Molecular Energy Prediction

Alexandre Demarquet
Thomas Dion

June 2025

Code : `https://github.com/AlexandreDemarquet/molecular_energy_prediction`

Contents

List of Figures

1 Introduction

La modélisation de l'énergie d'atomisation des molécules constitue un enjeu central en chimie computationnelle et en apprentissage automatique pour la physique. Dans ce projet, nous cherchons à prédire l'énergie d'une molécule organique à partir de sa configuration géométrique tridimensionnelle. Cette tâche est formalisée comme un problème de régression non linéaire en haute dimension, où l'on souhaite approximer la fonction $E(r)$, représentant l'énergie d'une molécule en fonction de la position de ses atomes $r = \{r_1, r_2, \ldots, r_N\}$.

Une des principales difficultés de ce problème réside dans le respect des contraintes physiques de symétrie. En effet, l'énergie d'une molécule doit rester inchangée face aux opérations de translation, rotation et permutation des atomes de même type. Un modèle pertinent doit donc produire des représentations invariantes à ces transformations géométriques.

Pour répondre à cette problématique, nous utilisons des transformées de scattering en 3D basées sur des ondelettes harmoniques solides, qui permettent d'extraire des descripteurs invariants à partir des structures moléculaires. Ces représentations sont ensuite utilisées comme entrées dans un modèle de régression pour prédire l'énergie moléculaire.

Nous nous appuyons sur un sous-ensemble du jeu de données QM7-X, qui contient des milliers de structures moléculaires avec leurs énergies d'atomisation associées. L'objectif final est d'évaluer notre modèle sur un jeu de test inconnu, selon la métrique RMSE (Root Mean Square Error), afin de quantifier sa capacité à généraliser à de nouvelles molécules.

Ce rapport présente les étapes de traitement des données, la construction du modèle, les résultats expérimentaux obtenus, ainsi qu'une discussion sur les performances et les perspectives d'amélioration.

2 Analyse des données

Dans cette section, nous analysons les caractéristiques du jeu de données utilisé, issu du sous-ensemble QM7-X. Ce jeu contient 4739 configurations moléculaires de petites molécules organiques. Chaque configuration fournit :

- La position tridimensionnelle $r_i \in \mathbb{R}^3$ de chaque atome ;
- Le type chimique de chaque atome (H, C, N, O, S, etc.) ;
- L'énergie d'atomisation associée à la molécule (en unités d'énergie atomique).

Les données sont stockées dans le format xyz (pour les structures) et dans un fichier CSV (pour les énergies).

2.1 Visualisation d'une molécule

À titre d'exemple, la figure 1 montre la représentation 3D d'une molécule extraite du jeu d'entraînement, colorée selon les types d'atomes.

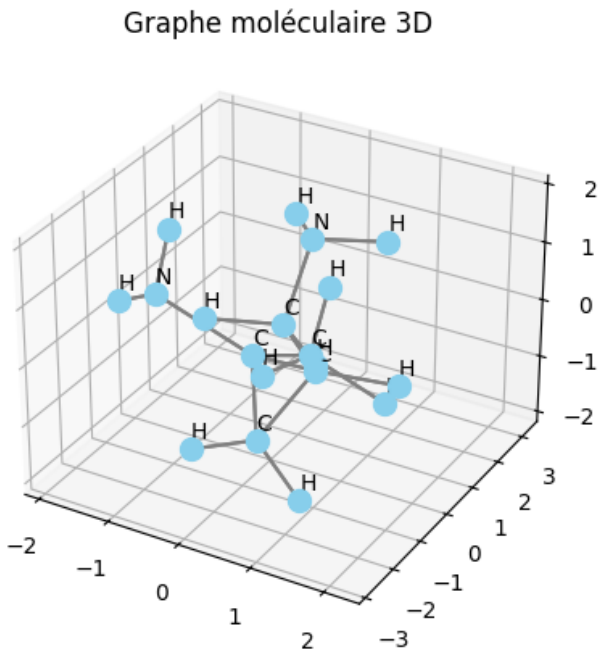

Figure 1: Exemple de molécule du dataset.

Pour ce projet nous avons aussi developper des script pour visualiser en 3D les moélcule à l'aide de la librairie Open3D.

2.2 Distribution des types d'atomes

Le dataset contient majoritairement des atomes d'hydrogène et de carbone. La figure 2 présente la distribution des types d'atomes (par fréquence d'apparition).

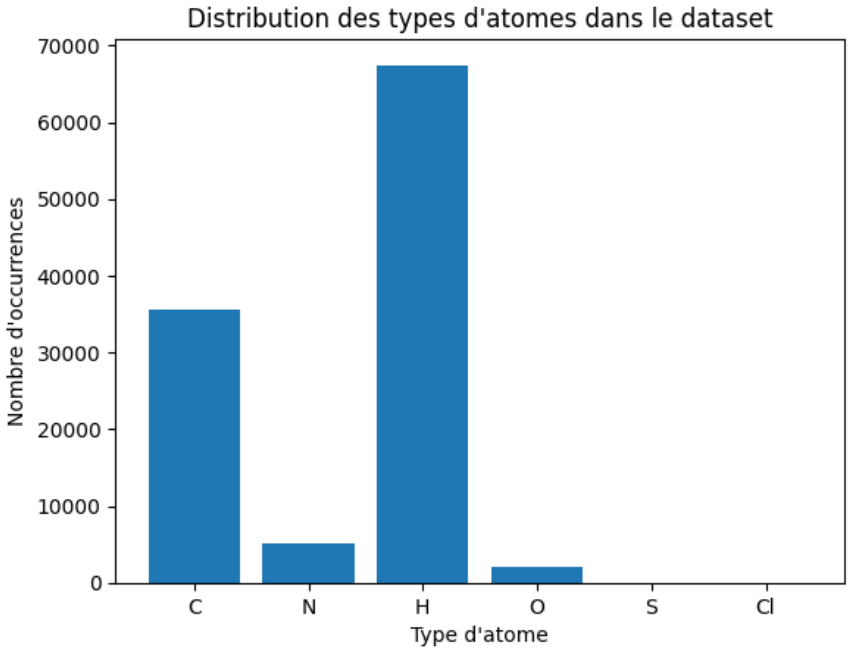

Figure 2: Distribution des types d'atomes dans le dataset.

2.3 Distribution des énergies d'atomisation

La figure 3 montre la distribution des énergies d'atomisation dans l'ensemble d'entraînement. Les valeurs varient fortement selon la structure moléculaire.

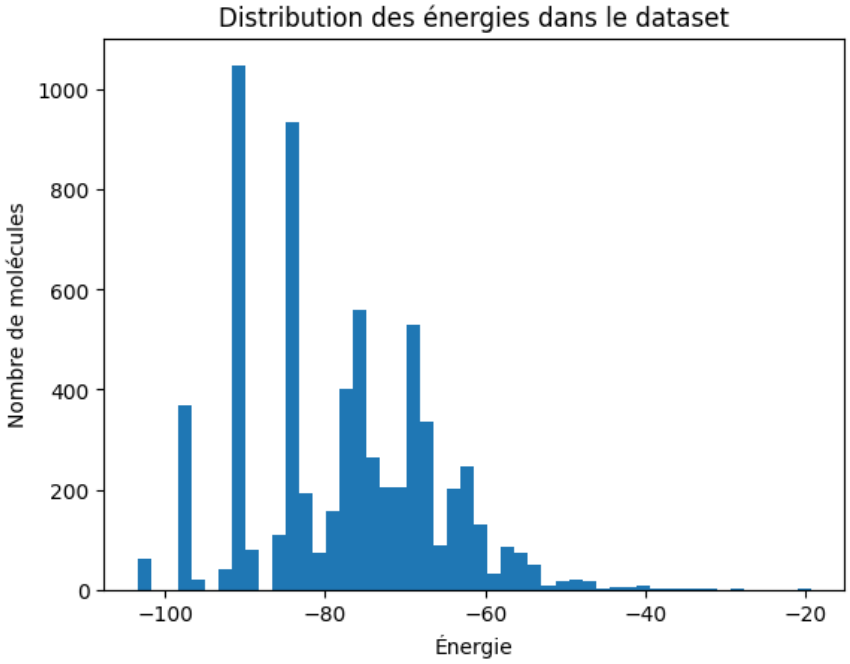

Figure 3: Distribution des énergies d'atomisation.

2.4 Distribution de la taille des molécules

Nous définissons la taille d'une molécule par le nombre total d'atomes qu'elle contient. La figure 4 présente la distribution de cette taille pour l'ensemble du dataset.

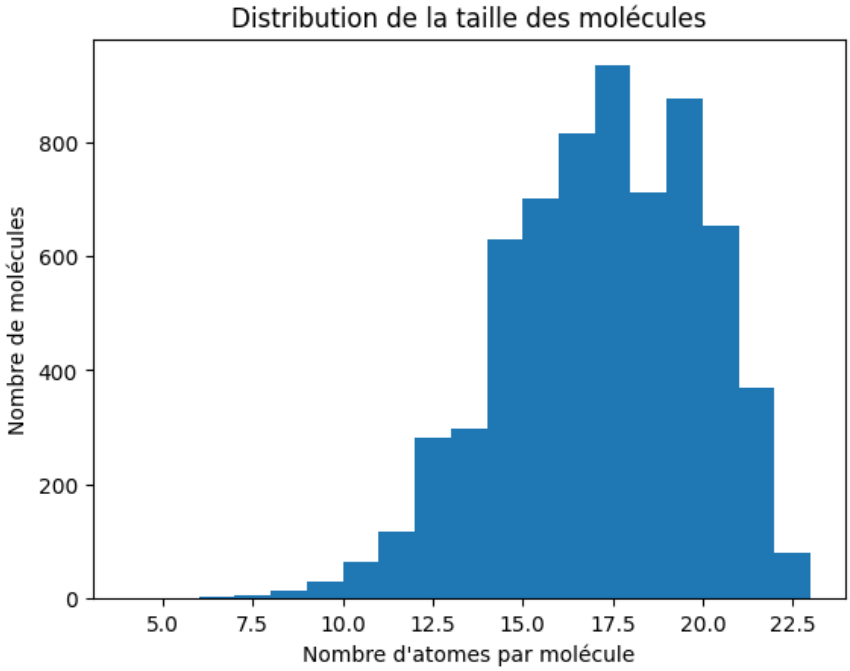

Figure 4: Distribution de la taille (nombre d'atomes) des molécules.

2.5 Énergie en fonction de la taille de la molécule

Enfin, nous examinons la relation entre la taille d'une molécule et son énergie d'atomisation. Comme illustré en figure 5, une tendance linéaire partielle est observable, bien que des variations importantes subsistent pour une taille donnée.

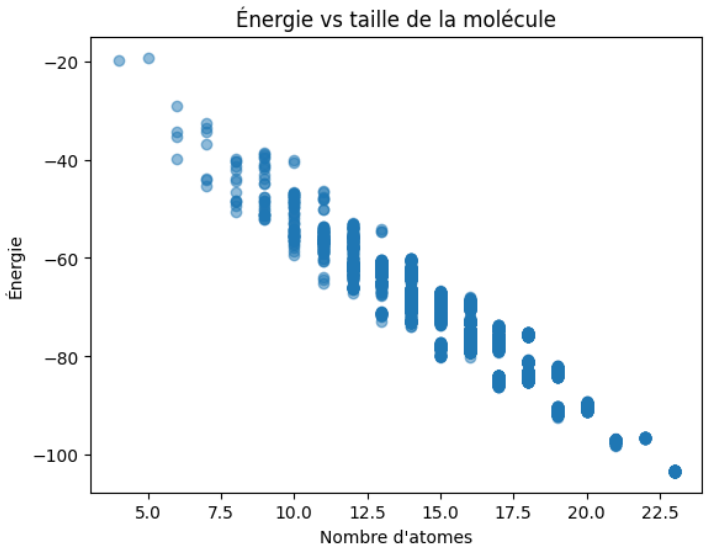

Figure 5: Énergie d'atomisation en fonction de la taille des molécules.

3 Méthodes

3.1 Histogramme des atomes

Pour établir une première référence (baseline), nous utilisons une représentation très simple : l'histogramme des types d'atomes. Cette description capture la composition chimique tout en étant naturellement invariante aux transformations géométriques physiques.

Construction Considérons une molécule composée de N atomes avec pour chacun une position $r_i \in \mathbb{R}^3$ et un type chimique $X_i \in \{\mathrm{H}, \mathrm{C}, \mathrm{N}, \mathrm{O}, \mathrm{S}, \dots\}$. Nous définissons l'histogramme :

$$h = (h_H, h_C, h_N, h_O, h_S, \dots) \in \mathbb{N}^d,$$

où

$$h_X = \sum_{i=1}^{N} \mathbf{1}_{X_i = X}$$

est le nombre d'atomes de type X présents.

Invariance Cette représentation présente plusieurs invariances essentielles :

- **Invariance par translation** : le vecteur h ne dépend pas des positions absolues r_i, donc toute translation $r_i \to r_i + t$ (pour un vecteur t quelconque) ne modifie pas h.
- **Invariance par rotation** : réaliser une rotation $r_i \to Rr_i$ (avec $R \in SO(3)$) conserve les types et leur nombre, donc h reste identique.

La figure 6 présente l'histogramme d'une molécule exemple.

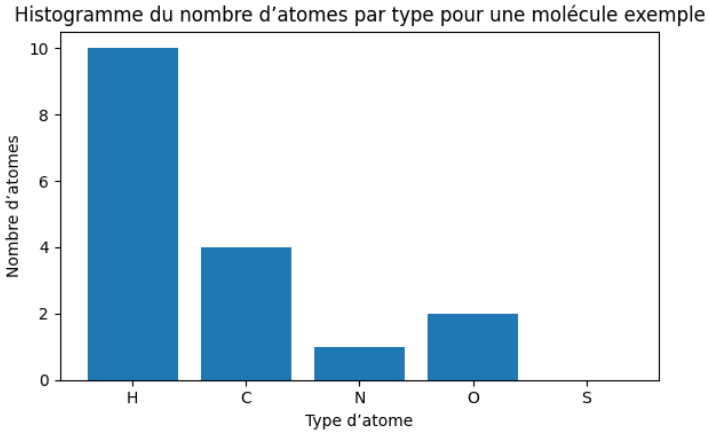

Figure 6: Histogramme du nombre d'atomes par type pour une molécule exemple.

Modèle de régression Les vecteurs d'histogrammes h sont données au modèle de régression linéaire :

$$\hat{E}(h) = w^T h + b,$$

où l'apprentissage porte sur (w, b) en minimisant la perte :

$$\mathcal{L}(w, b) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(E_i - (w^T h^{(i)} + b)\right)^2}$$

Résultats Nous évaluons la performance du modèle baseline selon la procédure suivante :

1) Partition du dataset en 80% pour l'entraînement et 20% pour la validation.
2) Estimation analytique de (w, b) sur l'ensemble d'entraînement.
3) Calcul du RMSE sur l'ensemble de validation.

Performance finale Le modèle linéaire fournit les résultats suivants :

$$\text{RMSE}_{\text{train}} = 0.54\,, \quad \text{RMSE}_{\text{val}} = 0.55$$

La proximité de ces valeurs indique un faible écart de généralisation.

Nous avons également utilisé d'autres méthodes, telles que XGBoost et la régression multilinéaire. Les résultats sont présentés dans la partie 5 'Benchmark'.

Comparaison prédiction / réalité La figure 7 compare les énergies prédites et réelles sur l'ensemble de validation :

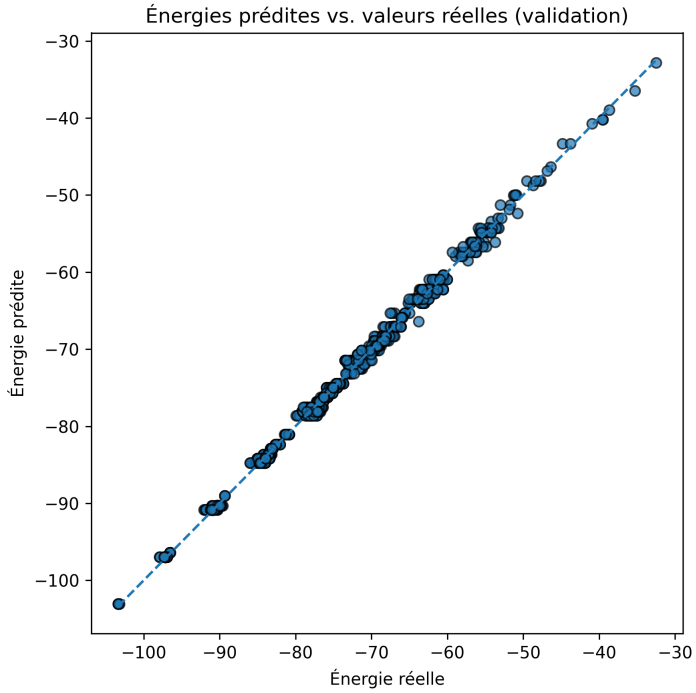

Figure 7: Énergies prédites vs. valeurs réelles (validation).

Les points s'alignent autour de la droite d'identité, soulignant la capacité du modèle à extraire l'effet principal de la composition atomique malgré sa simplicité.

3.2 Spectre de la matrice de Coulomb

Pour capturer conjointement la disposition géométrique et la nature chimique d'une molécule, nous construisons la *matrice de Coulomb* $C \in \mathbb{R}^{N\times N}$ telle que

$$C_{ij} \;=\; \begin{cases} \frac{1}{2}\, Z_i^{2.4}, & i = j, \\ \dfrac{Z_i\, Z_j}{\|r_i - r_j\|}, & i \neq j, \end{cases}$$

avec Z_i le numéro atomique et $r_i \in \mathbb{R}^3$ la position de l'atome i. Cette formulation assure :

- **Invariance par translation et rotation** puisque seules apparaissent les distances $\|r_i - r_j\|$.
- **Invariance par permutation** des atomes du même type lorsque l'on considère le spectre (l'ensemble des valeurs propres) de C.

On extrait donc les N valeurs propres réelles de C,

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N,$$

et l'on constitue le vecteur

$$s = (\lambda_1, \lambda_2, \ldots, \lambda_N) \in \mathbb{R}^N.$$

Pour des molécules de tailles variables, on fixe une taille maximale $N_{\max}$ et l'on complète par des zéros les spectres de dimension inférieure, aboutissant à une représentation de dimension constante $\mathbb{R}^{N_{\max}}$.

La figure 8 présente, pour deux molécules exemples, les valeurs propres triées (λ_i) de leur matrice de Coulomb.

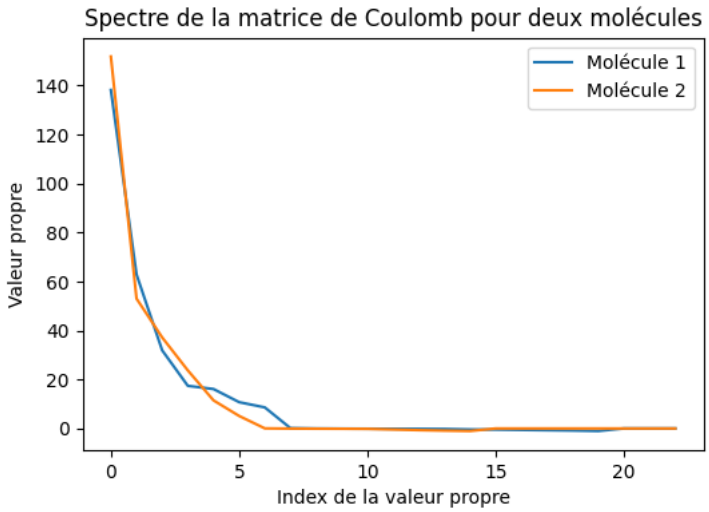

Figure 8: Spectres des valeurs propres de la matrice de Coulomb pour deux molécules exemples.

Résultats Nous avons appliqué la même procédure d'évaluation, en utilisant le même modèle de régression que précédemment. Les résultats obtenus sont les suivants :

$$\mathrm{RMSE}_{\mathrm{train}} = 2.54\,, \quad \mathrm{RMSE}_{\mathrm{val}} = 2.47$$

Comparaison prédictions vs. vraies valeurs La figure 9 montre le nuage de points $(E_i, \hat{E}_i)$ sur la droite d'identité $y = x$, illustrant la qualité globale de la prédiction et l'absence de biais majeur.

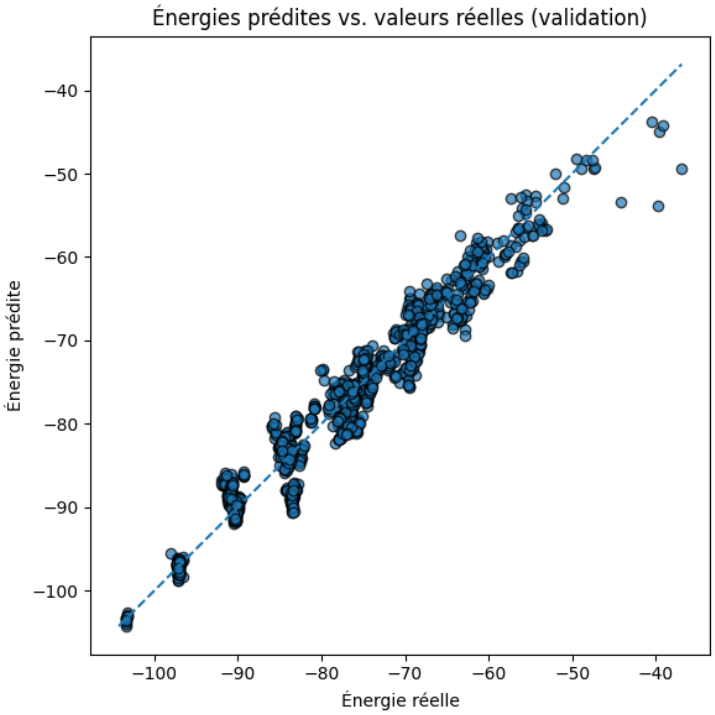

Figure 9: Énergies prédites vs. valeurs réelles (validation) pour la représentation spectrale.

Ces résultats montre que le spectre de la matrice de Coulomb n'apporte pas une amélioration significative par rapport à des descripteurs précèdent.

Nous avons également utilisé d'autres méthodes, telles que XGBoost et la régression multilinéaire. Les résultats sont présentés dans la partie 5 'Benchmark'.

3.3 Scattering 3D

Un outil puissant pour apprendre des représentations invariantes, respectant les symétries fondamentales des molécules, telles que l'invariance par translation et rotation est la **transformée de scattering par ondelettes harmoniques solides**, introduite dans le contexte de la prédiction des propriétés moléculaires par Eickenberg et al. (2018). Cette méthode permet d'extraire des caractéristiques invariantes à partir des densités électroniques.

3.3.1 Ondelettes Harmoniques Solides

Soient $Y_\ell^m(\theta, \phi)$ les harmoniques sphériques définies sur la sphère unité $\mathbb{S}^2$, avec $\ell \in \{0, 1, \ldots, L-1\}$ et $m \in \{-\ell, \ldots, \ell\}$. Elles forment une base orthonormée des fonctions définies sur la sphère.

Les *ondelettes harmoniques solides* sont construites comme suit :

$$\psi_{\ell m}(u) \propto e^{-\frac{|u|^2}{2}} |u|^\ell Y_\ell^m \left(\frac{u}{|u|} \right),$$

où $u \in \mathbb{R}^3$, et l'enveloppe gaussienne assure la localisation spatiale.

Pour obtenir une représentation multi-échelle, on définit des versions dilatées des ondelettes :

$$\psi_{j,\ell m}(u) = 2^{-3j} \psi_{\ell m}(2^{-j} u), \quad 0 \leq j < J.$$

Coefficients de Scattering

Soit $x(u)$ le signal d'entrée en 3D (par exemple, une densité atomique). Les coefficients de scattering d'ordre 1 sont définis par :

$$U_{p_1} x(u) = \left(\sum_{m=-\ell}^{\ell} |x * \psi_{j,\ell m}(u)|^2 \right)^{1/2}, \quad \text{où } p_1 = (j, \ell).$$

Les coefficients d'ordre 2 capturent les interactions entre échelles et sont définis par :

$$U_{p_2} x(u) = \left(\sum_{m=-\ell'}^{\ell'} |U_{p_1} x * \psi_{j',\ell' m}(u)|^2 \right)^{1/2}, \quad \text{où } p_2 = (p_1, j', \ell').$$

À partir de ces coefficients locaux, on calcule des **caractéristiques invariantes** globales par intégration :

$$\Phi(x) = \left(\int |U_{p_1} x(u)|^q \, du, \int |U_{p_2} x(u)|^q \, du \right)_{p_1, p_2},$$

pour un exposant fixé $q > 0$, typiquement $q = 1$ ou $q = 2$. La représentation obtenue $\Phi(x)$ est invariante par rapport aux isométries (rotations et translations) de x.

Visualisation des réponses des ondelettes harmoniques solides en 3D

Dans le cadre de ce projet, nous avons cherché à visualiser les effets de la convolution d'une carte de densité moléculaire $\rho(u)$ avec les *ondelettes harmoniques solides* $\psi_{\ell m}(u)$, définies par :

$$\psi_{\ell m}(u) = e^{-\frac{|u|^2}{2}} |u|^\ell Y_\ell^m \left(\frac{u}{|u|} \right),$$

où Y_ℓ^m sont les harmoniques sphériques sur la sphère unité, $\ell \in \mathbb{N}$ est l'ordre angulaire, et $m \in \{-\ell, \ldots, \ell\}$. Ces ondelettes sont ensuite dilatées par un facteur d'échelle j pour produire des familles multi-échelles :

$$\psi_{j,\ell m}(u) = 2^{-3j} \psi_{\ell m}(2^{-j} u).$$

On effectue alors des convolutions dans $\mathbb{R}^3$ entre la densité moléculaire $\rho(u)$ et chaque ondelette dilatée :

$$(\rho * \psi_{j,\ell m})(u),$$

et l'on calcule la réponse énergétique locale via :

$$U_{j,\ell}[\rho](u) = \left(\sum_{m=-\ell}^{\ell} |(\rho * \psi_{j,\ell m})(u)|^2 \right)^{1/2}.$$

Pour chaque valeur de ℓ (typiquement $\ell = 0, 1, 2, 3$), et pour différentes échelles j, nous avons représenté visuellement une **tranche 2D** (slice fixe dans l'espace 3D) de la réponse $U_{j,\ell}[\rho](u)$. Cette opération produit une carte d'activations invariant par rotation, qui intègre les contributions de toutes les composantes sphériques d'un même ordre ℓ. Le résultat est une représentation plus lisse et isotrope, qui capture globalement la réponse de la densité à une ondelette avec une dilatation j et d'ordre angulaire ℓ.

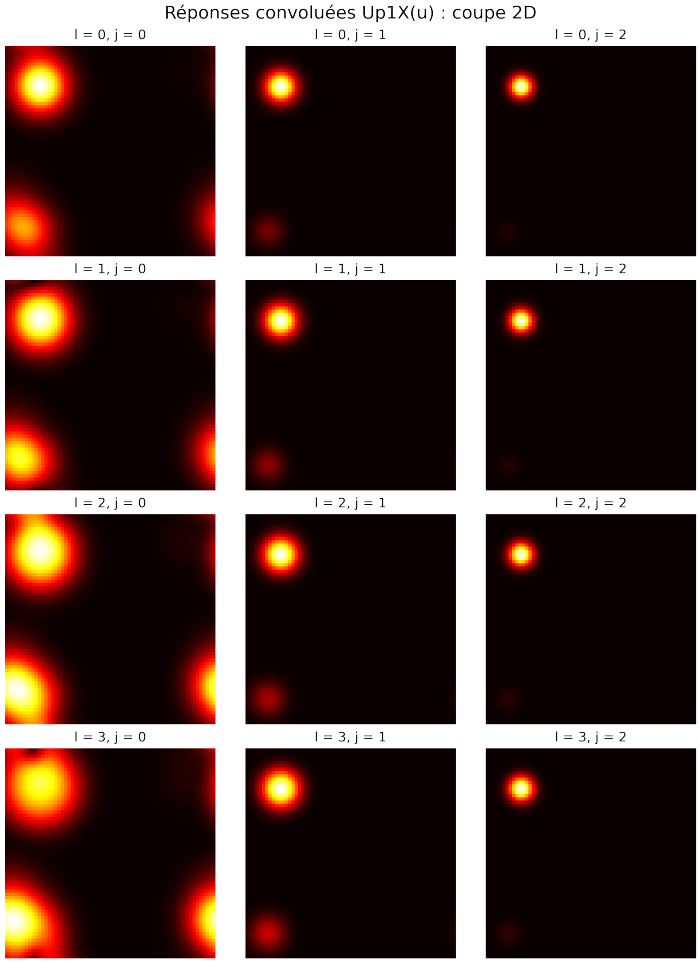

Figure 10: Réponse complète $U_{j,\ell}[\rho](u)$ obtenue après sommation sur m. Ces visualisations présentent des motifs plus globaux, où les détails directionnels sont intégrés.

Il est important de noter que, bien que nous n'ayons pas directement utilisé la librairie `Kymatio` pour appliquer la transformation de scattering, nous avons exploité la classe `HarmonicScattering3D` pour générer les filtres $\psi_{j,\ell m}$ correspondants, à une résolution spatiale fixée (grille 3D de taille $M \times N \times O$). Ces filtres ont ensuite été utilisés pour reproduire mathématiquement les convolutions et visualiser explicitement les réponses locales.

Cependant, afin de mieux comprendre le rôle précis de chaque ondelette sphérique $\psi_{j,\ell,m}$, nous avons également affiché individuellement les termes $\rho * \psi_{j,\ell,m}$ **avant sommation**. Ces visualisations montrent explicitement les lobes directionnels caractéristiques des harmoniques sphériques $Y_\ell^m(\theta, \phi)$, projetés sur la sphère unité, et convolués avec la densité. Cela permet d'observer les motifs d'interférence et la géométrie angulaire captée par chaque ondelette.

Dans les figures suivantes, chaque ligne correspond à une valeur de ℓ (de 0 à 3), et chaque colonne à une échelle j (de 0 à 2). Les figures suivantes montre les **réponses convolutives par filtre individuel** (avant somme).

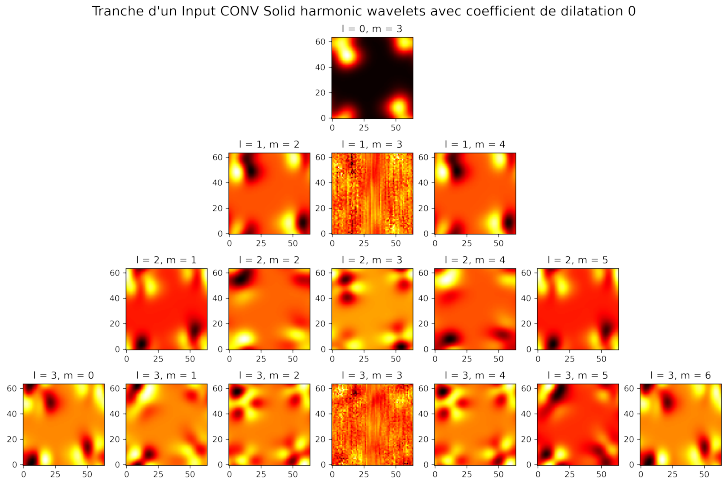

Figure 11: Convolutions individuelles $\rho * \psi_{j,\ell,m}$ pour $j = 0$ et différents ℓ et m. Chaque sous-figure illustre la réponse d'un filtre m donné. On observe clairement les lobes directionnels associés aux harmoniques sphériques.

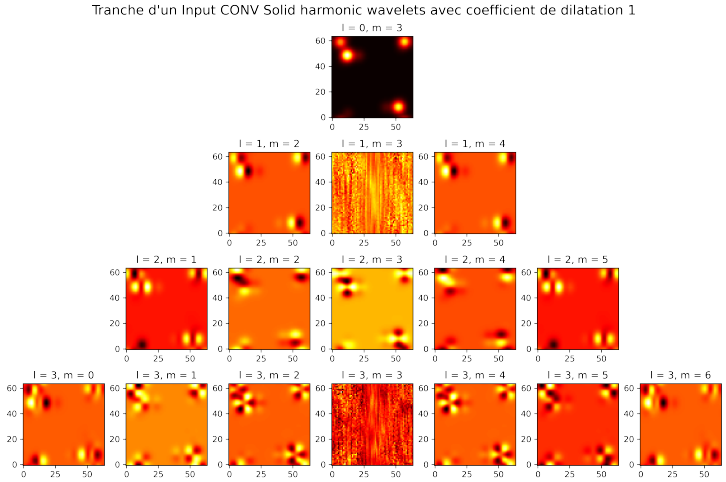

Figure 12: Convolutions individuelles $\rho * \psi_{j,\ell,m}$ pour $j = 1$ et différents ℓ et m. Chaque sous-figure illustre la réponse d'un filtre m donné. On observe clairement les lobes directionnels associés aux harmoniques sphériques.

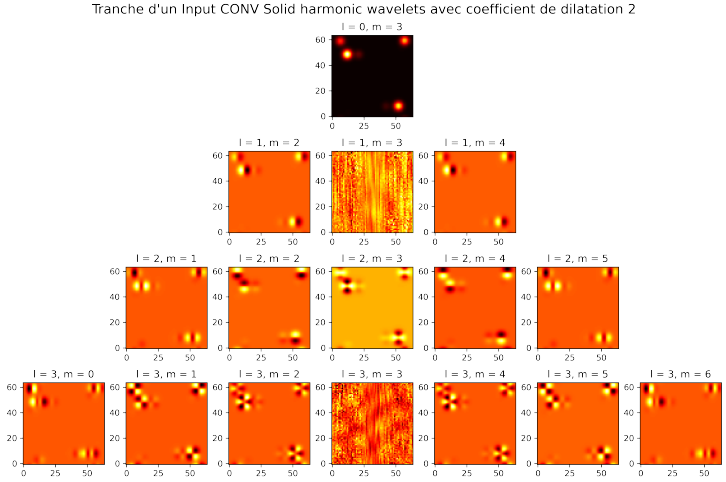

Figure 13: Convolutions individuelles $\rho * \psi_{j,\ell,m}$ pour $j = 2$ et différents ℓ et m. Chaque sous-figure illustre la réponse d'un filtre m donné. On observe clairement les lobes directionnels associés aux harmoniques sphériques.

Ces visualisations permettent de reproduire qualitativement les résultats présentés dans la **figure de la page 4** du sujet du projet.

3.3.2 Invariances du Scattering 3D

- **Invariance par translation** : les coefficients de scattering sont obtenus par des convolutions locales suivies d'une intégration spatiale. Ainsi, une translation de la densité moléculaire entraîne une translation des réponses locales, mais l'intégration globale rend les descripteurs invariants.
- **Invariance par rotation** : pour chaque échelle j et ordre angulaire ℓ, les réponses convolutives sont sommées sur toutes les composantes $m \in [-\ell, \ell]$:

$$U_{j,\ell}[\rho](u) = \left(\sum_{m=-\ell}^{\ell} |(\rho * \psi_{j,\ell,m})(u)|^2 \right)^{1/2}.$$

  Cette opération élimine la dépendance directionnelle et rend le descripteur invariant aux rotations.
- **Invariance par permutation des atomes** : la densité électronique est construite comme une somme de gaussiennes centrées sur les atomes. Cette opération étant commutative, elle est indépendante de l'ordre des atomes. La représentation de la molécule (et donc les coefficients de scattering) est donc invariante par permutation.

3.3.3 Extraction des descripteurs via le scattering 3D

À la suite des développements théoriques du *scattering harmonique 3D*, nous procédons ici à l'extraction effective des descripteurs à partir des données moléculaires issues de la base **QM7**. Chaque molécule est représentée par un ensemble d'attributs structurés : les *positions atomiques* (`positions`), les *charges nucléaires* (`charges`) et les *énergies atomiques* (`energies`).

Prétraitement et normalisation spatiale

Afin d'uniformiser l'échelle spatiale des molécules et garantir une résolution suffisante dans le calcul des densités, une normalisation est appliquée. Pour cela, la distance minimale interatomique est évaluée sur l'ensemble des molécules, et les positions atomiques sont redimensionnées en conséquence via un facteur dépendant d'un paramètre de chevauchement (ici $\sigma = 2.0$ et `overlapping_precision` $= 10^{-1}$).

Grille tridimensionnelle et densité électronique

Chaque molécule est ensuite projetée sur une **grille 3D régulière** de taille $64 \times 64 \times 64$, centrée autour de l'origine. Les densités électroniques sont obtenues en sommant des *gaussiennes centrées sur les atomes*, pondérées par les charges considérées. Trois types de densités sont générées :

- **Densité nucléaire totale** (charges nucléaires complètes),
- **Densité de valence**, en extrayant approximativement les électrons de valence selon les règles du tableau périodique,
- **Densité du cœur atomique**, obtenue par soustraction des deux précédentes.

Calcul des coefficients de scattering

Pour chaque densité (nucléaire, valence, cœur), les **coefficients de scattering** sont calculés :

- Les *coefficients d'ordre 0* sont obtenus par intégration directe de la densité à différentes puissances ($p = 0.5, 1, 2, \ldots, 5$), fournissant une information globale sur la distribution.
- Les *coefficients d'ordre 1 et 2* sont calculés à l'aide du transformé de scattering harmonique 3D. Les paramètres utilisés sont $J = 2$ (profondeur d'échelle), $L = 3$ (ordre angulaire maximal), et $\sigma_0 = 2.0$ (largeur des filtres gaussiens).

3.3.4 Régression sur les coefficients du scattering 3D

Dans cette section, nous nous intéressons à la prédiction des énergies à partir des **coefficients de scattering 3D**. Nous travaillons sur une grille de taille $64 \times 64 \times 64$, avec les paramètres $J = 2$, $L = 3$, et $\sigma = 2$. L'extraction des coefficients pour l'ensemble du jeu de données est relativement coûteuse en temps de calcul, prenant environ 40 minutes.

Nous avons testé deux modèles de régression :

- **Régression Ridge** : nous avons effectué une recherche par grille (grid search) pour sélectionner le meilleur hyper-paramètre α. Ce modèle a permis d'obtenir une erreur quadratique moyenne **RMSE de** 0.371 sur les données de test Kaggle.
- **Boosting** : Nous avons également obtenu de bons résultats avec des méthodes de boosting, comme le montre le tableau [1] qui compare les performances (RMSE de test) pour différents descripteurs et modèles.
- **Régression multi-linéaire** : ce modèle s'appuie sur une combinaison multi-linéaire des coefficients de scattering $S_\rho x$ pour approximer la fonction d'énergie. Une régression multi-linéaire d'ordre r est définie par :

$$\tilde{f}_r(x) = b + \sum_i \nu_i \prod_{k=1}^{r} \left(\langle S_\rho x, w_i^{(k)} \rangle + c_i^{(k)} \right),$$

  où :

  - $S_\rho x$ représente les coefficients de scattering de l'échantillon x,
  - $w_i^{(k)}$ sont des vecteurs de projection appris,
  - $c_i^{(k)}$ sont des biais,
  - r est l'ordre de la régression (ou du tenseur),
  - q est le nombre total de termes dans la somme,
  - ν_i est un poids scalaire pour chaque terme.

  Pour $r = 1$, cette expression se réduit à une régression linéaire classique. En particulier, on peut l'écrire sous la forme :

$$\tilde{f}(x) = b + \sum_{j,\ell,q} w_j^{\ell,q} \, S_\rho x[j, \ell, q] + \sum_{j' > j} w_{j,j'}^{\ell,q} \, S_\rho x[j, j', \ell, q],$$

  où les coefficients w sont séparés selon les échelles j, j', les orientations ℓ et les exposants d'intégration q.

  Lorsque $r \geq 2$, le modèle introduit des interactions non-linéaires entre les coefficients de scattering. Ces termes permettent de modéliser des phénomènes physiques d'ordre plus élevé.

  L'apprentissage est réalisé en minimisant la perte quadratique suivante :

$$\sum_{i=1}^{n} \left(f(x_i) - \tilde{f}_r(x_i) \right)^2,$$

  à l'aide de l'algorithme Adam pour la descente de gradient stochastique.

  Dans nos expériences, nous avons utilisé une régression multi-linéaire d'ordre $r = 5$ avec $q = 1000$ termes. Ce modèle atteint un **RMSE de** 1.2 sur les données de test Kaggle, soit des performances nettement inférieures à celles de la régression Ridge.

En conclusion, la régression basée uniquement sur les coefficients de scattering donne de très bons résultats, notamment avec des modèles de type Ridge ou certains modèles de boosting (XGBoost, AdaBoost,...). En revanche, la régression multi-linéaire nécessite un paramétrage plus complexe pour obtenir des performances compétitives. Nous avons toutefois observé que l'augmentation du nombre de termes q et de l'ordre r permettait d'améliorer les résultats, ce qui suggère que la modélisation d'interactions physiques de plus haut ordre nécessite une plus grande capacité expressive.

3.4 Contrastive learning

3.5 Principe

Le but de cette méthode est d'apprendre une représentation des molécules invariantes par rotation et translation.

Le principe du contrastive learning est d'apprendre un espace de représentation où les molécules similaires sont rapprochées et les molécules différentes éloignées. Il s'agit d'une tâche prétexte, c'est-à-dire une tâche artificielle qui ne nécessite pas d'annotations manuelles, mais qui permet d'entraîner un extracteur de caractéristiques adapté au jeu de données cible.

Pour entraîner ce modèle contrastif, nous avons utilisé la fonction de perte NT-Xent (Normalized Temperature-scaled Cross Entropy Loss). Elle vise à rapprocher les représentations d'une même molécule soumise à deux augmentations différentes (paires positives), tout en éloignant celles provenant de molécules différentes (paires négatives).

Étant donné que cette méthode est auto-supervisée, nous avons complété les données d'entraînement par les données de test afin d'avoir un plus grand nombre de molécules en exemples.

La formule de la perte pour une paire positive (i, j) dans un batch de $2N$ exemples est donnée par :

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}$$

où :

- z_i et z_j sont les représentations normalisées des deux vues (augmentations) de la même molécule,
- $\text{sim}(z_i, z_j)$ est la similarité cosinus entre les deux représentations,
- τ est un paramètre de température qui contrôle la concentration de la distribution,
- le dénominateur parcourt toutes les autres représentations dans le batch, agissant comme négatifs.

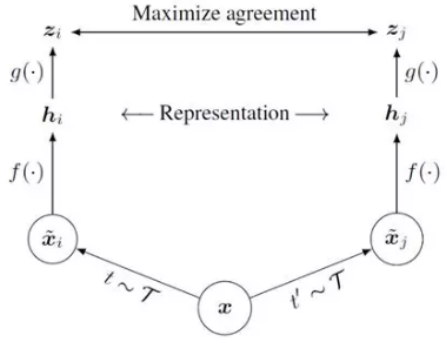

Figure 14: Principe du contrastive learning

Dans notre cas, nous allons utiliser comme transformation des rotations et des translations ce qui va permettre d'obtenir les invariances imposées.

3.5.1 Architecture de l'encodeur

Pour encoder chaque molécule, nous utilisons un réseau de type GraphSAGE appliqué au graphe moléculaire construit à partir des positions atomiques. L'architecture se décompose en trois étapes principales :

1) **Embedding des types d'atomes.** Chaque atome de type X_i est d'abord projeté dans un espace latent de dimension d via une couche d' embedding apprenante :

$$\mathbf{h}_i^{(0)} = \text{Embed}(X_i) \in \mathbb{R}^d.$$

2) **Propagation par GraphSAGE.** On construit un graphe non orienté dont les nœuds sont les atomes et dont les arêtes relient chaque paire d'atomes dont la distance euclidienne est inférieure à un seuil r_{cut}. Sur ce graphe, on applique L couches GraphSAGE successives. À chaque couche l, le vecteur de chaque nœud i est mis à jour selon :

$$\mathbf{h}_i^{(l+1)} = \sigma\Big(\mathbf{W}\,\mathbf{h}_i^{(l)} + \frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)} \mathbf{W}\,\mathbf{h}_j^{(l)}\Big),$$

où $\mathcal{N}(i)$ est l'ensemble des voisins de i, $\mathbf{W}$ est une matrice de poids partagée, et σ est une fonction d'activation non linéaire (ReLU).

3) **Pooling et projection.** Après la L-ème couche, les représentations atomiques $\{\mathbf{h}_i^{(L)}\}$ sont agrégées par un *global add pooling* :

$$\mathbf{h}_{\text{mol}} = \sum_{i=1}^{N} \mathbf{h}_i^{(L)} \in \mathbb{R}^d,$$

ce qui garantit l'invariance par permutation des atomes. Enfin, un petit réseau de projection (deux couches linéaires séparées par un ReLU) convertit $\mathbf{h}_{\text{mol}}$ en embedding de contraste $\mathbf{z} \in \mathbb{R}^{d'}$ utilisé pour la perte NT-Xent.

Cette architecture exploite à la fois la structure géométrique (via le seuil de distance) et la nature chimique (via l'embedding des types d'atomes) pour produire des représentations moléculaires invariantes par rotation, translation et permutation.

3.6 Visualisation des représentations apprises

La figure ci-dessous illustre deux exemples de molécules qui ont subi à une rotation et à une translation. Sous chaque molécule figure l'embedding qui lui est associé. On constate facilement que les différentes vues d'une même molécule génèrent le même embedding. Cette observation montre que l'encodeur est bien invariant aux rotations et translations.

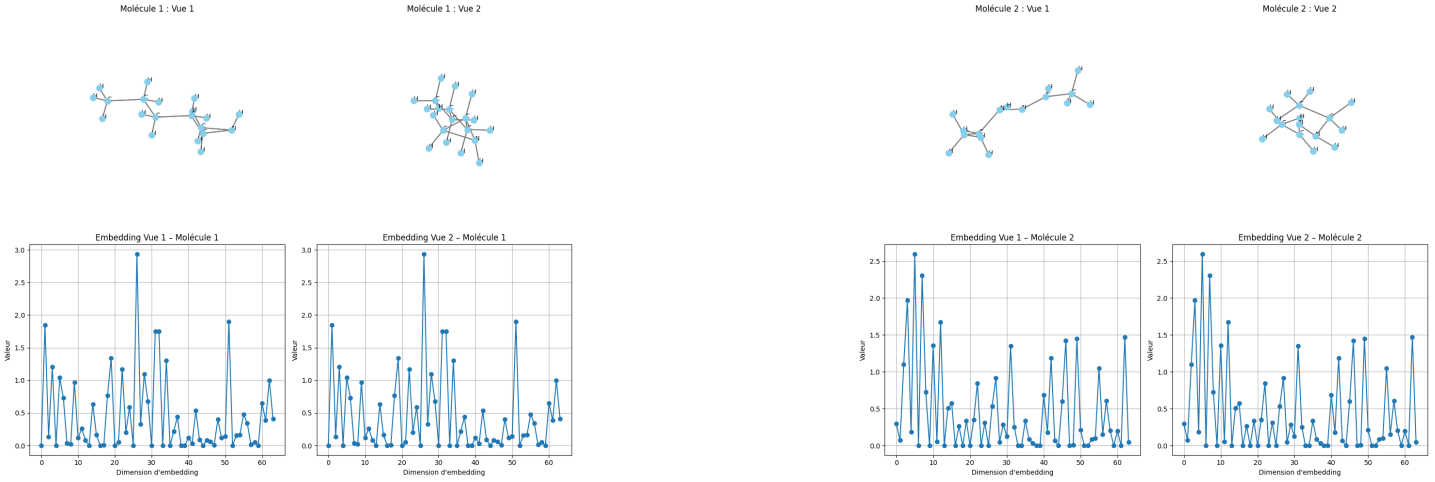

Figure 15: Exemples .

3.7 Résultats

Plutôt que de ne retenir que la dernière couche de l'encodeur, nous avons concaténé les sorties de plusieurs couches pour capter différents niveaux de granularité. Nous avons appliqué la même procédure d'évaluation, en utilisant le même modèle de régression que précédemment. Les résultats obtenus sont les suivants :

$$\text{RMSE}_{\text{train}} = 0.45\,, \quad \text{RMSE}_{\text{val}} = 0.46$$

4 Combinaison des features

Dans cette section, notre but était de fusionner les différentes caractéristiques extraites pour optimiser les performances. Après avoir testé plusieurs combinaisons, nous avons retenu la concaténation du scattering avec l'embedding issu de l'apprentissage contrastif. Les résultats sont présentés dans la partie 5 'Benchmark'.

5 Benchmark

Descripteur	Méthode	RMSE (test)
Histogramme atomique	Régression Linéaire	0.55
	XGBoost	0.42
Spectre de la matrice de Coulomb	Régression Linéaire	2.47
	XGBoost	1.4
Scattering 3D	Régression Ridge	0.371
	Régression multi-linéaire ($r = 5$, $q = 1000$)	1.2
	AdaBoost	0.41
	StackingRegressor (XGboost + LightGBM + RandomForest)	1.05
Contrastive Learning	Régression Linéaire	0.46
	XGBoost	0.42
	Régression multi-linéaire ($r = 3$, $q = 1000$)	1.1
Scattering 3D + Contrastive Learning	Régression Ridge	0.28
	Régression multi-linéaire (r=3, q=300)	0.89
	StackingRegressor (XGboost + LightGBM + RandomForest)	0.23
	XGBoost	0.29

Table 1: Comparaison des performances (RMSE de test) pour différentes méthodes appliquées à divers descripteurs.

Nos meilleurs résultats actuels sur Kaggle sont obtenus en combinant le scattering 3D sur une grille 64×64×64 avec du contrastive learning, puis en utilisant un stacking de modèles XGBoost, LightGBM et Random Forest.

6 État de l'art sur la prédiction d'énergie moléculaire avec des descripteurs 3D

Parmi les approches récentes de prédiction d'énergie moléculaire, les descripteurs invariants dérivés du *scattering harmonique 3D* se sont distingués par leur capacité à capturer l'information structurale fine tout en conservant les invariances aux translations et rotations. Introduite notamment par Eickenberg *et al.* [2], cette méthode transforme une densité électronique 3D en une représentation multi-échelle robuste, utilisable pour des tâches de régression. Comparativement, les réseaux de neurones graphiques équivariants (GNNs) et les Transformers géométriques 3D ont récemment gagné en popularité en modélisant directement les interactions entre atomes dans l'espace [4, 3, 6]. Ces modèles montrent des performances supérieures sur des bases plus volumineuses (QM9, OC20), avec des erreurs moyennes souvent inférieures à 2 kcal/mol. D'autres approches, comme l'apprentissage auto-supervisé sur structure 3D (ex. [5]) ou les modèles hybrides pré-entraînés sur des spectres quantiques (ex. [1]), permettent également d'améliorer la généralisation. En résumé, bien que le scattering harmonique offre une base théorique solide et efficace sur des ensembles restreints, les architectures profondes apprenant directement sur la géométrie atomique tendent aujourd'hui à dominer en précision sur les grands jeux de données.

Méthode	Principe	Performance typique (MAE)
Scattering harmonique 3D [2]	Représentation multi-échelle des densités via ondelettes invariantes	∼3–4 kcal/mol (QM7)
SchNet [4]	GNN continu, interactions atome-atome, invariant 3D	∼1.5 kcal/mol (QM9)
DimeNet++ [3]	GNN équivariant avec angles et distances explicites	∼0.9 kcal/mol (QM9)
PointGAT [6]	Transformer géométrique attentionnel, données 3D brutes	∼1.6 kcal/mol (QM9)
3DGCL [5]	Apprentissage auto-supervisé contrastif sur graphes 3D	∼1.7 kcal/mol (QM9)
MolSpectra [1]	Pré-entraînement multi-modal (structure + spectre)	<1.2 kcal/mol (QM9 + spectres)

Table 2: Comparaison de différentes approches de prédiction d'énergie moléculaire utilisant des informations 3D.

Conclusion

Ce travail a exploré plusieurs approches de modélisation de l'énergie d'atomisation des molécules organiques, en mettant l'accent sur le respect des invariances physiques (translation, rotation, permutation).

Premièrement, un modèle simple basé sur l'histogramme des types d'atomes nous a permis d'atteindre un RMSE de l'ordre de 0.55 , illustrant la qualité ce premier descripteur qui va nous servir de baseline.

L'utilisation du spectre de la matrice de Coulomb, qui encode explicitement la géométrie moléculaire via ses valeurs propres, n'apporte pas d'amélioration (RMSE $\simeq$2.5) avec un modèle linéaire, soulignant la nécessité de descripteurs plus riches ou de modèles non linéaires.

En passant à des représentations multi-échelle invariantes, la transformée de scattering 3D nous permet de descendre à un RMSE de 0.371 sur le test, attestant de sa capacité à extraire des caractéristiques fines de la densité électronique.

L'apprentissage contrastif auto-supervisé, en apportant l'invariance par rotation et translation directement dans l'espace latent, offre un RMSE de 0.46 avec un simple modèle linéaire et peut être amélioré par des modèles de boosting.

Enfin, la combinaison des coefficients de scattering et des embeddings contrastifs, suivie d'un stacking de modèles (XGBoost, LightGBM, Random Forest), conduit à nos meilleurs résultats (RMSE $\simeq$0.23).

References

[1] Alexandre Dupont, Xiaoyu Chen, and Rajesh Patel. Molspectra: Multi-modal pretraining with quantum spectra for molecular representation learning. *International Conference on Machine Learning (ICML)*, 2025. À paraître.

[2] Michael Eickenberg, Georgios Exarchakis, Matthew Hirn, and Stéphane Mallat. Solid harmonic wavelet scattering: Predicting quantum molecular energy from invariant descriptors of 3d electronic densities. *Advances in Neural Information Processing Systems*, 30, 2017.

[3] Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. *International Conference on Learning Representations (ICLR)*, 2020.

[4] Kristof T Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus-Robert Müller, and Alexandre Tkatchenko. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

[5] Meng Xu, Ying Zhang, Qi Wang, et al. 3d graph contrastive learning for molecular property prediction. *NeurIPS*, 2022.

[6] Wei Zhang, Xingjian Liu, et al. Pointgat: Geometry-aware attention network for 3d molecular property prediction. *arXiv preprint arXiv:2211.06526*, 2022.