Prédiction de l'énergie moléculaire

Karima Ghamnia, Cassandra Mussard

20/06/2024

Contexte

Prédire l'énergie d'une molécule à l'aide de la position des atomes.

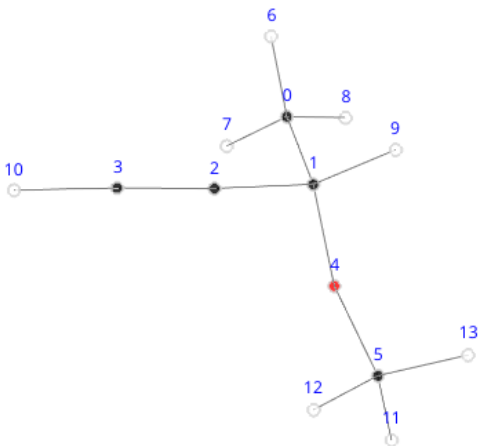
2 méthodes envisagées :

- **Matrice de Coulomb.**
- **Scattering 3D.**

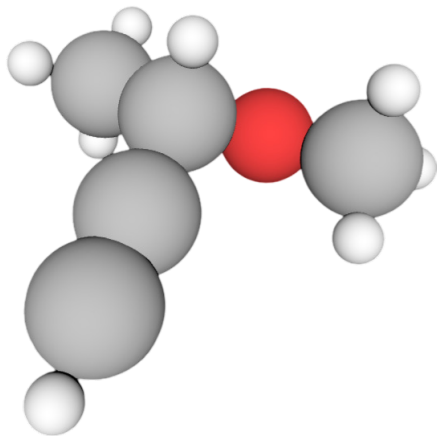
Présentation du jeu de données

Jeu de données

- Sous ensemble de QM7-X.
- Un fichier .xyz avec position des atomes.
- Un fichier .csv pour les énergies.



(a) Avec plotly



(b) Avec Ase

Qu'est ce que la matrice de Coulomb ? [1]

Représentation matricielle de la molécule.

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

avec Z_i les charges nucléaires et R_i les coordonnées cartésiennes.

- Élements de la diagonale : Energie d'un atome isolé.
- Élements hors diagonale : Répulsion de Coulomb entre toutes les paires de noyaux.

=> **Comment faire une classification de l'énergie à l'aide de la matrice de Coulomb ?**

S.Fazli1 M.Rupp G.Montavon, K.Hansen. Learning invariant representations of molecules for atomization energy prediction.
2012

Problèmes rencontrés

Comment rendre la matrice de Coulomb invariante aux permutations ?

Sorted Coulomb Matrix

$$\forall i,\ \|C_i\| \geq \|C_{i+1}\|$$

Comment avoir la même taille de la matrice pour chaque molécule ?

Zéro padding

Calcul du nombre max d'atomes dans chaque molécule.

Zéro padding pour avoir des matrices de taille 23x23 pour chaque molécule.

Augmentation de données

Sur la 3ème dimension de la matrice

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

avec $\theta = 1$

Autre méthode d'augmentation de données

- Ajout de bruit sur les positions.
- Rendre le modèle plus robuste, ajouter plus d'exemples d'entraînement.

Modèle

Multi-Layer Perceptron (MLP)

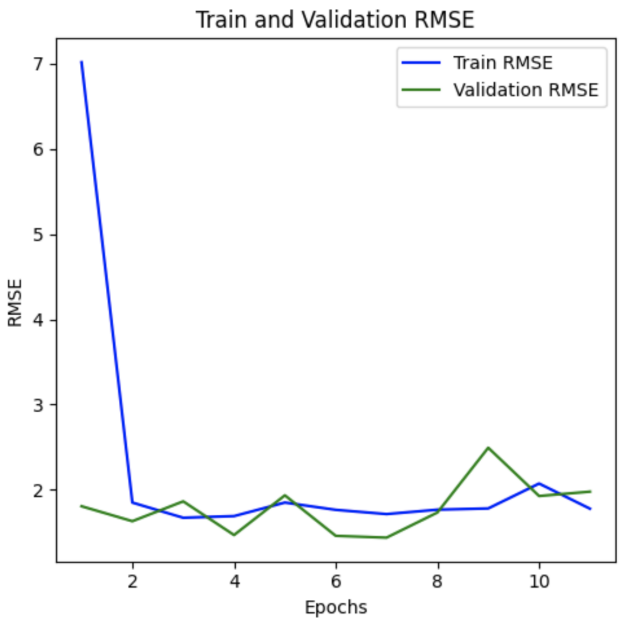
- 4 couches linéaires + leakyReLU.
- Initialisation des poids avec une distribution normale.

Entraînement

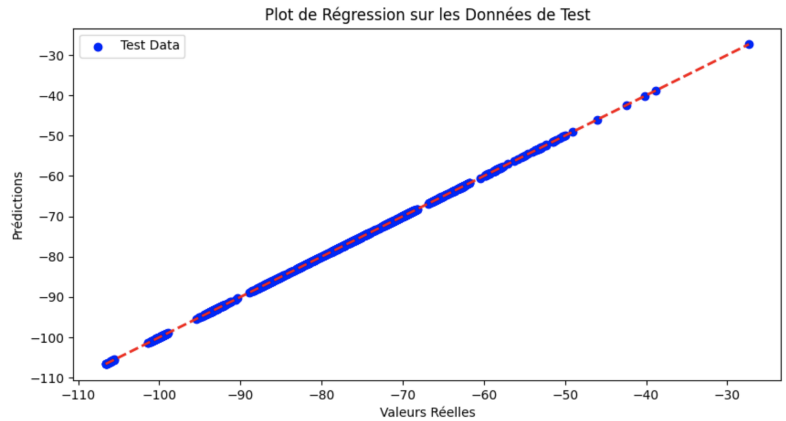
Hyperparamètres pour l'entraînement du modèle

- 8 epochs.
- Optimiseur Adam.
- Learning rate de 0.01.
- Terme de régularisation L2 à $1e-4$ dans l'optimiseur.
- Fonction de coût : RMSE.

Résultats



(a) Loss de d'entraînement et de validation



(b) Régression sur le dataset de test

Figure – Graphiques de la Loss et de la regression

Loss de test sur Kaggle : 1.56

Qu'est ce que le 3D Scattering ? [3]

Coefficients de scattering

- Représentation invariante aux rotations et translations tridimensionnelles des molécules.
- Multiéchelle avec 3 ordres (0 : propriétés globales, 1 : intéractions locales, 2 : intéractions quadratiques)

M.Hirn S.Mallat L.Thiry M.Eickenberg, G.Exarchakis. Solid harmonic wavelet scattering for predictions of molecule properties.
physics.chem-p, *2018*

Scattering 3D

Etapes

- Cartes de densité.
- Ondelette de solid harmonique.
- Ajout d'une non-linéarité.

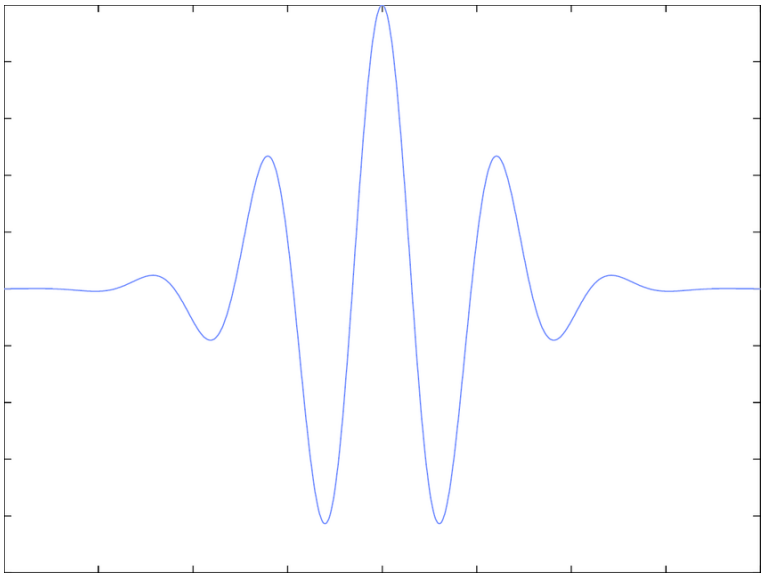


Figure – Ondelette

Pré-processing des données

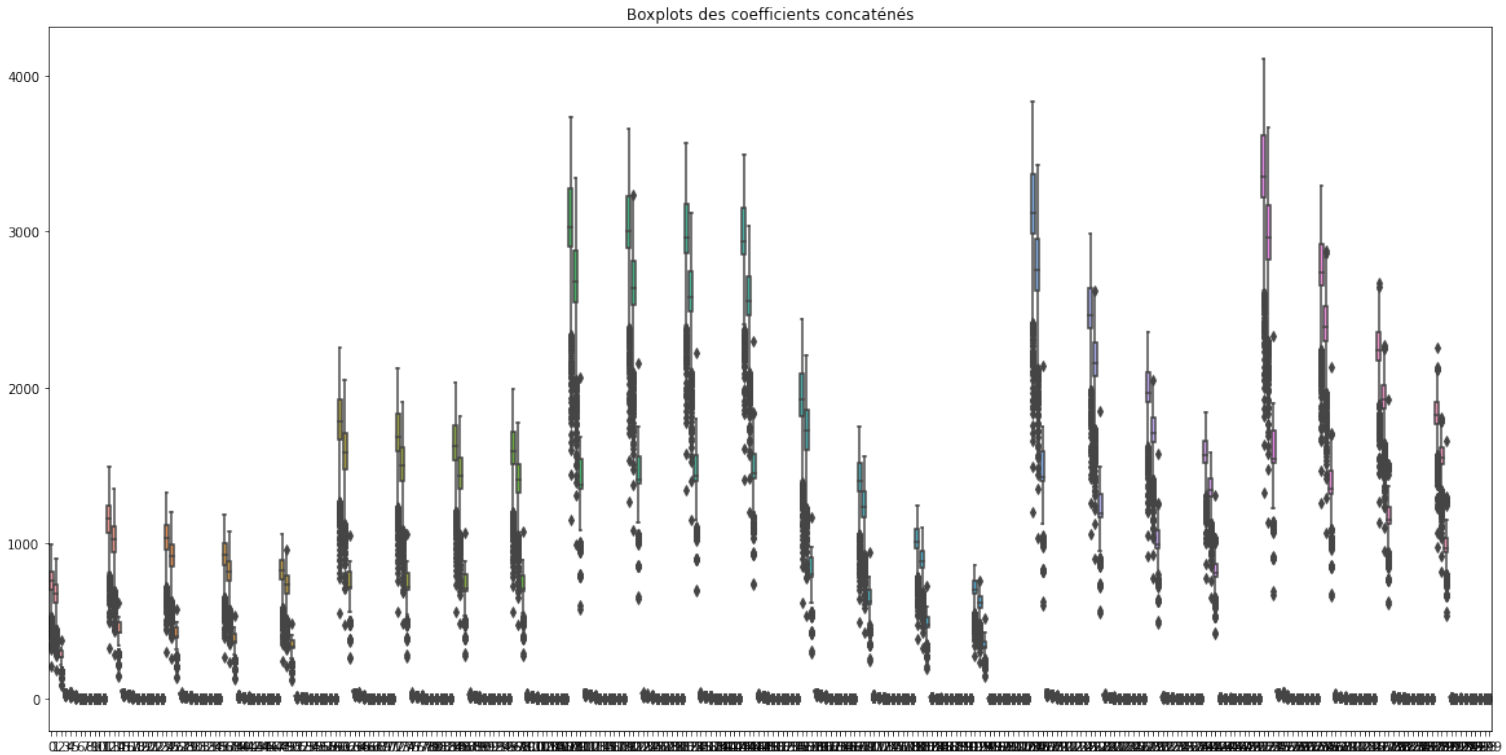


Figure – Boxplot des coefficients de scattering

=> Utilité de centrer-réduire les coefficients de scattering

Quels modèles tester ?

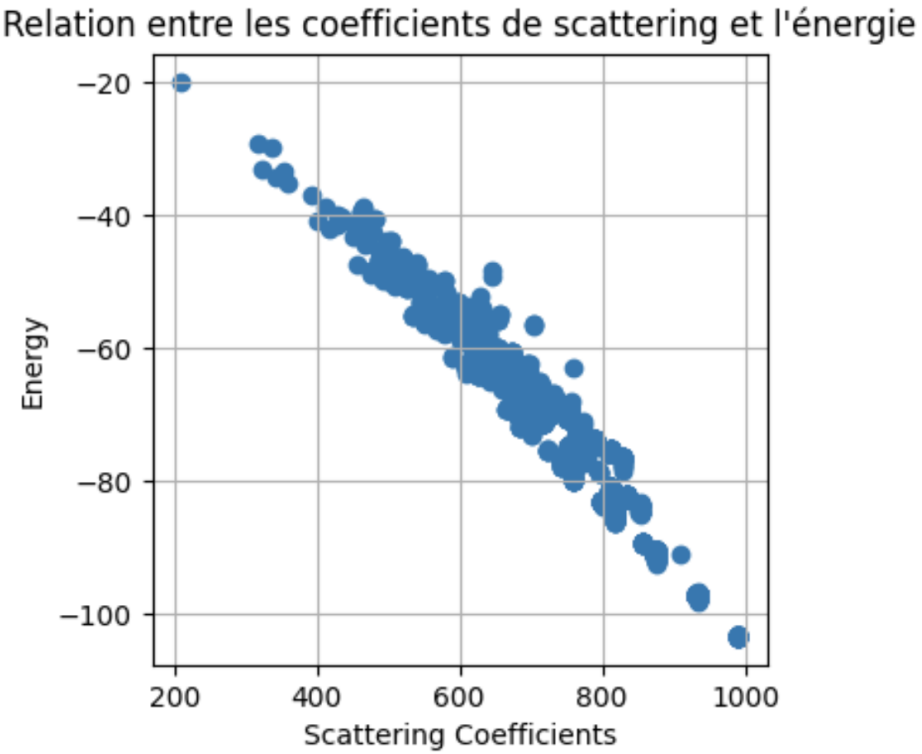


Figure – Relation entre les coefficients de scattering et l'energie

Relation linéaire => Modèle de régression linéaire

Résultats

Grid-search (minimiser MAE, RMSE) avec les hyperparamètres de la grille (M,N,O) = (192, 128, 96), (J, L) = (2, 3) :

Méthode	Ridge	Régression PLS	Ridge + Moyenne	Lasso	MLP	SVR
Score	0.003231 / 0.0018	0.016	3.6	0.11	2.18	2.88

Table – Résultats obtenus avec différentes méthodes de régression

Meilleur résultat : Ridge avec les hyperparamètres de la grille (M,N,O) = (250, 150, 100), (J, L) = (5,4), alpha = 10^{-9}, solveur = Cholesky.

Pour aller plus loin

Perspectives d'amélioration

- Data augmentation pour la matrice de Coulomb ? KAN ?
- GNN [4] ?

H.Wang C.Lu C.Lee Z.ZHANG, Q.Liu. Motif-based graph self-supervised learning for molecular property prediction properties.
2021

Bibliographie

[1] S.Fazli1 M.Rupp G.Montavon, K.Hansen. Learning invariant representations of molecules for atomization energy prediction. 2012.

[2] R.Spezialetti P.Ramirez S.Salti L.Stefano L.Luigi, A.Cardace. Deep learning on implicit neural representations of shapes. 2023.

[3] M.Hirn S.Mallat L.Thiry M.Eickenberg, G.Exarchakis. Solid harmonic wavelet scattering for predictions of molecule properties. *physics.chem-p*, 2018.

[4] H.Wang C.Lu C.Lee Z.ZHANG, Q.Liu. Motif-based graph self-supervised learning for molecular property prediction properties. 2021.