Fine-Tuning Foundation Models with MACE

BCN: 4HUY8

Word Count: 4555

Cavendish Laboratory, Department of Physics, J J Thomson Avenue, Cambridge. CB3 0HE

Abstract

Foundation models, pre-trained on vast datasets, are revolutionising the field of atomistic simulation with their broad applicability and impressive qualitative accuracy. However, these models struggle to achieve quantitative accuracy. Recently, it has been shown that by training a foundation model on a small set of additional system-specific reference calculations through a process known as fine-tuning, these pre-trained foundation models can achieve quantitative accuracy with much less data than training a model from scratch. Here, this is demonstrated for a $CaCO_3/H_2O$ system by comparing fine-tuned MACE-MP-0 foundation models to models trained with the same architecture from scratch. It is found that fine-tuned models with as little as 10 additional reference calculations can outperform scratch models trained on 300 configurations. This is shown through rigorous validation of numerical energy and force errors, along with static and dynamic properties from molecular dynamics simulations, where the fine-tuned models substantially outperform models trained from scratch. This investigation provides further confirmation of MACE-MP-0's suitability as a pre-trained model for fine-tuning, along with a new protocol for re-evaluating problematic atomic reference energies.

1. Introduction

Historically, atomistic simulation has always involved trade-offs: ab initio methods are computationally expensive but can achieve high accuracies, whereas molecular dynamics using classical force fields offers an efficient alternative at substantially reduced accuracy [1]. In recent years, machine-learned interatomic potentials (MLIPs) have become promising candidates to achieve ab initio level accuracy with classical force field efficiency [2]. MLIPs are used to predict energies, forces and stresses from an atomic configuration, or a representation of that configuration using chemical descriptors such as the Smooth Overlap of Atomic Positions (SOAP) descriptor [3]. To do this, they are trained using a minimal number of representative ab initio calculations [4]. Although this approach has laid the foundation for MLIPs, it has numerous limitations, such as relying on hand-crafted and often not-sensitive enough descriptors and poor generalisability, requiring a new set of computationally expensive DFT calculations for each new system [5].

Recently, these issues are being tackled primarily through two approaches. Firstly, using graph neural network (GNN) based architectures to create molecular graphs and learn the descriptors [6]. Graphs are naturally suited to describing molecules as they are made up of nodes and edges, where nodes are atoms and edges connect the nodes [7]. Secondly, the development of foundation models for improved generalisation, such as CHGNet [8], M3GNet [9], GNoME [10] and more. Notably, the MACE-MP-0 foundation model combines these two concepts by training an equivariant message passing neural network (MPNN), a type of GNN, on a large Materials Project (MP) dataset, MPTrj [11, 12]. The dataset used consists of 89 elements across $\sim$1.6 million static and relaxed bulk inorganic crystal trajectories, using the Perdew-Burke-Enzerhof (PBE) exchange-correlation functional [8].

MACE-MP-0 has shown remarkable "out of the box" qualitative performance across a wide range of systems from batteries to ice and water, without any additional system-specific training data. However, the foundation model often lacks quantitive accuracy, especially when the system at hand is far outside of the training set [11]. To address this,

a type of transfer learning known as fine-tuning can be used [13]. In fine-tuning the pre-trained MACE-MP-0 model is used as a starting point, requiring only a small amount of additional system-specific reference calculations to achieve substantially improved quantitative accuracy. This approach has the potential to produce highly accurate models at a dramatically reduced computational cost, making it a vital tool in atomistic simulations. Previously, transfer learning has been applied to competing pre-trained models such as CHGNet where an RMSE of 100 KJ/mol was achieved by fine-tuning on a dataset of nearly 200,000 structures [8]. Recently, fine-tuning of MACE-MP-0 has been successfully applied in calculating sublimation enthalpies to RMSEs of <1 KJ/mol using only tens of additional structures [14]. This data efficiency makes MACE-MP-0 a very promising pre-trained model for fine-tuning. However, its transferability to a wide range of systems, along with incorporating reference data of different levels of theory, must be more thoroughly tested before a conclusion can be made.

In this investigation, fine-tuning of the MACE-MP-0 foundation model on a $CaCO_3/H_2O$ dataset will be compared to training on an identical dataset from "scratch", which uses the MACE architecture but without pre-training data. As the training set for MACE-MP-0 uses the PBE functional, the model often struggles to reach quantitative accuracy for systems where dispersion interactions are key, such as in systems involving water like $CaCO_3/H_2O$, hence the requirement for fine-tuning [11]. Understanding the behaviour of Ca^{2+} ions in water is vital for a wide range of applications, from their interactions with proteins [15] and involvement in cell function [16, 17] to calcium ion batteries [18]. For MACE-MP-0 to be effective as a foundation model for fine-tuning then for the same dataset size, a fine-tuned model should have a higher performance, such as in the form of lower energy and force RMSEs, as well as secondary properties. This would allow specialised models to be trained very cheaply compared to training from scratch. Furthermore, revPBE-D3, revPBE0-D3 and MP2 datasets will be used to separately fine-tune the foundation model to compare how different levels of theory affect performance on the $CaCO_3/H_2O$ system.

2. Methodology

The backbone behind the MACE-MP-0 foundation model is the MACE architecture, using an equivariant Message Passing Neural Network (MPNN). The key advancement with MACE was achieving high accuracy via the use of four-body terms whilst only using two message passing layers [19]. This allows MACE-based models to capture complex interactions between many particles, expand their receptive field to probe beyond purely local features and remain efficient [12].

Fine-tuning is a form of transfer learning where additional task-specific training data is provided to a pre-trained/foundation model, here MACE-MP-0 [20]. This investigation will follow and expand upon the fine-tuning protocol outlined by Kaur et al. where the pre-trained model parameters are used as a starting point for fine-tuning [14]. Firstly, three identical datasets were provided, each with a different level of theory: revPBE-D3, revPBE0-D3 and MP2. The datasets contained energies and forces for 423 configurations of 98 water molecules and one $CaCO_3$. For each dataset, 400 configurations were randomly selected and then the test set was created by randomly selecting 100 configurations from this set of 400. This test set is a hold-out set for final model evaluation and is not used in model training. Ideally, one would sample 400 configurations from a larger initial dataset to ensure training set diversity. Training sets of size 10, 30, 50, 100, 200 and 300 were then created from the remaining 300 configurations, where the larger datasets contained configurations from the smaller datasets, as in the protocol from Kaur et al. [14]. After preparing these datasets, it is vital to assign the correct force and energy keys to the Atomic Simulation Environment Atoms objects, as these are required as inputs in the training script for MACE so that forces and energies can be accessed.

For comparison, models are trained in two ways: from scratch and fine-tuned from MACE-MP-0. Scratch training uses the MACE architecture without pre-training data, therefore the model is only learning from the training sets sized 10-300. Scratch models can also be trained on different levels of theory simultaneously via multihead training. Here, there is a different "head" or branch of the model for each level of theory. After training, each head has its own specialisation and the user can choose whichever head is most appropriate for the task.

Fine-tuning can be performed in two different ways. Firstly, naive fine-tuning is when the foundation model is simply trained on the additional training set. However, this method risks "catastrophic forgetting", where the fine-tuned model overwrites or "forgets" knowledge from its pre-training, eliminating the advantage of fine-tuning [21, 22]. To tackle this, an additional training set including a subset of the original pre-training data is included alongside the fine-tuning training set. This process is known as multihead replay fine-tuning, where one head is used to reshow or "replay" a subset of the original pre-training data, and the other head is used to learn the new fine-tuning dataset. Both heads contribute to training, allowing the model to retain its original knowledge whilst adapting to the new task. However, this replay dataset often requires 10k-100k configurations from the pre-training set, which substantially increases the computational cost when fine-tuning on small datasets. To reduce this cost whilst retaining as much of the information from the foundation model as possible, the lower end of 10,000 pre-training samples was used from the original MPTrj dataset. Model parameters will be kept consistent between all trained models. Initially, 20 epochs will be used and the models tested for convergence. Other parameters included $r_{\max} = 6$ Å, energy weight of 1, forces weight of 10, and a batch size of 2. The $r_{\max}$ value was chosen as a trade-off between computational cost and accuracy. For more densely packed systems, the $r_{\max}$ should be carefully considered, as the number of neighbours in the local environment can suddenly become very large and costly to compute. In this aqueous system, using a $r_{\max}$ of 6 Å should not be overly costly, whilst providing a large amount of information as the receptive field with two message passing layers will be 12 Å.

The medium foundation model will be used for fine-tuning, which contains $L = 2$ order message passing equivariance [12]. The large model could be used for higher accuracy but at a higher computational cost. For model validation, initially, the focus will be on energy and force RMSEs, before more rigorous validation on secondary properties derived from the potential energy surface. These will include relevant observables for the $CaCO_3/H_2O$ system from molecular dynamics simulations such as the radial distribution function and the coordination number of Ca^{2+} in solution.

3. Results Part 1: Re-estimation of atomic reference energies (E0s)

During initial model training, high initial validation energy RMSEs >3000 meV were observed. As stated in the MACE documentation, high initial validation RMSEs can be problematic and are often linked to issues with the isolated atom energies (E0s) [11]. E0 values are so crucial as MACE learns to predict *relative* energies instead of the total energy, as do many other MLIPs. This relative energy is the atomisation energy:

$$E^{\text{atm}} = E^{\text{tot}} - \sum_{i}^{N} E^{0}. \tag{1}$$

There are currently two choices for E0s: computed isolated atom energies using the same reference method as the training set, or using MACE's "average" argument which re-estimates the E0s using averaging. In this investigation, isolated atom energies are provided but have shown problematic behaviour, however, the alternative of using a dataset average is a very approximate and unreliable method. A linear system can be formulated to provide a more robust approach for E0 re-estimation. The energy prediction error ϵ_i for a configuration i is defined as:

$$\epsilon_i = E_i^{\text{true}} - E_i^{\text{predicted}}. \tag{2}$$

We assume this error can be systematically corrected for each element j by adjusting its value of E_0:

$$\epsilon_i = \sum_j N_{ij} \times c_j, \tag{3}$$

where N_{ij} is the number of atoms of element j in configuration i and c_j is the correction for element j. In matrix notation, we can write this as $N\mathbf{c} = \boldsymbol{\epsilon}$ [23]:

$$\begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1n} \\ n_{21} & n_{22} & \cdots & n_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ n_{m1} & n_{m2} & \cdots & n_{mn} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{bmatrix} \tag{4}$$

Since we are adjusting the E0 values of different elements, which are present in many configurations with different energies, we therefore have an overdetermined system with more equations than

unknowns. Since an exact solution may not exist, we can instead minimise the sum of squared residuals, $\min_{\mathbf{x}} ||A\mathbf{x} - \mathbf{b}||_2^2$, and efficiently solve this via the least squares method [24]. The solution is given by rearranging the normal equation [25]:

$$\mathbf{c} = (N^T N)^{-1} N^T \boldsymbol{\epsilon}, \tag{5}$$

which can be easily computed by leveraging SciPy [26]. Now the E0 values can be re-estimated:

$$E0_j^{\text{new}} = E0_j^{\text{old}} + c_j. \tag{6}$$

This method should provide more accurate estimations for E0 values, and be especially useful when fine-tuning across different levels of theory. However, the assumption of a linear system may not always be appropriate. For systems with large amounts of data, re-estimation could be extended to use neural networks which which would accurately learn the correction even if it is non-linear [27]. After applying this re-estimation method, initial validation energy RMSEs ~ 40 meV were observed. This significant reduction in error compared to the original E0 values suggests that this method effectively compensates for problematic E0 values.

4. Results Part 2: Training from scratch and fine-tuning MACE-MP-0

Now the E0 values have been re-estimated, scratch MACE models and multihead replay fine-tuned MACE-MP-0 models were first trained separately on the three different levels of theory (revPBE-D3, revPBE0-D3 and MP2). Figure 1 shows the loss curves for the scratch and finetuned models during training, where the models were trained until convergence, determined when the decrease in loss dropped below a predefined threshold.

The training loss is calculated as the average loss over all the batches during each epoch. A training/validation split of 90/10 was used, where this validation set was not used in training, but used for evaluation at the end of each epoch to give the validation loss curve. For small training sets, such as a set of size 10, this metric is less reliable as only one sample would be in the validation set. In terms of the scratch models (Figures 1a-c), increasing the training set size decreases both the training and validation loss for the same epoch number, showing that more data helps the model fit to the training set, as expected, and generalise to the validation set more effectively. Furthermore, no increase in the validation loss is seen, indicating no overfitting of the model to the training set. For a typical neural network model, the validation loss would be larger than the training loss, however, here we see the opposite. This is likely due to regularisation, such as weight decay, applied during training but not during validation, therefore increasing the training loss. Furthermore, the training and validation loss are computed in different ways, making them not directly comparable.

In the case of fine-tuning (Figures 1d-f), we have an additional validation loss for the pre-training (PT) head. This PT head validation remains approximately constant whilst increasing the training set size. This is very important as it shows that the foundation model is not experiencing catastrophic forgetting. The validation loss begins higher than the PT validation loss for training set size 30, but as the training set size is increased, the validation loss drops below the PT validation loss. This shows that the additional training configurations are producing model with a greater performance on the $CaCO_3/H_2O$ system than the out of the box foundation model, at least on the small validation set. This is very promising as it provides early confirmation of the success of fine-tuning. Overall, the loss curves for fine-tuning are less steep than for scratch training as the model begins at a much lower loss due to its pre-training.

To test the performance of the models, initial evaluations focused on numerical energy and force errors. The root mean squared error (RMSE) was used for its higher sensitivity to outliers compared to other metrics such as mean absolute deviation (MAD) [4]. This evaluation was performed on the hold-out test set of 100 configurations, which was not used in training.

Figures 2a and 2b show the correlation plots for a scratch model and a fine-tuned model respectively, providing a clear visual comparison of their performance. The RMSE quantifies the deviation of the predictions from the reference values. Therefore, a low RMSE would show a nearly straight line, indicating great agreement of the predictions with the reference energies or forces. Knowing this, it is immediately obvious that the scratch MACE model (Figure 2a) performs significantly worse than the fine-tuned model (Figure 2b) for the same amount of training data, showing RMSE values up to an order of magnitude larger than fine-tuned models.

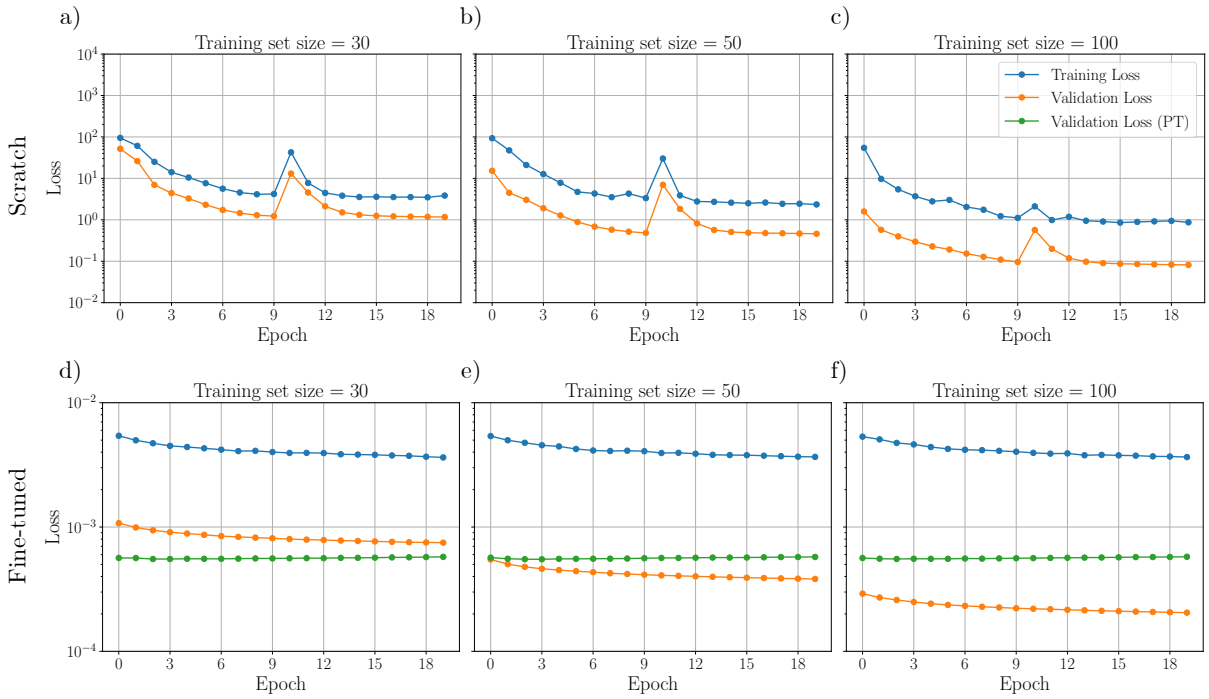

Figure 1: Loss curves during training of scratch MACE models (a-c) and fine-tuned MACE-MP-0 models (d-f) for training set sizes 30, 50 and 100. (d-f) additionally shows the pre-training (PT) validation loss from replaying the pre-training data. The training loss for epoch n is calculated as the averages loss over all batches during that epoch. Both validation losses are calculated on the validation set (90/10 training/validation split) at the end of each epoch.

Figures 2c and 2d show energy and force RMSEs as a function of training set size, where each data point has been calculated from its corresponding correlation plot. Figure 2c shows that, for the same-sized training set, fine-tuned models achieve significantly lower energy and force RMSEs, and therefore show better performance. This is true for all levels of theory and confirms the original hypothesis of the investigation. Even fine-tuning with a dataset size of 10 performs better than a scratch model with 300 configurations in its training set. Comparing reference methods, consistent performance is seen for force RMSEs. For energy RMSEs, there is a wider variation in performance. revPBE-D3 achieves the lowest RMSEs at every training set size for fine-tuned models, however, for scratch models we see revPBE0-D3 achieve lower RMSEs than the other reference methods at larger training set sizes. This is likely due to the larger training set for revPBE0-D3 containing training samples that are more representative of the hold-out test set.

Figure 2d compares scratch training to multihead scratch training. For example, for a training set of size 10, there is one head for each level of theory and each head is trained on 10 configurations before the heads are evaluated on the corresponding hold-out test set for their respective levels of theory. For numerical force errors, scratch models trained on a single level of theory show lower errors than the corresponding head of the mutltihead scratch model. For energy errors, the same is largely true but there is less of an obvious divide. However, by training a multihead scratch model, one is trading a slight decrease in performance for a much more versatile model. For well-studied systems where an optimal functional or reference method is known, this is not particularly useful. However, for complex or less studied systems, having a robust model can allow easy experimentation with different reference methods to help understanding. As mutltihead scratch models show lower performance than conventional scratch models, they will not be taken forward for further validation in this investigation.

Overall, fine-tuned models with very few additional training examples can achieve lower numerical energy and force errors than models trained

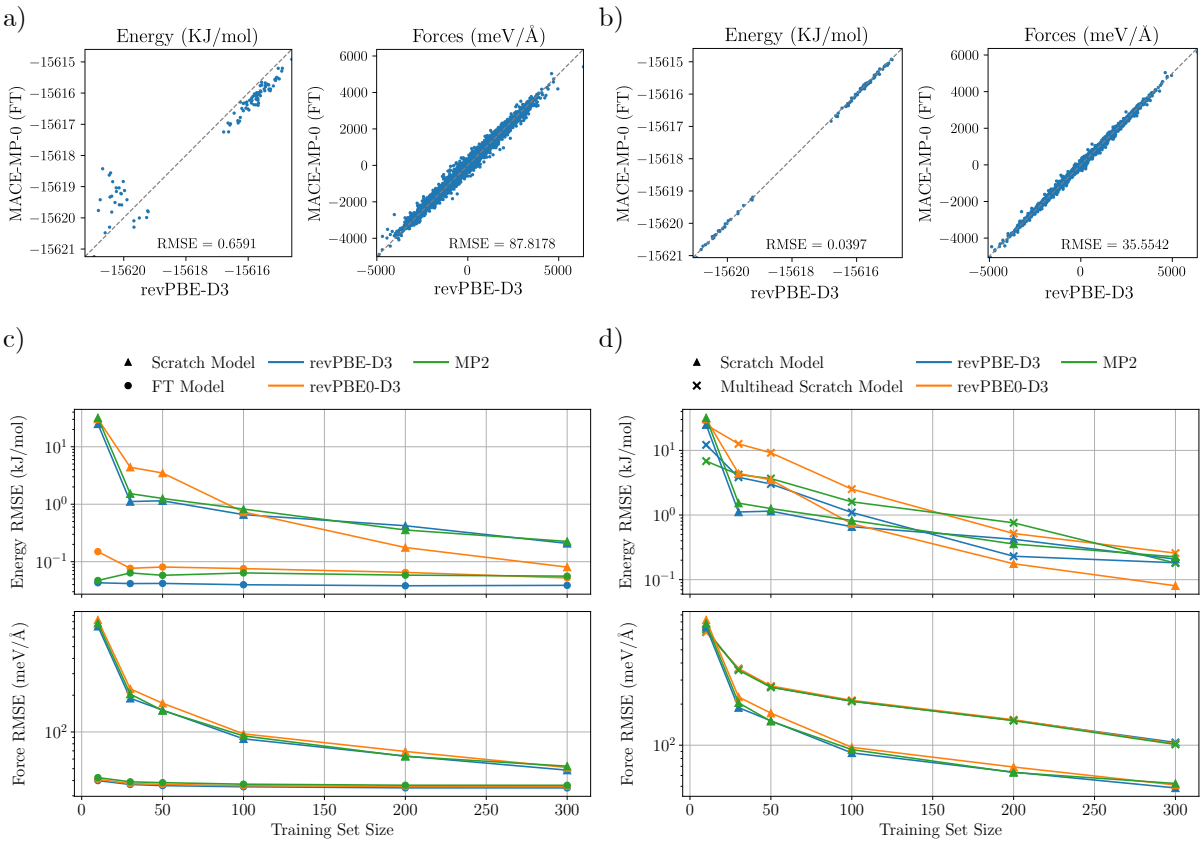

Figure 2: Numerical energy and force error evaluation of scratch MACE, multihead scratch MACE and multihead replay fine-tuned MACE-MP-0 models. Energy and force RMSEs are extracted from the correlation plots, where (a) is a scratch model and (b) is a fine-tuned model. Both models are trained on 100 configurations and tested on the hold-out test set. (c) and (d) show the effect of training procedure, training set size and level of training set theory on numerical force and energy errors (RMSEs). Each RMSE value in (c) and (d) is calculated from its own corresponding correlation plot. (c) compares scratch MACE models to multihead replay fine-tuned MACE-MP-0 models. (d) compares scratch MACE models to multihead scratch MACE models.

from scratch, which utilise at least an order of magnitude more training configurations. However, this hold-out test set only has 100 structures and may not be representative of actual configurations that are important for calculating observables and running molecular dynamics. Further validation of structural and dynamic properties is required to provide a more well-rounded perspective.

5. Results Part 3: Validation of secondary properties

Validation of secondary properties will be performed using scratch and fine-tuned models trained on revPBE-D3 reference data. NVT molecular dynamics simulations of the $CaCO_3/H_2O$ system were run using the Atomic Simulation Environment (ASE) [28] and performed at 300 K for 120 ps, using a randomly selected configuration from the hold-out test set as the initial configuration (Figure 3a). The Bussi thermostat, an extension of the Berendsen thermostat, was used as more consistent temperature control was observed compared to alternative thermostats tested, such as Langevin [29].

Literature studying similar systems commonly use Becke, Lee, Yang, and Parr (BLYP)-based exchange-correlation functionals, meaning for exactly comparable revPBE-D3 data one would need to run their own DFT calculations [30, 31]. However, as the aim of this investigation is to compare scratch and fine-tuned models, it is more important to focus on the comparison between the models, and

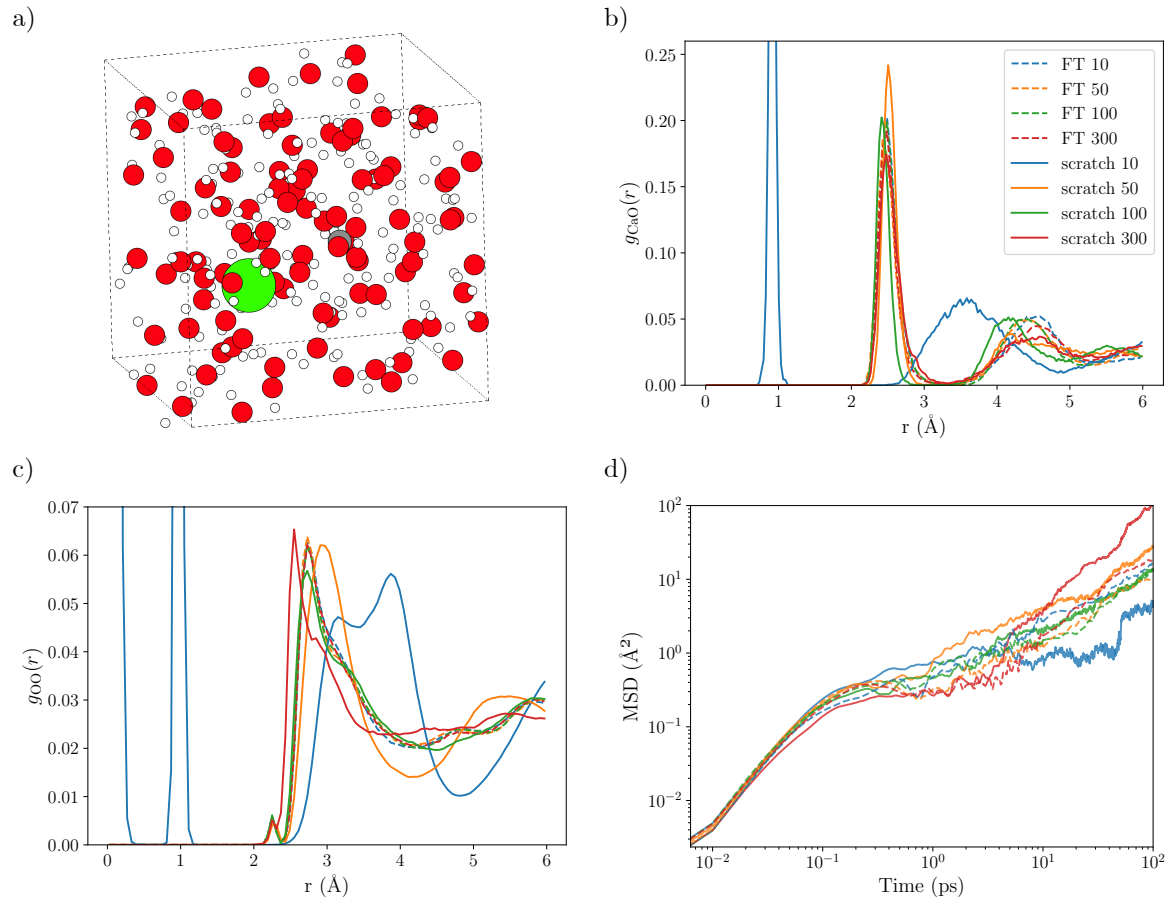

Figure 3: Comparison of static (b,c) and dynamic (d) properties calculated from NVT MD simulations at 300 K using scratch and fine-tuned models of different sizes. (a) shows the randomly selected starting configuration of the $CaCO_3/H_2O$ system from the hold-out test set used for MD. (b) and (c) show the radial distribution functions for Ca-O and O-O respectively, produced from scratch and fine-tuned models of 4 different training set sizes. (d) shows a log-log plot of the averaged MSD of oxygen calculated over the last 100 ps of the 120 ps run.

to look at which training set sizes the models converge at, if they do at all.

Firstly, radial distribution functions for Calcium-Oxygen ($g_{CaO}(r)$) and Oxygen-Oxygen ($g_{OO}(r)$) were calculated for scratch and fine-tuned models of size 10, 50, 100 and 300. These four sizes were chosen as a representative subset of the six available training set sizes to reduce compute time. Figure 3b shows $g_{CaO}(r)$, where all models show good agreement except for the scratch model trained on 10 configurations. Here, an unphysical, sharp peak is seen around 1 Å which is an artefact of an insufficient training set. This small training set has resulted in poor energy and force predictions, making it possible for oxygen atoms to be present at an unphysical distance from the Ca ion. This is expected as this model produced poor force and energy RMSEs (Figure 2), with a substantial decrease in errors upon expanding the training set size to 50 configurations, as seen for the RDF. For the larger training sets, all models produce a sharp first maximum at ~ 2.5 Å corresponding to the ordered first coordination sphere, and a second maximum at $\sim 4.3-4.5$ Å corresponding to the second sphere. For all of these models, a minimum between the two peaks is seen ~ 3.3 Å. Increasing the training set size for scratch models overall shows closer agreement with the fine-tuned models for both peaks. The second peak is less well-defined as the water molecules in the second coordination shell are bound more weakly and are less ordered than in the first, hence the less well-defined peak and lower consistency between model

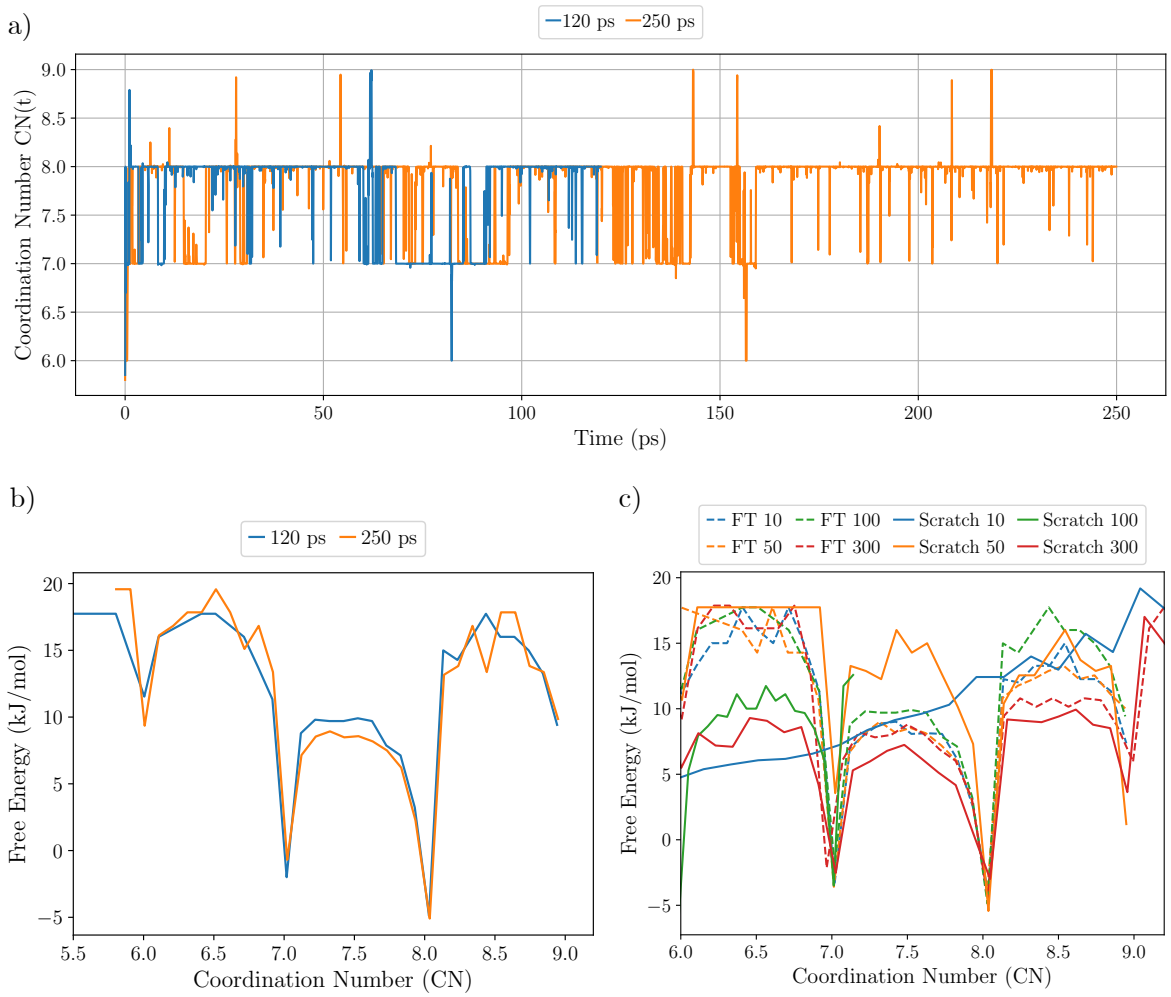

Figure 4: Coordination number of the Ca^{2+} ion. (a) shows the coordination number as a function of time, using the fine-tuned model with a size 100 revPBE-D3 training set. Simulations of length 120 ps and 250 ps at 300 K are compared. (b) shows the free energy as a function of the coordination number for the fine-tuned model (with a revPBE-D3 training set of 100 configurations) at simulation durations 120 ps and 250 ps to test for convergence. This is calculated by using the probability of the system being in coordination number n (Eq. 8). (c) also shows the free energy but testing scratch vs fine-tuned models of different revPBE-D3 training set sizes, all simulated at 300 K for 120 ps. The legend contains the training method (FT or scratch) followed by the training set size (10, 50, 100, 300).

sizes. Simulating for longer times would likely converge this second peak.

For $g_{OO}(r)$ (Figure 3c), we see similar unphysical behaviour for the first scratch model. Two hydration shells are present, where the second one is much less well defined due to the increasing disorder with distance. The amplitude of the maximum $g_{OO}(r)$ is much lower than $g_{CaO}(r)$ as strong electrostatic interactions are no longer involved, reducing the order of the hydration shells. Again, the fine-tuned models show very consistent performance, whereas the scratch models generally become more closely aligned to the fine-tuned models as more data is added to their training set, as before. However, the scratch model trained on 300 configurations appears to perform worse than 100 configurations on this observable. This suggests that the additional training examples were less representative of the dominant configurations in the MD simulation used to calculate the $g_{OO}(r)$, leading to lower accuracy.

For a dynamic property, the mean squared displacement (MSD) of oxygen atoms was calculated. All models show a ballistic regime at short times ($t < 1$ ps) and a diffusive regime at longer times. Again, the fine-tuned models show much more consistent performance compared to the scratch models. Here, the scratch model with training set size 10 appears to have a physically reasonable MSD, showcasing the importance of validating models on multiple observables, static and dynamic. In general, increasing the training set size moves towards convergence of the scratch model with the fine-tuned models. However, we again see the scratch model with 300 configurations shows a decrease in performance compared to 100 configurations. The MSD was calculated from the same MD run as $g_{OO}(r)$, therefore the same hypothesis regarding unfavourable conflagrations during simulation applies.

Calculation of the coordination number (CN) of Ca^{2+} can be challenging due to the long simulation time required to explore all the possible coordination numbers and reach convergence [30]. Many reactions are highly dependent on the geometry of the calcium-water complexes and hence the coordination number [32], therefore it is vital to correctly model. The coordination number was determined as outlined by Moison et al. [30], using the following analytical expression:

$$\mathrm{CN}(t) = \sum_{I=1}^{N_O} \frac{1 - \left(\frac{R_I(t)}{r_{CaO}^{\min,1}}\right)^n}{1 - \left(\frac{R_I(t)}{r_{CaO}^{\min,1}}\right)^m} \tag{7}$$

where $R_I(t)$ is the distance of oxygen atom I from the Ca ion, N_O is the total number of oxygen atoms, and n and m are parameters set to 60 and 200 respectively. Using this expression, coupled with long simulation times, can help address the convergence issue by tracking the coordination number throughout the simulation and mitigating its rapid fluctuations. This ensures a smooth transition as oxygen atoms enter and exit the first coordination sphere. Furthermore, the free energy is able to be calculated as a function of the probability p of different coordination numbers:

$$F = -k_B T \ln p. \tag{8}$$

Firstly, the fine-tuned model with training set size 100 was chosen as a high-performing model to probe Ca^{2+} coordination numbers. Figure 4a shows the time evolution of the CN whilst simulating at 300 K for 120 ps and 250 ps. The free energy was then calculated using Eq. 8, shown in Figure 4b. Free energy minima are located at each coordination number, where coordination number 8 is the most stable as it has the lowest free energy. Due to the stability of each coordination number, there are energy barriers between them. These energy barriers make resolving all the coordination numbers difficult as the system can be stuck in local minima for tens of picoseconds. Increasing simulation time from 120 ps to 250 ps does not show a significant difference in the free energy plot, suggesting that 120 ps is sufficient for exploring and converging the coordination numbers of Ca^{2+}. Therefore, using an MLIP has achieved the same conclusions regarding the convergence of the coordination number made by Moison et al. [30], but at a fraction of computational cost compared to their ab initio molecular dynamics simulations.

Now it has been shown that 120 ps is enough to adequately explore the different coordination numbers of Ca^{2+}, Figure 8c compares the free energy as a function of CN for scratch and fine-tuned models of different training set sizes. As before, the scratch model trained on 10 configurations results in a non-physical free energy curve with no minima. All fine-tuned models perform similarly, however, this observable shows us the largest variation in fine-tuned model performance in this investigation, likely due to the system being in local minima for extended periods of time. A possible hypothesis is that increasing the temperature would result in better convergence of the free energy profiles of all the fine-tuned models as local minima and energy barriers will have less effect. Furthermore, the exploration of coordination numbers will be enhanced with the additional thermal energy.

Scratch models see increased performance with increased training set size from coordination number 7 and higher. For lower coordination numbers, training set size 50 better matches the fine-tuned curves in terms of free energy values, however, the observed plateau appears anomalous and is due to a lack of data points in between coordination numbers 6 and 7. As this is a relatively high energy barrier, the model struggles to resolve the transition state, where the energy landscape is steep. A potential solution could be to increase the temperature or perform enhanced sampling around the transition state [33]. The scratch model trained on 100 configurations was only able to explore coordination

numbers 6 and 7, which would be solved by using a higher temperature. The scratch model trained on 300 configurations performs well around the free energy minima but underpredicts the energy barriers. Overall, fine-tuned models are clearly superior for the static and dynamic observables tested, as well as numerical force and energy errors tested in the previous section.

6. Conclusions

In summary, it has been successfully shown that fine-tuning using the MACE-MP-0 foundation model with only tens of training configurations is able to achieve lower numerical energy and force errors, as well as more consistent prediction of secondary properties, than models trained from scratch using the same architecture. Remarkably, training from scratch with 300 configurations produced higher force and energy errors than fine-tuning with just 10 configurations. This showcases the data efficiency of fine-tuning, where after the initial high computational cost of training the foundation model, it is possible to reach quantitative accuracy with just tens of additional reference calculations. This powerful technique will have many uses in applications where reference data is expensive, such as modelling large materials or biomolecules [34]. Furthermore, the ability to reach quantitative accuracy at such a low computational cost will lower the entry barrier to performing large-scale simulations, therefore driving innovation. Additionally, the improved method of recalibrating problematic E0 values showed a reduction in initial energy RMSEs by two orders of magnitude, and will be integrated into MACE in the near future.

To extend this research, it would be interesting to fine-tune the more recently released MACE foundation models, MACE-MPA-0 and MACE-OMAT-0, which are pre-trained on substantially more data using the sAlex and OMAT datasets [35]. Potentially under 10 representative configurations may be required to reach quantitative accuracy, which would be an extremely powerful tool. Furthermore, a full exploration of multihead scratch models would be interesting to further research and benchmark their performance. Additionally, investigating the coordination number of Ca^{2+} at higher temperatures would provide a more in-depth understanding, as local minima will be easily escaped and more coordination states explored.

Acknowledgements

I would like to thank Dr Christoph Schran for his supervision throughout the project and Ilyes Batatia for a helpful discussion on E0s.

Supplementary Information

Training and analysis scripts can be found at:
`https://github.com/joehart2001/WA_MACE_FT`

References

[1] R. Martin-Barrios, E. Navas-Conyedo, X. Zhang, Y. Chen, J. Gulín-González, An overview about neural networks potentials in molecular dynamics simulation, International Journal of Quantum Chemistry 124 (11) (2024) e27389.

[2] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schutt, A. Tkatchenko, K.-R. Müller, Machine learning force fields, Chemical Reviews 121 (16) (2021) 10142–10186.

[3] A. P. Bartók, R. Kondor, G. Csányi, On representing chemical environments, Physical Review B—Condensed Matter and Materials Physics 87 (18) (2013) 184115.

[4] F. L. Thiemann, N. O'neill, V. Kapil, A. Michaelides, C. Schran, Introduction to machine learning potentials for atomistic simulations, Journal of Physics: Condensed Matter 37 (7) (2024) 073002.

[5] I. Poltavsky, A. Tkatchenko, Machine learning force fields: Recent advances and remaining challenges, The journal of physical chemistry letters 12 (28) (2021) 6551–6564.

[6] Y. Wang, Z. Li, A. Barati Farimani, Graph neural networks for molecules, in: Machine learning in molecular sciences, Springer, 2023, pp. 21–66.

[7] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, B. Kozinsky, E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, Nature communications 13 (1) (2022) 2453.

[8] B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel, G. Ceder, Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling, Nature Machine Intelligence 5 (9) (2023) 1031–1041.

[9] C. Chen, S. P. Ong, A universal graph deep learning interatomic potential for the periodic table, Nature Computational Science 2 (11) (2022) 718–728.

[10] A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, E. D. Cubuk, Scaling deep learning for materials discovery, Nature 624 (7990) (2023) 80–85.

[11] I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, M. Avaylon, W. J. Baldwin, et al., A foundation model for atomistic materials chemistry, arXiv preprint arXiv:2401.00096 (2023).

[12] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, G. Csányi, Mace: Higher order equivariant message passing neural networks for fast and accurate force fields, Advances in neural information processing systems 35 (2022) 11423–11436.

[13] G. Vrbančič, V. Podgorelec, Transfer learning with adaptive fine-tuning, IEEE Access 8 (2020) 196197–196211.
[14] H. Kaur, F. Della Pia, I. Batatia, X. R. Advincula, B. X. Shi, J. Lan, G. Csányi, A. Michaelides, V. Kapil, Data-efficient fine-tuning of foundational models for first-principles quality sublimation enthalpies, Faraday Discussions 256 (2025) 120–138.
[15] F. Lightstone, E. Schwegler, M. Allesch, F. Gygi, G. Galli, A first principles molecular dynamics study of calcium ion in water, ChemPhysChem 6 (UCRL-JRNL-209868) (2005).
[16] K. Suzuki, H. Sorimachi, A novel aspect of calpain activation, FEBS letters 433 (1-2) (1998) 1–4.
[17] C. S. Adams, K. Mansfield, R. L. Perlot, I. M. Shapiro, Matrix regulation of skeletal cell apoptosis: role of calcium and phosphate ions, Journal of Biological Chemistry 276 (23) (2001) 20316–20322.
[18] S. Gheytani, Y. Liang, F. Wu, Y. Jing, H. Dong, K. K. Rao, X. Chi, F. Fang, Y. Yao, An aqueous ca-ion battery, Advanced Science 4 (12) (2017) 1700465.
[19] N. Bernstein, From gap to ace to mace, arXiv preprint arXiv:2410.06354 (2024).
[20] K. W. Church, Z. Chen, Y. Ma, Emerging trends: A gentle introduction to fine-tuning, Natural Language Engineering 27 (6) (2021) 763–778.
[21] T. L. Hayes, K. Kafle, R. Shrestha, M. Acharya, C. Kanan, Remind your neural network to prevent catastrophic forgetting, in: European conference on computer vision, Springer, 2020, pp. 466–483.
[22] Y. Lim, H. Park, A. Walsh, J. Kim, Accelerating co direct air capture screening for metal-organic frameworks with a transferable machine learning force field (2024).
[23] G. Strang, Linear algebra and its applications, Chapter 3, Thomson, Brooks/Cole, 2000.
[24] S. Weisberg, Applied linear regression, Chapter 2, Vol. 528, John Wiley & Sons, 2005.
[25] Å. Björck, Numerical methods for least squares problems, SIAM, 2024.
[26] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 Contributors, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods 17 (2020) 261–272. doi:10.1038/s41592-019-0686-2.
[27] J. S. Almeida, Predictive non-linear modeling of complex data by artificial neural networks, Current opinion in biotechnology 13 (1) (2002) 72–76.
[28] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, K. W. Jacobsen, The atomic simulation environment—a python library for working with atoms, Journal of Physics: Condensed Matter 29 (27) (2017) 273002.
[29] G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling, The Journal of chemical physics 126 (1) (2007).
[30] H. Moison, J. Aufort, M. Benoit, M. Méheut, On local structure equilibration of ca2+ in solution by ab initio molecular dynamics, The Journal of Physical Chemistry B 128 (13) (2024) 3167–3181.
[31] W. A. Adeagbo, N. L. Doltsinis, M. Burchard, W. V. Maresch, T. Fockenberg, Ca2+ solvation as a function of p, t, and ph from ab initio simulation, The Journal of Chemical Physics 137 (12) (2012).
[32] A. K. Katz, J. P. Glusker, S. A. Beebe, C. W. Bock, Calcium ion coordination: a comparison with that of beryllium, magnesium, and zinc, Journal of the American Chemical Society 118 (24) (1996) 5752–5763.
[33] J. Debnath, M. Invernizzi, M. Parrinello, Enhanced sampling of transition states, Journal of chemical theory and computation 15 (4) (2019) 2454–2459.
[34] M. Elstner, T. Frauenheim, S. Suhai, An approximate dft method for qm/mm simulations of biological structures and processes, Journal of Molecular Structure: THEOCHEM 632 (1-3) (2003) 29–41.
[35] L. Barroso-Luque, M. Shuaibi, X. Fu, B. M. Wood, M. Dzamba, M. Gao, A. Rizvi, C. L. Zitnick, Z. W. Ulissi, Open materials 2024 (omat24) inorganic materials dataset and models, arXiv preprint arXiv:2410.12771 (2024).