5A ModIA

Project : Molecular Energy Prediction

Auteurs :
Rémi Colin
Issam El Kadiri

Superviseur :
Sixin Zhang

16 janvier 2025

1 Introduction

The goal of this project is to model the inter-atomic potential energy surface of small organic molecules.

2 Available data

We use a subset of QM7-X, which contains more than 6000 structures of molecules, with various number of particles (atoms). They are stored in a .xyz file format, that we manipulated using the python package ase. The xyz file format contains spatial information about the atoms contained in a given molecule, and the ase library also provide tools to compute some basic chemical propreties of a given molecule such as the atomic numbers of its atoms.

3 Method 1 : Coulomb Matrix

3.1 Description of the Coulomb Matrix

The Coulomb Matrix representation is a way to encode molecular structures for the prediction of atomization energies. This method utilizes the raw molecular geometry directly, embedding the spatial and atomic number information into a matrix format that captures the interatomic Coulombic interactions.

The Coulomb matrix, C, for a molecule is defined as follows :

$$C_{ij} = \begin{cases} 0.5 \times Z_i^{2.4} & \text{si } i = j \\ \frac{Z_i Z_j}{\|R_i - R_j\|} & \text{si } i \neq j \end{cases} \tag{1}$$

where Z_i and Z_j are the nuclear charges of atoms i and j, and R_i and R_j are their spatial coordinates.

— The matrix is symmetrical, since the Coulombic interaction is symmetrical.
— The diagonal elements represent the potential energy of the free atom.
— Off-diagonal elements represent Coulombic repulsion between pairs of nuclei.

A major problem with the Coulomb matrix is its lack of invariance to the permutation of atom indices, which can lead to an explosion in the dimensionality of the problem. We will see several approaches to overcome this issue in the following section of this report.

3.2 Handling Permutation Variance

To address the issue of permutation variance in the Coulomb matrix, we explored three approaches :

— **Eigenspectrum Representation** : Calculation of the eigenspectrum of the Coulomb matrix, which is invariant to permutations of atoms.

— **Sorted Coulomb Matrix** : Sorting the rows and columns of the Coulomb matrix by norm, ensuring a consistent ordering of atoms across different matrices.

— **Randomly Sorted Coulomb Matrices** : Introducing randomness in the sorting of the Coulomb matrix to generate a variety of representations for the same molecule, effectively augmenting the dataset. This last approaches provides the best results.

3.3 Computational Implementation

For each molecule in our dataset, a Coulomb matrix was computed using the described methods. The Coulomb matrix contains crucial details within its elements. Thus, using these real quantities directly can lead to optimization challenges. Instead, we transform each Coulomb matrix dimension into a three-dimensional tensor of binary predicates for better conditioning as follow :

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

We will use $\theta = 1$ in the following, as it is the value that gives the best results.

These tensors served as the input features for a multilayer perceptron and a Convolutional Neural Network to predict molecular atomization energies. Also, we handled the matrix dimension issue by appling zero padding for molecules which add fewer atoms to match the largest molecule in the dataset.

3.4 Model Training and Evaluation

The models were trained using a combination of kernel ridge regression, multilayer perceptron and convolutional neural network, with a focus on exploring how different representations of the Coulomb matrix affect prediction accuracy. The performance of these models was evaluated using cross-validation techniques and a MSE loss. The results also demonstrated that incorporating permutation invariance directly into the machine learning models led to significant improvements in prediction accuracy.

$$\text{MSE} = \frac{1}{D}\sum_{id=1}^{D}\Big(E(\text{rid}) - \tilde{E}(\text{rid})\Big)^2$$

where :

— D represents the total number of test configurations,

— E(rid) is the actual atomization energy for the molecule configuration identified by rid,

— $\tilde{E}$(rid) is the predicted atomization.

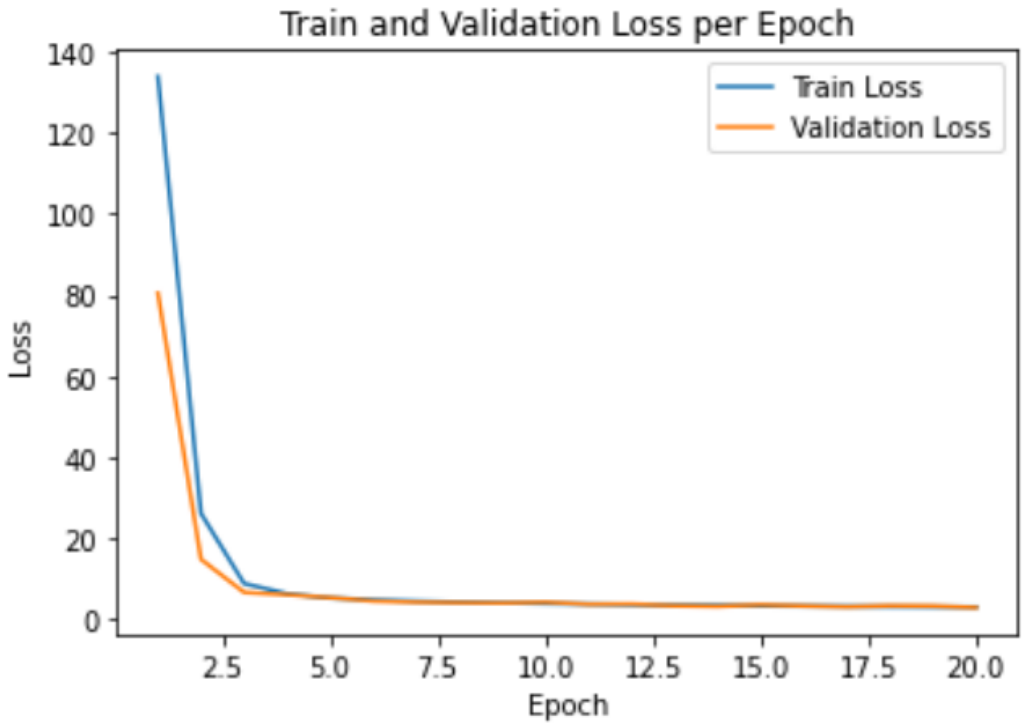

FIGURE 1 – Train and validation losses per epoch

The figure loss shows training and validation losses initially decreasing rapidly, indicating effective learning. After around 20 epochs, the training loss slows while the validation loss stabilizes, suggesting potential overfitting. This leads to a MSE score of 1.69 on the test datas.

3.5 Conclusion

The use of the Coulomb matrix as a molecular descriptor allowed for direct learning from the molecular structure without the need for hand-crafted features.

This method proved effective in capturing the essential quantum mechanical properties required for accurate energy predictions. Decreasing training and validation losses show our model learns well.

Yet, stabilized validation loss suggests limits to improvement without overfitting. Furthermore, even with cross validation, this approaches does not seem to be optimal. Fine-tuning parameters could be a good approache to optimize performance and generalization but we opted for another approache.

4 Method 2 : 3D wavelet scattering

We will now explore the 3D Wavelet Scattering method to enhance predictive accuracy and model generalization beyond the promising results of the Coulomb Matrix in modeling molecular energies.

4.1 Data preprocessing

The first step was to compute the maximum number of atoms contained in a molecule over the training and test dataset in order to standarize the format of each molecule. We found a maximum number of atoms of 23.

Each molecule is an element $x = \{(z_k, r_k) \in \mathbf{Z} \times \mathbf{R}^3\}_k$ where $\{z_k\}_k$ are the nuclear charges, and $\{r_k\}_k$ are the positions.

So each molecule is represented by a $(23, 3)$ matrix encoding positions, and a $(23,)$ array encoding charges. For molecules containing less than 23 elements (atoms), zero-padding was employed to ensure the coherence of this representation (?alongside ?) the training and test dataset.

4.1.1 Computing valence charges

Valence charges are the electrons in the outermost shell of an atom that participate in chemical bonding.

For each atom in each molecule, we compute its valence charges, deducing its core charges as well.

4.1.2 Normalisation of the positions

We then normalize the positions of the atoms. Specifically, the positions are rescaled so that two Gaussian functions, each with a width of sigma, placed at those positions will overlap with an amplitude that is less than the specified overlapping precision.

4.2 The 3D wavelet scattering process

4.2.1 Invariance

Since the ground-state energy of a molecule has the following invariance properties :

- Invariance to permutations energies do not depend on the indexation order k of each nuclei
- Isometry invariance Energies are invariance to rigid translations, rotations, and reflections of the molecule
- Deformation stability The energy is Lipschitz continuous with respect to scaling of distances between atoms.
- Multiscale interactions The energy has a multiscale structure, with highly energetic bonds between neighboring atoms, and weaker interactions at larger distances, such as Van-der-Waals interactions.

The 3D scattering method is invariant to translations and rotations [1] :

- Translation invariance : Convolutions with wavelets, inherently translation-invariant, capture features across different scales without changing their values when the signal is translated.
- Rotation invariance : The use of spherical harmonics, which are rotationally invariant functions, ensures that the wavelet coefficients remain consistent under rotations. By summing the energy over all indices, the transformation maintains rotation covariance, making it invariant to any rotation in 3D space.

In [4], Mallat proved that the scattering method is Lipschitz continuous to deformations.

The representation of the electrical density we use is invariant to permutations [1].

The multiscale interactions proprety is achived by computing order_0 and order_1 scattering coefficients.

4.2.2 The process

Given the rescaled positions and charges, we computed the density maps by placing Gaussians at the different positions weighted by the appropriate charge. These are fed into the 3D solid harmonic scattering transform to obtain features that are used to regress the energies.

The steps for this are :

- we first define a grid.
- We then define the scattering transform using the HarmonicScattering3D class

— The maps computed for each molecule are quite large, so the computation has to be done by batches. Here we select a small batch size to ensure that we have enough memory when running on the GPU. Dividing the number of molecules by the batch size then gives us the number of batches.

We are now ready to compute the scattering transforms. In the following loop, each batch of molecules is transformed into three maps using Gaussians centered at the atomic positions, one for the nuclear charges, one for the valence charges, and one with their difference (called the "core" charges). For each map, we compute its scattering transform up to order two and store the results.

4.3 Regression of energy

We apply a linear regression with Tikhonov, regularization pipeline on the concatenated zero-th and 1-2th order scattering coefficients and we evaluate the performance of the regression using four-fold cross-validation. The best regularization parameter we found was $\alpha = 10^{-10}$

We finally apply the same preprocessing steps on the test dataset, compute the scattering coefficients of its molecule, to predict their energy.

The linear regression model gave a public MSE of 0.0038. By using boosting (adaboost), we obtained a public MSE of 0.0030.

We get a final score of 0.0034 (MSE).

5 Perspective

5.1 State of the Art

In molecular energy prediction, the **Coulomb Matrix** method addresses permutation sensitivity via eigenspectrum and sorted matrices to enhance robustness as explain in [5]. The paper [1] argue that **3D Wavelet Scattering** leverages spherical harmonics and 3D wavelets for translational and rotational invariance, improving accuracy and generalization. As explained in [3],**Atomic Vectors combined with Transfer Learning** accelerate learning and enhance prediction by leveraging pre-trained models on specific tasks. Moreover, **Physics-Informed Neural Networks (PINNs)** ([2])incorporate physical laws into neural training, ensuring predictions conform to physical principles, beneficial in data-scarce scenarios.

5.2 Future work

In future extensions of this project, several approaches could be explored to enhance the predictive accuracy and robustness of molecular energy models. These include :

— Data Augmentation : Employing techniques to expand the training dataset, such as synthetic data generation.
— Transfer Learning : Utilizing pre-trained models on large datasets to accelerate learning and improve prediction capabilities for molecular properties.
— Advanced Representations : Investigating novel molecular structure representations.
— Other Neural Network Architectures : Experimenting with other neural network architectures could capture other data dynamics.

These strategies aim to further refine the models and push the boundaries of current predictive techniques in molecular modeling.

6 Conclusion

In this project, we modeled the inter-atomic potential energy surface of small organic molecules using two distinct methods : the Coulomb Matrix and 3D Wavelet Scattering Transform.

Coulomb Matrix Method :

— We successfully encoded molecular structures into matrices capturing interatomic Coulombic interactions.
— Permutation variance was mitigated using eigenspectrum representation, sorted matrices, and randomly sorted matrices, with the latter providing the best results.
— Applied transformations improved prediction accuracy but showed limited scope for further optimization.

3D Wavelet Scattering Transform :

— Standardized molecular data to ensure uniform representation, followed by computation of valence and core charges.
— Achieved translation and rotation invariance, crucial for accurate molecular energy predictions, through the use of spherical harmonics and solid harmonic wavelets.
— Computed scattering transforms to extract invariant features, enabling efficient regression. Performance

Linear regression and Tikhonov regularization pipelines were employed, yielding a public MSE of 0.0038. Boosting with AdaBoost improved the public MSE to 0.0030. The final score was a MSE of 0.0034.

Overall, the 3D Wavelet Scattering Transform demonstrated superior performance and generalization capabilities compared to the Coulomb Matrix approach, highlighting its potential for more accurate and robust energy predictions in molecular modeling. Future work may focus on further optimization and exploring additional machine learning techniques to enhance prediction accuracy.

Références

[1] Michael Eickenberg, Georgios Exarchakis, Matthew Hirn, and Stephane Mallat. Advances in neural information processing systems. volume 30. Curran Associates, Inc., 2017.

[2] Weiqiang Fu, Yujie Mo, Yi Xiao, Chang Liu, Feng Zhou, Yang Wang, Jielong Zhou, and Yingsheng Zhang. Enhancing molecular energy predictions with physically constrained modifications to the neural network potential. *Journal of chemical theory and computation*, 20, 06 2024.

[3] Jianing Lu, Cheng Wang, and Yingkai Zhang. Predicting molecular energy using force-field optimized geometries and atomic vector representations learned from an improved deep tensor neural network. *Journal of Chemical Theory and Computation*, 15(7) :4113–4121, 2019. Epub 2019 Jun 12.

[4] Stéphane Mallat. Group invariant scattering, 2012.

[5] Grégoire Montavon, Katja Hansen, Siamac Fazli, Matthias Rupp, Franziska Biegler, Andreas Ziehe, Alexandre Tkatchenko, Anatole Lilienfeld, and Klaus-Robert Müller. Advances in neural information processing systems. volume 25. Curran Associates, Inc., 2012.