MU4PY115: Artificial Intelligence for Physics

Molecular Dynamic : Data Analysis on Zundel and Mo2S4 molecules

Abstract

This report discusses a project that uses ab initio molecular dynamics simulations to create predictive models for molecular energy levels. Two datasets are examined: one from simulations of the Zundel ion (H2O-H-H2O) and another from simulations of a Molybdenum-Sulfur (Mo2S4) aggregate, each comprising approximately 10,000 atomic configurations, along with their respective potential energy values. The main objective is to integrate this intricate physical data into models capable of accurately predicting energy levels. To achieve this, Linear and Neural Network models are employed alongside with Principal Component Analysis (PCA). While the results are promising for the Molybdenum-Sulfur dataset, challenges arise due to the lack of suitable descriptors for the Zundel ions, resulting in less favourable outcomes.

Lucas BOISTAY

December 2023

Contents

Introduction

In recent years, the scientific community has shifted its focus towards Molybdenum-Sulfur (MoS) materials, owing to their diverse applications, ranging from hydrodesulfurization catalysts to cutting-edge transistors. This class of materials presents intriguing variations, from monolayers to intricate nanostructures such as inorganic fullerenes and nanoplatelets.

Research is now focusing on understanding the intricate formation mechanisms of MoS nanostructures and their interplay with extended forms. To unravel the complexities surrounding these materials, extensive experimental and theoretical investigations have been conducted, leading to significant insights ([1]).

In contrast to conventional approaches, which often rely on heuristic reasoning, a novel sampling strategy has emerged, spearheaded by ab initio molecular dynamics (MD).

Another interesting part of this analysis is the quantum description of Zundel ions energies, simulated and modelled with Variational Quantum Monte Carlo with Path Integral Langevin Dynamics ([2]).

This report presents a project that uses results from ab initio MD simulations to build a model for predicting the energy of certain molecular configurations. Here two data sets will be analysed: one from simulations of the Zundel ion (H2O-H-H2O) and another from simulations of a Molybdenum-Sulfur (Mo2S4) aggregate. Each data set has about 10,000 atomic configurations and a file with the potential energy for each configuration. The aim is to blend this complex physical data into a model that can predict energies accurately. This introduction outlines our approach to handling the data, creating the model, and checking how well it predicts in these complex molecular scenarios.

1 Dataset Description and Preprocessing

1.1 Overview of the data

1.1.1 File content

The atomic configurations are stored in a *.xyz* file, and every configuration is represented like :

```
6 # Number of atom in the molecule configuration
 generated by VMD
 Mo        0.746648        1.869656        0.279349 # Element and x, y and z position of the atom
 Mo       -1.188737        0.558299       -0.089274
 S         1.795806        0.234278       -1.878226
 S        -1.139974        2.551034        1.381834
 S        -0.410638        2.128456       -1.909369
 S         1.230796       -2.537868        1.275581
```

Listing 1: First Mo2S4 configuration

The same listing is present in the Zundel ion file for one configuration :

```
7
Properties=species:S:1:pos:R:3 Zundel=T ion=T with=T CCSD(T)=T potential=T pbc="F F F"
O        2.24643327       0.01217091       0.00847437
O       -2.25776180       0.00115926       0.00268762
```

```
H         0.00790166       -0.00138440        0.12917076
H        -3.01260594        1.47988102       -0.73707970
H        -3.20407528       -0.47042698        1.48716647
H         2.99892402       -1.48736113       -0.74511091
H         3.20703656        0.45930786        1.50818372
```

Listing 2: First Zundel configuration

Because the atomic element number will not be taking care of, and only the data on each different molecule separately will be analysed, the properties are not taken into account.

The total number of components for one configuration of Mo2S4 is $3 \times 6 = 18$ and $3 \times 7 = 21$ for Zundel ions. Multiplied by the number of available configurations, and $11,001$ configuration $\times 18 = 198,018$ components for Mo2S4 and $10,000 \times 21 = 210,000$ components for Zundel ions.

1.1.2 3D Visualisation

From the 3D coordinates of a configuration, it is possible to plot the molecule :

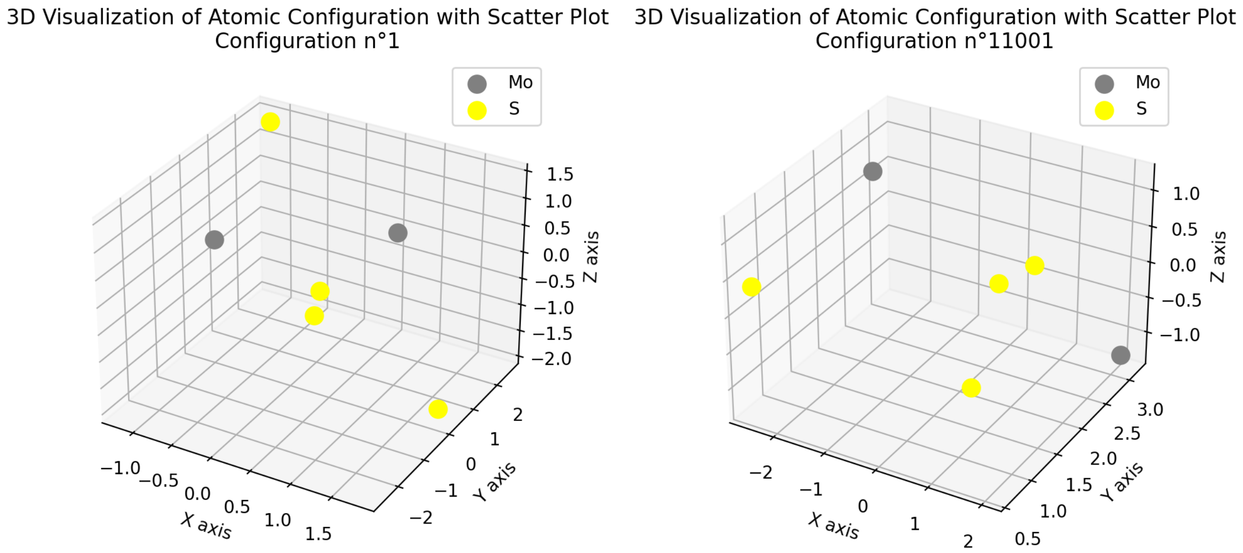

Figure 1: 3D Visualization of Atomic configuration with the coordinates in *.xyz* file.

As no further information are available with this data, the units of the positions are in the order of magnitude of the Å$(10^{-10}$m). Another way of visualising the atomic configuration is to use the *Jmol* software. Here are the two previous configurations plotted with *Jmol* :

The *Jmol* positions of atoms can be seen in the Figure 1, the 3D visualisation is working. Again for Zundel ions :

For the energies, there are available in another file with the extension *.out*. For Mo2S4, two separated files give 11,001 energies, but after a quick verification, they give the same energies for each configuration. Every line of energy correspond to the same index configuration. For Zundel, only one file gives the energies.

Again, even if the units are not given, the energies are in the order of magnitude of eV for Mo2S4 [1], so the energies units would be around 10^2 eV for Mo2S4. No direct order of magnitude was found for Zundel ions but we can consider a unit around the Hartree (cf [2]).

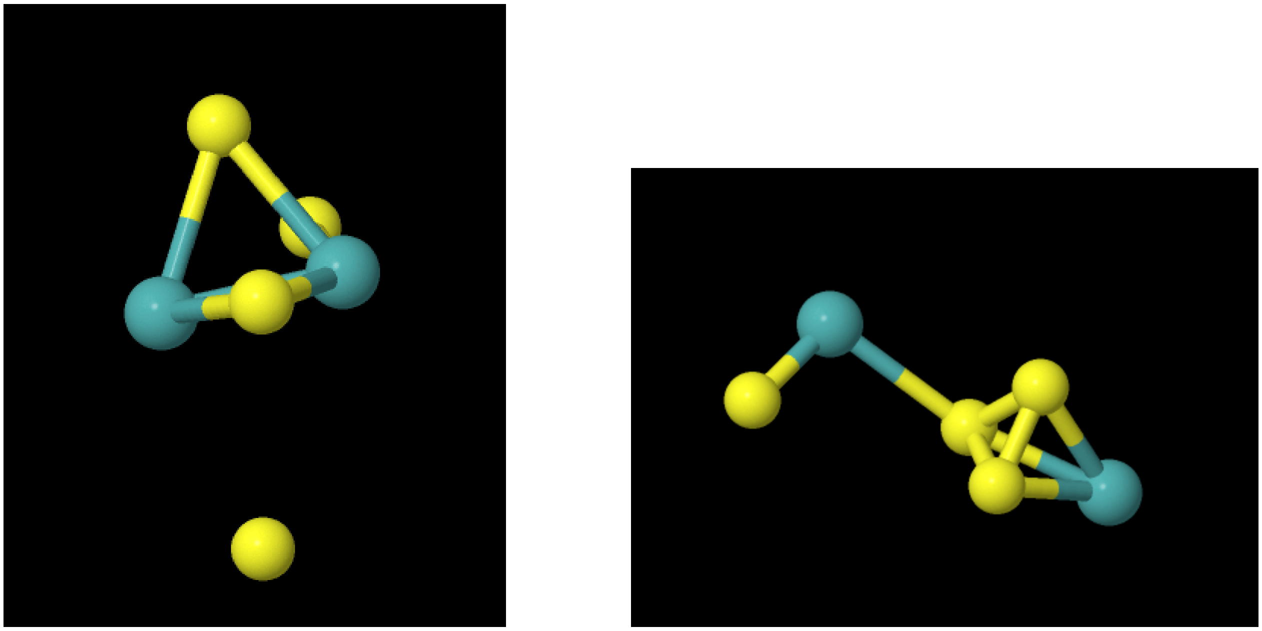

Figure 2: First (a) and last (b) configuration plot with *Jmol*. Molybdenum atoms are represented in blue and Sulfur atoms in yellow.

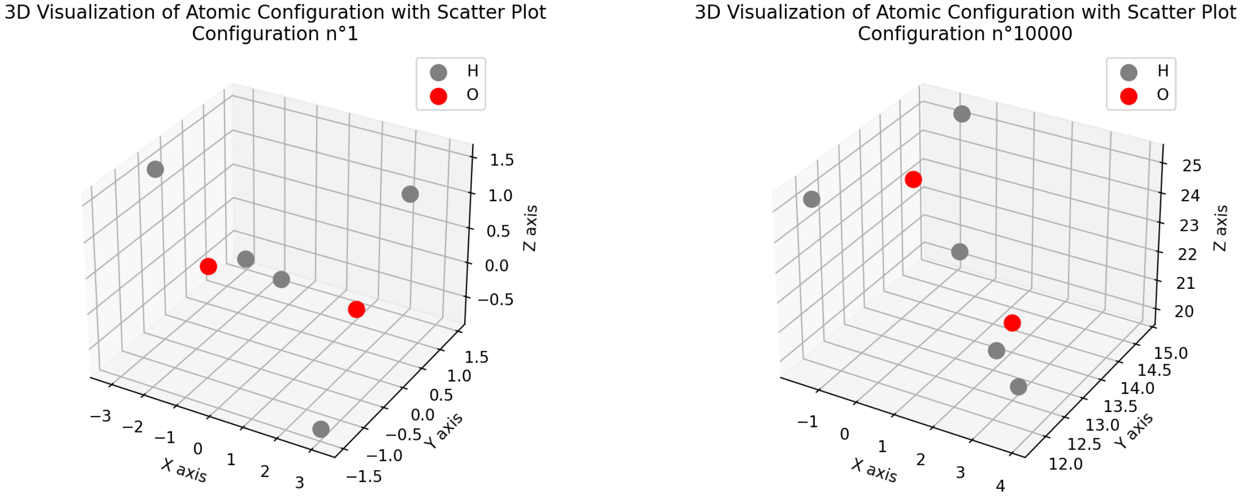

Figure 3: 3D Visualization of Atomic configuration with the coordinates in *.xyz* file.

1.2 Analysing datasets

Now, let us dig into the data to check how they are distributed. First, there are the energies for Mo2S4 and Zundel ions.

From Table 1, the maximum energies for Mo2S4 are way further from the mean than the rest of the data. Further investigation will be done later to see the impact on the analysis, but from now a separated set of data will be created with only the energies distant of 1σ from μ.

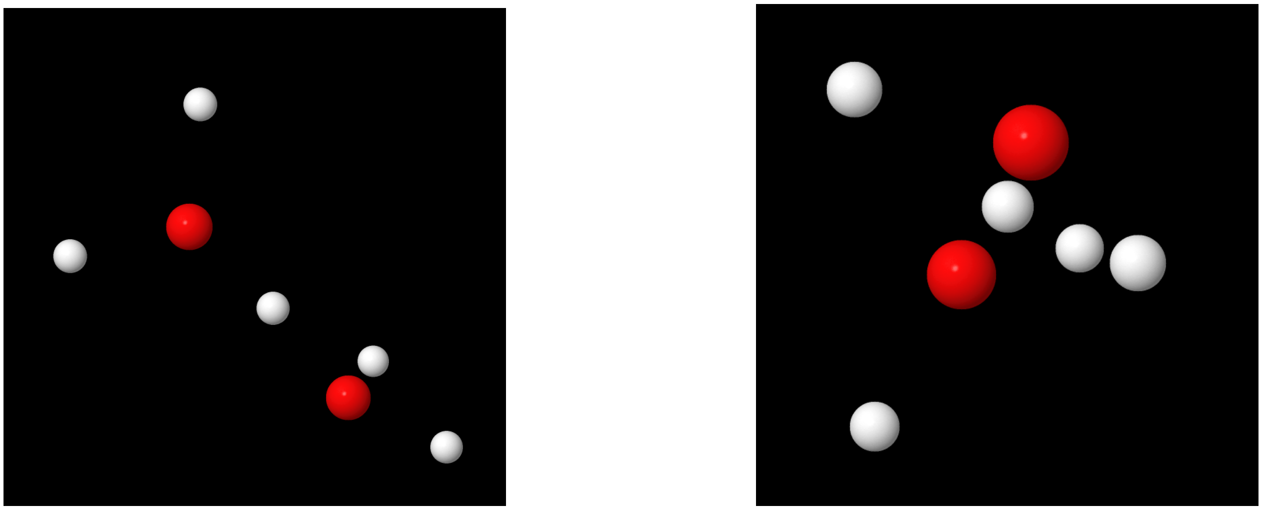

Figure 4: First (a) and last (b) configuration plot with *Jmol*. Hydrogen atoms are represented in white and Oxygen atoms in red.

Energies	Mo2S4	Zundel ions
Count	11,001	10,000
Mean (μ)	-174.78	2.54e-03
Std (σ)	32.38	9.10e-04
Min	-177.23	9.59e-05
25%	-176.91	1.88e-03
50%	-176.65	2.43e-03
75%	-176.29	3.08e-03
Max	633.49	7.21e-03

Table 1: Dataset summary for energies. Again, the units are not defined here.

The same analysis of the positions for Zundel and Mo2S4 was done[1], and no real issues were found with the distribution (the max/min positions being $\sim 2\sigma$ for Zundel and Mo2S4).

1.3 Defining Objectives and Evaluation Metrics

The objectives of this data analysis will be to predict the energy associated with a configuration (position of each atom) for one defined molecule (here either Zundel ions or Mo2S4). This is a regression on the energies of the molecule.

1.3.1 Variable Descriptions and Definitions

Let's represent the data. Let's say N is the number of configuration and M the number of atom inside a configuration. X is the vector representing all the positions for every configuration $X^{(i)}$ for $i \in [1, N]$. $X^{(i)}$ is then represented with $3 \times M$ components : x, y and z for each atom represented by

[1] For obvious visualisation matters, the table for all values of 18 and 21 components are not shown here. Another way to see the distances would have been to give the value of $r = |\vec{r}| = \sqrt{x^2 + y^2 + z^2}$ but this was still 13 values to check. We encourage you to check the *main.ipynb* file to see the analysis on the coordinates.

a number from 1 to M. $x_j^{(i)}$ representing the x coordinates of the j^{th} atom in the i^{th} configuration. For y, the energy of each configuration $E^{(i)}$

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \\ \vdots \\ X^{(i)} \\ \vdots \\ X^{(N-2)} \\ X^{(N-1)} \\ X^{(N)} \end{pmatrix} \quad \text{and} \quad y = \begin{pmatrix} E^{(1)} \\ E^{(2)} \\ \vdots \\ E^{(i)} \\ \vdots \\ E^{(N-1)} \\ E^{(N)} \end{pmatrix} \tag{1}$$

$$\text{with} \qquad X^{(i)} = \begin{pmatrix} x_1^{(i)} & y_1^{(i)} & z_1^{(i)} & x_2^{(i)} & \cdots & z_{M-1}^{(i)} & x_M^{(i)} & y_M^{(i)} & z_M^{(i)} \end{pmatrix} \tag{2}$$

Each dataset have a different unit and range, but a Standard Scale operation will be done to the input data X.

1.3.2 Metrics

The metrics used here will be the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE).

Another way to evaluate how good the data is will be to give the Pearson's correlation coefficient[2], as done in [3], defined by :

$$r_{pe} = \frac{\sum_{i=1}^{n}(y^{(i)} - \bar{y})(\tilde{y}^{(i)} - \bar{\tilde{y}})}{\sqrt{\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2 \sum_{i=1}^{n}(\tilde{y}^{(i)} - \bar{\tilde{y}})^2}} \tag{3}$$

with $\tilde{y}$ the predicted value of y, $\bar{y}$ the mean of y and $\bar{\tilde{y}}$ the mean of the predicted y. This coefficient can be interpreted as the R coefficient for a linear regression between the predicted value of y ($\tilde{y}$) and the real y. The data will be splited into a training set and a test set[3].

However, the paper [3] explains that the direct positions of atoms are not a good representation of the data. A way more convenient way to describe it are Coulomb Matrices.

1.4 Coulomb Representation

1.4.1 Coulomb

The Coulomb matrix encodes the atomic species and interatomic distances of a finite system in a pair-wise, two-body matrix inspired by the form of the Coulomb potential. The elements of this matrix are given by :

$$M_{ij} = \begin{cases} \frac{1}{2} Z_i^{2.4} & \text{if } i = j \\ \frac{Z_i Z_j}{|R_i - R_j|} & \text{if } i \neq j \end{cases}$$

[2] Here, we are not using the Pearson coefficient as a metric for the loss function, it's just a hint for how good our model is.

[3] Test set is size $0.1N$ and train set size $0.9N$.

where Z_i is the atomic number of the atom i and $|R_i - R_j|$ is the Euclidean distance between atoms i and j.

This matrix size is $M \times M$, then 36 for Mo2S4 and 49 for Zundel ions.

The Coulomb Matrix is supposed to be a way better representation of a molecular configuration, as it contains information about the Coulomb potential and the distance between each atom. The *Dscribe* [4] package was used to create Coulomb Matrix.

Both these representations will be used for our machine learning.

1.4.2 Other representations

Other representations commonly used in MD are the Many-Body Tensor Representation (MBTR) or the Smooth Overlap of Atomic Positions (SOAP), which is a local representation that can be used to describe the local environment of atoms in a system. It is based on the idea that the local environment around an atom can be expressed as a sum of spherical harmonics around that atom. The SOAP descriptor is a histogram of the spherical harmonics coefficients.

However, this is difficult in practise to use and the Coulomb Matrix will be the only one used.

2 Model Learning: Methods, Parameters, and Results

This section will dive into the different learning models, from a simple Naive Model (Section 2.1), then using a Linear Model on the data to see if this will be enough (Section 2.2), and finally using a Neural Network Model (Section 2.3). Moreover, a Principal Component Analysis (PCA) will be used to see how the size of the data can be reduced with a PCA, or at least prevent over-fitting (Section 2.4). The representation of the molecular configuration will be done both with raw data and with Coulomb Matrices (see Section 1.4).

2.1 Naive Model

A Naive Model is a starting point for comparing metric results. It will set a baseline so that, any new more advanced model can be compared to it, and must surpass it to be considered improvement.

In a regression, a simple basic naive model is the mean of the target variable. For instance, in our case the naive model will be the mean of the energies $\bar{y}$.

Our model $\tilde{f}(X)$ is then :

$$\tilde{f}(X) = \tilde{y} = \bar{y}$$

The metrics[4] given by this model are :

	Mo2S4	Zundel
RMSE	32.38	9.10e-04
MAE	3.5	7.19e-04
r_{pe}	NaN	NaN

Table 2: Metrics of the naive model for raw dataset.

[4] The Pearson's correlation coefficient is a NaN because the $(\tilde{y}^{(i)} - \bar{\tilde{y}})$ factor is obviously equals to 0 here.

The RMSE of each dataset is equal to the standard deviation σ of each dataset in the Table 1. This comes from the formula of the RMSE which is directly equal to σ in case of $\tilde{y} = \bar{y}$.

2.2 Linear Model

A slightly more complex model will be used, using the atom positions as features and the potential energies as the target. A linear regression model from *Scikit-learn* will calculate the MSE and RMSE of this model.

2.2.1 Results with Linear Model

	Mo2S4	Zundel
RMSE	28.66	9.16e-04
MAE	3.88	7.24e-04
r_{pe}	0.103	0.004

Table 3: Metrics of the linear model for raw dataset.

For Mo2S4 the RMSE is just a bit better than the naive model, but somehow it is worse for Zundel ions. This might be an insight that our model is way too simple, and even maybe that the raw configuration position files are not a good way of describing our data.

Another way to see graphically the Pearson Correlation Coefficient is to plot the predicted energies output $\tilde{f}(X_{test}) = \tilde{y}_{test}$ vs the real data for y_{test}.

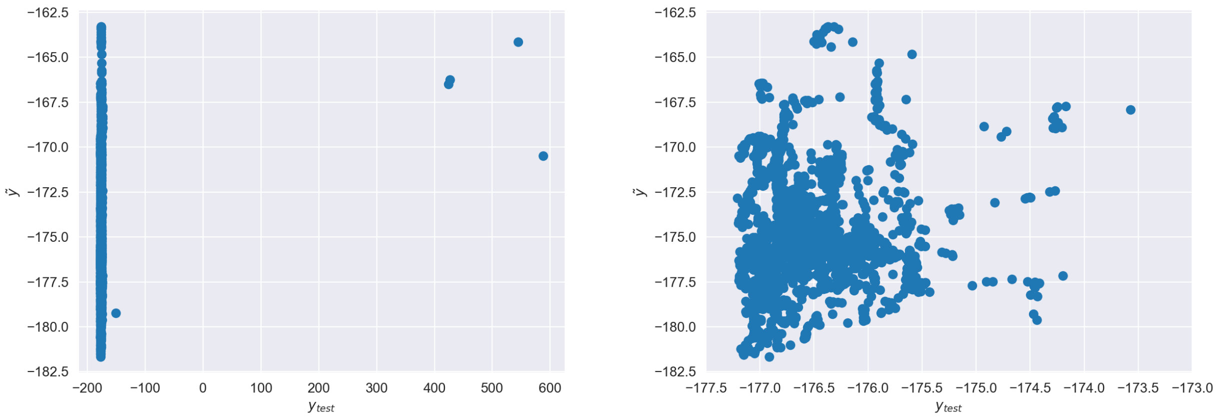

Figure 5: Plot of $\tilde{y}_{test}$ vs y_{test} for Mo2S4 in a linear regression model. The graph on the right is the zoomed version to not take into account higher energies.

Figure 5 implies that the high energies are taken into account and are not correctly predicted with the model[5]. This long distance between $\tilde{y}_{test}$ and y_{test} are causing a huge difference in the RMSE and r_{pe} (increased by the power of 2). Moreover, the lack of data in the high energy region, coupled with the fact that these energies are positive while the rest are negative, leads us to disregard them and only consider the data around the cluster of low energies. A clustering approach, such

[5] If it was, they would be displayed way further on the y-axis around 600, which is not the case here.

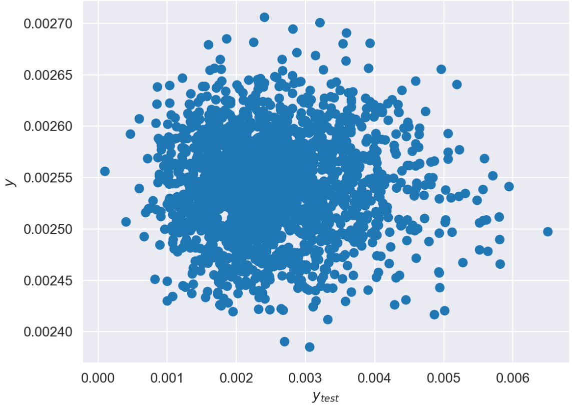

Figure 6: Plot of $\tilde{y}_{test}$ vs y_{test} for Zundel ions in a linear regression model.

as using the K-means algorithm, could have been considered for this analysis. However, in this specific case, it's more straightforward and effective to simply apply a filter around the mean value ($\bar{y} - \sigma/2 \leq y \leq \bar{y} + \sigma/2$). This decision is based on the specific characteristics and distribution of our dataset, where a simple filter efficiently isolates the relevant data segment.

This seems not to be the case for Zundel ions (Figure 6), therefore this filter will not be used on it.

2.2.2 Results with Linear Model and filter on Mo2S4

Now let's give the new dataset summary for energies with the filter applied, and redo all previous processes.

Energies	Mo2S4
Count	10962
Mean (μ)	-176.53
Std (σ)	0.521
Min	-177.23
25%	-176.91
50%	-176.91
75%	-176.30
Max	-169.95

Table 4: Dataset summary for filtered Mo2S4 energies. Again, the units are not defined here.

	Naive	Linear
RMSE	0.52	0.42
MAE	0.39	0.30
r_{pe}	NaN	0.585
$(r_{pe})^2$	NaN	0.342

Table 5: Metrics of Naive and Linear model for filtered Mo2S4 dataset.

Table 4 shows that the number of configuration inside the filter is almost the same as before (39 configuration less). Again, even if these simulations of molecule simulation are correct, there are not enough data out of the filter zone to have a good regression in these places.

Table 5 shows that the Linear model is really doing pretty well for explaining the data, tho σ changed, so this is taken into account (from ~ 32 to ~ 0.6). The r_{pe} squared is displayed from now on to increase readability on how good is our model.

However, both Naive and Linear model are very bad for Zundel ions. Due to inherent potentials inside Zundel ions, the energy is not easily explainable by the positions of atoms inside the molecule.

	Mo2S4	Filtered Mo2S4	Zundel
RMSE	28.76	0.272	7.98e-04
MAE	3.44	0.2	6.32e-04
r_{pe}	0.062	0.853	0.489
$(r_{pe})^2$	0.004	0.728	0.239

Table 6: Metrics of Linear model for Mo2S4 and raw Zundel with Coulomb matrices.

The filtered Mo2S4 will now always be used. Here are same graphs as before :

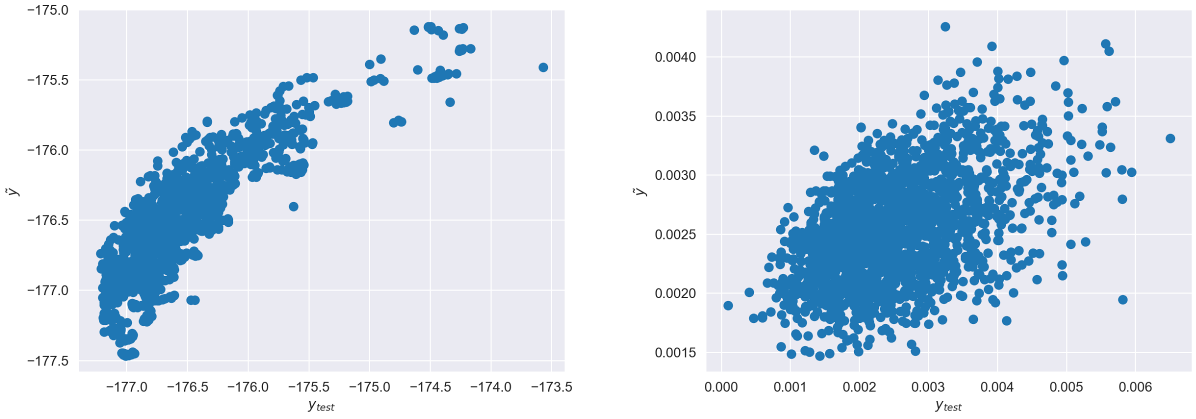

Figure 7: Plot of $\tilde{y}_{test}$ vs y_{test} for Mo2S4 (left) and Zundel (right) in a linear regression model with Coulomb Matrices.

From Table 6 and Figure 7, it appears that a good prediction on the data from filtered Mo2S4 and Zundel datasets with Linear models has been achieved.This was reached by computing the Coulomb matrices of each configuration and fitting our model on these. Now, the fitting model will try to be maximized over some Neural Network model (Section 2.3) and then try a Principal Component Analysis (Section 2.4).

2.3 Neural Network Model

2.3.1 Neural Network Architecture

There are no clear and rational approach to choose the size of a Neural Network (NN) in the literature (cf [3]). A 3 hidden layer architecture (18, 18 and 8 neurons) with *ReLu* for all hidden layers and *linear* for the output, for a total of 1,169 trainable parameters (see Figure 8)[6].

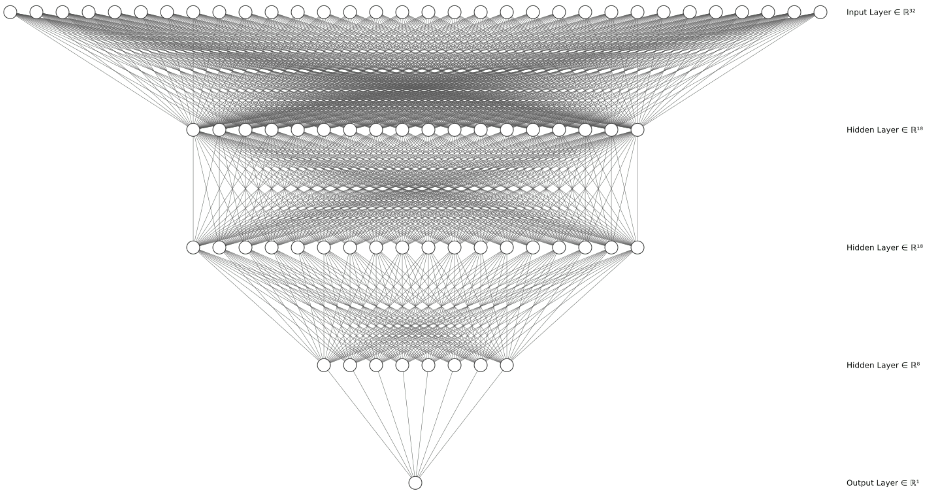

Figure 8: Structure of our Neural Network (here with Mo2S4). 3 hidden layers with *ReLu* activation, and resp. 18, 18 and 8 neurons per layer and 1 output layer with one neuron.

	Filtered Mo2S4	Zundel ($\times 10^5$)	Zundel
RMSE	0.114	64.97	6.50e-4
MAE	0.086	51.63	5.16e-4
r_{pe}	0.979	0.703	0.703
$(r_{pe})^2$	0.958	0.494	0.494

Table 7: Metrics of Linear model for Mo2S4 and raw Zundel with Coulomb matrices.

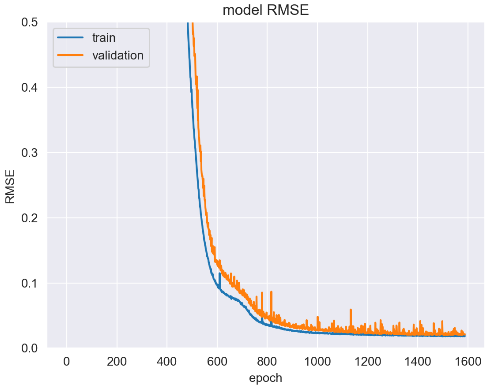

Figure 9: The MSE vs epoch of the NN for Mo2S4 zoomed.

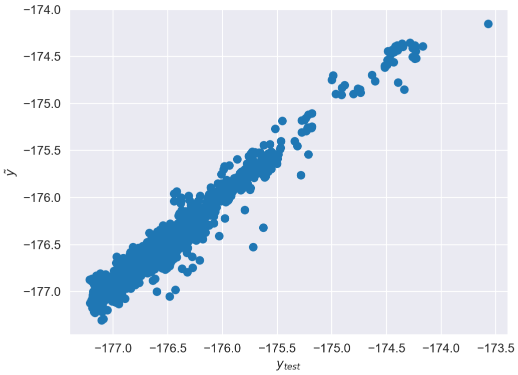

Figure 10: Plot of $\tilde{y}_{test}$ vs y_{test} for Mo2S4 in the NN with Coulomb Matrices.

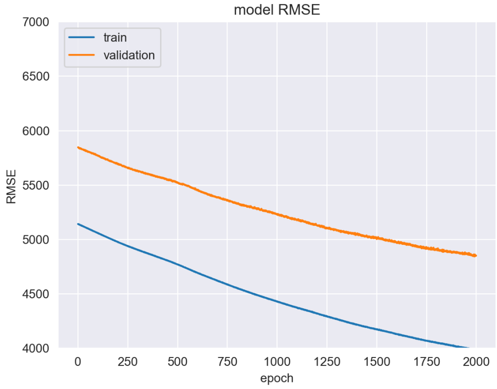

Figure 11: The MSE vs epoch of the NN for Zundel zoomed.

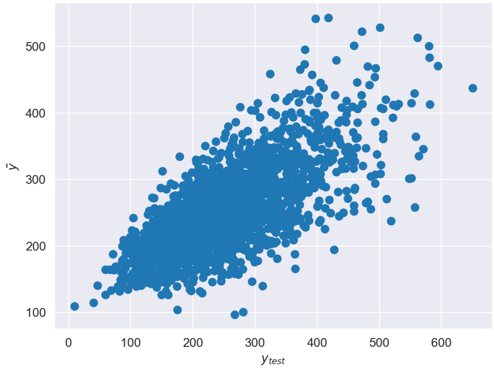

Figure 12: Plot of $\tilde{y}_{test}$ vs y_{test} for Zundel in the NN with Coulomb Matrices.

2.3.2 Results with Neural Network

As visible in Table 7[7], the filtered Mo2S4 (Figure 9) and Zundel (Figure 11) are way more accurate than before. However, because of the range value of Zundel energies ($\sim 10^{-3}$), the NN had some issues fitting a good model. To prevent this, all Zundel energies will be multiplied by a factor 10^5 (to be around the values of Mo2S4). This is visible on the RMSE and MAE, but the values of these metrics in a Linear model can just be multiplied by the same factor to get the new one.

The Pearson's correlation coefficient is much better (Figure 10 and 12). It seems like our model has improved from NN.

However, even if predictions are very good for Mo2S4, some issue to attain a good prediction for Zundel were reached. Figure 11 shows that the fitting could be developed further, but the distance between train and validation data is increasing. This may result in an overfitting. To level this, a Principal Component Analysis will be implemented on the data (especially for Coulomb matrices of Zundel ions).

2.4 Principal Component Analysis

2.4.1 PCA Principle

Principal Component Analysis (PCA) aims to represent data in a new basis and assess the importance of each new component in explaining the data. If the original data has M components, PCA transforms it into a new space with M new components, which are linear combinations of the original ones. Some of these new components may not carry much variance and therefore may not explain the data well, which is intentional. These less informative components can be discarded, reducing the data's dimensionality while retaining a significant portion of the original data variance. Typically, the goal is to select the right number of components to explain around 99% of the data variance.

[6] The model parameters were : Adam optimiser with 0.00005 learning rate, batch-size=32, verbose=1, validation-split=0.2 and an EarlyStopping callback depending on the data.

[7] On every graph of NN, the MSE is displayed instead of the RMSE. However, *Keras* does not keep the history of a model. Just keep in mind that every graph with RMSE is in reality the MSE, but every values in the table are indeed the RMSE.

2.4.2 PCA Cumulative explained variance ratio

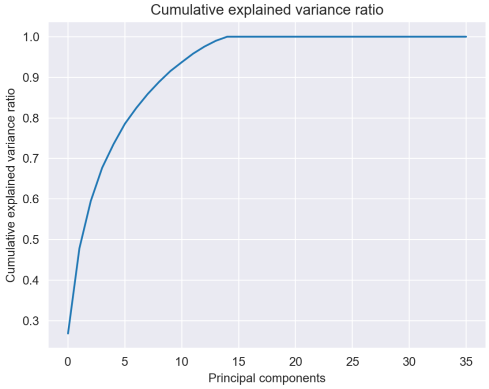

Figure 13: Cumulative explained variance ratio for Mo2S4 after PCA transformation (Sum of the explained variance by each component on every component).

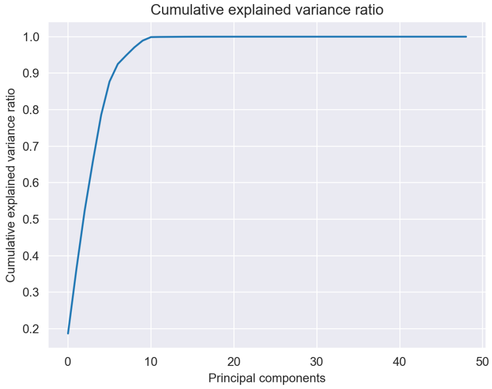

Figure 14: Cumulative explained variance ratio for Zundel ions after PCA transformation (Sum of the explained variance by each component on every component).

From Figures 13 and 14, the number of principal components explaining the majority of the data variance is way less than M (resp. 36 and 49). The majority of these new PCA components can be discarded, to explain 99% of the data variance. This means our data pass from 36 and 49 components to 15 and 11.

A way to see how much the data changed with this process is to get the RMSE of the inverse PCA transformation of the data (meaning the data is getting back to the original space but with only the components explaining 99% of the variance) vs the real data with all original components. This will be our metric of how good the truncation of components on PCA transformation is, allowing us to lower the dimension of the data while remaining the information.

The RMSE for Mo2S4 is $7.23 \cdot 10^{-11}$ and for Zundel $3.12 \cdot 10^{-2}$. It is way better for explaining the Mo2S4, but can still be considered good for Zundel ions.

Another good part of the PCA transformation is to lower the dependency of linear combination in the data and decreasing the exactness, allowing models to be better for generalisation and decreasing the possible over-fitting.

2.4.3 Results with PCA on Naive Model and Linear Model

Table 8 gives our final results from naive and linear models to conclude on this analysis[8].

[8] The NN model was implemented but due to a lack of time, it was not thoroughly parametrized (results may be found in the *main.ipynb*).

Model Metric		Filtered Mo2S4	Zundel ($\times 10^5$)
Naive Model	RMSE	5.21e-01	9.10e+01
	MAE	3.89e-01	7.19e+01
	r_{pe}	NaN	NaN
	$(r_{pe})^2$	NaN	NaN
Linear Model	RMSE	2.71e-01	9.00e+01
	MAE	2.01e-01	7.11e+01
	r_{pe}	0.854	0.181
	$(r_{pe})^2$	0,729	0.033

Table 8: Metrics of Naive and Linear models for Mo2S4 and Zundel with PCA transformation from Coulomb matrices and truncation of principal component (99% of data variance).

3 Conclusion

This project was aimed at analysing the energies of Zundel and Molybdenum-Sulfur atomic configurations. Starting from analysing the raw data and filtering it, to develop models as Linear and even more complex as Neural Network. These datasets, being the position of every atom in the molecule, were then translated to another representation (here Coulomb Matrices). The Principal Component Analysis described a new space to carry our data and gave us a chance to lower the dimension of it.

To conclude, the results are very promising for the Mo2S4 dataset, reaching a Pearson's correlation coefficient squared $(r_{pe})^2 = 0.958$. However, the Zundel dataset seems to not be easily describe by the Coulomb Matrix, certainly due to quantum mechanics playing a major role in the energy of a configuration. A forward task would be to try to describe it with Many-Body Tensor Representation (MBTR) or Smooth Overlap of Atomic Positions (SOAP) to overcome the lack of interesting results.

References

[1] M. Moog, S. Schaack, F. Pietrucci, and A. M. Saitta. Unsupervised exploration of MoS2 nanocluster configurations: structures, energetics, and electronic properties. *The Journal of Physical Chemistry C*, 123:22564–22569, 2019.

[2] F. Mouhat, S. Sorella, R. Vuilleumier, A. M. Saitta, and M. Casula. Fully quantum description of the zundel ion: combining variational quantum monte carlo with path integral langevin dynamics. *Journal of Chemical Theory and Computation*, 13:2400–2417, 2017.

[3] A. B. Tchagang and J. J. Valdés. Prediction of the atomization energy of molecules using Coulomb Matrix and atomic composition in a bayesian regularized neural networks, journal = Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions. pages 793–803, 2019.

[4] L. Himanen. Dscribe: library of descriptors for machine learning in materials science. 2022.