Rapport projet : apprentissage sous contraintes physiques

Kssim Aymane, Song Mickaël

ENSEEIHT & INSA Toulouse - 5ModIA
2023-2024

1 Introduction

Ce rapport présente notre projet de prédiction de l'énergie moléculaire utilisant le jeu de données QM7-X. L'objectif principal est de modéliser la surface d'énergie potentielle interatomique de petites molécules organiques, en se basant sur les positions des atomes en 3D et des informations supplémentaires sur ces atomes c'est-à-dire prédire l'énergie d'atomisation d'une molécule donnée.

Notre objectif est de produire un fichier CSV contenant les prédictions d'énergie pour le jeu de données de test comprenant 1155 molécules.

Le principal défi de ce projet est de respecter la contrainte d'invariance, impliquant les opérations de translation, rotation et permutation des atomes. Nous allons appliquer une méthode se basant sur les matrices de coulomb puis une méthode de scattering 3D, parmi d'autres approches, pour résoudre ce problème. La précision de notre modèle sera évaluée en utilisant l'erreur quadratique moyenne (MSE).

2 Pré-traitement des données

2.1 Visualisation des molécules

Nous avons utilisé la librairie MolGraph afin d'effectuer la lecture et la visualisation des molécules stockées dans des fichiers .xyz

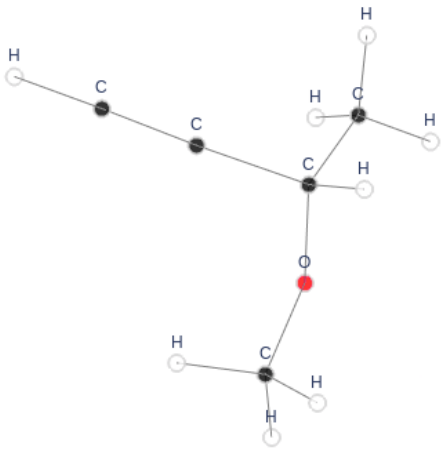

FIGURE 1 – Visualisation d'une molécule du jeu de données de train avec MolGraph

2.2 Visualisation du dataset d'entraînement

Le jeu de données d'entraînement comprend 4739 structures de molécules avec un nombre variable d'atomes. En étudiant notre jeu de données nous voyons que les molécules des jeux de données sont composées d'un maximum de 23 atomes, il s'agit donc de petites molécules.

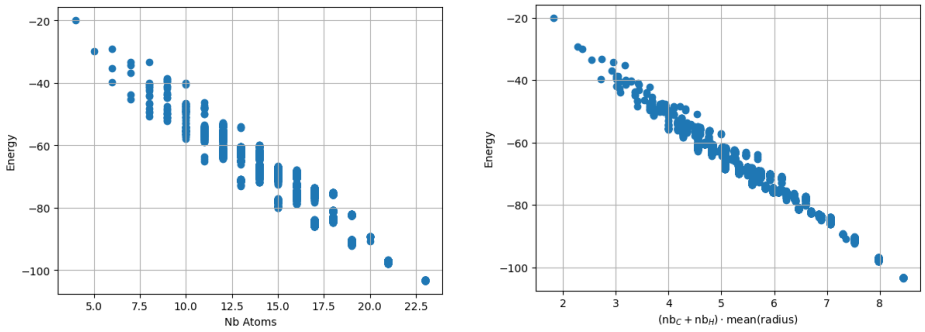

FIGURE 2 – **À droite :** Énergie d'atomisation en fonction du nombre d'atomes. **À gauche :** énergie d'atomisation en fonction du produit du rayon moyen des atomes et de la somme du nombre d'éléments H et d'éléments C.

La figure 2 montre qu'il y a une relation linéaire négative entre le nombre d'atomes dans une molécule et l'énergie d'atomisation. Cela est expliqué par le fait que puisque les molécules sont plus grandes, elles ont plus de liaisons chimiques, et donc la somme des énergies de toutes ces liaisons est plus grande en valeur absolue. Puisque l'énergie de liaison est négative, l'énergie d'atomisation totale devient plus négative.

2.3 Dataframe des features

Nous avons stocké dans un dataframe toutes les molécules avec leur id, leur nombre d'atomes totaux, chacun de leurs atomes avec leurs coordonnées cartésiennes ainsi que leur numéro atomique associé. Cela va nous servir notamment dans la construction de la matrice de Coulomb.

3 Description des modèles

3.1 Méthode se basant sur la matrice de Coulomb

3.1.1 Description et nouveauté de cette méthode

Nous proposons dans cette partie d'entraîner un réseau de neurones simple qui prend en entrée la matrice de Coulomb associée à la molécule. Cette méthode est inspirée de l'article [1].

La nouveauté de l'approche utilisant la matrice de Coulomb réside dans sa capacité à capturer les interactions atomiques d'une molécule de manière invariante aux translations et rotations dans l'espace 3D, tout en conservant une représentation structurée et riche en informations. Ceci est réalisé grâce au fait que l'on calcule ces matrices grâce aux distances relatives des atomes les uns avec les autres, permettant d'ignorer le repère absolu. Chaque élément de la matrice quantifie soit l'énergie potentielle d'un atome isolé (éléments diagonaux), soit la force de répulsion entre paires d'atomes (éléments hors diagonaux).

Cependant, la matrice de Coulomb n'est pas invariante aux permutations des indices atomiques, ce qui signifie que de nombreuses matrices de Coulomb peuvent être associées à la même molécule en permutant simplement les colonnes où les lignes. Pour pallier ce problème, on introduit la matrice de Coulomb aléatoire qui consiste à ordonner les colonnes de la matrice suivant le module de chacune puis de perturber la matrice initiale en ajoutant du bruit et en appliquant des permutations aléatoires aux lignes et aux colonnes.

3.1.2 Construction de la matrice de Coulomb aléatoire

Comme vu dans la partie précédente, nos molécules sont composées d'un maximum de 23 atomes. Nos matrices de Coulomb sont donc de taille (23x23), en ajoutant des "atomes invisibles" c'est-à-dire du zero-padding aux molécules ayant un nombre d'atomes inférieur à 23.

La matrice de Coulomb a été construite selon cette formule :

$$C_{ij} = \begin{cases} 0.5 Z_i^{2.4} & \text{si } i = j \\ \frac{Z_i Z_j}{|R_i - R_j|} & \text{si } i \neq j \end{cases}$$

et la matrice de Coulomb aléatoire selon ces étapes :

— Calcul de la norme par ligne de la matrice de Coulomb $||C|| = (||C_1||, \ldots, ||C_d||)$.

— Tirage de $n \sim \mathcal{N}(0, \sigma I)$ et recherche d'une permutation telle que permutation$(||C|| + n) =$ sort$(||C|| + n)$.

— Permutation de la matrice de Coulomb ligne par ligne et colonne par colonne avec la même permutation, c'est-à-dire $C_{\text{random}} = \text{permutecols}_P(\text{permuterows}_P(C))$.

3.1.3 Entraînement du modèle

Nous utilisons un réseau de neurones simple pour notre tâche de régression composé de deux couches cachées comprenant respectivement 400 et 100 unités avec des fonctions d'activation sigmoïdales. Pour éviter un problème d'optimisation mal conditionné, nous décidons de convertir chaque dimension de la matrice de Coulomb aléatoire pour obtenir un tenseur tridimensionnel de prédicats essentiellement binaires pour notre entrée du réseau, définit comme suit :

$$x = \left[\tanh\left(\frac{C - \theta}{\theta}\right), \tanh\left(\frac{C}{\theta}\right), \tanh\left(\frac{C + \theta}{\theta}\right) \right]$$

Les paramètres utilisés pour l'entraînement sont les suivantes :

— $\theta = 1$: défini en haut

— $learning\_rate = \frac{\gamma_0}{\sqrt{m}}$ avec $\gamma_0 = 0.01$: taux d'apprentissage adaptatif. Le taux d'apprentissage γ est ajusté en fonction du nombre d'unités d'entrée m, ce qui aide à optimiser l'apprentissage du réseau de neurones

— nb_epochs = 80

— loss : Mean Square Error

— optimiseur : SGD avec un momentum = 0.9

Nous avons également testé d'autres valeurs de paramètres et une autre architecture, mais celles-ci ont donné une MSE plus élévée :

— Taux d'apprentissage constant : lr = 0.01 ou lr = 0.001

— Nombre d'epochs : 200

— Régularisation L2 avec alpha = 0.01

— Fonctions d'activation ReLU au lieu de sigmoïdes

— Ajout de 2 couches cachés, augmentant la profondeur du réseau

3.1.4 Résultat

Nous voyons sur la courbe 3 que notre modèle est bien fitté au vu de la courbe de validation.

Nous obtenons une MSE de 3.006 sur la moitié du jeu de données de test avec les hyperparamètres précédents. Cette erreur peut être diminué, notamment avec la méthode du scattering 3D, montrant ainsi les limites de la matrice de Coulomb.

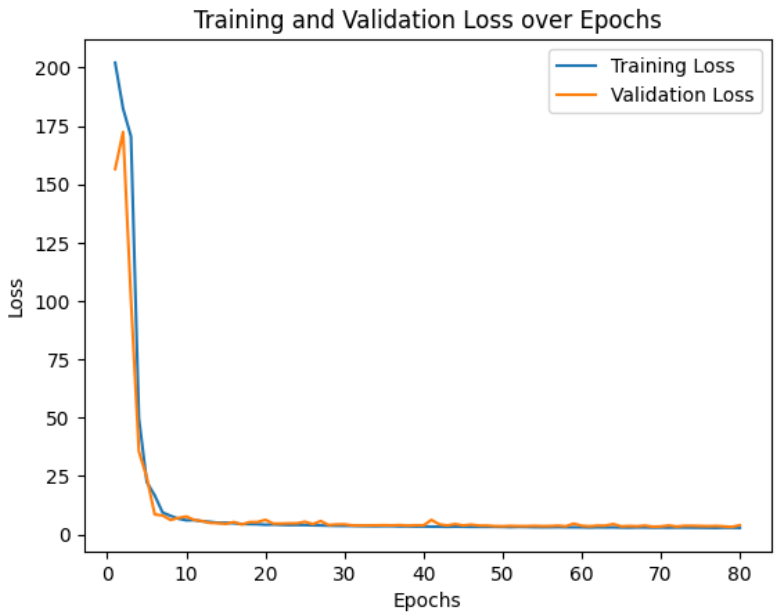

FIGURE 3 – Évolution de l'erreur à travers les épochs d'entraînement

3.2 Méthode du scattering 3D

3.2.1 Description

La méthode de scattering consiste à calculer de nouvelles caractéristiques des molécules, invariantes par rotation et translation au vu de la nature du problème posé. On commence d'abord par représenter les molécules du dataset dans l'espace tridimensionnel par la somme de gaussiennes pondérées par le numéro atomique de chaque atome présent dans les molécules étudiées, d'où résulte un tenseur tridimensionnel qui couvre l'espace occupé par l'atome, comme représenté dans la figure 4.

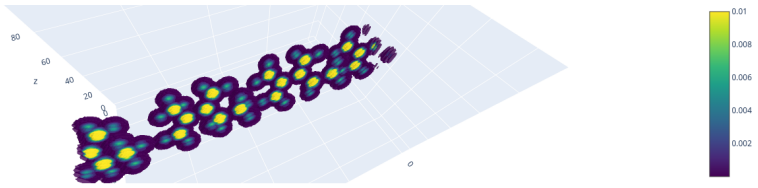

FIGURE 4 – Représentation de la molécules en utilisant une somme de gaussiennes pondérées par les numéros atomique de la molécule 2903 de l'ensemble d'entraînement. On peut remarquer les atomes de Carbone et les atomes d'Hydrogène grâce à leur différence de poids, les atomes de Carbone étants plus lourds.

Ensuite, la méthode consiste à réaliser des convolutions avec les différentes ondelettes harmoniques sphériques montrées dans la figure 5 ainsi que de calculer différents moments des représentation gaussiennes des molécules afin d'en extraire les caractéristiques spatiales. Grâce à leur symétrie sphérique, les caractéristiques extraites grâce à ces méthodes sont invariantes par translation et par rotation[2]. Lors du calcul de ces coefficients on applique la fonction module aux

coefficients ce qui rajoute une non-linéarité, ce qui diffère du calcul d'ondelettes classique.

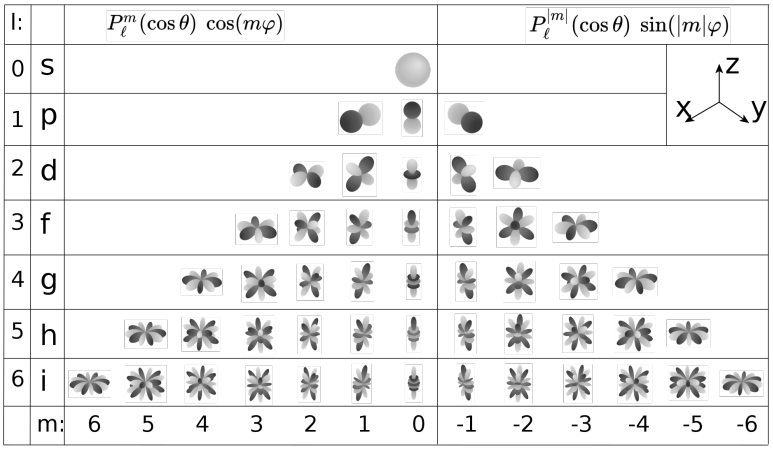

FIGURE 5 – Ondelettes harmoniques sphériques pour calucler le scattering

3.2.2 Calcul des coefficients

On peut remarquer que le calcul des coefficients de Scattering est long et coûteux. Le calcul des coefficients pour l'ensemble d'entraînement prend 3 heures et 3 minutes sur une carte graphique de type RTX3060, et 44 minutes pour l'ensemble de test sur le même GPU. Cependant on peut aussi voir que le calcul des coefficients d'une molécule est indépendant des autres molécules ce qui veut dire qu'il est parallélisable sur plusieurs machines. Nous avons pris l'initiative de paralléliser ce calcul sur 18 machines de TP à l'ENSEEIHT en essayant d'exploiter les GPUs présents sur celles-ci grâce à la bibliothèque MPI4Py. Cependant, à cause des limitations imposées par la direction informatique de l'école et du quota de GPU par utilisateur et non par machine imposé, nous avons uniquement pu effectuer cette parallélisation en utilisant le processeur ce qui a impacté les performances de calcul. En effet, avec 18 machines nous arrivons à peine à réduire le temps de calcul et passant à 2 heures et 30 minutes estimées. Nous avons alors décidé de ne pas exécuter le code par préoccupation écologique.

3.2.3 Entraînement et résultats

Nous obtenons à l'issue du calcul des coefficients de Scattering des vecteurs de taille $300 = 3 \cdot (6 \cdot 4 \cdot 4 + 4)$ où :

- 3 vecteurs auxquels on applique le scattering. Ces vecteurs étant le vecteur de Scattering des molécules complètes, le Scattering des molécules en ne prenant en compte que les éléctrons de valence et le Scattering des molécules en ne prenant en compte que les éléctrons de cœur ;
- $(6, 4, 4) = (1 + J + \frac{J(J+1)}{2}, L + 1, P)$ la taille de la sortie de l'ordre 1 et 2 du Scattering ;
- $4 = \#Q$ le nombre de coefficients d'ordre 0.

Nous avons utilisé une régression linéaire de Ridge afin d'avoir de la régularisation $L2$. La figure 6 montre que la MAE est croissante en fonction de α ce qui voudrait dire que la valeur optimale de ce paramétre serait $\alpha = 0.0$ ce qui veut dire qu'on n'effectue plus de régularisation et nous obtenons une MSE de $2.52 \cdot 10^{-3}$.

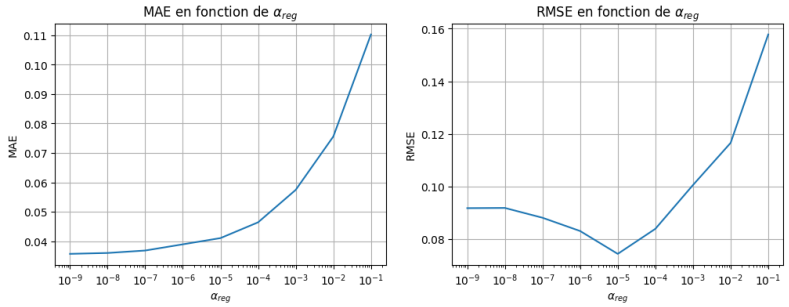

FIGURE 6 – RMSE et MAE en fonction du paramétre de régularisation α

4 Conclusion

Nous remarquons que la précision des prédictions est 3 ordres de grandeurs meilleure grâce à l'extraction des caractéristiques des molécules par Scattering comparée aux matrices de Coulomb.

5 Perspective

5.1 Etat de l'art

L'article [3] propose une méthode innovante combinant apprentissage automatique classique et quantique pour prédire les énergies d'atomisation des molécules. Un autoencodeur convolutif extrait les caractéristiques des matrices de Coulomb et des positions atomiques, utilisées ensuite dans divers algorithmes de régression quantique. Bien que les modèles quantiques n'aient pas montré une accélération significative, cette approche ouvre des perspectives intéressantes pour l'application des technologies quantiques en chimie computationnelle.

En parallèle, l'article [4] utilise des réseaux à paramètres naturels (NPN) pour prédire les énergies d'atomisation des molécules. En représentant les molécules avec des matrices de Coulomb et en appliquant un modèle de réseau neuronal probabiliste, cette méthode atteint une précision supérieure aux travaux existants, avec un temps de prédiction accéléré.

5.2 Travaux futurs

Nous souhaitons implémenter la méthode décrite dans le deuxième article qui utilise des réseaux à paramètres naturels. Cette approche semble prometteuse en raison de sa précision et de sa rapidité. Nous prévoyons de comparer cette méthode avec notre propre méthode de scattering 3D, qui a montré d'excellents résultats jusqu'à présent.

Références

[1] Montavon G., Hansen K., Fazli S., Rupp M., Biegler F., Ziehe A., Tkatchenko A., Lilienfeld A. V., Müller K.-R. Learning Invariant Representations of Molecules for Atomization Energy Prediction. NIPS 2012, 25, 440–448.

[2] Eickenberg M, Exarchakis G, Hirn MJ, Mallat S. Solid Harmonic Wavelet Scattering : Predicting Quantum Molecular Energy from Invariant Descriptors of 3D Electronic Densities. Neural Information Processing Systems. 2017 ;30 :6540-6549. https ://papers.nips.cc/paper/7232-solid-harmonic-wavelet-scattering-predicting-quantum-molecular-energy-from-invariant-descriptors-of-3d-electronic-densities.pdf.

[3] Reddy P, Bhattacherjee AB. A hybrid quantum regression model for the prediction of molecular atomization energies. Machine Learning : Science and Technology. 2021 ;2(2) :025019. https ://doi.org/10.1088/2632-2153/abd486.

[4] Chu C, Xiao Q, He C, Chen C, Li L, Zhao J, Zheng J, Zhang Y. A novel method for atomization energy prediction based on natural-parameter network. Computational and Theoretical Chemistry. 2023 ;1224 :114207. https ://doi.org/10.1016/j.comptc.2023.114207.