Prediction of molecular energy

Machine learning under physical constraints: Kaggle projet

Table of Contents.

01 > OBJECTIVE AND DATASET

We will talk about the objectives of the project.

02 > DATA MANIPULATION

We will talk about how we manipulated datas.

03 > Coulomb matrix

Then, we will talk about the Coulomb Matrix representation to predict atom energy.

04 > 3D wavelet scattering

After that we will talk about the wavelet scattering approach.

05 > Perspective

We will then talk about the perspective of the project.

06 > Conclusion

And we will conclude.

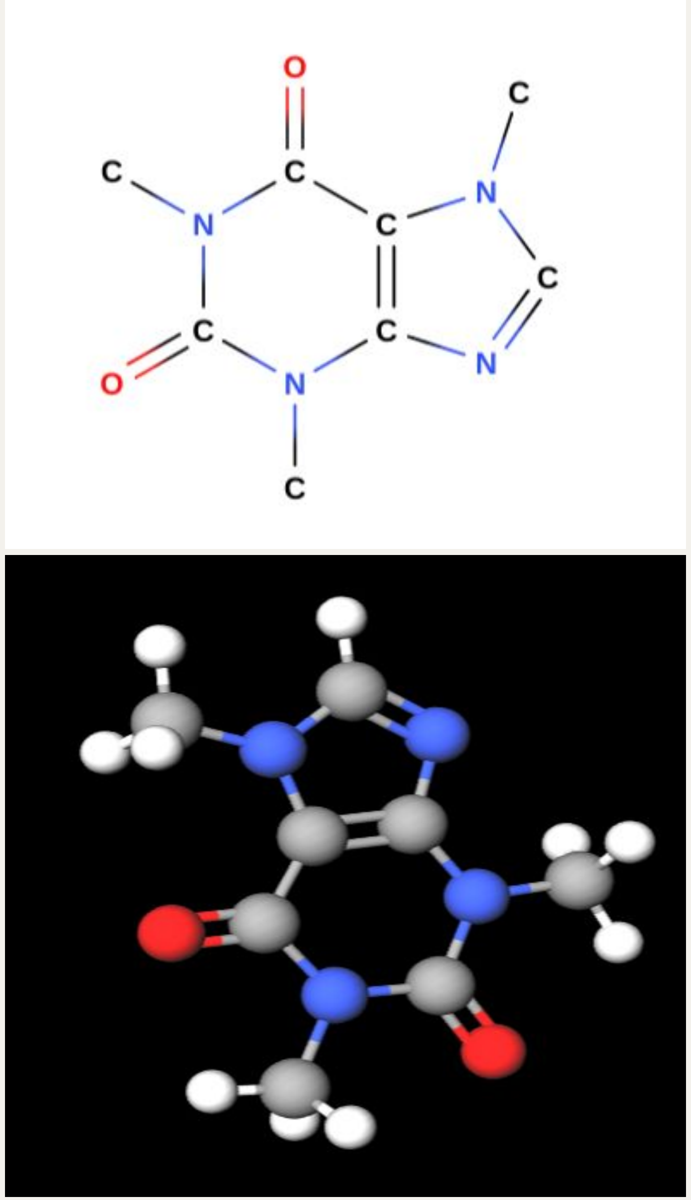
01

Objective and dataset

Objective and dataset

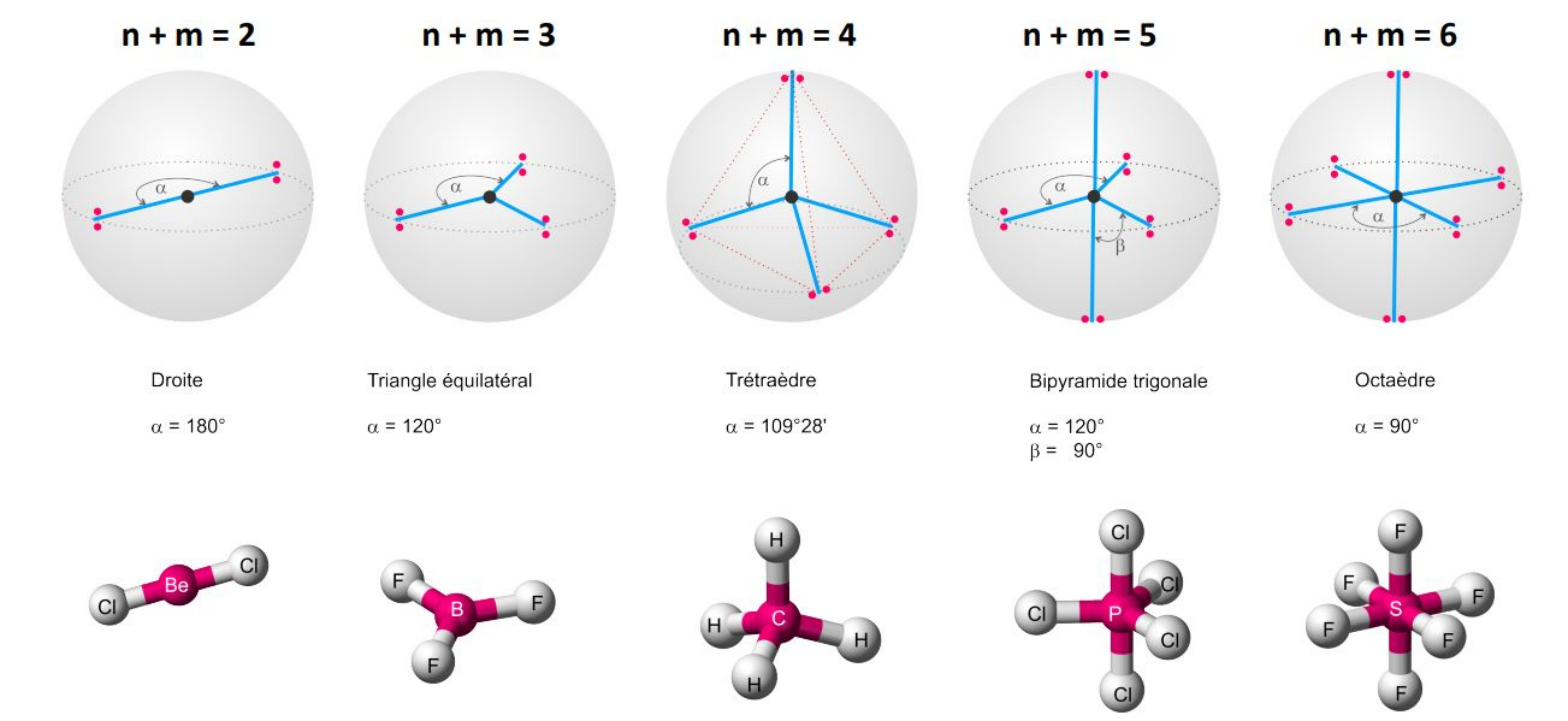
- Predict the molecular energy in 3d space based on its geometric structure.

- Dataset: subset of QM7-X, which contains more than 5000 structures of molecules, with various number of particles (atoms).



A molecule and its 3D representation (molview.org)

VSEPR Model

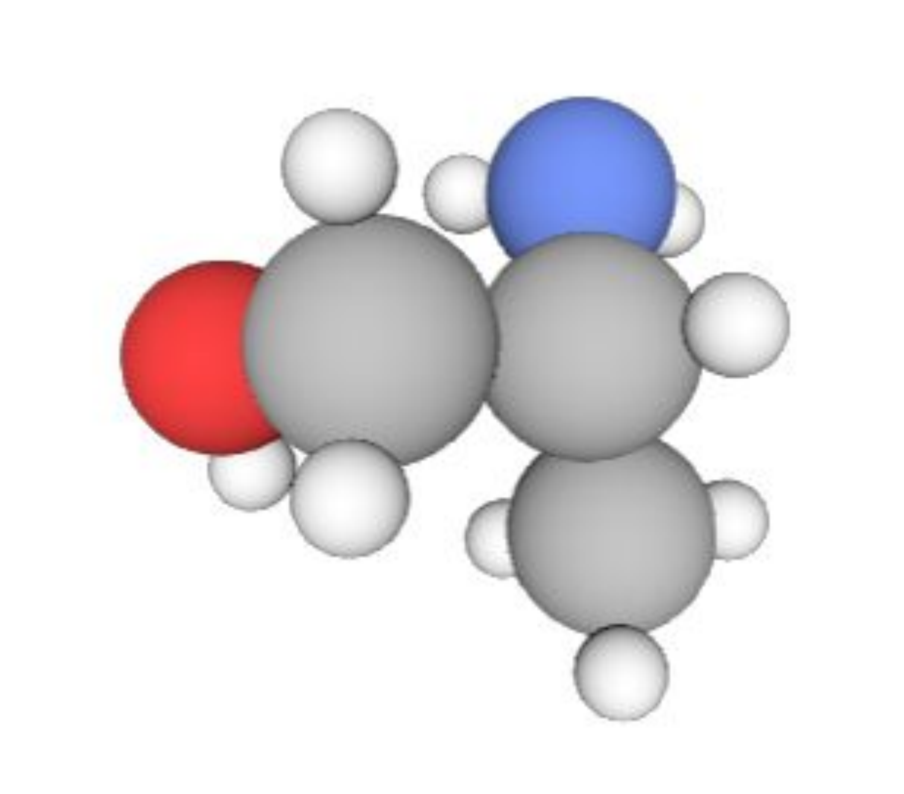

n + m = 2
n + m = 3
n + m = 4
n + m = 5
n + m = 6
α
α
α
α
β
α
Droite
α = 180°
Triangle équilatéral
α = 120°
Trétraèdre
α = 109°28'
Bipyramide trigonale
α = 120°
β = 90°
Octaèdre
α = 90°
Cl
Be
Cl
F
B
F
F
H
H
C
H
H
Cl
Cl
Cl
P
Cl
Cl
F
F
F
S
F
F
F

2

Data manipulation

Python package ase

- We used ase python package to manipulate our .xyz files.
- With ase, we can extract information about the positions of atoms in each molecules, atomic numbers, chemical formula...
- Visualization is easy with ase



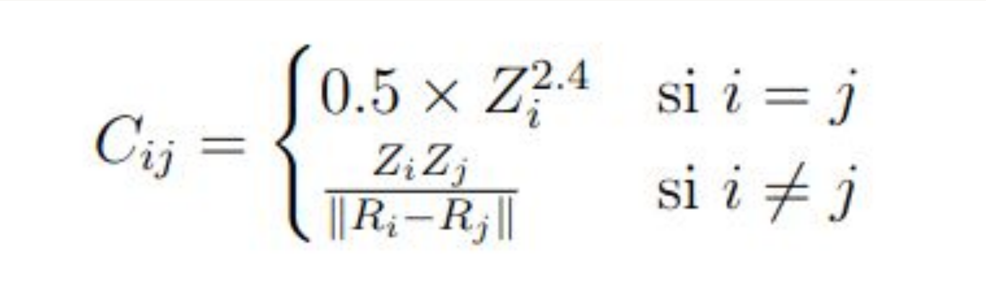
Molecule 3D visualization with ase

3

Method 1: Coulomb matrix

Coulomb matrix description

The Coulomb Matrix is a new way to encode molecular structures for the prediction of atomization energies.



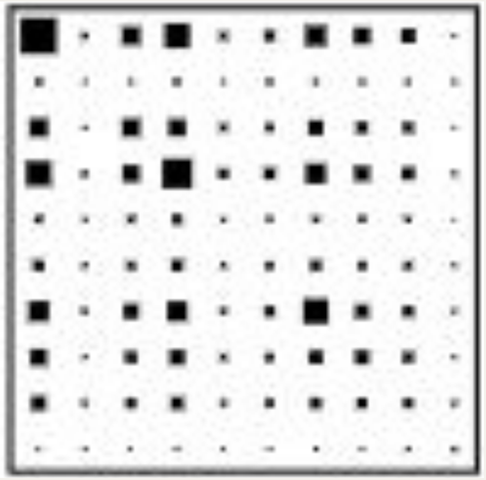
$$C_{ij} = \begin{cases} 0.5 \times Z_i^{2.4} & \text{si } i = j \\ \frac{Z_i Z_j}{\|R_i - R_j\|} & \text{si } i \neq j \end{cases}$$

With Z the nuclear charges of atoms and R their spatial coordinates.

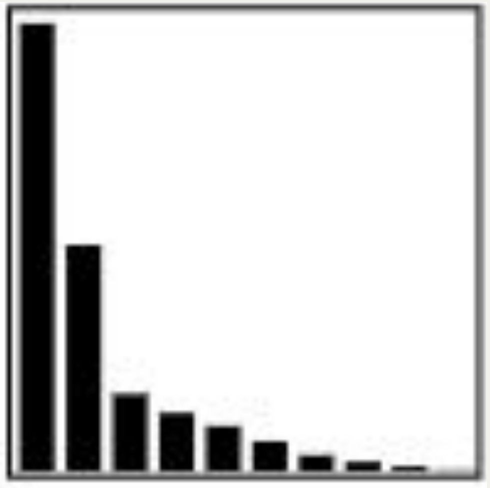
- The matrix is symmetrical, since the Coulombic interaction is symmetrical.
- The diagonal elements represent the potential energy of the free atom.
- Off-diagonal elements represent Coulombic repulsion between pairs of nuclei.

Handling Permutation Variance

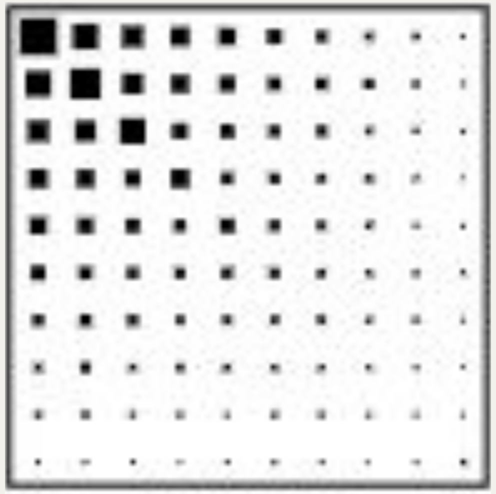
Coulomb Matrix



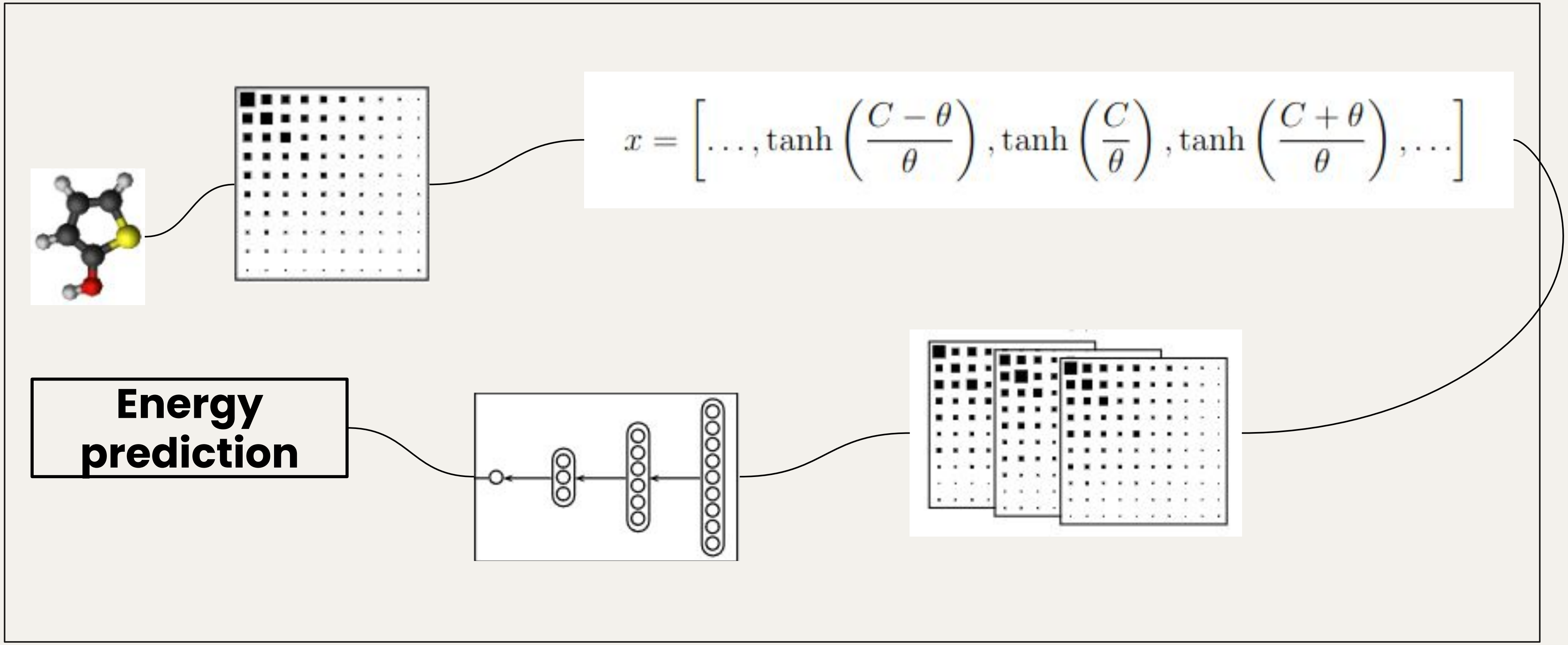
Eigenspectrum Representation



(Randomly) Sorted Coulomb Matrices



Computational implementation

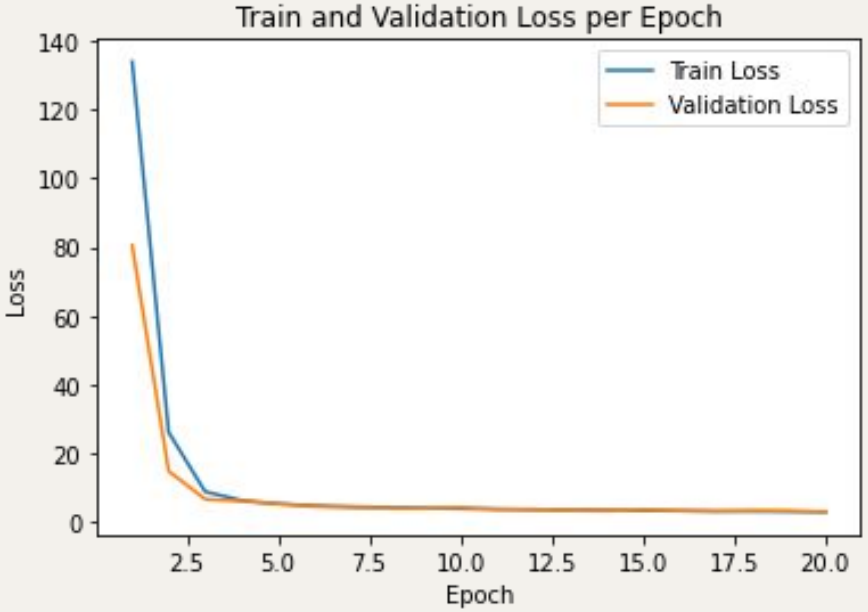


$$x = \left[\ldots, \tanh\left(\frac{C-\theta}{\theta}\right), \tanh\left(\frac{C}{\theta}\right), \tanh\left(\frac{C+\theta}{\theta}\right), \ldots\right]$$

Energy prediction

Model training and results

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



Loss on test:
1.69

4

Method 2: 3D wavelet scattering

General description

- 3D wavelet scattering is an advanced mathematical approach that combines wavelet theory and scattering transforms to analyze and process data with three-dimensional structures, such as molecular configurations.
- This method is particularly relevant in predicting molecular energy.
- Pioneered by Michael Eickenberg et al. in **2017** for 3D molecular structures, and by Matthew Hirn and Stéphane Mallat et al. in **2016** for planar molecules.
- The wavelet theory itself was pioneered by french scientists : Jean **Morlet**, Alex **Grossmann**, Yves **Meyer** and Stéphane **Mallat** himself.

Data preprocessing: Positions

- We then normalize the positions of the atoms:
- Each atom's position is represented by a Gaussian function, which spreads out from the atom's center

Data preprocessing

- For each molecule, we compute valence charges and deduce core charges,
- We use Gaussian to modele these electrical densities, weighted by the number of electrons at atom location rk.

$$\rho_x(u) = \sum_k \gamma_k \, g(u - r_k),$$

- Core and valence densities are obtained by setting γk to be the number of core electrons or the number of valence electrons of atom k

$$\rho_x^{\text{total}}(u) = \rho_x^{\text{core}}(u) + \rho_x^{\text{valence}}(u).$$

Invariance properties

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

1. Permutation Invariance: Energy levels remain constant regardless of the sequencing of nuclei indices.
2. sometry Invariance: Molecular energies are consistent under rigid body transformations such as translations, rotations.
3. Deformation stability The energy is Lipschitz continuous with respect to scaling of distances between atoms
4. Multiscale interactions The energy has a multiscale structure

Invariance properties conserved

1. The representation of the electrical density we use is invariant to permutations.
2. The 3D scattering method is invariant to translations and rotations.
3. Mallat proved that the scattering method is Lipschitz continuous to deformations.
4. The multiscale interactions property is achieved by computing order 0 and order 1&2.

Scattering coefficients

After the preprocessing steps, we compute order 0 and order 1&2 scattering coefficients that we use in the energy regression step.

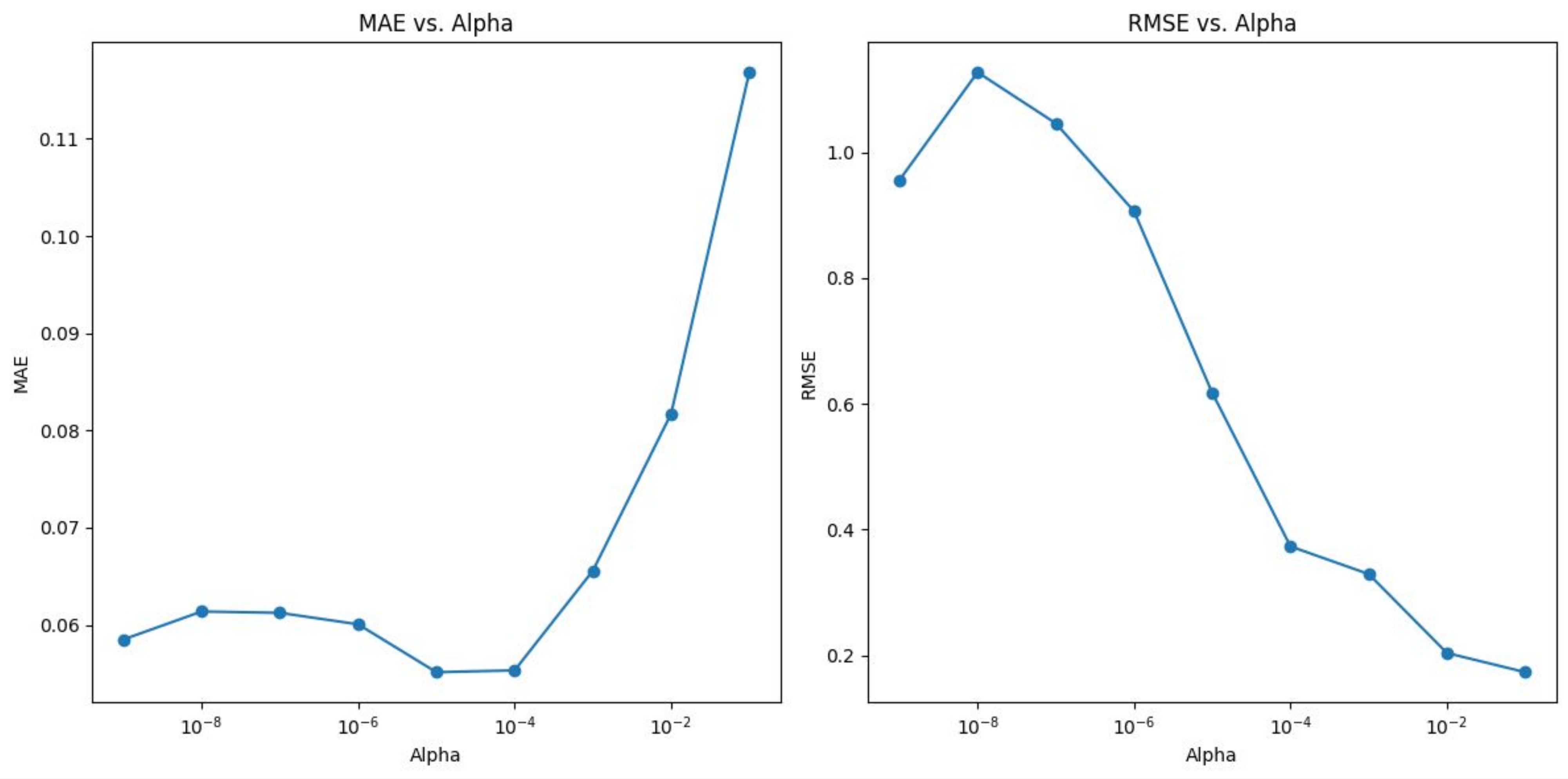
Regression : Linear model

We apply a linear regression with Tikhonov (parameter alpha), regularization pipeline on the concatenated 0th and 1-2th order scattering coefficients and we evaluate the performance of the regression using four-fold cross-validation.

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

We obtained better results with a simple linear model that a bilinear model : a public MSE score of approx. 0.0038.

MAE and RMSE as a function of alpha



Results

Method	Linear model	Boosting of linear model (adaboost)
Public score (MSE)	0.0038	0.0031

5

Perspective

State of the art | Future work

Coulomb Matrix approach

3D Wavelet scattering

Atomic vectors + Transfert learning

PINNs

Data Augmentation

Transfert Learning

Advanced Representations

Other Neural Network architectures

Algorithmic Improvements

6

Conclusion

Conclusion

Prediction of molecular energy under physical contraints

Coulomb Matrix

3D Scattering

References

(1) Project Report:
R. Colin, I. El Kadiri

(2) Coulomb Matrix representation article:
https://proceedings.neurips.cc/paper_files/paper/2012/file/115f89503138416a242f40fb7d7f338e-Paper.pdf

(3) 3D Wavelet Scattering representation:
https://arxiv.org/pdf/1805.00571

(4) Atomic vectors approach:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6615995/

(5) PINN 1:
https://www.researchgate.net/publication/381124452_Enhancing_Molecular_Energy_Predictions_with_Physically_Constrained_Modifications_to_the_Neural_Network_Potential

(6) PINN 2:
https://pubs.acs.org/doi/10.1021/acs.jctc.3c01181