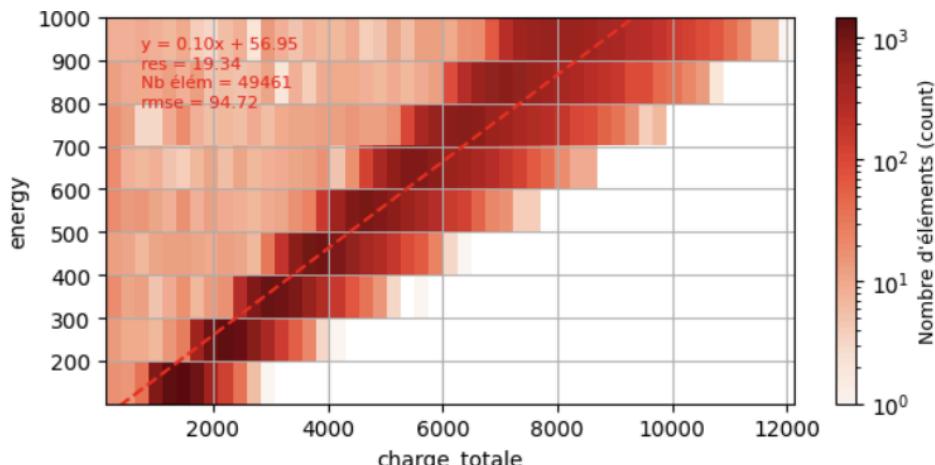


Fig. 17. – Régression linéaire de l'énergie en fonction de la charge totale.

Avec ce premier modèle on constate directement que la densité d'événements autour de la droite est conséquente, seulement on observe aussi que pour chaque bin d'énergie on a des événements dont la charge totale ne correspond pas du tout avec l'énergie de l'événement. On a comme résultat :

- **Root Mean Squared Error (RMSE) :** 94.72 MeV
- **Résolution globale moyenne :** environ 19%

Il est intéressant aussi de regarder la résolution moyenne par bin d'énergie (ce qui diffère de la résolution globale) :

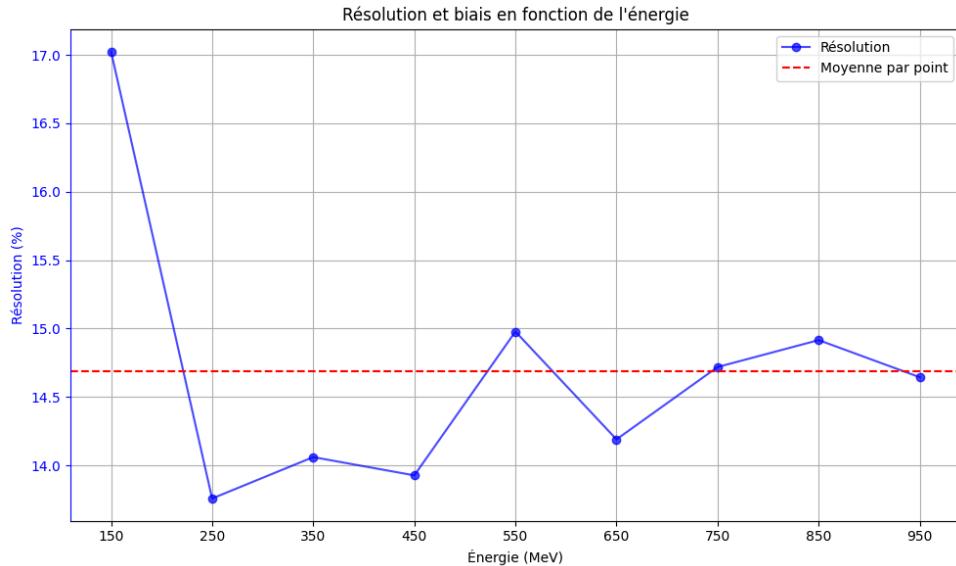


Fig. 18. – Résolution par tranche d'énergie pour le modèle basé sur la charge totale.

On observe une résolution stable avec cependant une résolution plus haute pour le bin d'énergie [100 - 200] Mev.

```
\# to do interprétation hausse de la résolution
- le modèle se fit sur l'ensemble des données et pas particulièrement sur les basses énergie, et ici il semble que la droite ne se soit pas ajuster correctement pour les basses énergies (densité élevée de points par sur la droite mais plutôt à sa droite) donc c'est peut être pour ça qu'on observe une hausse de la resolution.
- indicateur devient plus punitif sur les écarts pour les petites énergi*
```

### 5.2.c. Modélisation avec le nombre de hits

Un second modèle a été testé en utilisant `n_hits` comme variable explicative.

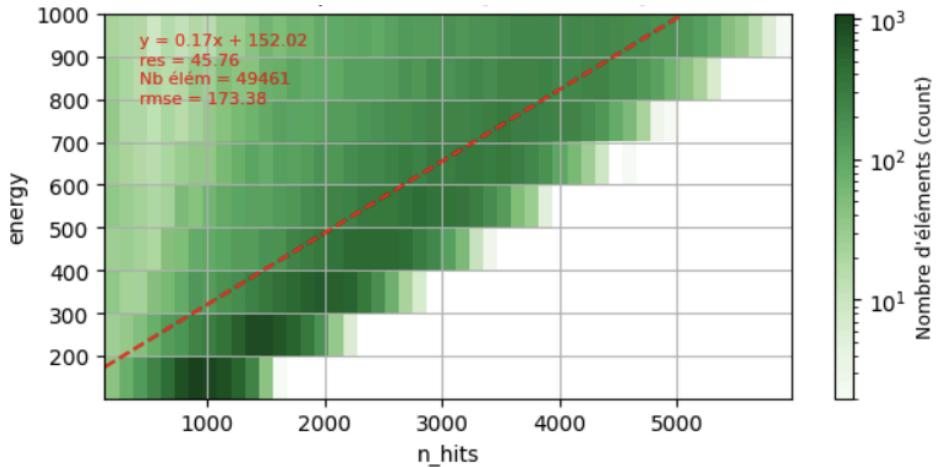


Fig. 19. – Régression linéaire de l'énergie en fonction du nombre de hits.

Les résultats sont moins bons que ceux obtenus avec la charge totale :

- **RMSE** : 173.38 MeV
- **Résolution globale moyenne** : environ 45%

Cependant, la résolution moyenne par bin d'énergie reste intéressante, autour de 23%, ce qui confirme que `n_hits` est une variable pertinente.

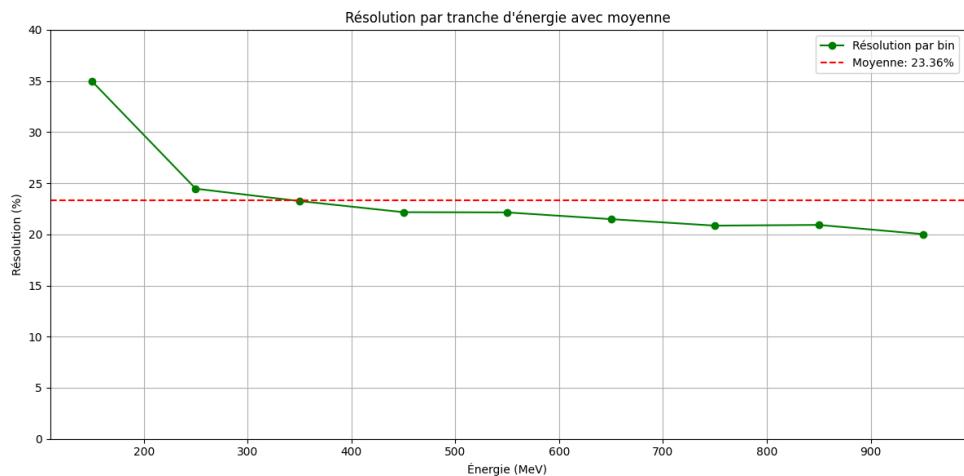


Fig. 20. – Résolution par tranche d'énergie pour le modèle basé sur `n_hits`.

Donc les résultats obtenus à l'aide du nombre de hits sont nettement moins bon que ceux obtenus avec la charge totale, on pouvait s'y attendre car comme on l'a vu leur corrélation respective avec l'énergie suivent la même logique.

#### 5.2.d. Impact de la variable `towall`

Bien que la variable `towall` (distance au mur) soit faiblement corrélée à l'énergie, son influence est indirecte. Elle affecte la quantité de lumière collectée et donc la relation entre `charge_totale/n_hits` et l'énergie. Pour rappel, Hyper-Kamiokande est un cylindre de dimension ... Ici on a affiché la corrélation de deux variables avec l'énergie en fonction de `Towall` (distance au mur selon la direction de la particule).

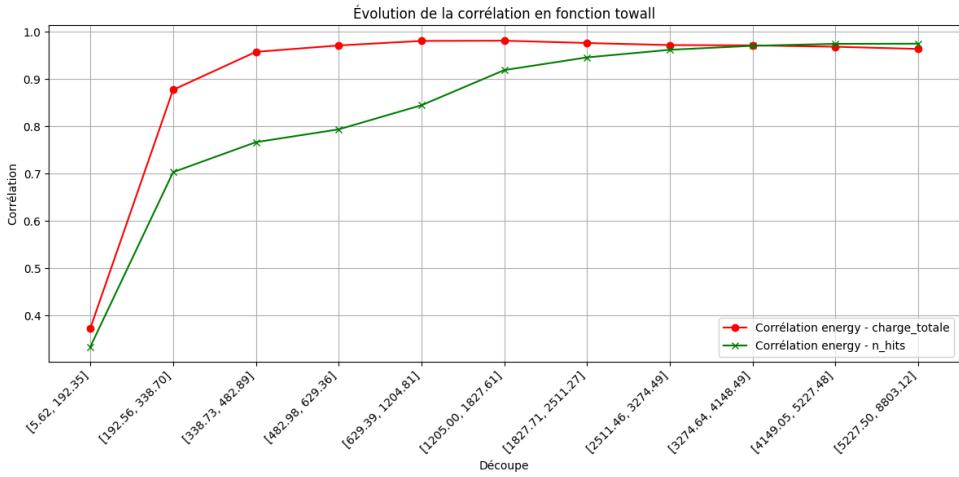


Fig. 21. – Évolution de la corrélation (Énergie-Charge et Énergie-n\_hits) en fonction de towall.

On observe que la corrélation s'améliore nettement pour les événements plus éloignés des parois du détecteur (grand towall). Les graphiques suivants montrent bien cette dépendance.

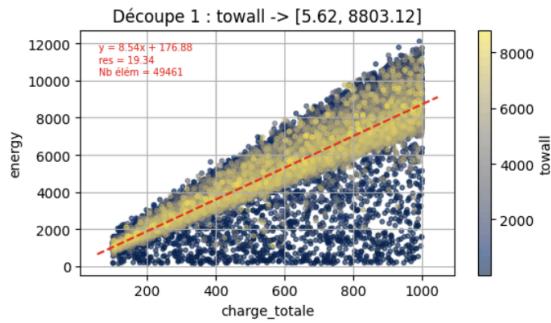


Fig. 22. – Distribution des événements (Énergie vs Charge totale) colorée par towall.

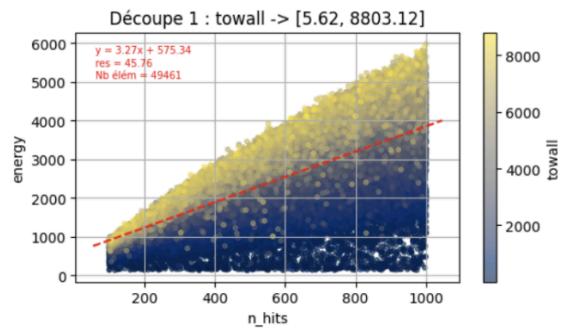


Fig. 23. – Distribution des événements (Énergie vs nombre de hits) colorée par towall.

Ce sont des nuages de points colorés en fonction de leur valeur towall. On constate pour chacun des graphiques que les points avec un towall élevé (points clairs) sont plutôt proche de la droite de régression. Alors que les points ayant un towall faible ne sont pas forcément proche de la droite, leur dispersion autour de la droite semble beaucoup plus élevée.

Notons tout de même que le plot de ces points a été fait avec un `dataframe` trié en fonction de towall. Les points avec un towall élevé apparaissent donc au dessus de ceux avec un towall plus bas. Cela ne change pas l'interprétation que l'on peut en tirer car même si des points avec un towall faible sont proche de la droite, ici on voit surtout qu'il y a peu de points clairs loin de la droite.

En segmentant les données par intervalles de towall et en appliquant une régression linéaire sur chaque segment, on constate que la performance du modèle s'améliore drastiquement pour les towall élevés (résolution d'environ 12%), tandis que pour les towall faibles, le modèle linéaire simple n'est plus adapté.

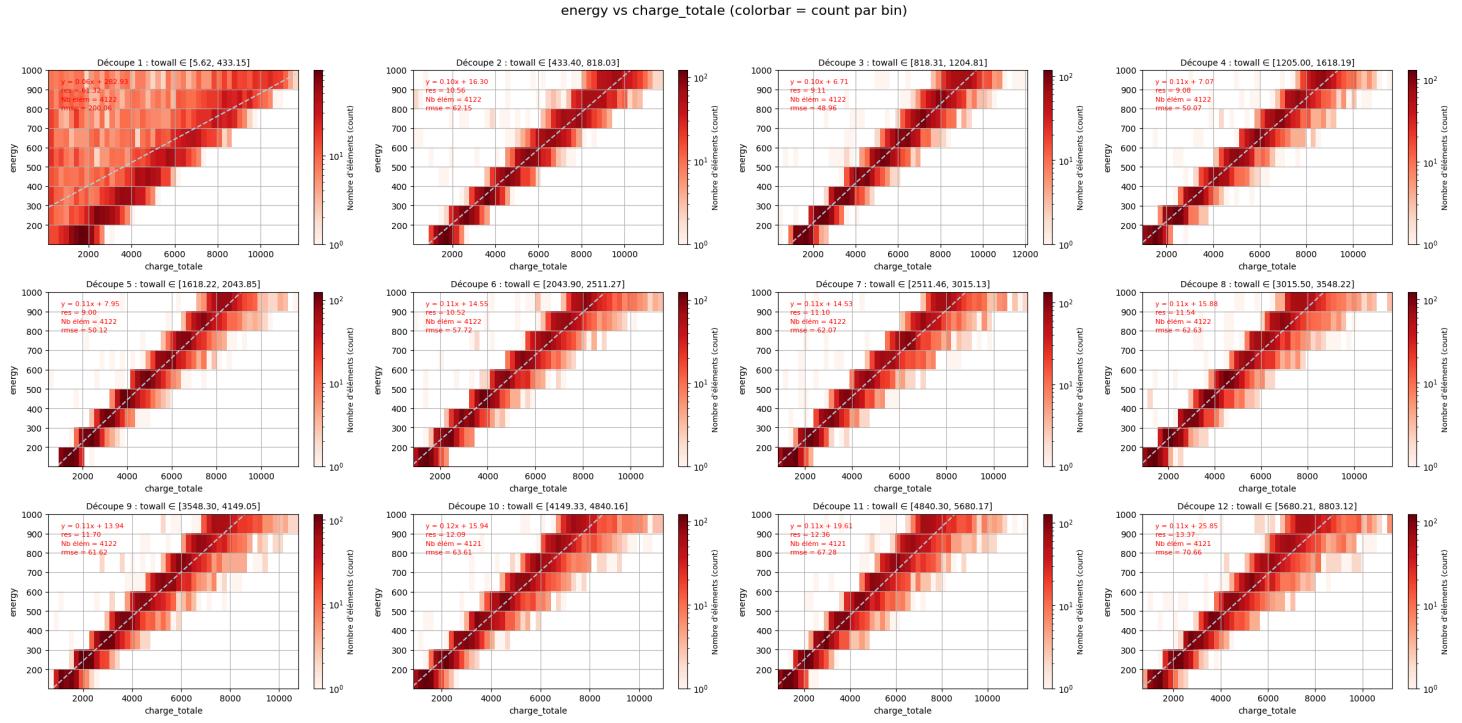


Fig. 24. – Grille de régressions (Énergie vs Charge totale) pour différents intervalles de towall.

Le panel de graphiques ci-dessus se lit de gauche à droite puis de haut en bas. Chaque graphique représente le nombre de points se trouvant dans chaque bin d'énergie ainsi que chaque bin de charge totale pour un intervalle de towall donné dans le titre du graphique.

Sur le 1er graphique : on observe tous les événements ayant un towall compris entre 5.62 et 433 centimètres. On voit que pour environ 4000 événements, un certain nombre d'entre eux sont loin de la droite de régression. Cependant les cases foncées sont tout le temps sur la droite de régression.

Contrairement aux premiers graphiques, les autres tranches de towall semblent suivre un modèle linéaire. On peut préciser que dans les bins d'énergie élevés on observe une plus grande dispersion (pas visible dans la résolution car l'indicateur est moins punitif).

### 5.2.e. Bilan Régression Linéaire Simple

La régression linéaire simple a permis de mettre en évidence la forte corrélation entre l'énergie des événements et certaines variables globales comme charge\_totale et n\_hits. Les résultats montrent que :

- La variable charge\_totale fournit de meilleures performances (RMSE plus faible, résolution globale autour de 19%) que n\_hits.

- Bien que `n_hits` soit aussi pertinent, son pouvoir prédictif reste inférieur, ce qui est cohérent avec l'analyse de corrélation initiale.
- L'introduction de la variable géométrique `towall` permet d'expliquer une partie des écarts observés : la performance du modèle est meilleure pour les événements éloigné du point de sortie du détecteur, tandis que pour ceux proche la relation linéaire se dégrade.

Ainsi, la régression linéaire simple constitue une première étape utile pour comprendre les relations entre variables, mais elle reste limitée pour capturer la complexité des événements, notamment dans les zones où la géométrie du détecteur joue un rôle important. Ces limites motivent naturellement le passage à des modèles de régression multiple qui utilisent plus de données.

## 5.3. Régression Linéaire Multiple

Afin d'améliorer la précision de l'estimation, on entraîne le modèle en intégrant plusieurs informations en même temps. L'objectif est de construire un modèle de régression linéaire multiple, c'est-à-dire une modélisation où l'énergie est prédite à partir de plusieurs variables explicatives combinées de manière optimale.

### 5.3.a. Théorie

La régression linéaire multiple est une extension de la régression linéaire simple, qui permet de modéliser la relation entre une variable dépendante (dans notre cas, l'énergie) et plusieurs variables explicatives (par exemple : `n_hits`, `charge_totale`, `max_charge`, etc.).

L'objectif est d'ajuster un modèle de la forme :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$

où :

- $y$  est la variable à prédire (énergie),
- $x_1, x_2, \dots, x_p$  sont les variables explicatives (observables sur les PMTs),
- $\beta_0$  est l'ordonnée à l'origine (terme constant),
- $\beta_i$  sont les coefficients associés aux variables explicatives,
- $\varepsilon$  représente l'erreur aléatoire (bruit).

On utilise la méthode des moindres carrés ordinaires pour estimer les coefficients  $\beta_i$ , en minimisant la somme des carrés des erreurs entre les valeurs observées et les valeurs prédites par le modèle.

Une hypothèse fondamentale de la régression linéaire multiple est la **linéarité** entre la variable dépendante et les variables explicatives. De plus, on suppose que les erreurs  $\varepsilon$  sont **indépendantes, de variance constante (homoscédasticité)** et **suivent une loi normale**.

Dans notre cas, on fait l'hypothèse que les variables choisies conservent une relation suffisamment linéaire avec l'énergie pour permettre une première estimation grossière mais efficace. On a cependant vu plus tôt avec les modèles simples que cette linéarité s'arrête pour certains des événements avec nos deux variables les plus corrélées (le nombre de hits et la charge totale).

### 5.3.b. Choix des variables explicatives

Pour intégrer plusieurs variables à notre modèle, nous choisissons dans un premier temps toutes les variables auxquelles nous avons accès dans le cas des événements réels, c'est-à-dire toutes les informations liées aux photomultiplicateurs (PMTs).

Cependant, afin d'éviter d'introduire des variables inutiles qui risqueraient de complexifier le modèle sans véritable gain en précision, nous procédon à une sélection itérative des variables en utilisant le critère d'information d'Akaike (AIC).

L'AIC est une mesure qui permet de comparer plusieurs modèles statistiques en prenant en compte à la fois la qualité de l'ajustement et la complexité du modèle. Il pénalise les modèles

utilisant un grand nombre de paramètres afin de prévenir le surapprentissage (overfitting). La formule de l'AIC est la suivante :

$$\text{AIC} = 2k - 2\ln(L)$$

où :

- $k$  est le nombre de paramètres estimés par le modèle,
- $L$  est la vraisemblance du modèle (la probabilité d'observer les données sachant le modèle).

L'objectif est de minimiser l'AIC : un AIC plus faible indique un meilleur compromis entre la fidélité de la prédiction et la simplicité du modèle.

Nous procérons ainsi à une sélection ascendante : en partant d'un modèle vide, nous ajoutons itérativement les variables qui permettent la plus forte diminution de l'AIC à chaque étape, jusqu'à ce qu'aucune amélioration significative ne soit obtenue.

À l'issue de cette démarche, les variables suivantes ont été retenues pour la régression :

- **charge\_totale**
- **n\_hits**
- **max\_charge**
- **min\_charge**

### 5.3.c. Analyse des résidus

Dans cette partie, on fait l'analyse des résidus une fois le modèle entraîné. Les résultats de ce dernier sont dans la partie suivante, car l'analyse des résidus met en évidence les événements qui pourrait poser problème à notre modèle linéaire, de plus elle sert à vérifier les hypothèses de la régression multiple (normalité des erreurs).

#### 5.3.c.i. Théorie

- Les Résidus Studentisés : En régression linéaire, le résidu d'une observation est la différence entre la valeur observée et la valeur prédite par le modèle.

Cependant, ces résidus bruts ne permettent pas toujours de juger correctement si une observation est « anormale » (outlier) car :

Ils ne tiennent pas compte de la variance locale de la prédiction.

Ils sont influencés par les points ayant un fort effet de levier.

Les résidus studentisés (ou « externally studentized residuals ») corrigent ce problème en standardisant les résidus selon leur variance locale :

$$t_i = \frac{e_i}{\hat{\sigma}_i \sqrt{1 - h_{i,i}}}$$

où :

- $e_i$  est le résidu brut de l'observation  $i$ .
- $h_{i,i}$  est l'élément diagonal de la matrice « Hat ».
- $\hat{\sigma}_i$  est l'écart-type des erreurs estimé sans l'observation  $i$ .

- La **Matrice Hat (H)** est donnée par :

$$H = X(X^T X)^{-1} X^T$$

Elle projette les valeurs observées  $y$  sur les valeurs prédictes  $\hat{y}$  :

$$\hat{y} = Hy$$

Les termes diagonaux  $h_{i,i}$ , appelés **levier (leverage)**, quantifient l'influence d'une observation sur sa propre prédiction :

- Un  $h_{i,i}$  élevé signifie que le point est situé loin du centre des données explicatives.
- La valeur moyenne de  $h_{i,i}$  est  $\frac{p+1}{n}$ , avec  $p$  le nombre de variables explicatives.

### 5.3.c.ii. Distribution des résidus

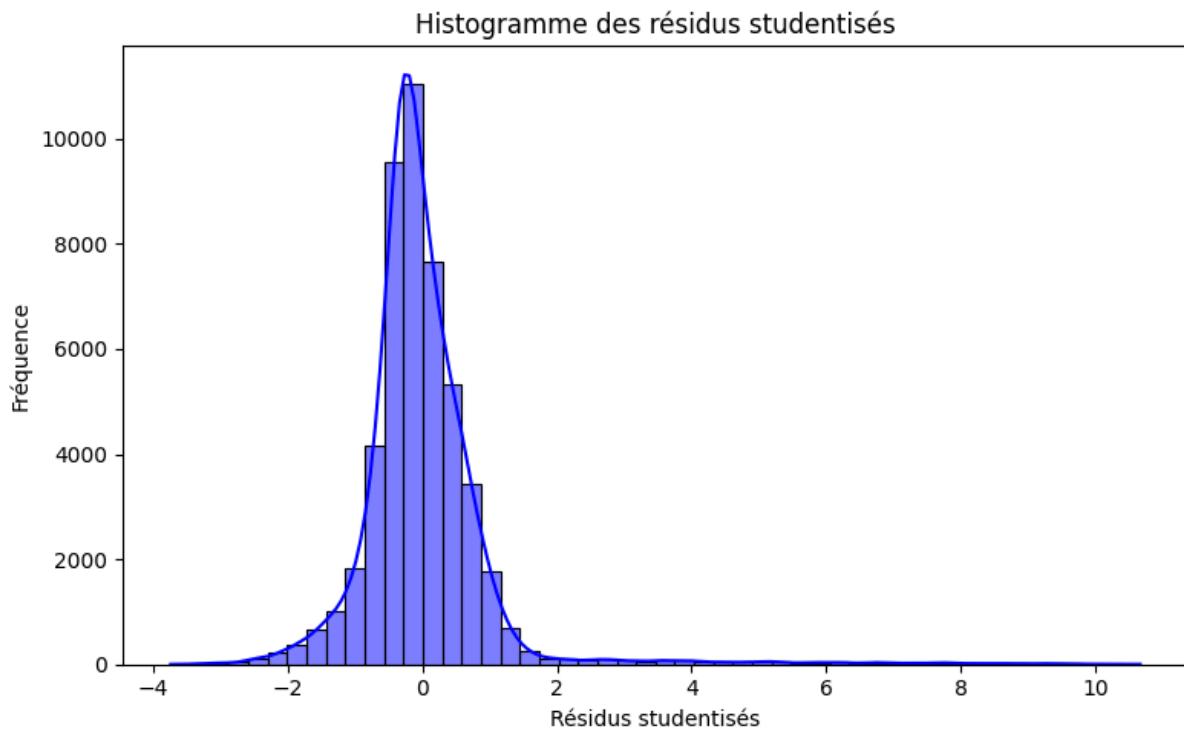


Fig. 25. – Distribution des résidus

On observe que la distribution suit à peu près une loi normale pour les résidus compris entre  $[-2, 2]$ . Cependant, on observe aussi que il y a des événements ayant une erreur studentisée très haute ce qui est anormale par rapport à l'hypothèse de normalité des erreurs.

Ces événements ayant un résidu studentisé très haut sont des événements qui sont différents de ce à quoi on peut s'attendre. Il serait donc intéressant de s'y intéresser.

### 5.3.c.iii. Événements leviers ayant un résidu élevé

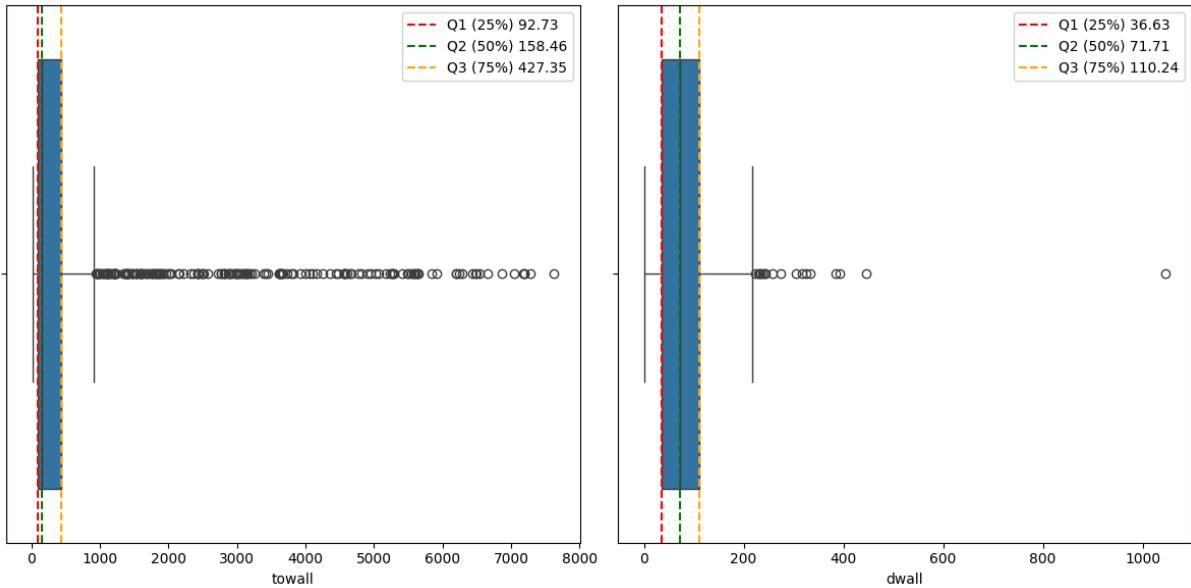


Fig. 26. – Distribution du `towall` des événements outliers et leviers

Comme on l'a vu dans nos différentes régressions linéaires simples, les événements qui posaient problème avaient un `towall` assez bas.

Ces boxplots affiche le `towall` de tout les événements ayant un fort résidus et étant des points leviers. On observe que 75% de ces points ont un `towall` < 427 et un `dwall` < 110. Autrement dit, ce sont ces points qui posent le plus de problèmes lors de l'entraînement de notre modèle. Ces résultats coïncident avec les parties précédentes, on peut alors essayer de visualiser à quoi ressemblent ces événements pour nous éclaircir.

Voici quelques exemples d'événements faisant partie de ces « outliers » (ci-dessous) :

### HK Event Display

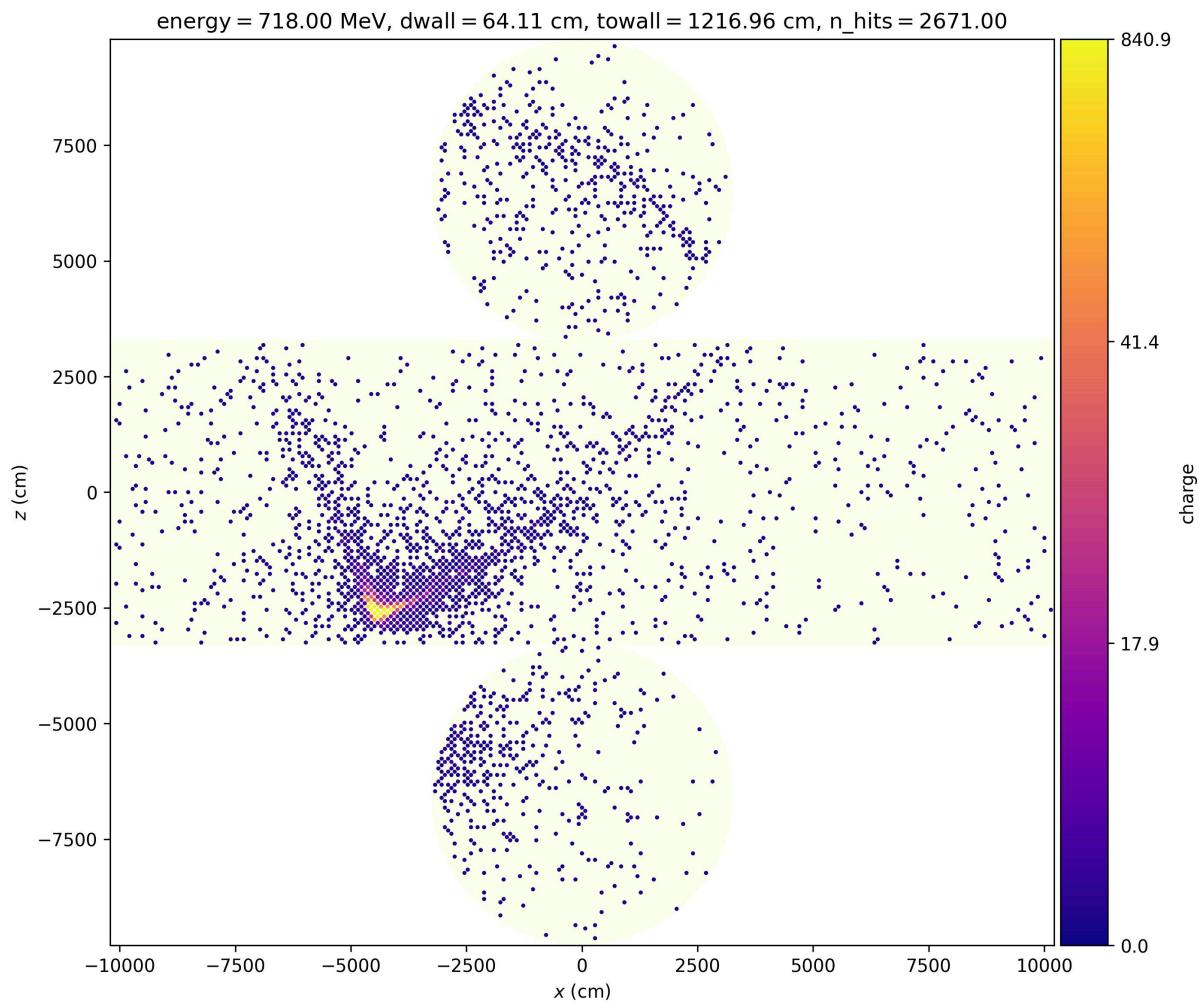


Fig. 27. – Événement n°35\_317

### HK Event Display

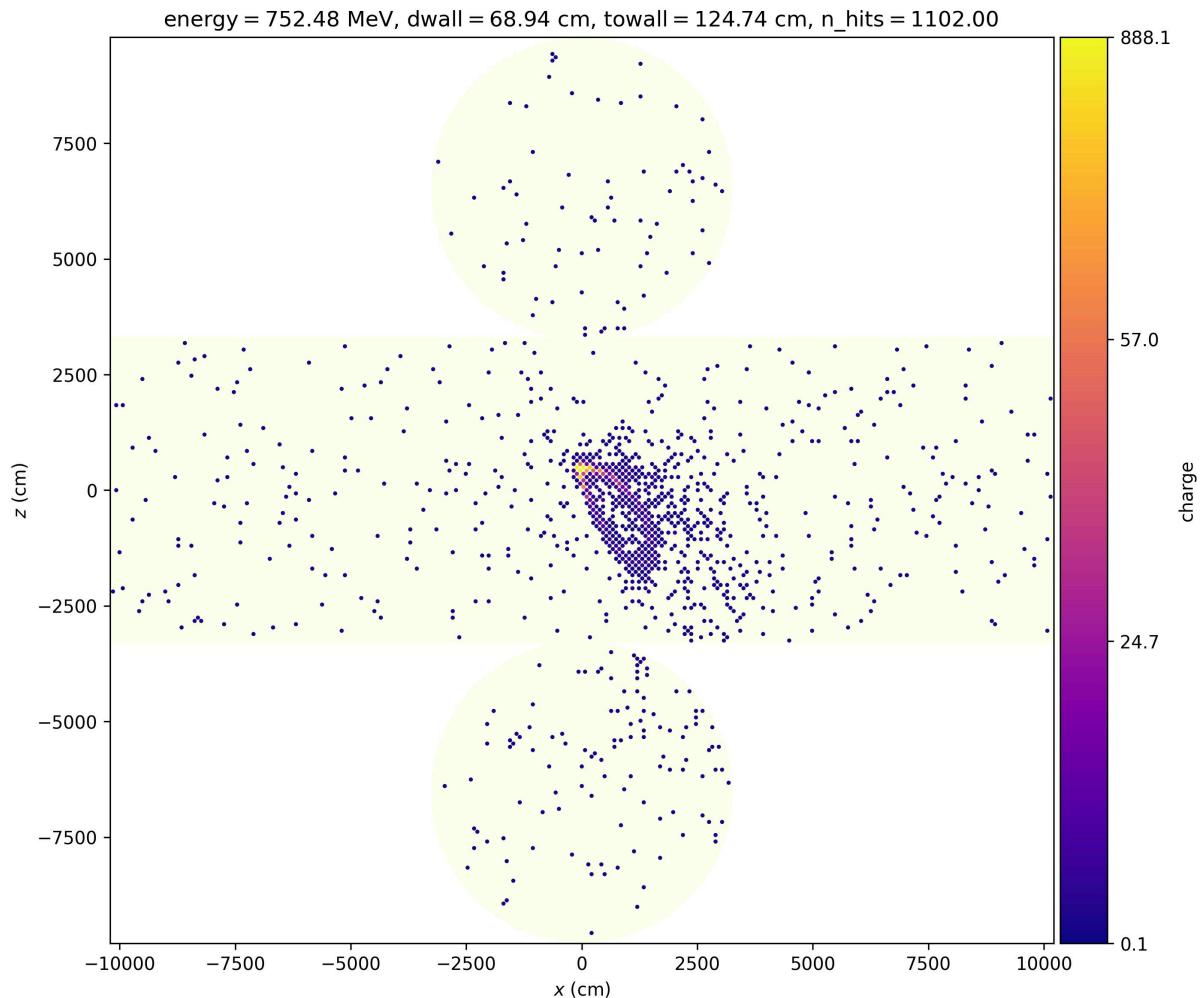


Fig. 28. – Événement n°46\_030

Comme on peut le voir, ces deux événements ont une énergie assez élevée mais une charge totale très faible (seulement quelques PMTs ont mesuré une charge très élevée). Cependant ce n'est que deux exemples, il serait sûrement intéressant d'en regarder plus pour peut être d'une part identifier des similitudes et d'autre part réussir à déterminer quelles autres types données on pourrait calculer et ajouter à notre modèle.

## 5.3.d. Résultats de performance

### 5.3.d.i. Résultats sur l'ensemble des données

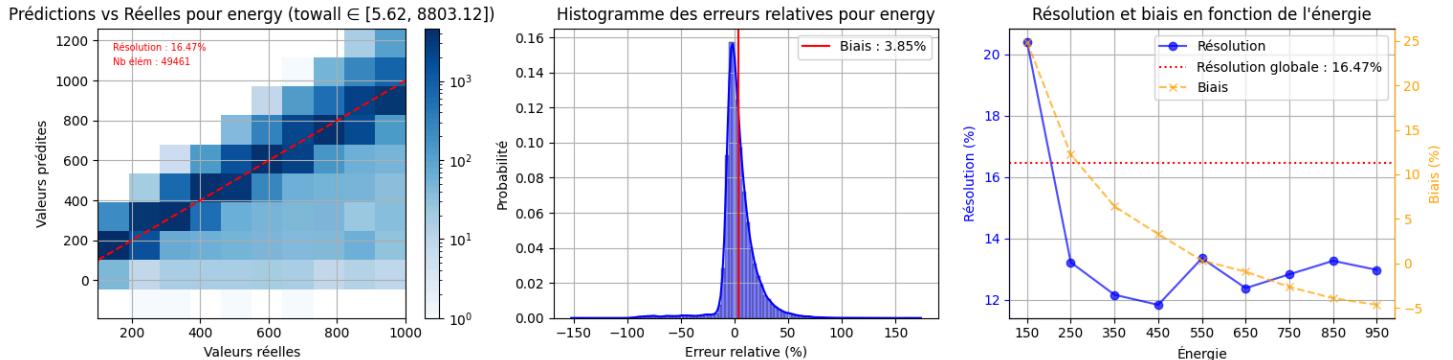


Fig. 29. – Résultats du modèle de régression linéaire sur toutes les données

Comme on peut le constater le modèle de régression linéaire multiple a une résolution globale meilleure que les modèles simple : **16.47 %** de résolution globale contre **19 %** pour le modèle simple.

On observe encore une fois des performances moins bonnes sur les événements à basse énergie qui sont encore une fois liées à la sensibilité de notre indicateur (la résolution).

De plus, on observe que le modèle a plus tendance à sous estimer de manière générale l'énergie réelle de l'événement, surtout sur les plus grandes bins d'énergie. En effet, on observe la présence d'événements pour lequel le modèle sous estime très fortement leur énergie.

### 5.3.d.ii. Résultats sur les événements ayant un $\text{towall} > 500$

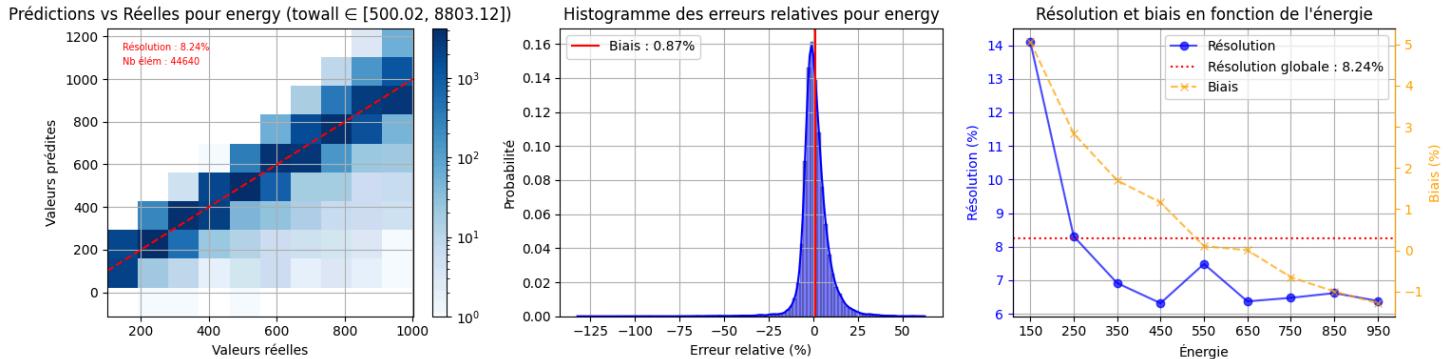


Fig. 30. – Résultats du modèle de régression linéaire sur les événements ayant un  $\text{towall} > 500$

Comme on l'avait vu dans l'analyse des résidus les événements ayant un  $\text{towall} < 500$  sont les principaux qui ne suivent pas le modèle linéaire. Supposons qu'on arrive à déterminer si un événement a un  $\text{towall}$  plus grand que 500, on pourrait alors obtenir de bien meilleurs résultats sur ces événements.

On obtient une résolution globale deux fois plus petite : **8.24 %**

### 5.3.d.iii. Résultats sur une segmentation de towall

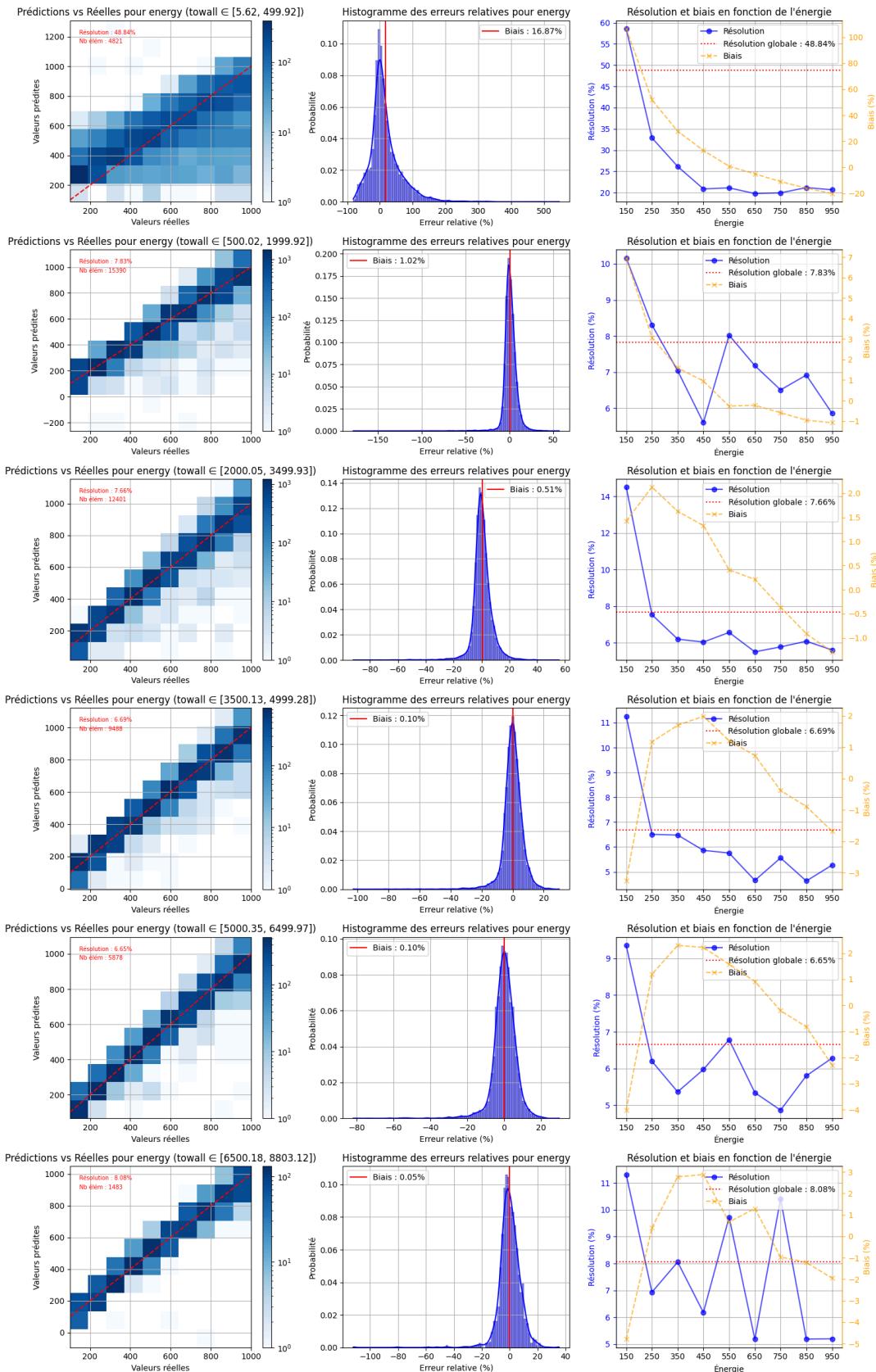


Fig. 31. – Résultats du modèle de régression linéaire sur les événements segmentés **towall**.

Comment lire ce panneau de graphiques : chaque ligne représente un certaine tranche de towall, il y a 3 graphiques : le 1er est la comparaison des prediction du modèle fit sur la tranche de towall de la ligne avec les énergies réelles, le 2eme est la distribution des erreurs relatives et le dernier est la résolution par bin d'énergie.

Comme on peut le voir sur ce grand panneau de graphiques, on obtient avec un modèle de régression linéaire multiple des résolutions globales aux alentours de 7% pour les tranches de towall supérieures à 500. De plus, pour les événements à basse énergie (- de 200 MeV) on a une résolution à environ 10%.

## 5.4. Modèle cyclique (Pipeline)

### 5.4.a. Corrélation entre nhits et towall

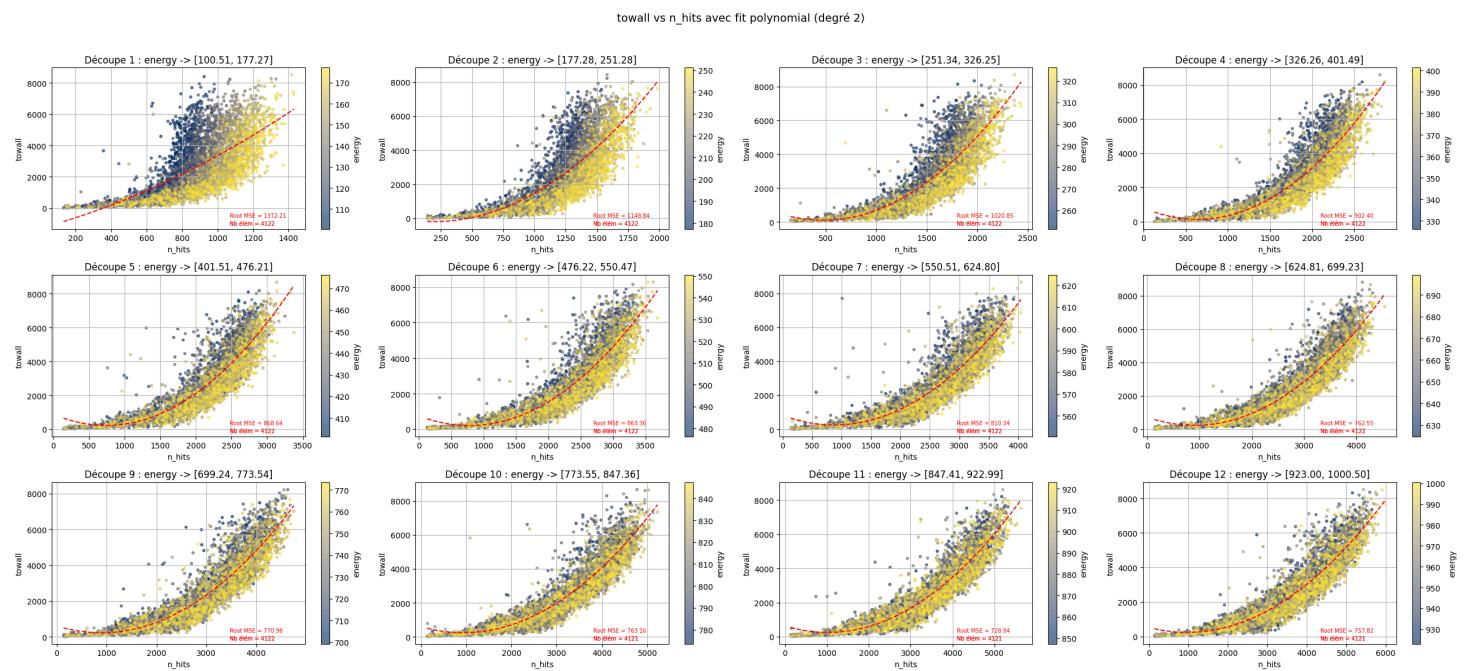


Fig. 32. – Résultats du modèle de régression linéaire sur toutes les données

On peut observer sur ces graphiques que le nombre de hits pourrait aider à prédire le towall à environ 1000cm près pour les énergies plus haute que 200 MeV (à l'aide d'une première prédition de l'énergie).

### 5.4.b. Architecture du pipeline

L'objectif de cette approche est d'améliorer la prédition de l'énergie des événements en prenant en compte la variable towall. Cette variable est importante car la topologie d'un événement (proche ou loin des parois) influence la quantité de lumière détectée et la manière dont elle est répartie (hypothèse).

Cependant, towall n'est pas directement disponible pour les événements réels. L'idée est donc de la prédire dans un premier temps à l'aide d'un modèle intermédiaire, basé sur une estimation initiale de l'énergie.

- Étape 1 : Prédiction initiale de l'énergie

Dans un premier temps, on effectue une prédiction de l'énergie en utilisant un modèle de régression multiple basé sur les variables globales disponibles dès la reconstruction : `charge_totale`, `n_hits`, `max_charge`, `min_charge`. Cette prédiction initiale n'est pas parfaite. Je n'ai pas eu le temps de vérifier à quelle point l'erreur de prédiction influencera la prédiction du `towall`.

- Étape 2 : Prédiction de la tranche de `towall`

À partir de l'énergie prédite, on cherche ensuite à estimer la valeur de `towall`. Pour cela, on segmente les données en tranches d'énergie (bins). Sur chacune de ces tranches, on ajuste un modèle de régression polynomiale reliant le nombre de hits détectés (`n_hits`) à la distance au mur (`towall`). Cette approche permet d'exploiter la corrélation locale entre `n_hits` et `towall`, qui dépend fortement de l'énergie (d'après les observations).

- Étape 3 : Prédiction finale de l'énergie conditionnelle à `towall`

Une fois la tranche de `towall` prédite, on utilise un modèle final de régression linéaire multiple, spécifique à cette tranche de `towall`, pour affiner la prédiction de l'énergie.

Ce pipeline peut être résumé de la manière suivante :

1. Prédiction brute de l'énergie à partir des variables PMTs.
2. Prédiction de la tranche de `towall` à l'aide de cette énergie.
3. Prédiction finale de l'énergie en conditionnant sur la tranche de `towall`.

### 5.4.c. Résultats

J'ai évalué les performances du pipeline à l'aide d'une validation croisée. Les résolutions obtenues pour la prédiction de l'énergie sont de  **$16.44 \pm 0.06$  %** pour la première estimation (modèle de régression multiple) et de  **$20.66 \pm 0.71$  %** pour la seconde estimation après la correction via `towall`. On obtient donc de moins bons résultats avec ce modèle. Concernant la prédiction de `towall`, la RMSE est de  **$1085.47 \pm 10.76$**  ce qui paraît correct.

Pour analyser plus finement les performances, j'ai également étudié la résolution en énergie en fonction des tranches d'énergie (bins). On observe que le modèle offre de meilleures résolutions pour les basses énergies (inférieures à 200 MeV). Cependant, un pic de dégradation des performances apparaît pour les énergies comprises entre 300 MeV et 600 MeV. Je n'ai pas eu le temps de chercher à expliquer ce « bump » de la résolution dans cette zone.

Le graphique ci-dessous présente la résolution en fonction de l'énergie, mettant en évidence cette variation selon les tranches :

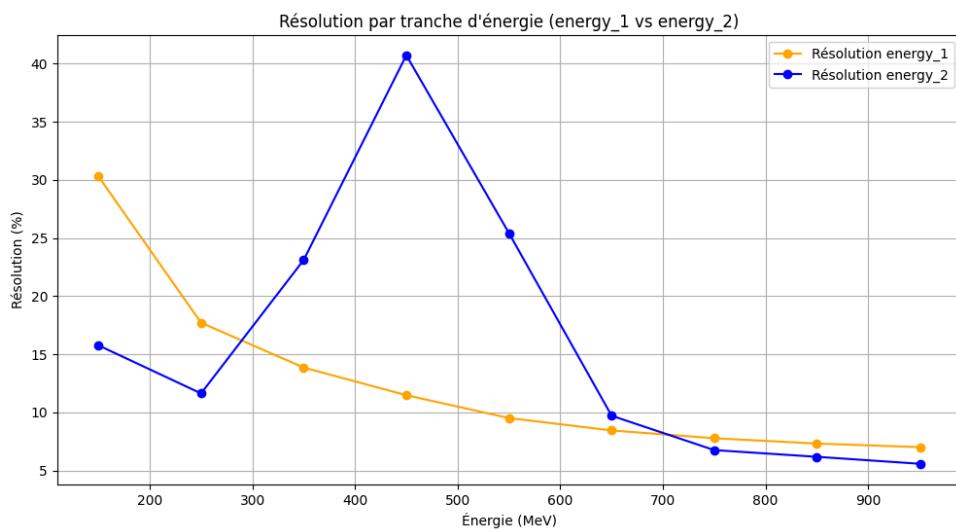


Fig. 33. – Évolution de la résolution en énergie en fonction des tranches d'énergie.

## 5.5. Résumé de résultats par modèles

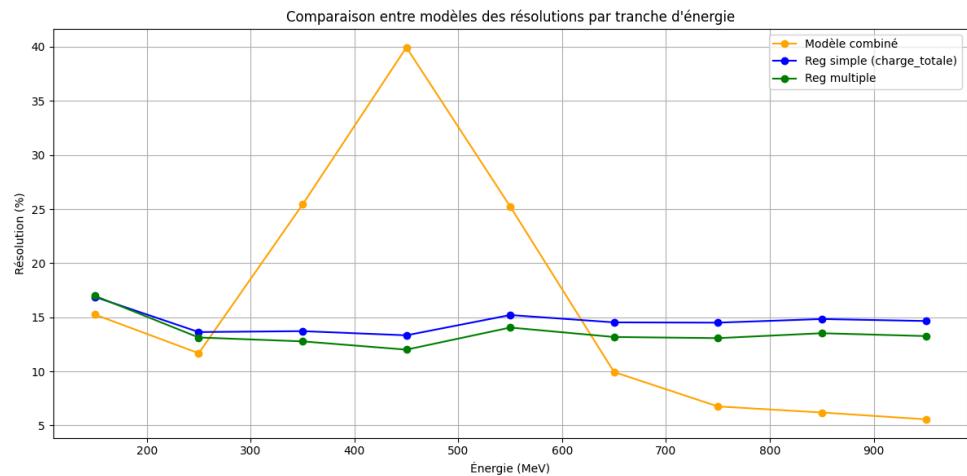


Fig. 34. – Évolution de la résolution en énergie en fonction des tranches d'énergie.

## 5.6. Régression Ridge

La régression Ridge est une extension de la régression linéaire classique, utilisée pour lutter contre les problèmes de multicolinéarité et de sur-apprentissage. Elle ajoute une pénalité sur la norme des coefficients pour réduire leur variance.

L'estimateur Ridge est défini par :

$$\hat{\beta}_{\text{Ridge}} = \arg \min_{\beta} \left( \sum_{i=1}^n (y_i - X_i \beta)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right)$$

où :

- $\lambda = 0$  est le paramètre de régularisation.
- Plus  $\lambda$  est grand, plus les coefficients  $\beta_j$  sont « rétrécis » vers zéro.

Contrairement à la régression classique, Ridge empêche les coefficients de devenir trop grands, surtout en présence de variables explicatives fortement corrélées.

Dans notre cas, utiliser Ridge aurait permis **d'atténuer l'effet des points à fort effet levier**, en les empêchant d'avoir une influence démesurée sur les coefficients.

## 5.7. Correction des charges captées

Dans cette partie, on cherche à augmenter la précision de notre modèle en corrigéant la charge captée par nos PMTs. Ici, on se concentre plus sur la qualité de nos données que la quantité.

### 5.7.a. Théorie sur l'absorption et dépendance angulaire

#### 5.7.a.i. Absorption des photons

Les photons traversent une certaine distance ( $d$ ) du point d'interaction (vertex) jusqu'au PMT avant d'être captés. Or, selon la loi de Beer-Lambert, la luminance  $L$  qui atteint le PMT n'est pas la même que la luminance initiale  $L_0$  émise :

$$L(d) = L_0 e^{-d\alpha}$$

Où  $\alpha$  est le coefficient d'absorption du milieu et  $d$  la distance parcourue.

J'ai considéré que la charge captée par le PMT est une même fonction ( $f$ ) pour chacun et donc que l'on peut remplacer la luminance ( $L_d$ ) par la charge captée ( $f(L_d)$ ) :

$$\text{Charge captée} = \text{Charge initiale} \times e^{-d\alpha}$$

On cherche donc ici à corriger la charge captée par chaque PMT pour chaque événement, ce qui représente environ 117 millions de points de données à traiter.

Cependant, cette formule dépend du milieu que traversent les photons. Il est donc nécessaire de connaître le coefficient d'absorption  $\alpha$  de l'eau ultrapure dans le détecteur pour appliquer cette correction.

### 5.7.a.ii. Dépendance angulaire

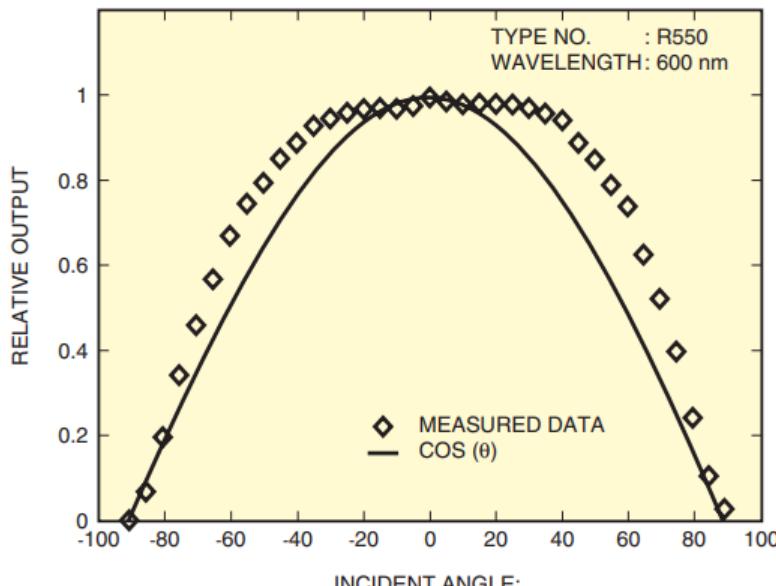


Fig. 35. – Dépendance angulaire de la réponse du capteur : comparaison avec la loi  $\cos(\theta)$

Pour corriger précisément la charge mesurée par les photomultiplicateurs (PMT), il est également nécessaire de prendre en compte la dépendance angulaire. En effet, la sensibilité des PMT varie en fonction de l’angle d’incidence de la lumière sur leur surface.

Cette dépendance angulaire est due au fait que la surface efficace du photocathode diminue lorsque l’angle d’incidence augmente (ou baisse, angle allant de  $-90$  à  $90$ ), ce qui conduit à une réduction apparente du signal collecté.

Cette correction angulaire doit être prise en compte dans le traitement des données afin d’obtenir une estimation précise de la charge initiale déposée par les photons, indépendamment de l’angle sous lequel ils atteignent les différents PMT.

Cependant sur ce graphique c’est pour une longueur d’onde de  $600\text{nm}$ , de plus l’expérience montre que cette dépendance angulaire ne suit pas totalement une loi cosinus.

### 5.7.b. Méthodes

#### 5.7.b.i. Coefficient d’absorption de photons

Pour trouver le coefficient d’absorption des photons dans l’eau ultra pure du détecteur j’ai effectué un fit d’une fonction exponentielle décroissante ( $e^{\frac{d}{\alpha}}$ ). On trouve le coefficient de la loi de Beer-Lambert  $\frac{1}{\alpha}$

Pour fit cette fonction, j’ai exprimé la charge capté par le PMT en fonction de sa distance euclidienne au vertex ( $\|\text{coordPMT} - \text{coordVertex}\|^2$ )

Cependant, j’ai encore un doute sur ma manière de procéder, en effet j’ai utilisé les données de 5 000 événements **sans distinction d’énergie**. Peut être serait-ce pertinent de regarder comment se comporte ces charges en fonction de la distance en ajoutant ce paramètre.

On trouve comme résultat un lambda très bas par rapport à ce qu’on pouvait attendre en regardant les points à haute charge :

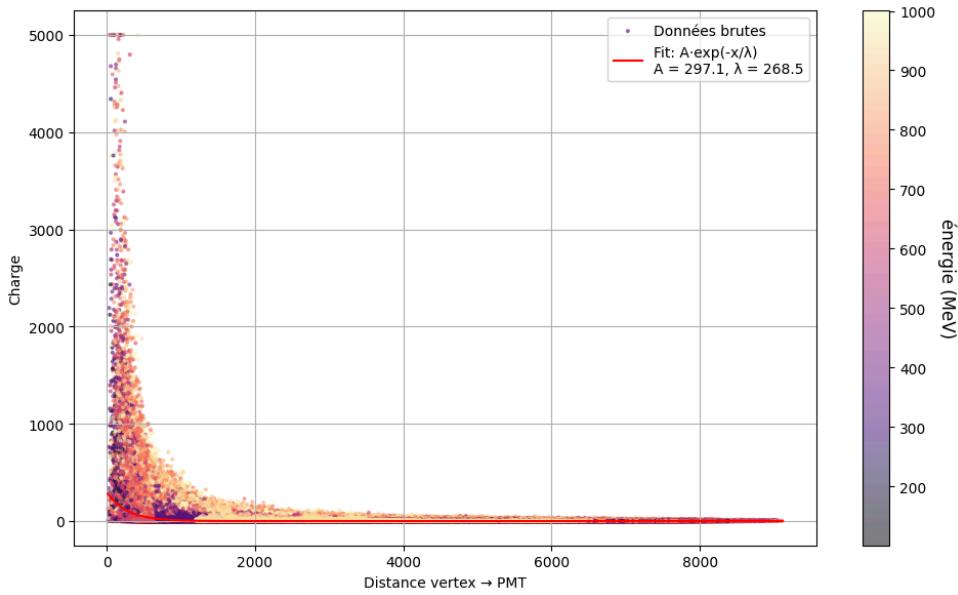


Fig. 36. – Courbe Exponentielle fit sur 5 000 événements toutes énergie.

J'ai aussi essayé de fit la courbe sur des événements tout seul et j'obtenais des  $\alpha$  beaucoup plus élevés ( $\sim 1200$ )

### 5.7.b.ii. Correction de la charge en fonction de la distance au vertex

Une fois après avoir obtenu le coefficient  $\alpha$  on peut alors corriger la charge de chaque PMT pour chaque événement. On obtient alors des charges multipliées par un facteur compris entre  $[1 - e^{-\frac{8000}{\alpha}}]$ , ce facteur grandissant quand la distance au vertex grandit.

C'est à dire pour  $\alpha = 268.5$  :  $[1 - 8.70e+12]$  dépendamment de la distance au vertex

Et  $\alpha = 1200$  :  $[1 - 785]$  dépendamment de la distance au vertex

Quelques exemples avec  $\alpha = 268.5$ :

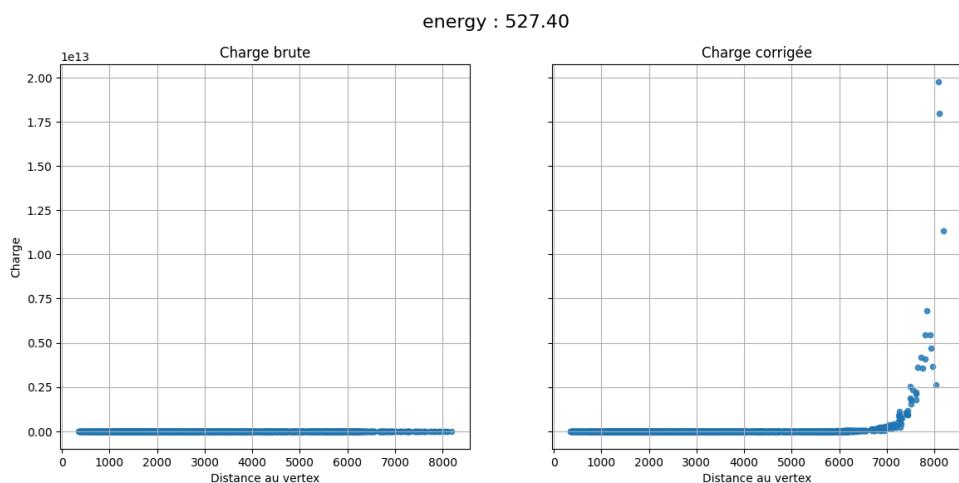


Fig. 37. – Comparaison des charges aux charges corrigées, event n°10\_000

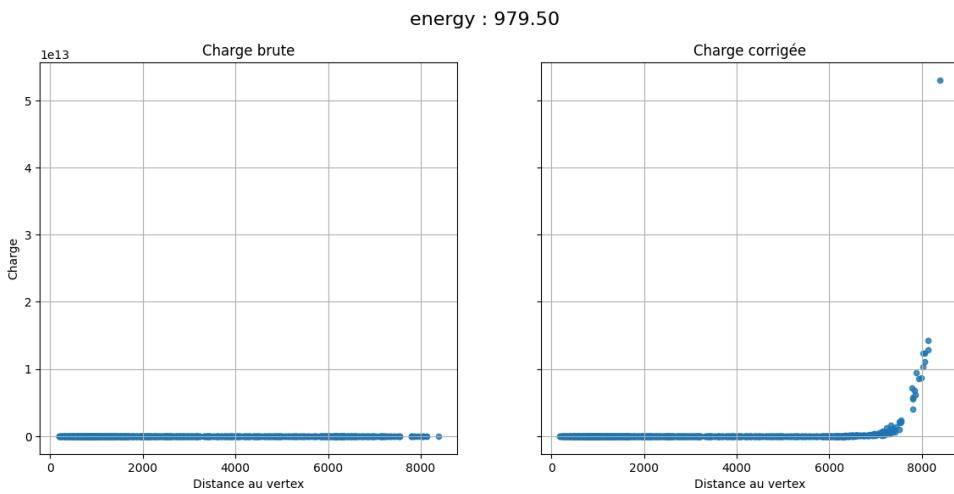


Fig. 38. – Comparaison des charges aux charges corrigées, event n°22\_869

Quelques exemples avec  $\alpha = 1200$ :

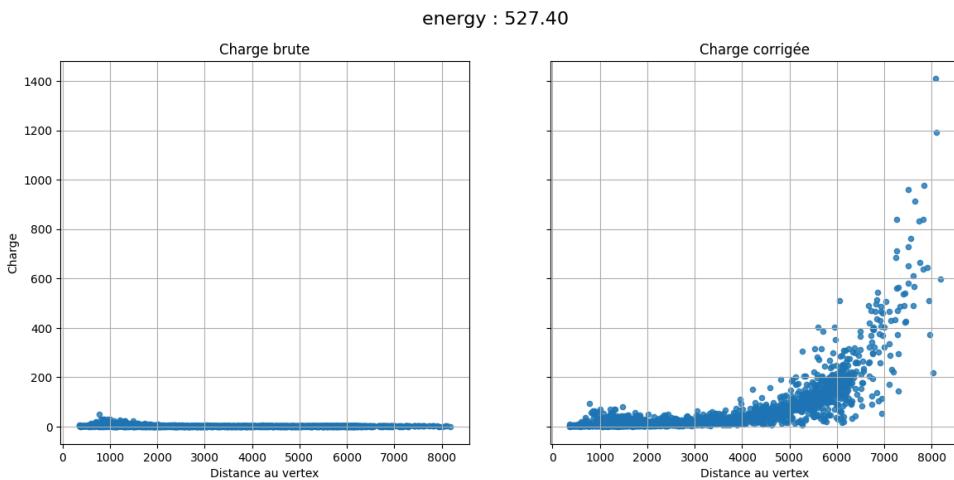


Fig. 39. – Comparaison des charges aux charges corrigées, event n°10\_000

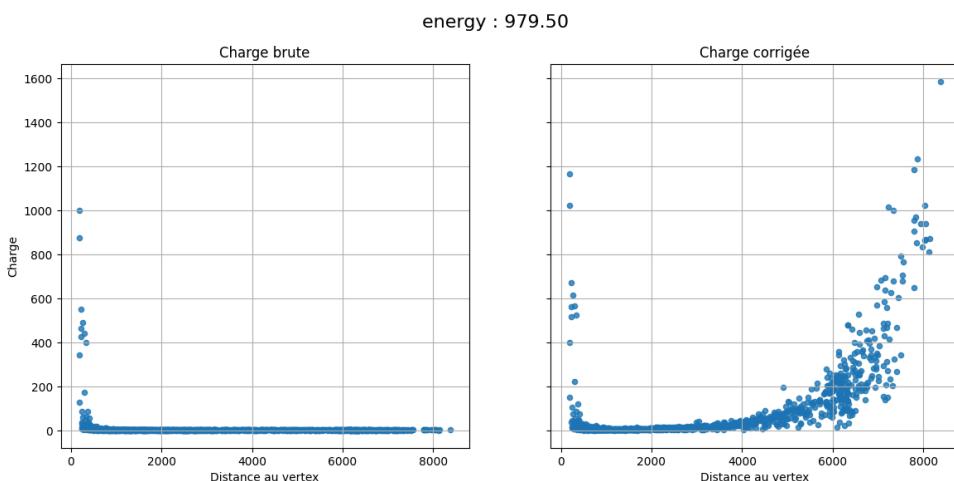


Fig. 40. – Comparaison des charges aux charges corrigées, event n°22\_869

Comme on peut le voir avec ces différents événements, la charge corrigée qu'on obtient n'est pas celle attendue avec le  $\alpha = 268.5$ . Cependant avec le  $\alpha = 1200$  on obtient des charges dans les même échelle ce qui est déjà mieux.

Cependant, il faudrait essayer de faire un fit par bin d'énergie pour voir ce que ça donne.

### 5.7.b.iii. Calcul de l'angle Incident

Tout d'abord j'ai calculé l'angle d'incidence pour chaque PMT. J'ai procédé de la manière suivante :

- On a le vecteur ParticleDir qui donne la direction de la particle
- Si notre PMT est sur un disque inférieur ou supérieur du détecteur, on prend alors comme vecteur normal  $(0, 0, z)$  avec  $z$  la coordonnée de hauteur du PMT. On a alors le vecteur normal de ce plan :

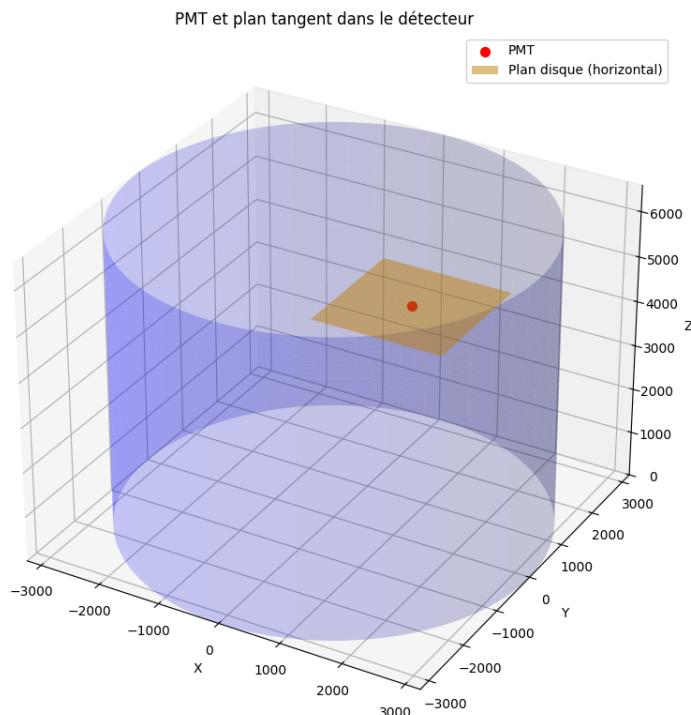


Fig. 41. – plan pris pour calculer l'angle incident lorsque le PMT est sur un disque du cylindre

Sinon notre PMT se situe sur coté du cylindre, on prend alors comme vecteur normal  $(x, y, 0)$  avec  $x$  et  $y$  les coordonnée du PMT. On a alors le vecteur normal de ce plan :

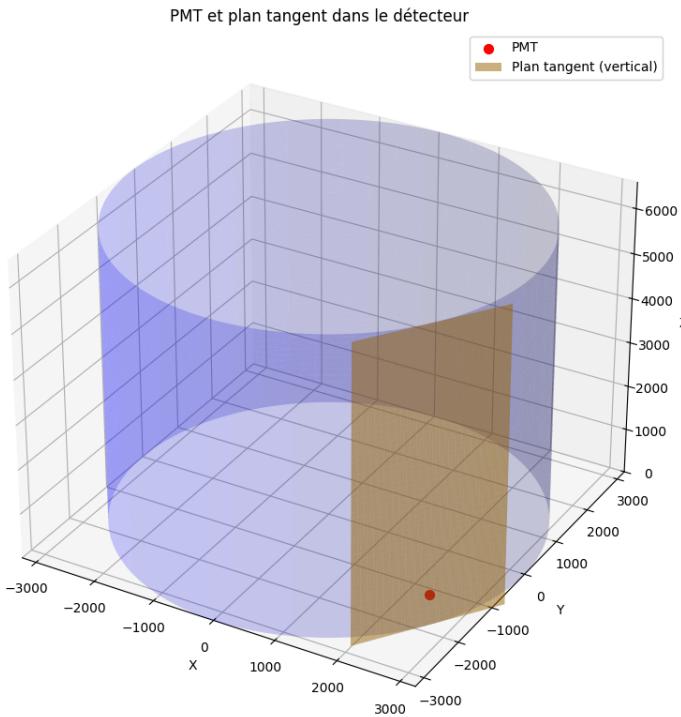


Fig. 42. – plan pris pour calculer l’angle incident lorsque le PMT est sur le coté

- Enfin avec ces deux vecteurs, On peut calculer le cosinus de l’angle qu’il forme avec cette formule :

$$(u) \cdot (v) = |u| |v| \cos(\theta)$$

- On en déduit l’angle

#### 5.7.b.iv. Correction de la charge en fonction de l’angle

Comme on la vu en partie 5.7.a.ii, la correction de notre charge en fonction de l’angle d’incidence a été effectué à l’aide d’un fonction cosinus :

$$\bullet \text{charge\_corr} = \frac{\text{charge}}{\cos(\theta)}$$

Cependant pour éviter que la charge corrigée s’emballe quand on a un angle vers les  $90^\circ$ , j’ai limiter l’augmentation maximum à 400%. Sans cette limitation j’obtenais des charge totale corrigée qui était aberrantes par rapport aux autres.

#### 5.7.c. Résultats

Avec ces corrections possible on obtient 3 nouvelles variables :

- la charge totale corrigée : on a effectué la correction de charge en fonction de l’angle et en fonction de l’absorption
- la charge totale corrigée en fonction de l’absorption
- la charge totale corrigée en fonction de l’angle

Avec la sélection par le critère AIC j’ai vérifier qu’elle apportait de l’information suffisamment conséquente par rapport à la complexité qu’elles apportaient.

J’obtiens alors les résultats suivant :

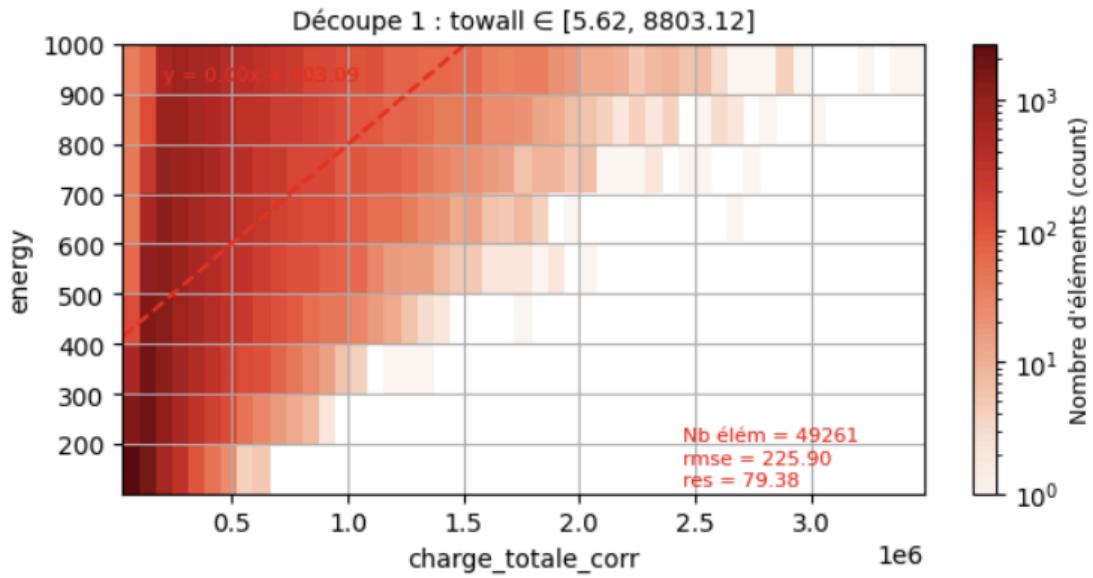


Fig. 43. – Régression linéaire simple en fonction de la charge totale corrigée

On constate que la charge totale corrigée a un pouvoir d'explication beaucoup moins précis que la charge totale d'origine. On observe que certaine charge totale s'envole mais pas toutes, ce qui augmente la dispersion des points de la charge totale pour une même énergie.

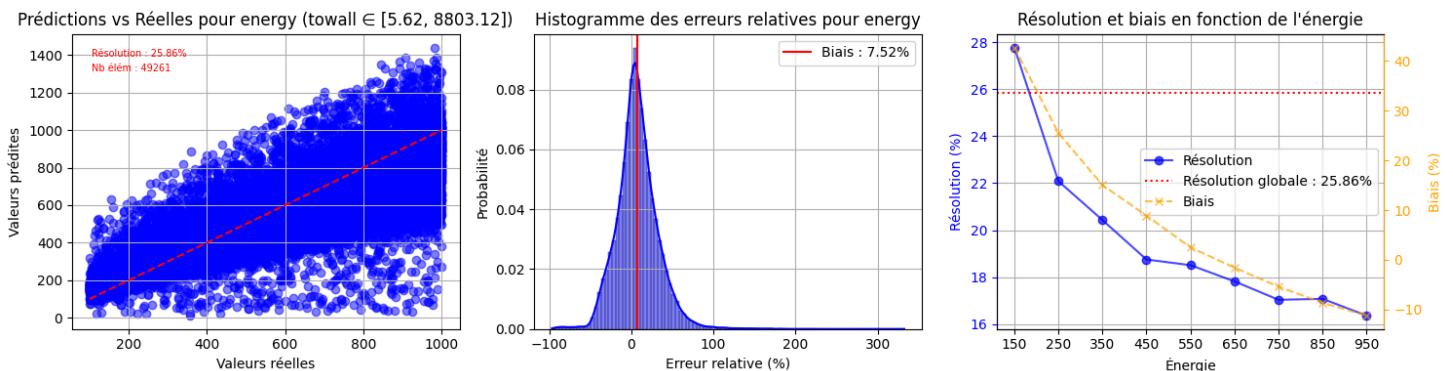


Fig. 44. – Régression linéaire multiple avec ces trois nouvelles variables (sans la charge totale non corrigée)

Ensuite, en ne gardant que les charges totales corrigées, notre modèle perd aussi en précision comparé au modèle utilisant la charge totale d'origine.

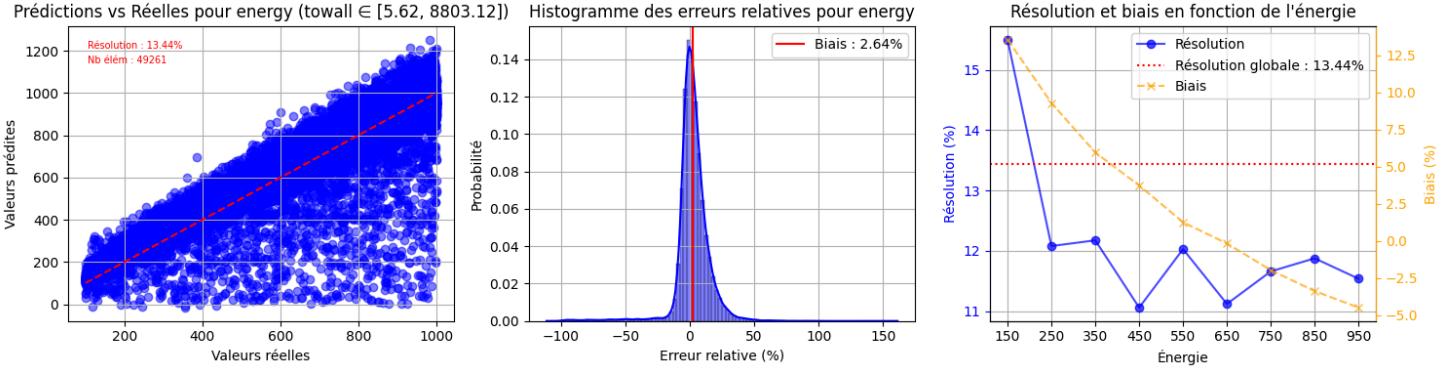


Fig. 45. – Régression linéaire multiple avec ces trois nouvelles variables (avec la charge totale non corrigée)

Enfin, le modèle prenant en variables les trois nouvelles charges corrigées ainsi que la charge totale d'origine possède un pouvoir d'explication plus haut. Cependant c'est dans la logique de rajouter des variables supplémentaires, en effet une variable en plus ne peut que rajouter de la précision ou ne rien faire (mais pas en enlever).

Comparé au modèle n'utilisant que la charge totale d'origine on ne gagne qu'environ 1 à 2% de précision par bin d'énergie, ce qui est peu par rapport au temps de calcul de cette charge corrigée (117 millions de hits à corriger)

## 5.8. Aspects non aboutis

Dans tout ce rapport certains aspects restent encore non aboutis.

Tout d'abord, l'analyses des résidus et des points leviers pourrait être approfondit pour déterminer quels paramètres d'un événement le modèle n'explique pas.

De plus, dans ce rapport j'ai seulement utilisé des modèles de régression(s) linéaire(s) simple et multiple. On pourrait alors imaginer utiliser d'autre types de modèle : régression Ridge, une forêt d'arbres aléatoires, etc...

Aussi, on pourrait réfléchir à ajouter des nouvelles variables explicatives pour décrire par exemple la distribution des charges, la dispersion des hits. On pourrait aussi exploiter la variable `time` qui décrit les timings des charges captées par les PMTs.

Enfin pour l'amélioration de la qualité des données (passant par la correction des charges captées), je pense que la méthode peut être améliorée notamment pour la sélection du coefficient d'absorption des photons.

## 6. Conclusion

Ce stage au Laboratoire Leprince-Ringuet a permis de contribuer à l'amélioration de la reconstruction des événements dans le détecteur Hyper-Kamiokande, en travaillant sur le traitement des données simulées sur le Hyper-Kamiokande et en développant des modèles de régression pour l'estimation de l'énergie des événements.

Les principales contributions incluent la modification de l'algorithme RootToGraph pour une gestion plus efficace des données, ainsi que l'exploration et l'analyse de différents modèles de régression pour estimer l'énergie des événements. Les résultats ont montré que la charge totale et le nombre de hits sont des variables clés pour cette estimation, avec des performances variables selon la distance au mur (towall) et l'énergie de l'événement.

Bien que des améliorations aient été apportées, certains aspects restent à approfondir, notamment l'analyse des résidus et des points leviers, ainsi que l'exploration d'autres modèles de régression et la correction des charges captées.

Je tiens à exprimer ma profonde gratitude à mon maître de stage, M. Erwan Le Blevec, pour son encadrement précieux. Je remercie également toute l'équipe travaillant sur les neutrinos pour leur accueil et leur disponibilité à répondre à mes questions !

Ce stage a été une expérience enrichissante, tant sur le plan professionnel que personnel, et a renforcé mon intérêt pour la recherche. L'expérience que j'ai acquises ici sera sans aucun doute un atout précieux pour la suite de mon parcours académique et professionnel.