MSc TCC Machine-learning in Materials Chemistry Assessed Problems

Candidate Number: 1091127

July 2025

Question 1

This problem is a practical application of the Gaussian Process Regression (GPR) theory learnt in the lecture material. The formula from the lectures for calculating the vector $\mathbf{c}$ (model parameters) used in the GPR fits in this question is:

$$\mathbf{c} = (\mathbf{K}_{NN} + \mathbf{\Sigma})^{-1} \mathbf{y}$$

Where $\mathbf{y}$ is the vector of data and $\mathbf{K}_{NN}$ is the matrix of kernel values where each matrix element is defined using the formula from the lecture and the *Chemical Reviews* article on which Lecture 2 is built (VLD et al., Chem. Rev. 121, 10073 (2021), herein after referred to as VLD et al., (2021)):

$$\begin{aligned} K_{ij} =& k(x_i, x_j) \\ =& e^{-\frac{|x_i - x_j|^2}{2\sigma^2}} \end{aligned}$$

Where sigma is the length scale of the data (a hyperparameter whose optimisation was also discussed in the lecture). The theory from the lectures was expanded upon using theory from VLD et al., (2021) to give a simplified formula for calculating a GPR fit which learnt from both function and derivative values:

$$\mathbf{c} = (\mathbf{K}_{DD} + \mathbf{\Sigma})^{-1} \mathbf{y}$$

Where $\mathbf{y}$ now contains both all the potential data points (first half of the vector) and the derivative (force) data points (second half of the vector). $\mathbf{K}_{DD}$ can be split into four matrices:

$$\begin{bmatrix} \mathbf{K}_{NN} & \frac{\partial \mathbf{K}_{NN}}{\partial x_i} \\ \frac{\partial \mathbf{K}_{NN}}{\partial x_i} & \frac{\partial^2 \mathbf{K}_{NN}}{\partial x_i^2} \end{bmatrix}$$

With the entries of each derivative matrix defined similarly to before:

$$
\begin{aligned}
\frac{\partial}{\partial x_i} K_{ij} =& \frac{\partial}{\partial x_i} k(x_i, x_j) \\
=& -\frac{x_i - x_j}{\sigma^2} e^{-\frac{|x_i - x_j|^2}{2\sigma^2}} \\
\frac{\partial^2}{\partial x_i^2} K_{ij} =& \frac{\partial^2}{\partial x_i^2} k(x_i, x_j) \\
=& -\frac{e^{-\frac{|x_i - x_j|^2}{2\sigma^2}}}{\sigma^2} - \frac{(x_i - x_j)^2}{\sigma^4} e^{-\frac{|x_i - x_j|^2}{2\sigma^2}}
\end{aligned}
$$

Now the function used in the practical application of this question was a Leonard Jones potential at a point one unit away from a string of atoms (originally it was planned that the point and the atoms lie on the same line, however for the sake of simplicity (avoiding infinite potentials), the point was offset), i.e. in the x-y plane a string of atoms lie along y = 0 and the potential is calculated along the line y = 1. Using this potential and the theory discussed previously the question 1 supporting code, gives the GPR fits shown in Figure 1. Note three different models were calculated aiming to correspond to the models calculated in VLD et al., (2021). One model fits the data points of potential energy values. One fits data points of force values (the calculated force function is then integrated to give the potential function). And the final model uses both force and potential energy data.

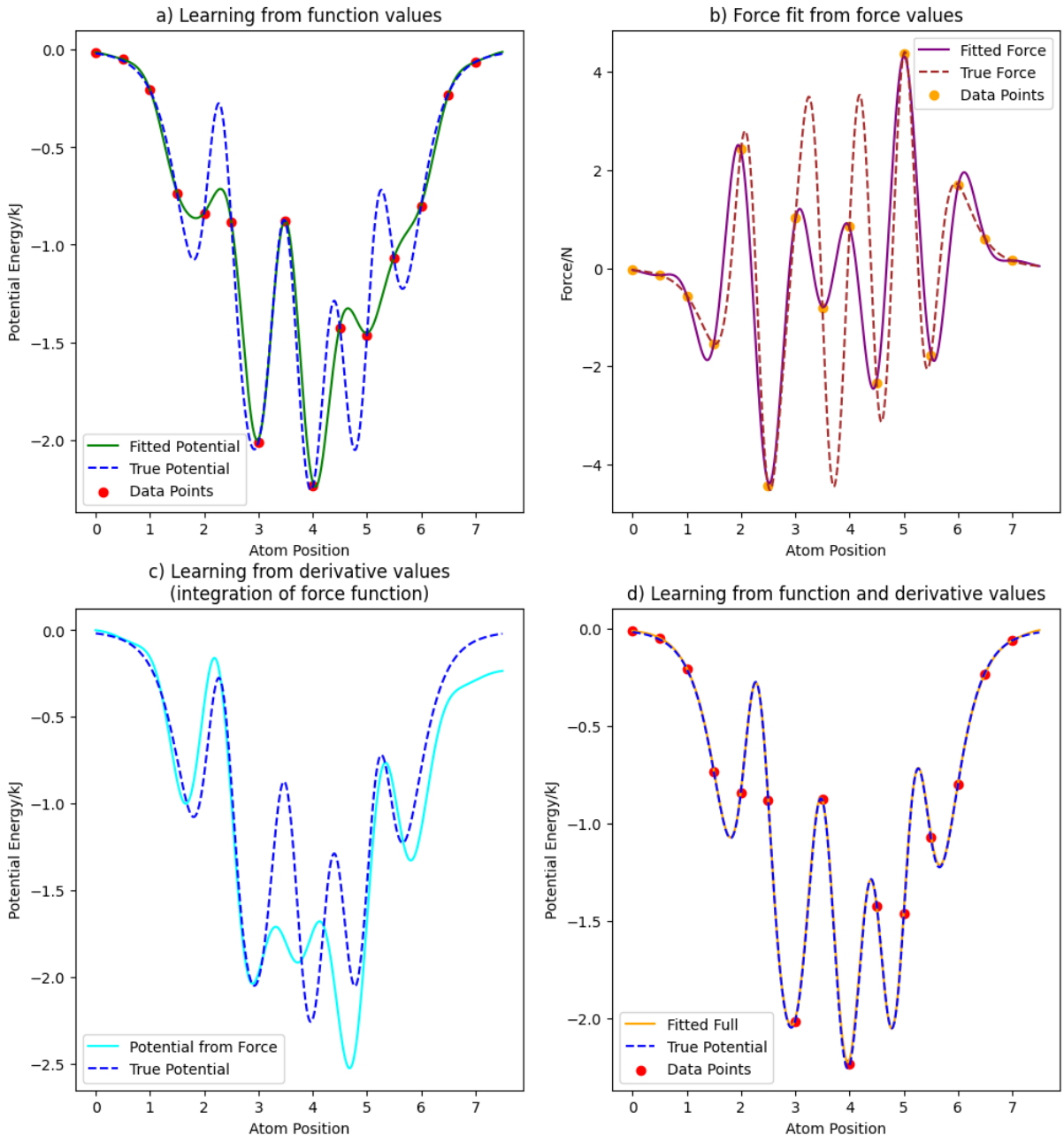


Figure 1: Results from question 1. In each case the solid line is the result from the model (calculated at 1000 points within the domain), the dotted line is what the function should look like (simply using the function defined to calculate the data points) and the red dots are the data points used to fit the models (the same in all graphs). a) is the result of using just the potential data points in the GPR fit. b) and c) are the result of using just the potential derivative (force) points in the GPR fit. And d) is the result of using both potential and derivative values in the GPR fit. Note a), c), and d) correspond to the leftmost column of Figure 6 from VLD et al., (2021) whereas b) is the fit derivative of the potential modelled using only the force values

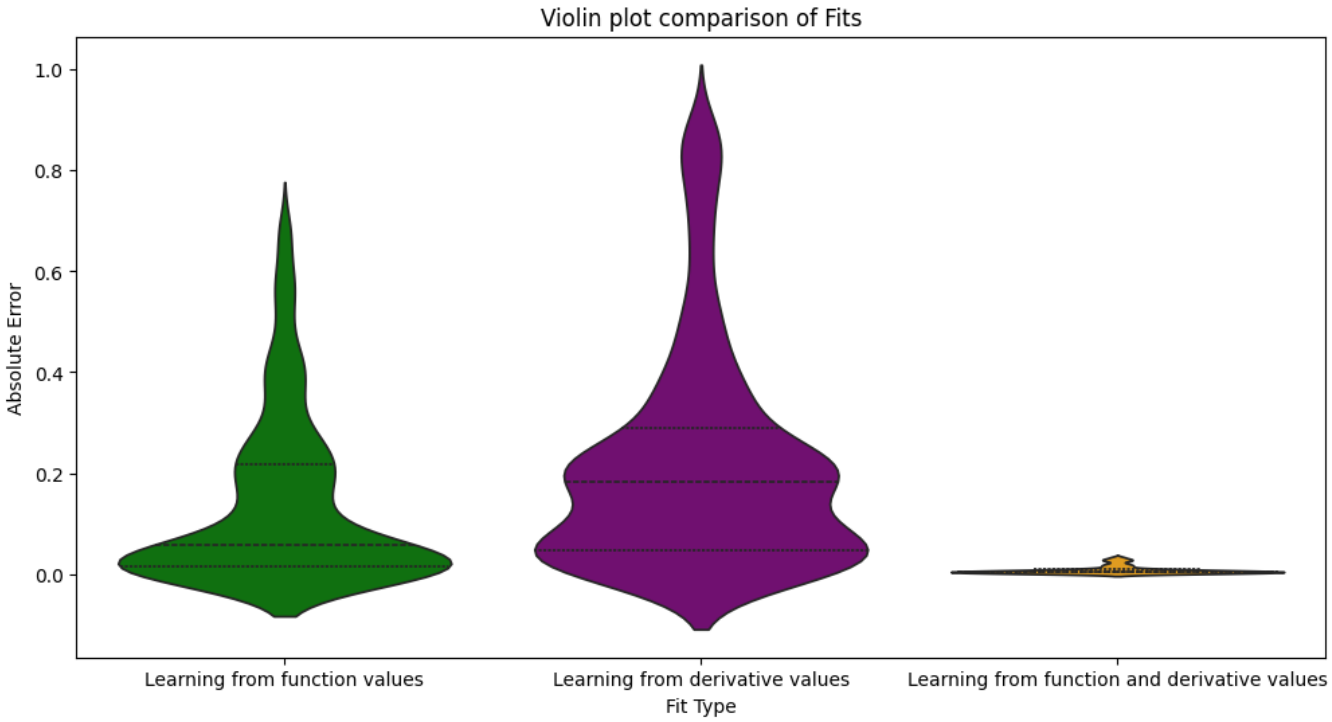


Figure 2: Violin plots of the absolute error for the different models

These results closely match the trend observed in VLD et al., (2021) as required for part (b). The GPR fit which only used potential values matches the data somewhat but outside the area of its immediate training points it misses several features. Similarly the fit which only used potential derivative values (force values) also missed some of the finer structure. Finally the fit which learnt from both derivative and potential values almost perfectly matches the true potential function. This visual performance is also quantitatively supported by the root mean squared deviation in the models of 0.2155, 1.4520 and 0.0108 for the models shown in a), c) and d) of Figure 1 respectively. Finally, the violin plot in Figure 2 further supports that learning from function and derivative values produces the best fit as the absolute errors appear to be small and without a significant trend. Note also that this potential can be easily changed and similar trends have been observed for similar potentials.

Finally note that the error is very low in the training points (0.01 units) and thus the regularisation strength is also low (identity matrix multiplied by 0.01) but this can be easily changed and was chosen to be so as it gives simple, easy to interpret results.

Question 4

Structure maps relate to the lecture material in two main ways. Firstly, they use the idea of descriptors, specifically Smooth Overlap of Atomic Positions (SOAP) to write the atomic information in a machine useful way. This is done by first smoothing the neighbour density (as described in question 1 and shown again in equation (1) below), then expanding out the resultant three dimensional function in terms of a local basis set (shown in equation (2) below). This allows many body interactions to be accounted for as both angles and distances are captured in the basis vector (compared to simple radial distribution functions capturing only two body terms). Finally, this basis vector is made

rotationally invariant giving a unique description of any atomic arrangement (shown in equation (3) below).

$$\rho^{i,a}(\mathbf{r}) = \sum_j \delta_{aa_j} e^{\frac{-|\mathbf{r}-\mathbf{r}_{ij}|^2}{2\sigma_a^2}} f_{cut}(r_{ij}) \tag{1}$$

$$\rho^{i,a}(\mathbf{r}) = \sum_{nlm} c_{nlm}^{i,a} R_n(r) Y_l^m(\hat{\mathbf{r}}) \tag{2}$$

$$\rho_{nn'l}^{i,aa'}(\mathbf{r}) = \frac{1}{\sqrt{2l+1}} \sum_m (c_{nlm}^{i,a})^* c_{n'lm}^{i,a'} \tag{3}$$

Where $\rho^{i,a}(\mathbf{r})$ is the function describing the environment around atom i, j is the surrounding atoms within the cutoff distance, $f_{cut}(r_{ij})$ is the cut off function (excluding atoms which are not near), $c_{nlm}^{i,a}$ are the vector coefficents in the new basis, $R_n(r)$ is the radial basis function, and $Y_l^m(\hat{\mathbf{r}})$ is the angular basis function.

Using this machine readable format, different structures can be compared. However, the vector resulting from the SOAP process is highly dimensional, so a dimensionality reduction technique such as principal component analysis (PCA) is applied to allow the differences in structures to be easily visualised. PCA aims to capture the greatest variance in a few dimension therefore attempting to retain as much information as possible in the dimensional reduction.

In addition to the methods applied in to calculation of structure maps, the lecture material also covers their importance. A models performance is expected to be highly dependent on the range of their training data. Outside their range of training data they are expected to perform significantly more poorly. Furthermore, it was highlighted that not only should lowest energy conformations be shown to models but rather a range to allow them to sample the potential energy landscape. The information on a structure map can therefore aim to convey the structures a model would be expected to perform accurately on and structures which would be expected to be outside it's scope.

Now the publication chosen for this question was 'arXiv:2401.00096 [physics.chem-ph]' (available at `https://doi.org/10.48550/arXiv.2401.00096`). While technically this is still a pre-publication, it was chosen for it's relevance and interestingness. Furthermore, as the full dataset for the 'MACE-MP-0' (the model the paper is about) model is >10GB, only a very small subset of the data was mapped, specifically, all structures containing xenon. This was chosen mostly because the size of data was manageable and because xenon's uniqueness may be relevant in the context of machine learning models. The calculated structure map of this data is shown in Figure 3. Although xenon does not have many applications, it may still be relevant to consider as this model seems to claim that it 'can be applied out of the box and as a starting or "foundation model" for any atomistic system of interest'.

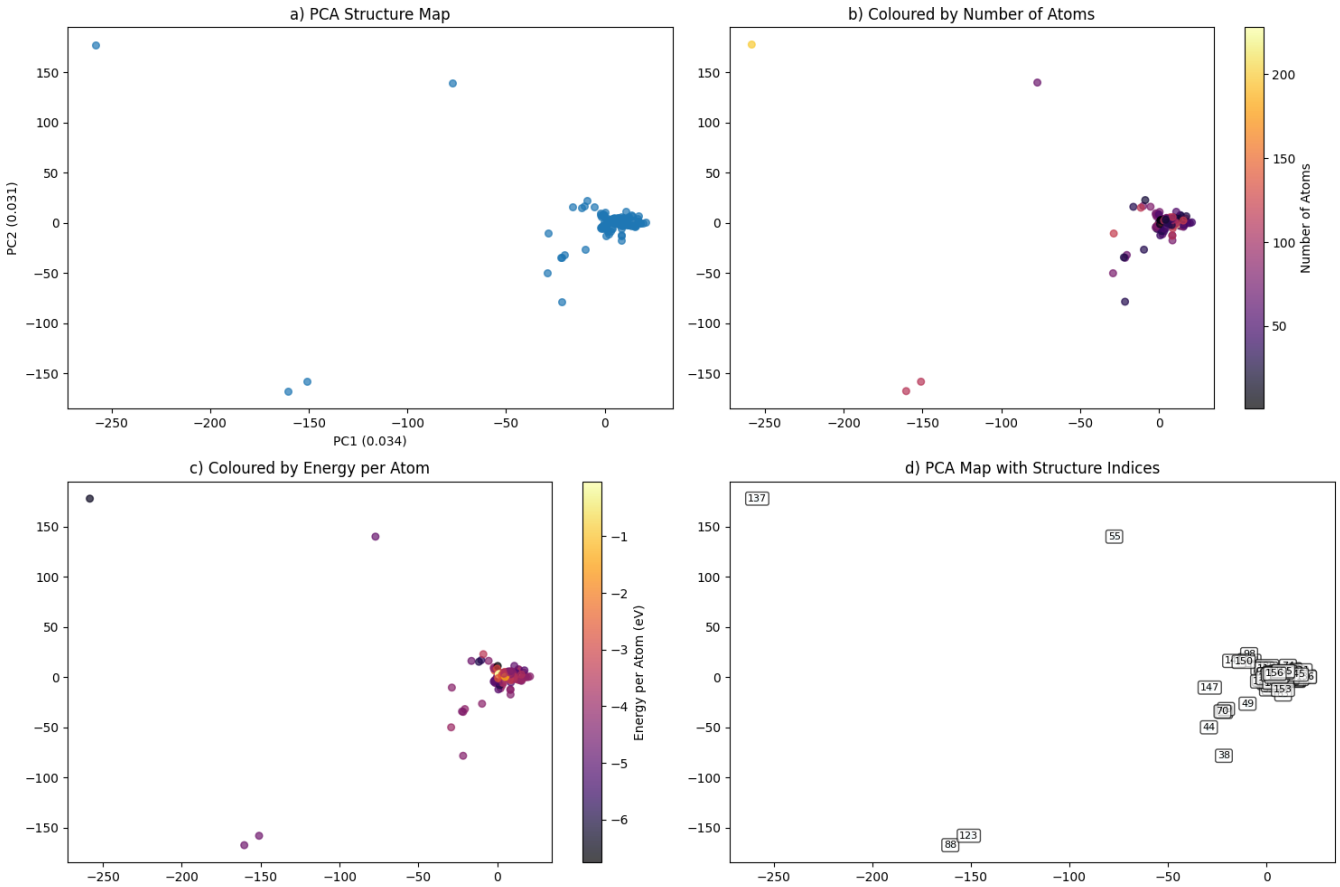
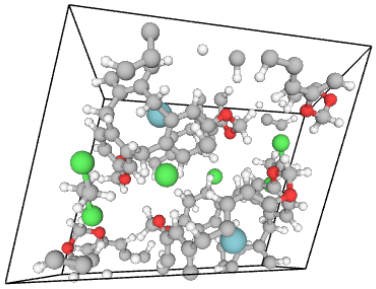
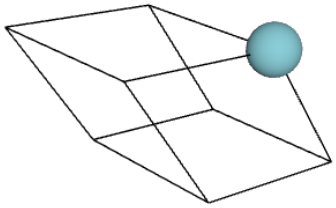


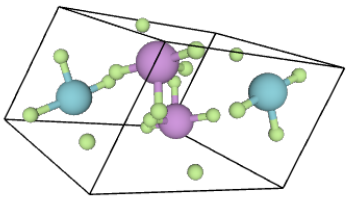
Figure 3: Variations of the same structure map calculated from the structures containing xenon in the data set of MACE-MP-0 (arXiv:2401.00096 [physics.chem-ph]). a) Standard structure map. b) structure map with the points colour coded by number of atoms in the crystal structure. c) structure map with the points colour coded by energy per atom in the crystal structure. d) Structure map with each point is labelled by the index of the corresponding structure to allow closer inspection



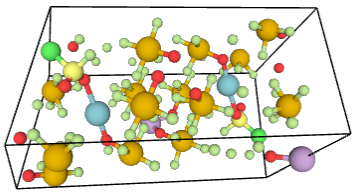
(a) Structure index 137



(b) Structure index 60



(c) Structure index 156



(d) Structure index 88

Figure 4: Unit cell of different structures. The atoms corresponding to each colour are: light blue - xenon, white - hydrogen, grey - carbon, red - oxygen, bright green - chlorine, dull green - fluorine, purple - bismuth (in structure 156), purple - antimony (in structure 88), light yellow - sulfur, and dark yellow - tellurium.

These structure maps and corresponding structures suggest an extremely broad dataset with very few data points relative to it's size. Now although the model has many data point's which are not in this set, the unique behaviour of xenon would be expected to prevent other data being applicable to xenon structures. Therefore, although this model appears to claim qualitative accuracy, it would be expected that performance on xenon structures would require significant fine tuning to handle these systems even qualitatively.

The way that even the points close together on the structure map (such as structure 60 and 156 in Figure 4), are very dissimilar (structure 60 is a pure xenon crystal but structure 156 is a crystal with xenon tetrafluoride and bismuth pentafluoride and therefore would be expected to have different properties), suggests that space of the dataset is very large and most points are very dissimilar. The points far away from the main cluster (such as structure 137 and 88 in Figure 4) also show strong dissimilarity and are further examples of the range of this data. Whilst this set may be significant in terms of the number of xenon structures publicly available, it does not seem sufficient to train a qualitatively general purpose model, as is claimed in the paper.

One positive aspect of this data set is the distribution of 'energies per atom' and 'number of atoms' as shown in Figure 5. By capturing many data points in a close distribution energies and no. atoms the model could handle similar xenon situations more accurately. Note also from (b) and (c) of Figure 3 that most of these points are quite close together and at least visually there does not seem to be significant trend in the distribution of energy and no. atom values within the structure map.

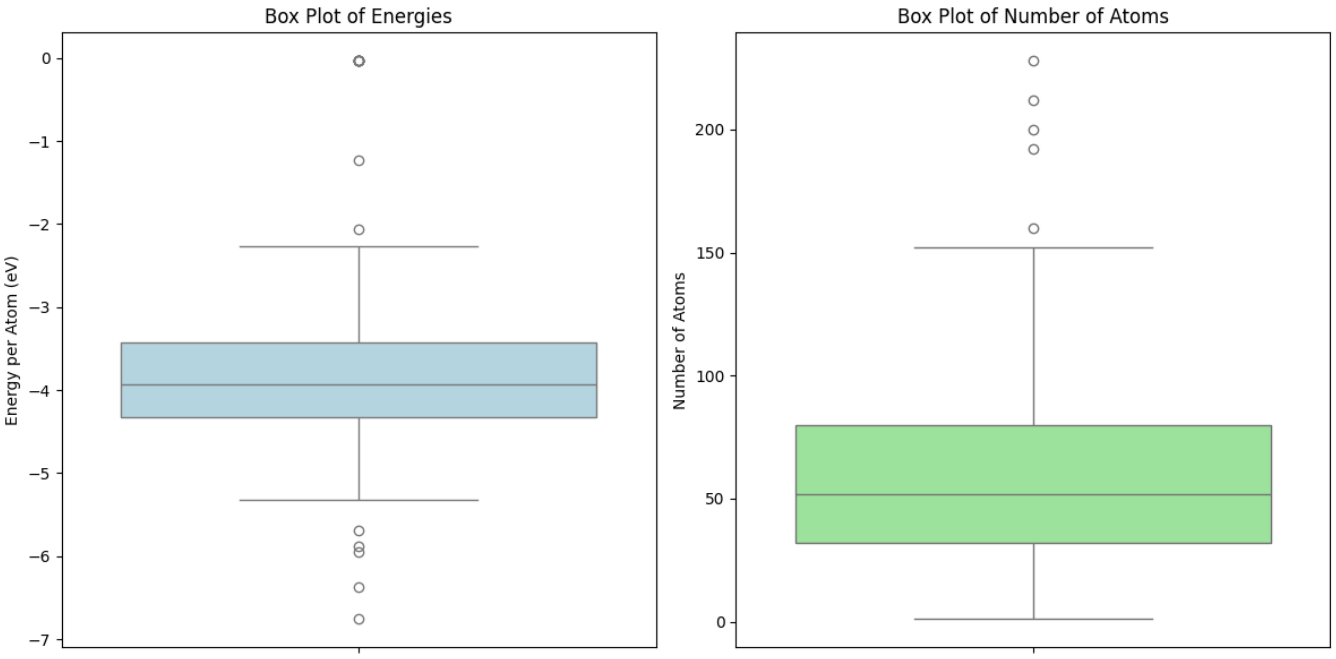


Figure 5: Box plot of the different 'energy per atom' and 'number of atoms' values

Note that in the program several different 'r_cutoff', n_max and l_max values were tested with but if there were too large it crashed the program and didn't seem to significantly change the result anyway. Overall, however, this is only a very small subset of the data so general claims about the MACE-MP-0 should not be extrapolated from this.