J. Phase Equilib. Diffus.
https://doi.org/10.1007/s11669-025-01225-z

ORIGINAL RESEARCH ARTICLE

Fine-Tuning Universal Machine-Learned Interatomic Potentials for Applications in the Science of Steels

Naveen K. Mohandas[1] · Sebastián Echeverri Restrepo[2] · Marcel H. F. Sluiter[1,3]

Submitted: 25 September 2025 / in revised form: 10 December 2025 / Accepted: 11 December 2025

Abstract Recycling steel at scale is hindered by tramp elements such as Cu and Sn, which degrade material properties. Atomistic simulations using foundational machine-learned interatomic potentials (MLIPs) trained on large databases, such as Materials Project, Alexandria, and OMAT, offer a promising approach to study the effects of these impurities. However, fine-tuning these models to specific systems can lead to catastrophic forgetting–the loss of general chemical knowledge acquired during pretraining. Here, we evaluate forgetting in three foundational MLIPs: CHGNet, SevenNet-O, and MACE, by fine-tuning on a data set of bcc-based structures, with Fe atoms only. When evaluated on a subset of the Materials Project data set with a learning rate of 0.0001, the fine-tuned MLIPs of CHGNet and SevenNet-O exhibited only a minor increase in RMSE of 0.047 and 0.022 eV/atom, respectively, indicating markedly minor forgetting. In contrast, fine-tuned MACE exhibited catastrophic forgetting, despite a range of additional strategies such as layer freezing and data set replay. We attribute the catastrophic forgetting to architectural sensitivity. These results highlight the importance of fine-tuning hyperparameters, model architecture, and data set design, with fine-tuned models of CHGnet and SevenNet-O showing some potential for efficient and transferable modeling of recycled steels.

Keywords CHGNet · fine-tuning · iron · machine learned interatomic potentials · MACE · SevenNet-O

✉ Marcel H. F. Sluiter
m.h.f.sluiter@tudelft.nl

Naveen K. Mohandas
n.k.mohandas@tudelft.nl

Sebastián Echeverri Restrepo
sebastian.echeverri.restrepo@skf.com

[1] Department of Materials Science and Engineering, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands

[2] SKF Research and Technology Development (RTD), SKF B.V., Meidoornkade 14, 3992 AE Houten, The Netherlands

[3] Metal Science and Technology, Department of Electromechanical, Systems and Metal Engineering, Ghent University, Technologiepark 46, 9052 Ghent, Belgium

1 Introduction

Transitioning to more sustainable steel manufacturing requires increased utilization of scrap steel. Scrap steel often introduces so-called tramp elements, such as Cu from electrical parts in car bodies, into the steel production process. Other common tramp elements that affect the quality of the final product are Sn, Cr, and Ni.[1] Currently, the detrimental effects of these elements are controlled by diluting them with pure iron. However, this strategy imposes a limitation on the volume of scrap steel that can be effectively recycled during steel production. Improving the use of scrap steel requires a deeper understanding of the detrimental effects of the tramp elements. Atomistic simulations are a promising method for gaining such insights.

Conventionally, atomistic simulations have been performed using density functional theory (DFT) or empirical interatomic potentials. These two methods come with their own advantages and disadvantages. DFT is accurate but is computationally expensive, practical systems are limited to only a few hundred atoms.[2–4] On the other hand, simulations employing empirical potentials can deal with millions of atoms, but these potentials are limited to the

Published online: 22 January 2026

specific systems for which they have been developed.[5] Each new impurity or alloying element in steel requires painstaking development of a new potential, a process that becomes even more complicated for multi-component systems. This is where machine-learned interatomic potentials (MLIP) show promise. MLIPs have been fit to data generated using density functional theory (DFT), approaching the level of accuracy of DFT. However, MLIPs, especially for systems with many electrons, can yield energies and forces orders of magnitude faster than DFT. Since these MLIPs are built on local atomic environments, they can be scaled up to model larger length scales in a manner comparable to empirical potentials. Over the years, various MLIPs have been developed each targeting higher accuracy and lower computational cost.[6–9]

Numerous machine-learning interatomic potentials (MLIPs) have been developed for iron and iron-based alloys.[10–18] Many of these MLIPs use explicitly constructed, physics-informed descriptors of the local atomic environment as inputs to the fitting procedure.[19] Some of these MLIPs also incorporate magnetism,[13,15,20] which is important for systems involving iron. However, such descriptor-based MLIPs (e.g., GAP, MTP, ACE) face scaling limitations with increasing number of chemical species due to the combinatorial growth of descriptor terms and the amount of required training data.[19] This combinatorial growth poses a practical challenge when modeling realistic steels that contain numerous alloying elements.

Recently, MLIPs utilizing graph-based neural networks (GNNs) have gained popularity owing to their broad applicability across a wide range of material systems. This is achieved by training on data from large databases, such as the Materials Project,[21] Alexandria[22] and OMAT,[23] each containing millions of structures, molecules and compounds. M3GNet,[24] CHGNet,[25] MACE,[26] SevenNet-O,[27] GRACE,[28] Mattersim[29] and eqV2[30] are some of the potentials in this category.

In materials science, MLIPs are used to study defects such as vacancies, dislocations, and grain boundaries, which are critical to material behavior and must be accurately captured in simulations. While MLIPs are trained on energy per atom to normalize the loss function, defect energies are calculated as energy differences between configurations–for example, the vacancy formation energy is obtained from $E^{f}_{vac} = E^{vac}_{n-1} - \frac{n-1}{n} E^{ref}_{n}$ (where E^{vac}_{n-1} is the energy of the supercell with one vacancy and E^{ref}_{n} is the bcc Fe reference supercell with n atoms). As a result, even small per-atom errors (~ 1 meV) can lead to large inaccuracies in defect energies, especially for larger supercells where $n \geq 100$.[31,32] Various studies have now reported this limited applicability of out-of-the-box universal potentials to defect properties, thus requiring fine-tuning.[33–35]

Fine-tuning is a transfer learning technique where a pre-trained model is adapted to a new task.[36,37] The model is first pretrained on a large data set containing a wide variety of elements and compounds to capture general element interactions and structure–property relationships. It is then fine-tuned using a data set tailored to the specific application to deliver accurate energy predictions for the target system. Fine-tuning has proven to be data-efficient;[34,38–40] Radova et al.[39] reported that using only 10–20% of the training samples was sufficient to achieve an accuracy comparable to that of a model trained on the full dataset. Despite these advantages, a major challenge during fine-tuning of foundational models is catastrophic forgetting.[36,41] This is particularly relevant when the application of an MLIP relies on knowledge acquired during its pretraining. For instance, in the case of steels, preserving the learned interactions with various alloying elements from the pretraining allows the model to handle impurities absent during fine-tuning.

In this study we evaluate catastrophic forgetting in foundational models, namely CHGNet,[25] MACE[26] and SevenNet-O,[27] by fine-tuning them to a data set containing pure Fe structures only (referred to as the "Fe data set" below). The performance of the fine-tuned MLIPs was then tested by predicting the binding energies of various elements in the bcc Fe matrix, a task that requires retention of information from the pretraining.

2 Methods

2.1 Machine Learned Interatomic Potentials

Graph-based neural networks (GNNs) represent material structures as graphs, with nodes corresponding to atoms and edges capturing their connections to neighbouring atoms. This allows a GNN to learn the representation of the material structure and its relationship to properties during training instead of relying on crafted features or fixed descriptors.[42] These models rely on message passing to extract information from the neighbouring atoms, allowing incorporation of more than two body interactions. Some models, such as MACE, further enforce many-body terms directly in the message passing while also accounting for invariance and equivariance.[26]

Here, we select the models CHGNet, MACE and SevenNet-O for fine-tuning (CHG2-FT, MACE-FT, Sevenn-FT). We choose CHGNet as it directly incorporates magnetic moments in the architecture, which is important for magnetic systems such as Fe.[43] Although, CHGNet is trained only on the absolute values of the magnetic

moments, Deng et al.[25] demonstrated its usefulness by identifying the oxidation state of transition metal ions in compounds. Specifically, V^{4+} and V^{3+} were identified in Na_2 $V_2(PO_4)_3$ through the magnetic moments.[25] CHGNet is then compared with MACE, and SevenNet-O, each of them trained on the same Materials Project (MPTRJ) data set. This choice is made to make a fair comparison with the CHGNet model although there exists MACE and SevenNet-O models trained on larger data sets.

2.1.1 CHGNet

Crystal Hamiltonian Graph neural network abbreviated as CHGNet[25] is a GNN based MLIP. It is constructed such that it consists of two graphs, the atom graph and the bond graph. The atom graph captures the non-directional bonding information, whereas the bond graph captures the directional bonding information. The message passing, referred to as the interaction block in the CHGNet, incorporates the many body interactions. Unlike other MLIPs, CHGNet predicts absolute magnetic moments after three layers of message passing from the atom embedding. This information is then used to infer local charge redistribution in ionic systems. In this study we use the '0.2.0' version of the CHGNet (referred to as CHG2) as it was observed that the '0.3.0' model (CHG3) performed poorly for the Fe system.[35]

2.1.2 MACE

MACE[26] is another GNN based MLIP which uses spherical harmonics and radial basis functions to generate descriptors for atomic environments. The design of MACE enables it to gather the important features from neighbours with just two message-passing layers, making it fast and scalable. Various versions of the model trained on different data sets are available. Here, we use the updated model trained on the MPTRJ data set named 'mace-medium-mp-0b3'.

2.1.3 SevenNet-O

SevenNet-O is a MLIP based on the NequIP[44] architecture with a focus on scalability. The E(3)-equivariant[1] representation in NequIp allows SevenNet-O to accurately capture the atomic interactions while respecting the equivariance constraints. SevenNet-O implements a parallelization strategy that allows simulation of large supercells using multiple GPUs. Here we use the 11July2024 model that was trained on the MPTRJ data set.

2.2 Fine-Tuning the MLIP

We use foundational models pretrained on large data sets covering 94 elements from the periodic table and fine-tune them to better fit the specific case of iron. We compare how different models, pretrained on the same data set, perform during fine-tuning. For proper comparison, all foundational models were fine-tuned using the same training, validation, and test datasets.

There are numerous hyperparameters that influence the fine-tuning. Here we focus on different learning rates as they dictate the change in trainable parameters. Three learning rates 0.01, 0.001 and 0.0001 were used. All other hyper-parameters were kept same as those used during the pretraining. The MLIPs were fit to energy and force during fine-tuning, for CHGNet magnetic moments were used as well.

Various strategies are available to minimize forgetting during fine-tuning,[41] such as freezing layers, replaying the original data set, employing sub-networks, or using dynamic architectures. Freezing layers reduces forgetting by keeping part of the trainable parameters fixed during subsequent fine-tuning,[39] while data set replay incorporates all or part of the original data set into the fine-tuning process.[45] Sub-networks restrict training to smaller parts of the model, and dynamic networks expand the architecture by adding new parameters to accommodate new tasks.[34] A key limitation of architecture-based methods is that they are best suited for clearly separated tasks; they cannot accommodate scenarios where both old and new knowledge must be combined for a single prediction. An alternative is the use of regularization-based approaches, such as elastic weight consolidation (EWC).[46] In this technique, fine-tuning is guided by penalizing changes to parameters deemed important for the original task, thereby preserving prior knowledge without explicitly having a need for the original training data. However, in comparison to other continual learning approaches regularization-based methods perform worse.[47]

Here, we examine two strategies: (1) with all parameters trainable, referred to as naive fine-tuning, and (2) freezing of layers. For MACE, we additionally evaluated as a third strategy the multi-head replay strategy.[48] This strategy allows for learning at multiple levels of theory while maintaining transferability across systems. Specifically, MACE employs two readout layers, or "heads": one dedicated to the new task and the other to the original task, with 99% of the trainable parameters shared between them. During fine-tuning, one head is trained on the new data set while the other is simultaneously trained on a subset of the

[1] E(3)-equivariance ensures that scalar quantities such as energies remain invariant, while vector quantities like forces transform exactly as they should under rotations, translations, and reflections.

original data set to minimize forgetting. This strategy was applied using the implementation available in the MACE PyPI release (version 0.3.12).[49]

2.2.1 Freezing of layers

The considered MLIPs contain numerous layers with trainable parameters. These layers are designed to emulate underlying physical laws; however, due to the black-box nature of the models, it is difficult to verify whether they capture such behavior. The MLIPs begin by encoding the element information in the first layer. It could be expected that the model learns to differentiate the elements in this layer and then during message passing learns the interactions with neighbouring atoms. The MLIP then predicts the energy for each atomic site and next sums it to predict the energy of the supercell.

By freezing some of the layers it would retain the information the model learned during the pretraining. Radova et al.[39] found freezing the first four layers to be optimal for MACE, hence we used the same strategy for MACE. For CHGNet, layers were frozen based on their function in the network, while the first embedding layer was always kept frozen. As training models for various conditions are quite computationally intensive, only MACE and CHGNet models were chosen for frozen learning.

2.2.2 Complementing MP data set

The training data set plays an important role during the training of the NNs. Various methods are used in the literature to generate data, such as active learning,[50,51] sampling configurations from ab-initio MD,[11] or designed configurations that target defect and other properties. Each strategy comes with its own advantages and limitations. Active learning involves performing molecular dynamics (MD) simulations with an uncertainty measure in the MLIP. The uncertainty measure is used to determine if a prediction is out-of-distribution. When an out-of-distribution structure is encountered, it is evaluated using DFT and subsequently incorporated into the training database. This is used for MLIPs such as ACE,[8] MTP[7] and GAP.[52] For graph based MLIPs, generating an uncertainty measure is not straightforward. The common approach is bootstrapping, where an ensemble of models are trained with different hyper-parameter initializations or using different training, validation and test sets.[53] The uncertainty is then quantified using the standard deviation of the predictions across the ensemble models. This is computationally expensive, as multiple models are required for predictions.

Another way is to generate data sets by running ab-initio MD simulations and extracting configurations from various time steps. The shortcoming with this approach is the absence of configurations that correspond to rare events such as vacancy formation or vacancy migration. In such cases, it is necessary to bias the system towards these rare events or explicitly incorporate them in the data set, as done by Meng et al.[11] for studying hydrogen in iron. This data set allows studying fracture in bcc-Fe in the presence of hydrogen because it contains various configurations covering a large sample space of vacancies, grain boundaries, free-surfaces and deformed cells.

Similar to Meng's data set, we generated a DFT data set corresponding to pure Fe with input parameters consistent with the MPTRJ data set. This approach was adopted to prevent errors that could result from energy discrepancies caused by varying DFT convergence criteria. *Vienna ab initio simulation package* (VASP)[54] was used to run the ab-intio simulations. For consistency, the input files were generated using *MPMetalRelaxSet* implemented in Pymatgen package.[55] The calculations were performed using the Perdew–Burke–Ernzerhof (PBE) functional within the generalized gradient approximation (GGA).[56] The number of valence electrons used for the Fe pseudopotential was 8.

A subset of the original MPTRJ data set consisting of 90000 structures ($\sim 10\%$) was used to evaluate the catastrophic forgetting in the MLIPs. It was generated by randomly sampling from the MPTRJ data set.[57] It is referred to as MPTRJ validation set from here on. Root Mean Square Error (RMSE) metric was then used as an estimate of the validation error.

2.3 Testing performance on selected Fe properties

Defects such as vacancies, dislocations, and grain boundaries are of particular interest to materials scientists because they play a crucial role in determining the macroscopic properties and overall behavior of materials. However, as discussed earlier, discrepancies between the fitting procedures of MLIPs and the prediction of defect energies often lead to significant errors in the latter.[31,32]

One approach to address this issue is to incorporate defect energies directly into the loss function, requiring the MLIP to be trained on defect energies rather than supercell energies–something not feasible when fine-tuning foundational models. Alternatively, defect structures can be assigned higher weights during training. In this work, we emphasize defects by including a large number of similar defect configurations in the training data set, effectively increasing their weight and improving the model's accuracy for the defects in the Fe system.

The MLIPs after fine-tuning were then validated on the properties of bcc Fe such as elastic properties, vacancy formation energy, vacancy migration energy, surface

energies and grain boundary energies. For all validation simulations the Atomic Simulation Environment (ASE)[58] package was used.

The elastic tensor was determined using the stress–strain method as per.[59,60] Six strain tensors (ϵ) with four magnitudes each were applied to a relaxed bcc Fe configuration. Then, with the predicted stress(σ), Hookes' law was used to calculate the elastic tensor ($\sigma = \mathbf{C}\epsilon$, where $\mathbf{C}$ is the stiffness tensor). For evaluating cubic elastic stability the bulk modulus(B), and C_{44} and C' shear moduli are considered.

The linear coefficient of thermal expansion (α) and the constant-pressure specific heat (C_P) were determined from MD simulations in the NPT ensemble. Due to computational cost of MLIPs a smaller supercell with 432 atoms was used and MD simulations were run for 100000 steps with a time step of 1 fs. The average volume and enthalpy were measured at temperatures between 200 K and 1000 K in 100 K increments. Cubic-spline interpolations of volume and enthalpy as function of temperature, $V(T)$ and $H(T)$, were constructed, and their derivatives at 300 K, $(\mathrm{d}V/\mathrm{d}T)|_P$ and $(\mathrm{d}H/\mathrm{d}T)|_P$, were used to evaluate α and C_P, respectively.

To calculate the vacancy formation energy, a supercell with 128 Fe atoms was generated corresponding to the equilibrium lattice parameter. An atom was then removed to introduce a vacancy, the atoms in the supercell were allowed to relax but the dimensions of the supercell were kept fixed. The vacancy formation energy was determined using:

$$E^f_{\square} = E[Fe_{n-1}\square] - \frac{n-1}{n}E^{ref}[Fe_n] \quad \text{(Eq 1)}$$

where $E[Fe_{n-1}\square]$ is the energy of the relaxed supercell containing a vacancy with $n-1$ Fe atoms and $E^{ref}[Fe_n]$ is the energy of the relaxed supercell containing n Fe atoms.

The climbing image nudged elastic band (CI-NEB) method, as implemented in the ASE package, was used to determine the nearest-neighbour vacancy migration energy.

The bcc surfaces (100), (110) and (111) for Fe were generated using the ASE package with the equilibrium lattice parameter predicted by each corresponding MLIP. A 10 Åvacuum gap is added to the simulation supercell to prevent interactions between surfaces caused by periodic boundary conditions (PBC). During relaxation, only the atom positions were relaxed keeping the dimensions of the supercell fixed. The surface energy is then calculated using:

$$E^{ijk}_s = \frac{E^{ijk}[Fe_n] - \frac{n}{m}E^{ref}[Fe_m]}{2A^{ijk}} \quad \text{(Eq 2)}$$

where $E^{ijk}[Fe_n]$ is the energy of the supercell with the surface with normal $[ijk]$ and n Fe atoms, $E^{ref}[Fe_m]$ is the energy of the bcc Fe supercell with m atoms and A^{ijk} is the area of the free surface.

Grain boundaries (GBs) are an important class of defects in polycrystalline materials. Mechanical properties such as hardness, yield strength and brittleness are directly influenced by GBs.[61] To study realistic systems using MLIPs it is important to evaluate their performance on grain boundary energies. However, ab-initio studies are limited to low Σ symmetric GBs due to the complexity of grain boundaries combined with the limitations of the supercell size and the PBC in ab-initio methods. We evaluate the MLIP on $\Sigma3$, $\Sigma5$, $\Sigma7$ and $\Sigma9$ symmetric grain boundaries and compare it with the values obtained in the literature.[62] The GB structures were extracted from the Materials Project Database, the atom positions in these structures were relaxed keeping the supercell fixed. Then the GB energy was determined using:

$$E_{GB} = \frac{E_{GB}[Fe_n] - \frac{n}{m}E^{ref}[Fe_m]}{2A_{GB}} \quad \text{(Eq 3)}$$

where $E_{GB}[Fe_n]$ is the energy of the supercell with the grain boundary containing n Fe atoms, $E^{ref}[Fe_m]$ is the energy of the bcc Fe supercell with m atoms, and A_{GB} is the area of the GB plane.

2.4 Evaluation of solute-solute interactions

Substitutional atoms affect the properties of steel through various mechanisms, such as grain size refinement, solid solution hardening, and precipitation hardening.[63] This is in turn affected by the tendency of these elements to segregate. Their segregation tendency is estimated using their nearest neighbour interactions. Here, we calculate the solute-solute interactions for the first five nearest neighbours for various substitutional atoms. In addition to Cu, Ni and Sn we also consider other commonly found elements in steel: Al, Ti, Zn, Mo, Nb and V. Equation 4 was used to calculate the binding energies for each of these configurations, where X and Y are substitutional atoms. $E[Fe_{n-1}X]$ and $E[Fe_{n-1}Y]$ is the energy of supercell containing 1 substitutional atom X or Y in the Fe supercell and $E^{ref}_n[Fe]$ is the reference energy of supercell containing n Fe atoms.

$$E^{XY}_{be} = E[Fe_{n-2}XY] + E^{ref}[Fe_n] - (E[Fe_{n-1}X] + E[Fe_{n-1}Y]) \quad \text{(Eq 4)}$$

Interstitial solutes such as C, N and O are important in steel and affect its properties, even at low concentrations. Due to their small size they occupy the octahedral interstices in the lattice. For simulations involving diffusion of these interstitial atoms, it is important that the MLIPs predict the octahedral site to be the stable configuration and the tetrahedral site as a saddle point during diffusion.[64,65]

Hence, we used the MLIP to predict the energy of configurations with a solute atom in the octahedral interstices and tetrahedral interstices. The difference in energy ($\Delta E_{oct-tet}$) was then used to identify the stable site in the lattice. In addition, the migration barrier was also determined using the CI-NEB for the elements C, O and N.

3 Results and Discussion

3.1 Fine-tuning the MLIPs

The performance of the foundational models was initially evaluated on the Fe data set, see Fig. 1. All foundational models reproduced the DFT data set poorly, under-predicting the energies for most of the structures. This is in accordance with previous work reporting that models trained only on near-equilibrium structures exhibit systematic softening.[34] Furthermore, even though the data set was generated using the MPMetalRelaxset, there is an offset in the energies for all the foundational models.

Figure 2 presents the performance of the models on the MPTRJ validation set introduced in Sect. 2.2.2, after being fine-tuned on the Fe data set. The foundational models CHG2, MACE, and SevenNet-O show good fit to the MPTRJ validation set with RMSEs of 0.063, 0.043 and 0.040 eV/atom respectively (Sect. A.1). After fine-tuning, the MLIPs show an increase in the RMSE values signifying forgetting in the models. The models were trained for 50 epochs. Possibly, the models do not need the full 50 epochs to achieve a good fit to the Fe data set. However, since the focus of this work is on catastrophic forgetting, it is beneficial to examine how extended training over more epochs influences forgetting.

After fine-tuning, both naive and replay MACE models (Sect. 2.2), exhibit high RMSE values of 4.63 eV/atom and 0.605 eV/atom on the MPTRJ validation set. The large error of the naive model indicates that transfer learning does not occur effectively and that the model largely forgets the knowledge gained during initial training.

The MACE model trained with the replay approach preserves most of the learned information; however, a few configurations show significant deviations. As RMSE was employed as the evaluation metric, these outlier configurations disproportionately influence the reported error. Both CHG2 and SevenNet-O show less forgetting than MACE, as evidenced by the minor increase in the RMSE to 0.111 eV/atom and 0.062 eV/atom, respectively. The majority of data lies clustered close to the ground truth prediction. For simulations involving structural relaxations, dynamical processes, and kinetic processes, MLIPs must accurately predict the forces. Figure 3 shows the performance of the CHG2-naive, SevenNet-O, MACE-naive and MACE-replay on the forces after fine-tuning. For CHG2 and SevenNet-O, similar to the energy predictions, the performance appears similar to the foundational model (Fig. 9) however there is an increase in the RMSE to 0.397 and 0.267 $eV/\text{Å}$, respectively. On the contrary, the performance of the MACE-replay model is poor with a drastic increase in the RMSE to 1.730 $eV/\text{Å}$. It is noted that the force errors for all the fine-tuned models are too large to be useful for simulations involving structural relaxations, kinetics, or dynamics.

Figure 4 shows the change in the MPTRJ RMSE as the fine-tuning progresses for CHG2-naive, SevenNet-O and MACE-FT-replay. The MPTRJ RMSE increases for all models, though only marginally for CHG2. Fine-tuning is influenced by numerous hyper-parameters, with the learning rate being a key factor. To investigate the impact of learning rate on forgetting, we examine three learning rates: 0.01, 0.001, and 0.0001. Both CHG2 and SevenNet-O exhibit a drastic increase in RMSE when trained with higher learning rates; in particular, at a learning rate of 0.01, the RMSE rises by orders of magnitude for both models. In contrast, at lower learning rates, the RMSE values plateau and remain stable with continued training.

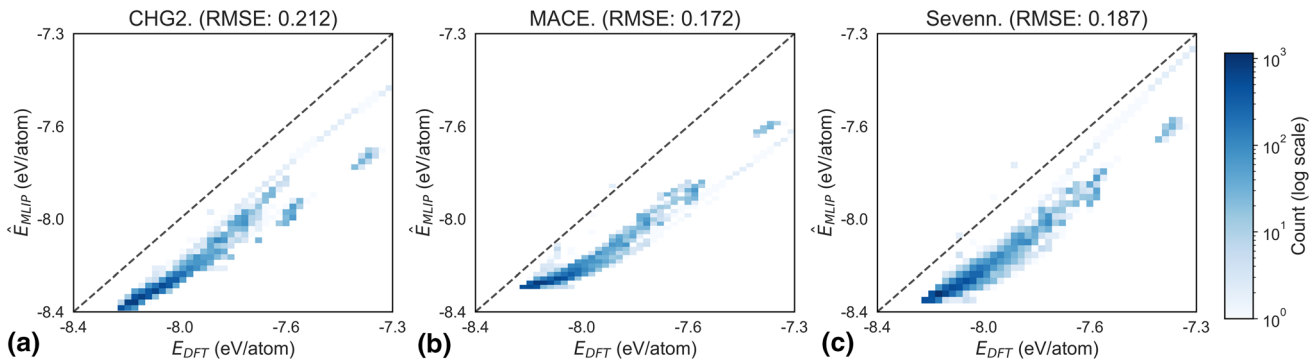

Fig. 1 Performance of the foundational models on the pure Fe data set before fine-tuning

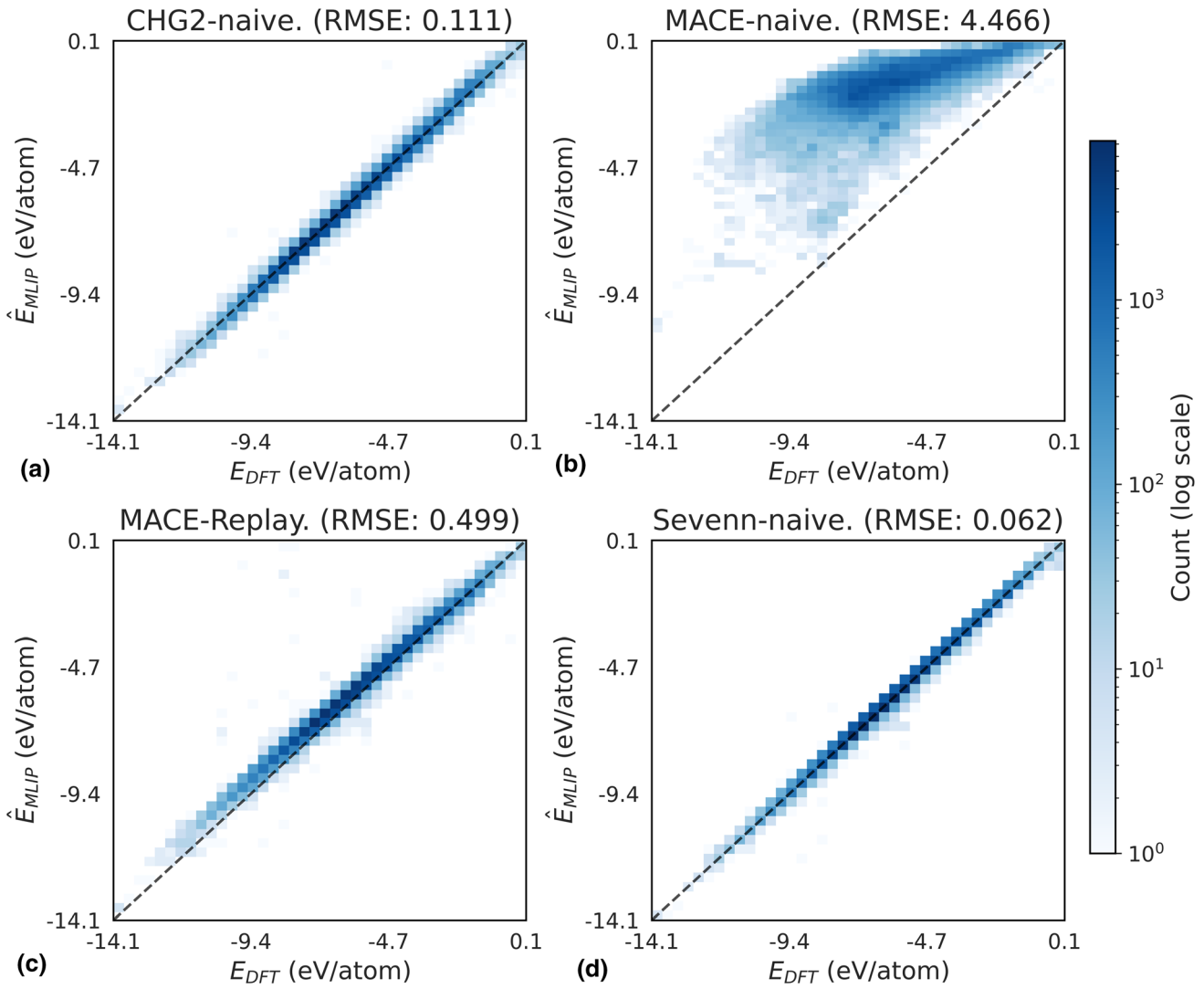

Fig. 2 The MLIP predicted and DFT energy per atom for the MPTRJ validation data set for MLIPs trained with learning rate of 0.0001, (a) CHG2 fine-tuned naive, (b) MACE fine-tuned naive, (c) MACE fine-tuned with replay, (d) SevenNet-O fine-tuned naive

For MACE-replay, as the training progresses the MPTRJ RMSE increases drastically reaching 0.6 eV/atom after 50 epochs. As for the learning rate, there does not appear to be an observable effect on the forgetting of the information for the MACE model. All models end with high RMSE values (0.4 eV/atom) after 50 epochs. For CHG2 the MPTRJ RMSE only increases to 0.111 eV/atom although the MPTRJ data set was not replayed.

For learning rates 0.001 and 0.0001 the Fe RMSE decreases for all models as the fine-tuning progresses (Fig. 12), signifying an improved fit to Fe data set. The models achieve RMSE values of less than 10 meV / atom after 10 epochs for the Fe data set. This aligns with the findings of,[39] which demonstrated that fine-tuning foundational models can be highly data-efficient, achieving high accuracy using only 664 training configurations. Here, as we use 10,400 training data, fewer epochs were required to reach an RMSE below 10 meV/atom for the Fe data set.

Freezing of layers during fine-tuning is one of the strategies for minimizing catastrophic forgetting. For CHG2 freezing the layers generally resulted in MPTRJ RMSE values comparable to or higher than those from the naive approach (Sect. A.2). The only exception occurred when only the convolution layers were trained–that is, when all layers were frozen except the atom and bond convolution layers. To further validate the observations, two additional fine-tunings were carried out with randomized training, validation and test set for the best model (only convolution layers trainable). The mean MPTRJ RMSE for the three models was 91 meV/atom with a standard deviation of 0.001 eV/atom. This confirms that the reduced forgetting seen in the model was not a random occurrence. For MACE, freezing layers did not lead to a

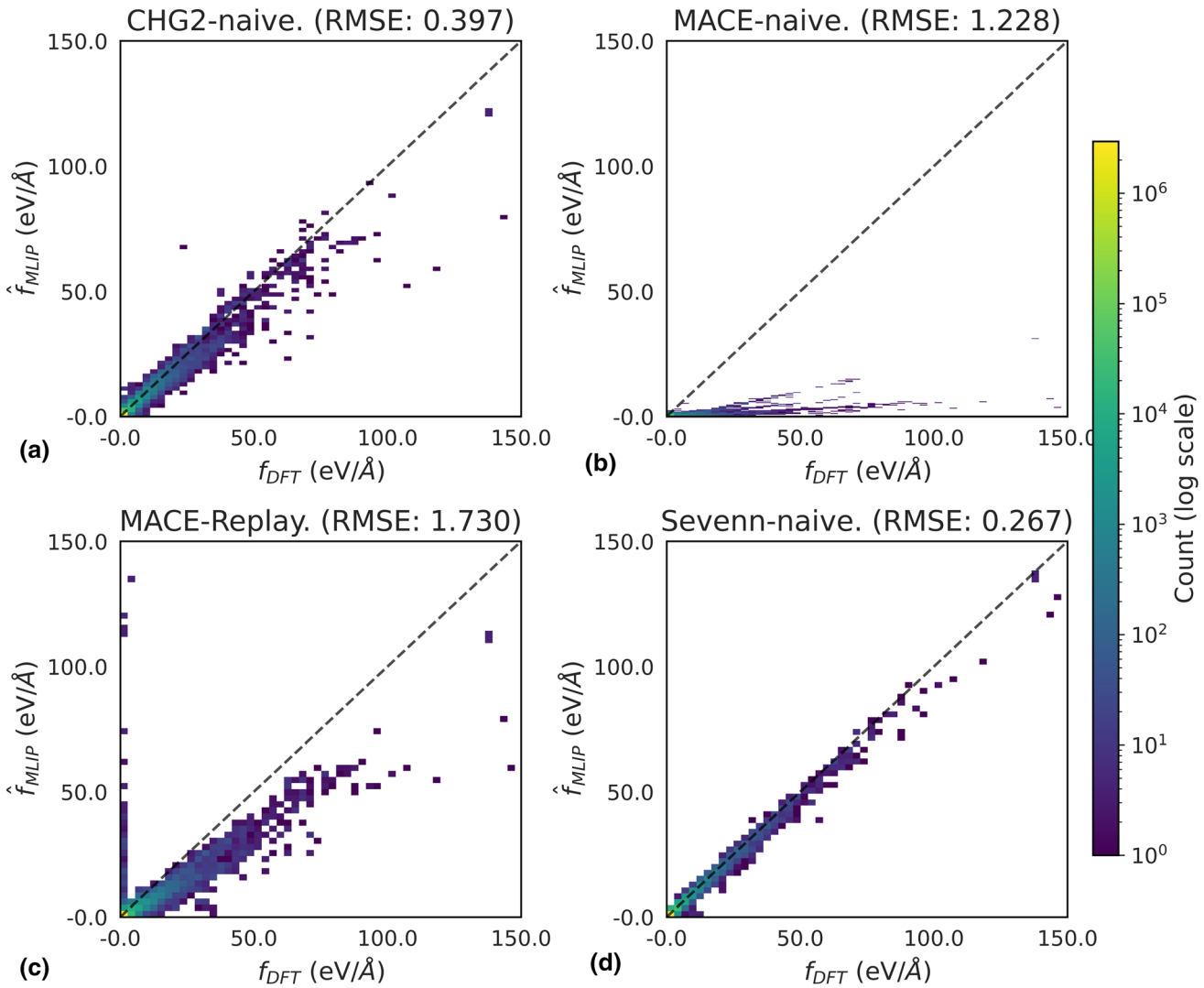

Fig. 3 The MLIP predicted force and DFT force for the MPTRJ validation data set, (a) CHG2-naive, (b) MACE-naive, (c) MACE-replay, and (d) SevenNet-naive, respectively

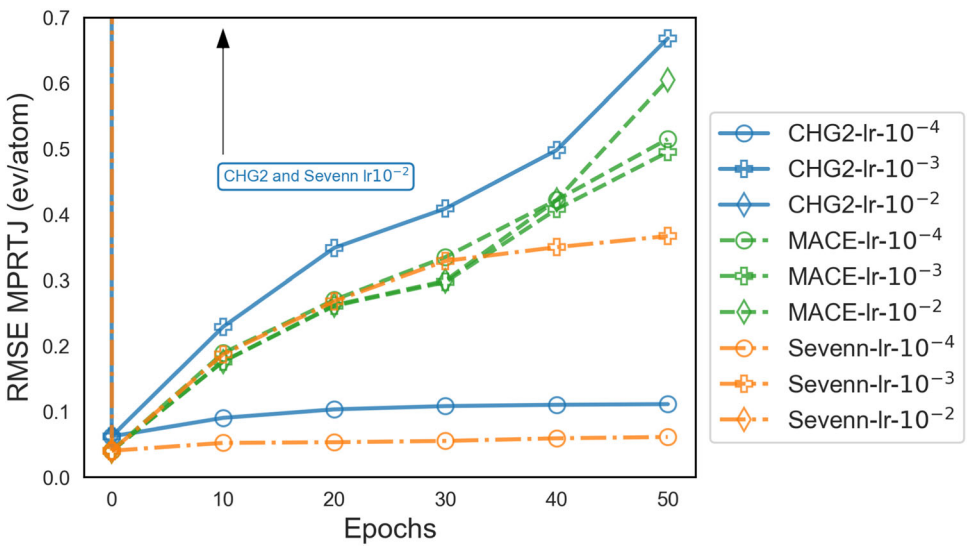

Fig. 4 Effect of training epochs on forgetting in MACE-replay, CHG2-naive, and SevenNet-O-naive models across three different learning rates, lr-10^{-2}, lr-10^{-3} and lr-10^{-4} correspond to learning rates 0.01, 0.001 and 0.0001 respectively. The figure is cropped at 0.7 as the errors from CHG2 and SevenNet-O with learning rate of 0.01 were orders of magnitude higher

notable reduction in RMSE, which remained high at 4.59 eV/atom on the MPTRJ validation set (Sect. A.2). This outcome may reflect suboptimal layer selection for freezing; however, unlike CHG2 and SevenNet-O naive models, the MACE models already exhibited high RMSE values even without freezing, showing lower applicability for the current transfer learning strategy. Consequently, this approach was not pursued further.

RMSE reflects the quality of a fit to a given data set, but it is not optimal for materials simulations, where accuracy in properties like vacancy formation and elastic tensor is critical. Hence, we further evaluated the MLIPs on bcc-Fe properties; the best models of CHG2 (frozen) and MACE (MACE-replay) after fine-tuning were considered. Since the fine-tuned CHG2 and SevenNet-O models exhibit comparable performance on the MPTRJ dataset, we focus on CHGNet in the main discussion. This choice is motivated by CHG2's additional capability to predict magnetic moments alongside energies and forces. For completeness, the results for SevenNet-O are provided in the Sect. A.5.1. The fine-tuned CHG2 and MACE models would be referred to as CHG2-FT and MACE-FT from here on. For both CHG2 and MACE, five models were fine-tuned with randomized train, test and validation sets to measure the standard deviation in predictions.

3.2 Validation on Fe Properties

Figure 5(a) shows the energy volume curve for bcc Fe as predicted by the MLIPs. All MLIPs, predict the equilibrium volume per atom for a bcc unit cell in the range of $11.3 - 11.4 \mathring{A}^3$ atom comparable to that of DFT ($11.3 \mathring{A}^3$ atom). As the foundational models are trained on near equilibrium structures they accurately predict the equilibrium volume. However, the curvature of the energy volume curve deviates largely for MACE, which is reflected in the predicted bulk modulus (Fig. 5b). After fine-tuning, the energy volume curves are comparable to that of DFT for both CHG2-FT and MACE-FT.

Table 1 shows the bulk properties of Fe as predicted by the MLIPs, they are compared with the DFT values calculated in the present study and from literature.[11] For simulations involving structural relaxations, MLIPs must predict the elastic properties accurately. The performance of foundational models on Fe elastic properties were evaluated in previous work.[35] It was observed that for elastic properties the foundational models performed poorly with CHGNet version '0.3.0' (CHG3) being the worst. CHG3 predicted bcc Fe to be mechanically unstable($C' < 0$), hence was not considered in this study. The MACE foundational model (mp-0b3) used in this study is a more recent model, but it performs worse than the version used by Echeverri Restrepo et al.[35] After fine-tuning, all models accurately predict the elastic properties with an error $< 10\%$. This is expected as the fine-tuning data set contains various deformed cells that allows the MLIP to capture the elastic deformation accurately. CHG2 and MACE both before and after fine-tuning reproduce well the DFT literature values for the coefficient of linear thermal expansion and the specific heat C_P. CHG2 yields a bulk magnetic moment of 2.35 μ_B, which decreases to 2.19 μ_B after fine-tuning bringing it closer to the reported literature value of 2.17 μ_B.[66]

Because the local atomic environments of defects differ from those in perfect crystals, vacancy structures represent an extrapolation for foundational models trained only on near-equilibrium configurations. This is seen in the vacancy formation energy predicted by the foundational models (Table 1). After fine-tuning with the Fe-data set, both CHG2 and MACE predict a vacancy formation energy of 2.13 eV and 2.23 eV respectively, comparable with the DFT value. Similar observations were also seen for vacancy migration barriers, with MACE's error reducing to 0.08 eV and 0.04 eV for CHG2.

Other defects important for materials simulation are surfaces and grain boundaries. Figure 5(c) shows the predicted energies for symmetric tilt GBs. Both foundational models under-predict the GB energies, although they capture the DFT trends. They predict $\Sigma 3(1\bar{1}\bar{2})$ as the lowest energy GB in accordance with DFT.[62] Similar observations were also seen for symmetric twist grain boundaries (Fig. 15). The quantitative predictions improve after fine-tuning with an error of 0.139 and 0.119 Jmm^{-2} for CHG2-FT and MACE-FT respectively. $\Sigma 3(111)$ and $\Sigma 7(111)$ twist grain boundaries show the largest deviation.

Additionally we evaluate the models on the (100), (110) and (111) bcc Fe surfaces. Figure 5(d), shows the surface energy predicted by the MLIPs. The surface energies predicted by the foundational models deviate by more than 1 Jmm^{-2}. However, after fine-tuning the error decreases to 0.027 Jmm^{-2} for CHG2-FT and 0.042 Jmm^{-2} for MACE-FT respectively. It was found that irrespective of the performance of the foundational models, after fine-tuning all MLIPs fit the Fe properties well.

3.3 Fe-impurity interactions

Substitutional impurities: As the MLIPs have only been fine-tuned on the Fe data set, they rely on the foundational models training to predict the solute-solute interactions. Here we look at the binding energy as predicted by the CHG2-FT and MACE-FT for the first five nearest neighbours for pairs of substitutional atoms. Figure 6 shows the binding energies for a subset of combination of elements Al, Cu, Ni, Nb and Sn. The binding energies predicted by

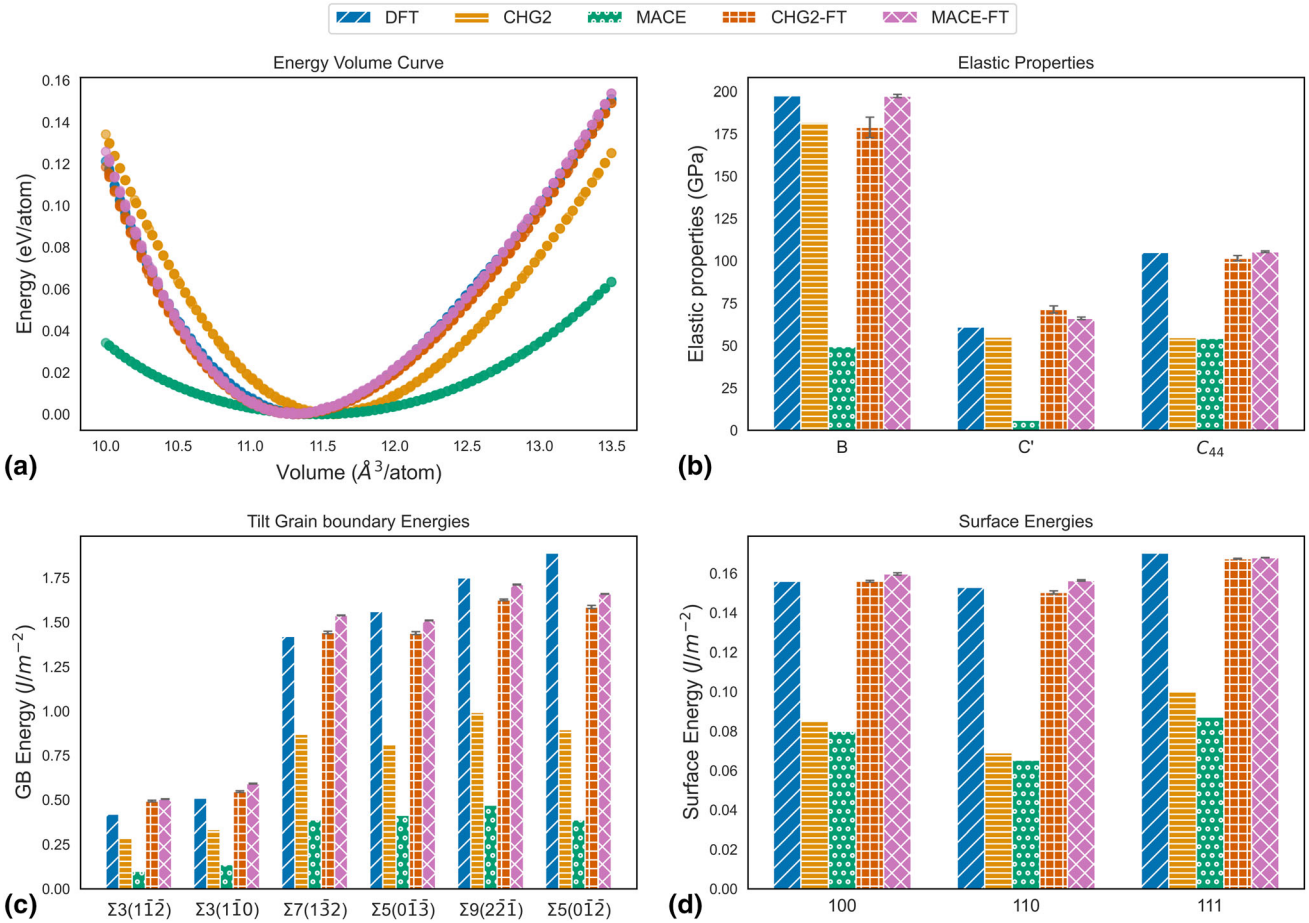

Fig. 5 Comparison of DFT, CHG2, MACE, CHG2-FT and MACE-FT predictions. (a) Energy volume curve for bcc Fe (b) Elastic properties of bcc Fe (c) Symmetric tilt grain boundary energies of bcc Fe (d) bcc Fe surface energies for (100) (110) and (111) surfaces. The error bars for fine-tuned models were determined with five different models fine-tuned on randomized dataset. The DFT values for GB and surfaces are taken from[62] and[67] respectively, the other DFT values are calculated in the present work

Table 1 Properties of bcc Fe as predicted by DFT, CHG2, MACE, CHG2-FT and MACE-FT

Properties		CHG2	CHG2-FT	MACE	MACE-FT	DFT (this study)	DFT
a_{lat}	Å	2.84	2.83	2.84	2.83	2.83	2.83[11]
B	GPa	182	189 ± 6	49	196 ± 1	197	199[11]
C'	GPa	55	74 ± 2	5	64 ± 1	67	73[11]
C_{44}	GPa	55	104 ± 1	54	105 ± 1	105	105[11]
α (300 K)[1]	10^{-5}	1.08	1.28± 0.10	1.10	1.33± 0.11		1.02[68]
C_P (300 K)	J/(mol K)	26.03	25.55 ± 0.89	22.40	25.71 ± 0.74		23.30[68]
μ^{bulk}	μ_B	2.35	2.14	-	-	2.18	2.18[59]
E^f_{vac}	eV	0.73	2.13 ± 0.01	0.49	2.23 ± 0.01	2.19	2.20[11]
E^m_{vac}	eV	0.65	0.67 ± 0.01	0.40	0.71 ± 0.01		0.65[69]

[1]Linear coefficient of thermal expansion

the MLIPs are compared with the DFT calculations performed in the present study. The binding energies for other combinations of elements are given in Sect. A.6.

For all elements, the MLIPs faithfully reproduce the trend of binding energies as a function of increasing nearest-neighbour distance. CHG2 and MACE have an RMSE of 0.073 and 0.117 eV, respectively, after fine-tuning the error slightly increases for CHG2-FT to 0.084 eV. However, the deviation of these predictions from the CHG2 model predictions is minimal.

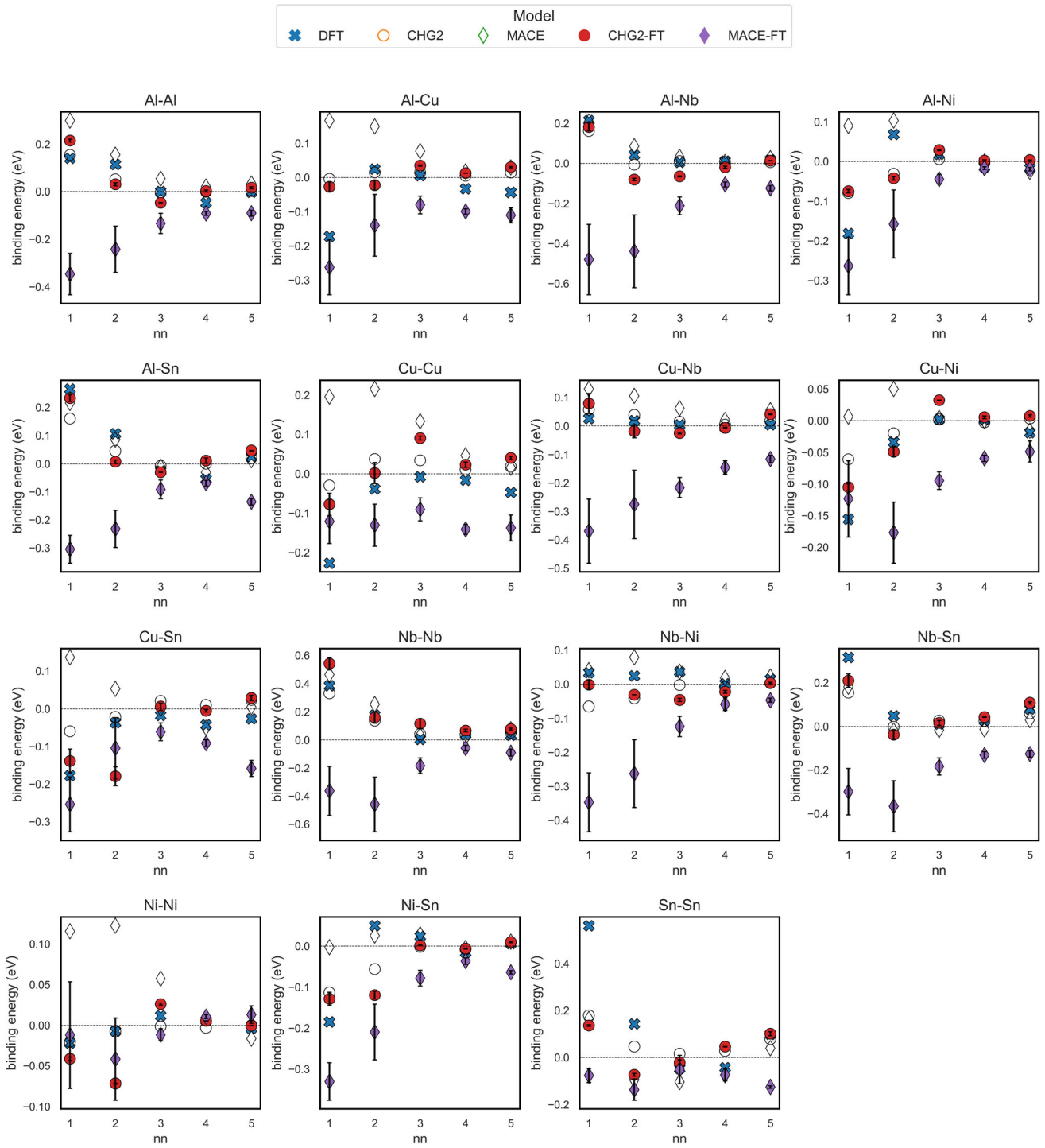

Fig. 6 Binding energies for the first five nearest neighbours subset of elements

Unlike CHG2-FT, the forgetting during fine-tuning observed for MACE is visible in the binding energy predictions of MACE-FT, with a RMSE of 0.361 eV. Further, for Al-Al, Al-Nb, Al-Sn, Cu-Nb, Nb-Nb, and Nb-Sn MACE-FT incorrectly predicts the interactions as attractive. This is despite replaying the MPTRJ data set during the fine-tuning of MACE.

Vacancy impurities interaction: Next we look at vacancy-solute interactions. It is expected that the addition of vacancy structures during fine-tuning improves the relaxation for defect structures. This is evident with the decrease in RMSE from 0.134 eV to 0.067 eV for CHG2-FT. As MACE-FT does not capture the interactions accurately, there was no improvement noted in the RMSE (Fig. 7).

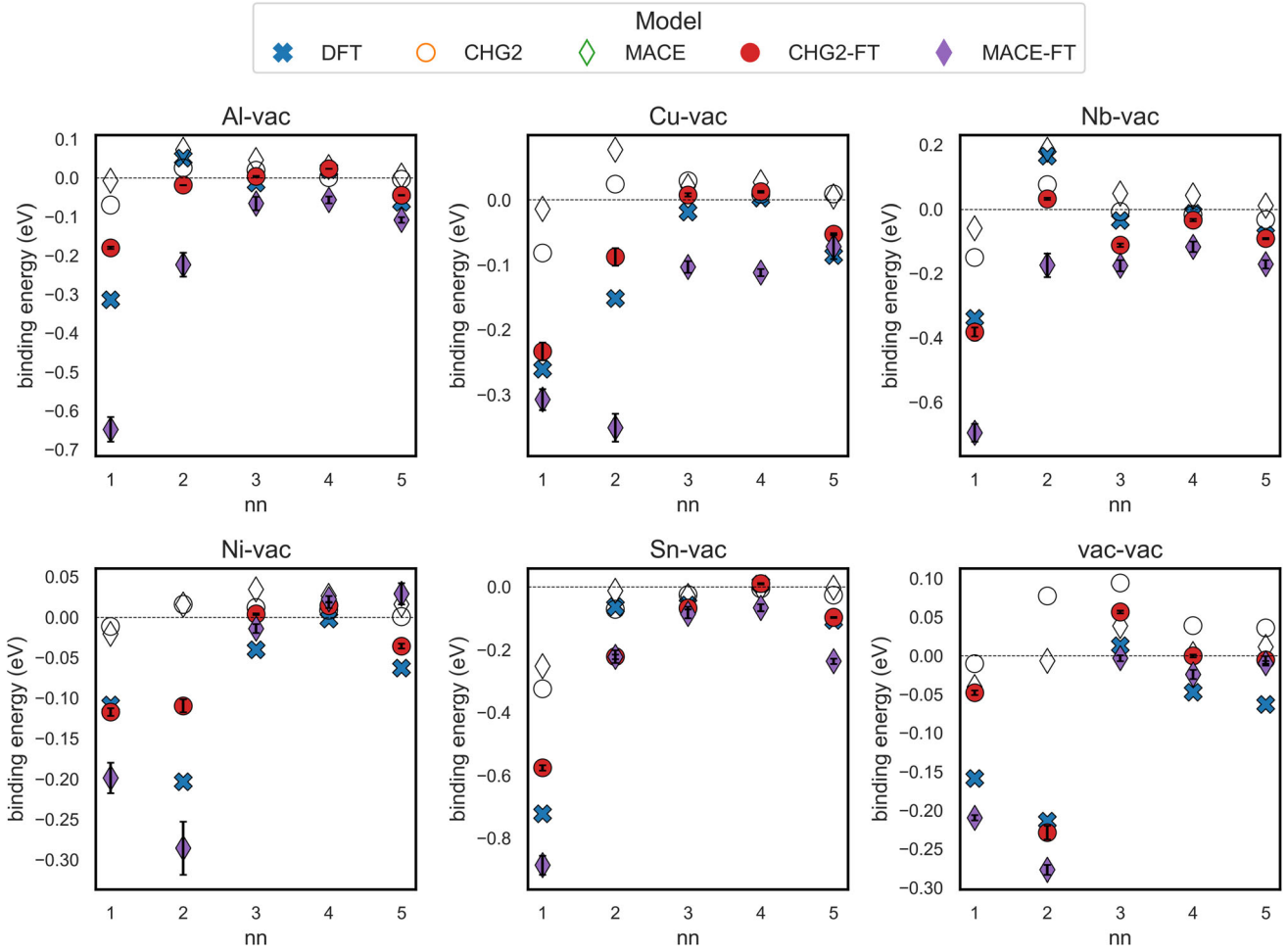

Fig. 7 Binding energies for the first five nearest neighbours for vacancy-solute interactions

Table 2 The difference in energy between octahedral and tetrahedral site for interstitial atoms C, N and O

	$\Delta E_{oct-tet}$				
Solute	CHG2	CHG2-FT	MACE	MACE-FT	DFT
C	−0.515	−0.262	−0.262	0.479	−0.86,[65] −0.94[70] −0.86[71]
N	−0.142	−0.0001	−0.000	−0.12	−0.8[70] −0.73[71]
O	0.224	−0.043	−0.043	0.747	−0.57[70] −0.52[71]

Interstitial impurities: In addition to substitutional impurities, steel also has interstitial impurities like C, O and N. To identify the preferred interstitial sites, the difference in energies of the tetrahedral and octahedral interstitial sites for these elements are determined using the MLIPs (Table 2). The interstitial atoms prefer octahedral sites if $E_{oct-tet} < 0$. For carbon, CHG2, MACE and CHG2-FT correctly predict the octahedral site as the stable site which is consistent with DFT. However, the predicted $\Delta E_{oct-tet}$ is significantly lower in magnitude. $\Delta E_{oct-tet}$ determines the energy barrier for diffusion, with the tetrahedral site acting as the saddle point. Thus, a significantly underestimated $\Delta E_{oct-tet}$ leads to extreme overestimation of diffusivities in simulations. MACE-FT wrongly predicts the tetrahedral site to be stable, a pathological feature. As with C, all MLIPs underestimate the $\Delta E_{oct-tet}$ for N, but consistently identify the octahedral site as the most stable. For O, however, CHG2 and MACE-FT incorrectly predict the tetrahedral site as stable.

3.4 Discussion

n this study, we observe that fine-tuning foundational models is not a straightforward task. The model

architecture plays an important role during fine-tuning, as seen in the case of CHGNet, MACE and SevenNet-O. The foundational models evaluated here were trained on the same MPTRJ data set and further fine-tuned on the same Fe data set. However, each model behaves differently after fine-tuning. All models display catastrophic forgetting though to different degrees. Both CHGNet and SevenNet-O show a reduced tendency to forget as the learning rate is lowered, with minimal forgetting observed at a learning rate of 0.0001. In contrast, for both the naive and replay strategy in MACE, lowering the learning rate does not lead to a reduction in forgetting, with the naive strategy showing the worst performance. It is possible that the training strategy used here is not the best suited for MACE, as numerous other hyper-parameters could influence the fine-tuning.[72]

IWhile the replay strategy in MACE provides a more effective mitigation of forgetting compared to naive training, its performance remains inferior to that of CHGNet and SevenNet-O. A downside of the replay strategy is the need for replaying the old data set every time the model is fine-tuned, this both increases the training cost and data set size for fine-tuning. It was seen that replaying does not guarantee a good fit to the forces after fine-tuning (RMSE 1.7 $eV/\text{Å}$), limiting its applicability to structural relaxations.

When the layers of CHGNet were frozen to mitigate catastrophic forgetting, it was observed that updating only the convolutional layers yielded the lowest error rates. In contrast, training other layers resulted in errors comparable to those from naive training. This is likely due to the key role convolutional layers play in extracting environmental features for each atom; selectively fine-tuning these parameters may enable the model to better adapt to new atomic environments. Verifying this however requires more in depth study which is not within the scope of the current article.

Irrespective of the performance of the foundational models on the Fe properties, the models after fine-tuning fit accurately to the properties of Fe. This was seen with the improvement in the vacancy formation energy, elastic properties, grain boundary energies and surface energies for all fine-tuned models.

Forgetting in models during fine-tuning directly influences the binding energy prediction of substitutional atoms. CHGNet predicted binding energies that were comparable to those of the foundational model, agreeing with the minimal catastrophic forgetting observed earlier. In the case of vacancy defect interactions, minor improvements were observed due to the presence of the Fe vacancy configurations in the training data set. This indicates that foundational models pre-trained on large data sets can be effectively fine-tuned using data specific to a new application, while maintaining their generalization to other systems. In the case of steel, this facilitates simulations of the combined effects of multiple elements–an area of study that was previously inaccessible through computational approaches.

In contrast to substitutional impurities, the model shows poor performance for interstitial atoms. This is likely due to the distinct atomic environments for interstitial solute atoms. Substitutional atoms occupy lattice sites where their surroundings resemble those of pure Fe, which are well represented in the training data. However, interstitial atoms are placed in between lattice sites, creating configurations not present in the fine-tuning data set. Moreover, while foundational models were trained on Fe-carbides, they lacked representations of carbon as an interstitial in a bcc Fe lattice. Thus, predictions for such configurations involve extrapolation, leading to significant errors.

4 Conclusions

This study investigates catastrophic forgetting in fine-tuning foundational machine learning interatomic potentials (MLIPs) for the Fe system, comparing CHGNet, SevenNet-O, and MACE. Our findings reveal that learning rates below 0.0001 significantly mitigate forgetting in CHGNet and SevenNet-O, enabling effective adaptation to system specific data while retaining broad prior knowledge. In contrast, MACE exhibits greater sensitivity to fine-tuning, with higher forgetting rates despite strategies like freezing and data set replay, likely due to architectural differences that limit its robustness and transferability. Additionally, all models showed poor performance for interstitial atoms, for steels a critical issue. It highlights the critical need for including relevant configurations in fine-tuning data sets. These findings are particularly relevant for industrial applications, where reliable prediction of impurity interactions is essential for designing steels with improved recyclability and performance. Although the present study is limited to single-phase bcc Fe and a restricted set of substitutional elements, the results demonstrate that carefully tuned MLIPs can preserve both data efficiency and transferability, while also highlighting directions for improving their applicability to more complex systems.

Appendix

Performance of foundational models on MPTRJ dataset

Figures 8 and 9 show the performance of foundational models for energies and forces evaluated on the MPTRJ dataset.

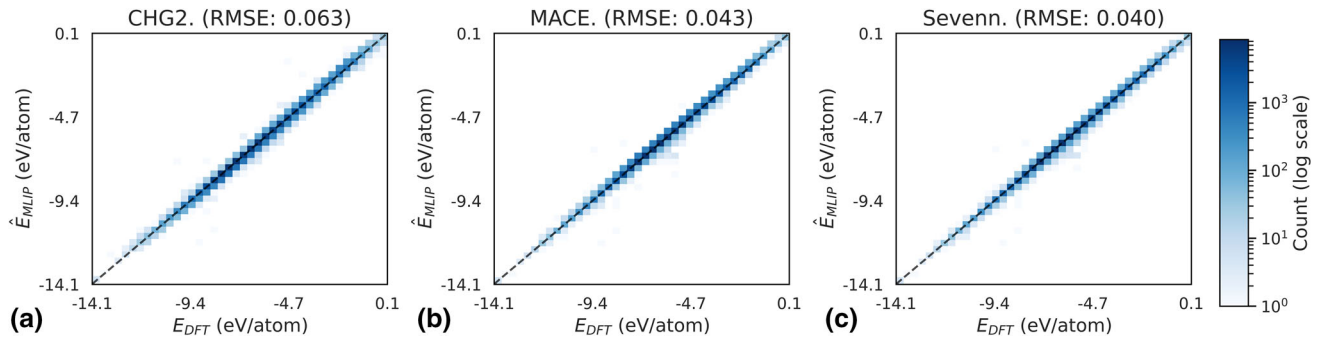

Fig. 8 Performance of foundational models on the MPTRJ validation set before (a) CHG2, (b) MACE, (c) SevenNet-O

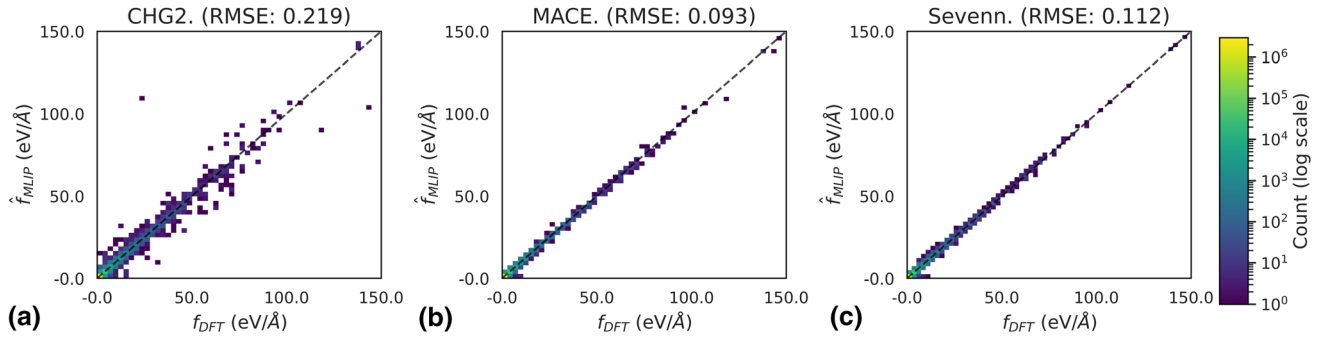

Fig. 9 Performance of foundational models to MPTRJ forces (a) CHG2, (b) MACE, (c) SevenNet-O

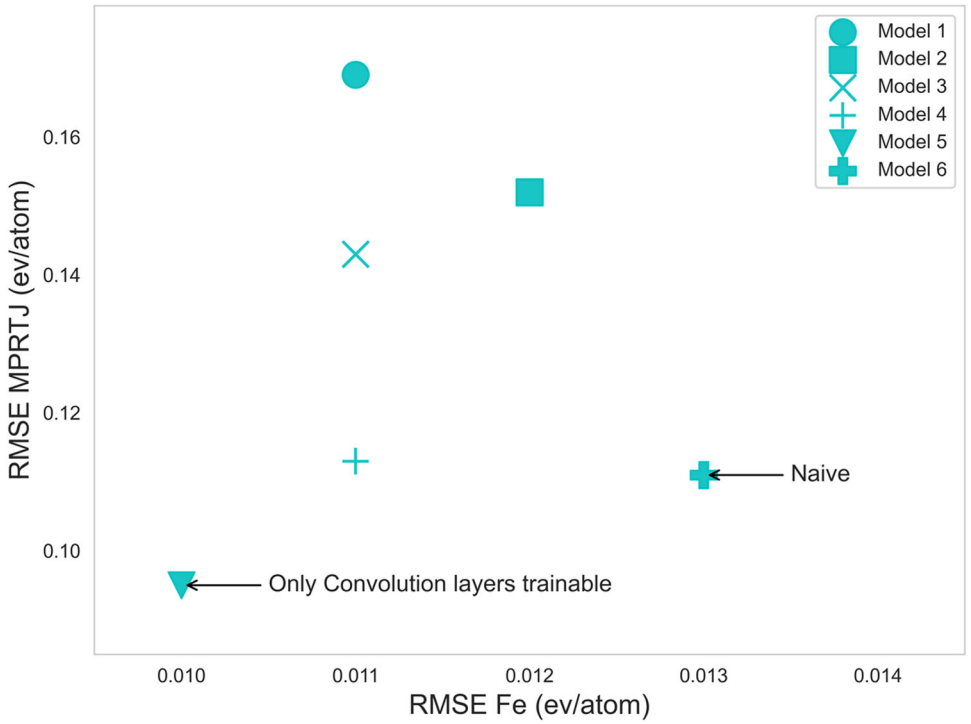

Fig. 10 RMSE for the MPTRJ and Fe data set set when layers were frozen in CHGNet. The details the layers frozen are given in Table 3

Table 3 The numbers correspond to the layers explained above. T indicates that the parameters in the layers are trainable, while F represents the parameters in the layers kept frozen

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Atom embedding	F	F	F	F	F	T
Atom convolution	F	F	F	T	T	T
Bond embedding	T	F	T	F	F	T
Bond convolution	T	T	T	T	T	T
Bond Basis	T	T	T	T	F	T
Angle embedding	T	F	F	T	F	T
Angle layers	T	T	F	T	F	T
Angle basis expansion	T	T	F	T	F	T

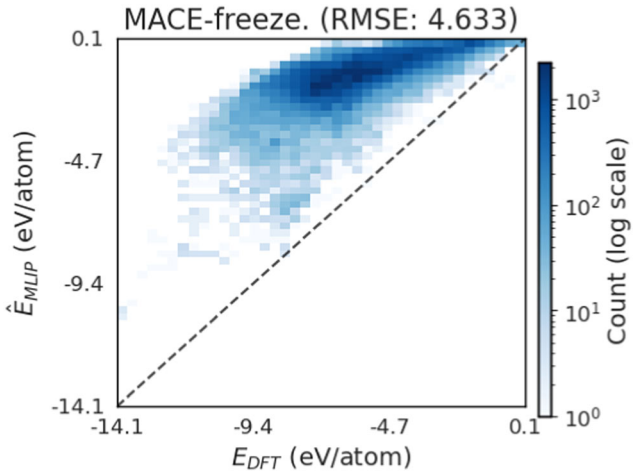

Fig. 11 Performance of MACE-freeze model for energy evaluated on the MPTRJ validation set

Freezing of layers

Figure 10 shows the performance of the CHG2 models with some of the layers frozen. The layers were frozen based on the named layers in CHGNet, the models corresponding to the layers frozen are given in Table 3. The models where only the convolution layers were allowed to be trained showed the least forgetting while also giving a better fit to the Fe data set.

Figure 11 shows the performance of the MACE-freeze model on the energies for the MPTRJ validation set. The model does not show any improvement upon freezing the layers. Hence it was not considered for further studies.

Fit during the training

Figure 12 shows the performance of the MLIPs fine-tuned to the full Fe data set during the training. For a learning rate of 0.01, CHGNet does not converge or fit to the Fe data set. SevenNet-O on the other hand does improve but saturates with a high error of 0.05 eV/atom. When the learning rate is reduced it is expected that more iterations are required to improve the fit, however, here we see that MACE and CHG2 with a learning rate or 0.0001 achieved good fit withing the first 2 epochs (Fig. 13). Thus demonstrating that fine-tuning is much quicker than training from scratch.

Figure 14 shows the performance on elastic tensor as the training progresses for CHGNet. 50 epochs was sufficient for having a good fit to C_{11}, C_{12} and C_{44} elastic properties.

Fe properties

Twist grain boundaries

Similar to the tilt grain boundaries, the MLIPs were also evaluated on twist GBs. Figure 15 shows the performance of the MLIPs.

Specific heat and Coefficient of Thermal Expansion

Figure 16 presents the temperature-dependent coefficients of thermal expansion and heat capacity, obtained through numerical differentiation of cubic spline interpolations. Initially, both MACE and SevenNet-O models show poor alignment with reference data, SevenNet-O in particular exhibits significant deviations. However, their accuracy improves substantially after fine-tuning. While the fine-tuned models successfully capture the general temperature-dependent trends predicted by DFT, their values tend to be slightly higher and more consistent with experimental

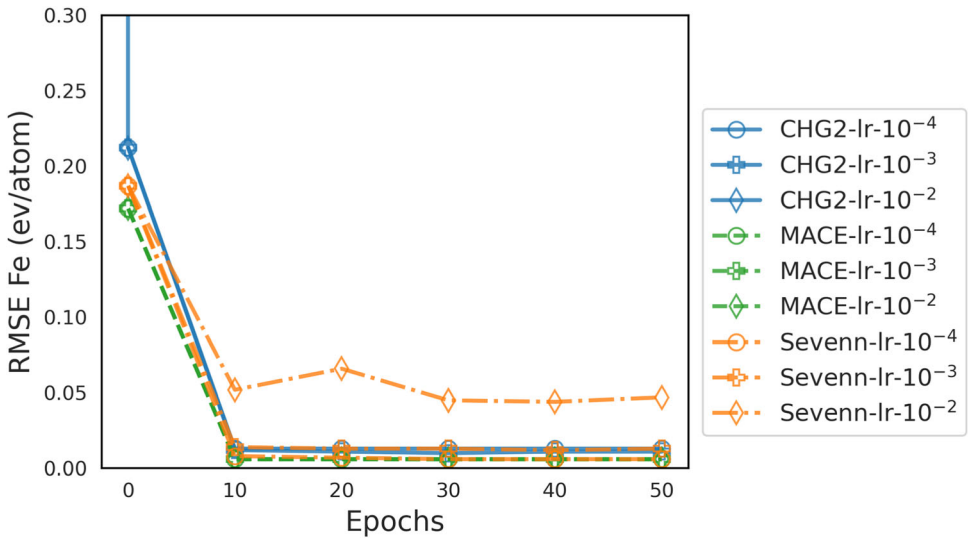

Fig. 12 RMSE for the whole Fe data set as the training progresses

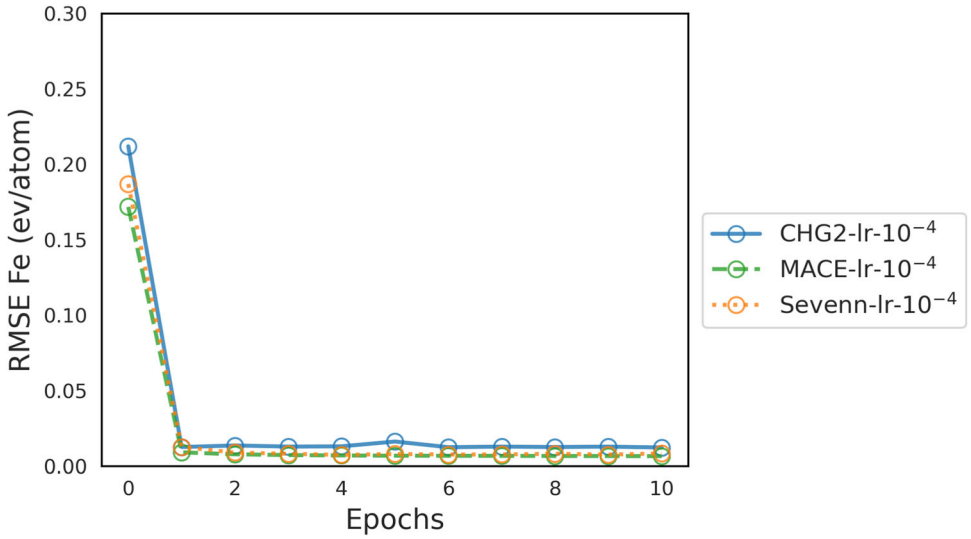

Fig. 13 RMSE for the whole Fe data set as the training progresses for epochs 0–10 for CHG2, MACE and SevenNet-O with a learning rate of 0.0001

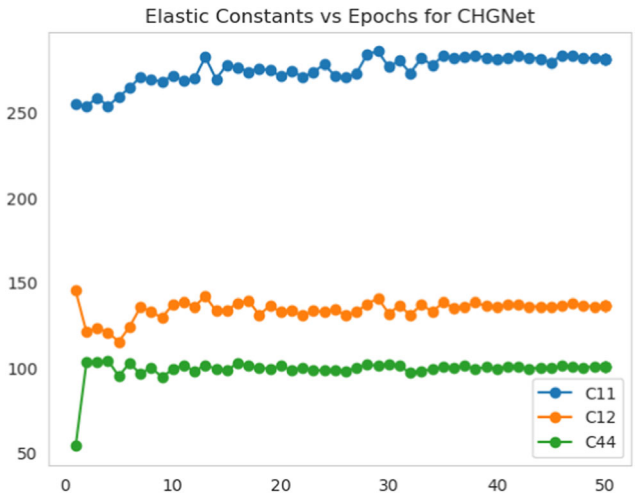

Fig. 14 Fit to Fe elastic tensor as the training progresses for CHGNet

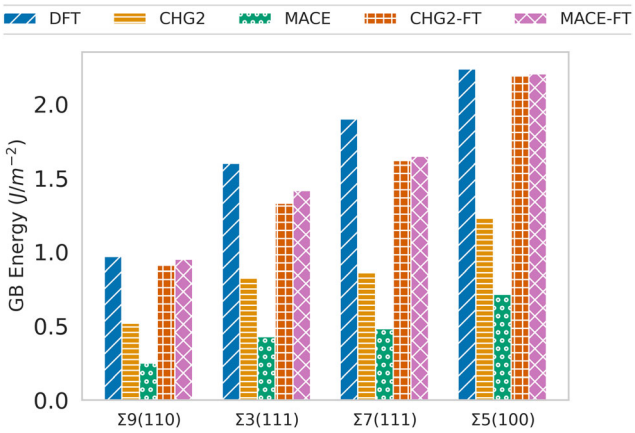

Fig. 15 MLIP predictions for twist grain boundaries with DFT values from[62]

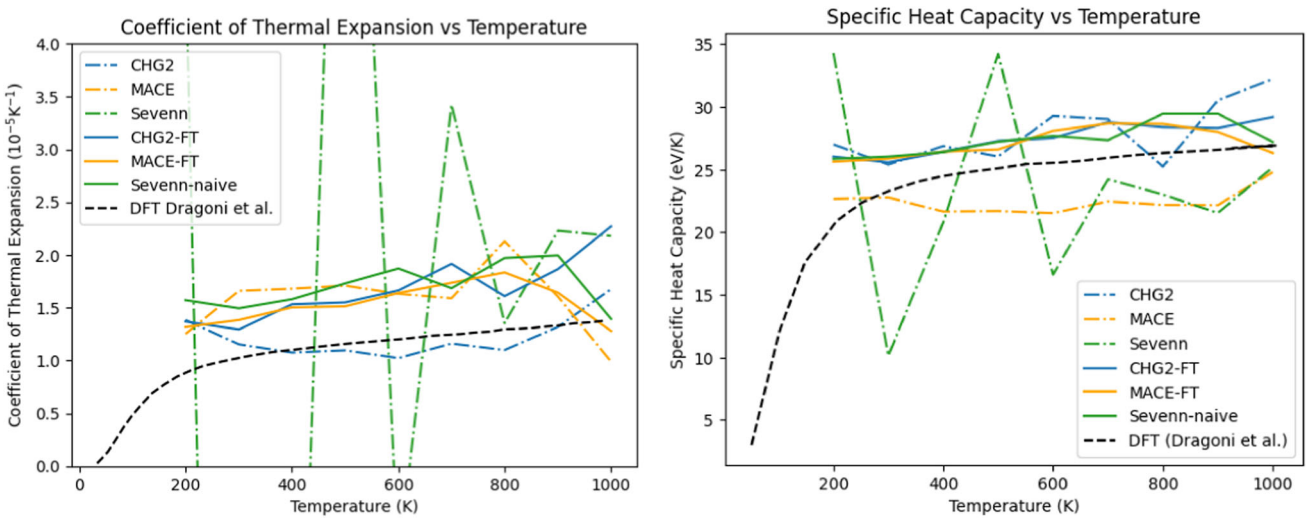

Fig. 16 (a) Coefficient of linear expansion as a function of temperature. (b) heat capacity at constant pressure as a function of temperature. DFT values taken from Ref.[68]

Table 4 Properties of bcc Fe as predicted by DFT, CHG3, SevenNet-O, CHG3-FT and Sevenn-FT

Properties		CHG3	CHG3-FT	SevenNet-O	Sevenn-FT	DFT (this study)	DFT[11]
a_{lat}	Å	2.84	2.83	2.84	2.83	2.83	2.83
B	GPa	110	208	97	189 ± 1	197	199
C'	GPa	−1.25	58	19	66 ± 1	67	73
C_{44}	GPa	96	91	110	97 ± 1	105	105
α (300 K)[1]	10^{-5}			−23.9	1.55 ± 0.26		1.02[68]
C_P (300 K)	J/(mol K)			13.13	26.05 ± 1.33		23.3[68]
μ^{bulk}	μ_B	2.35	2.16	-	-	2.18	
E^f_{vac}	eV	0.82	2.08	1.29	2.23±0.01	2.19	2.20

[1]Linear coefficient of thermal expansion

results.[68] Some noise is evident in the data, likely due to the limited size of the simulation cells and the relatively short molecular dynamics trajectories used in the MLIP evaluations.

SevenNet-O and CHG3

Table 4 shows the properties of Fe as predicted by the fine-tuned CHG3 and SevenNet-O. Though there are large errors in the naive model, after fine-tuning both models show good performance similar to MACE-FT and CHG2-FT. Similarly, Fig. 17 shows the energy volume curve, GB energies and surface energies as predicted by CHG3-naive and Sevenn-naive. The coefficient of thermal expansion and specific heat (C_P) deviate for SevenNet-O before fine-tuning. After fine-tuning the values are well within the accepted range.

Binding energies

The binding energies of different solute element combinations in bcc Fe are presented in Fig. 18 and 19. Comparisons are made with values reported in the

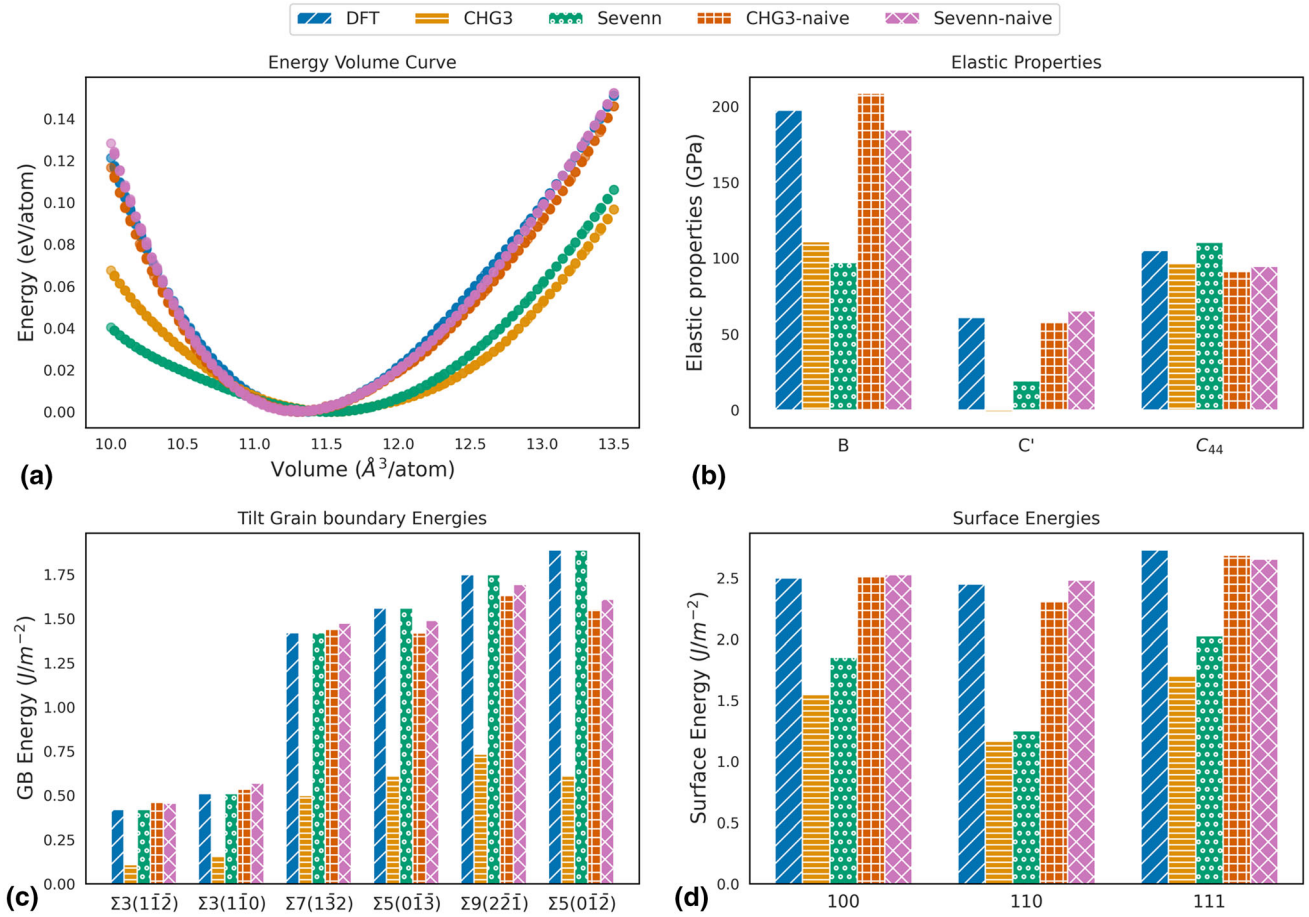

Fig. 17 Comparison of DFT, CHG3, SevenNet-O, CHG3-naive and SevenNet-naive predictions. (a) Energy volume curve for bcc Fe (b) Elastic properties of bcc Fe (c) Symmetric tilt grain boundary energies of bcc Fe (d) bcc Fe surface energies for (100) (110) and (111) surfaces. The DFT values for GB and surfaces are taken from[62] and[67] respectively, the other DFT values are calculated in the present work

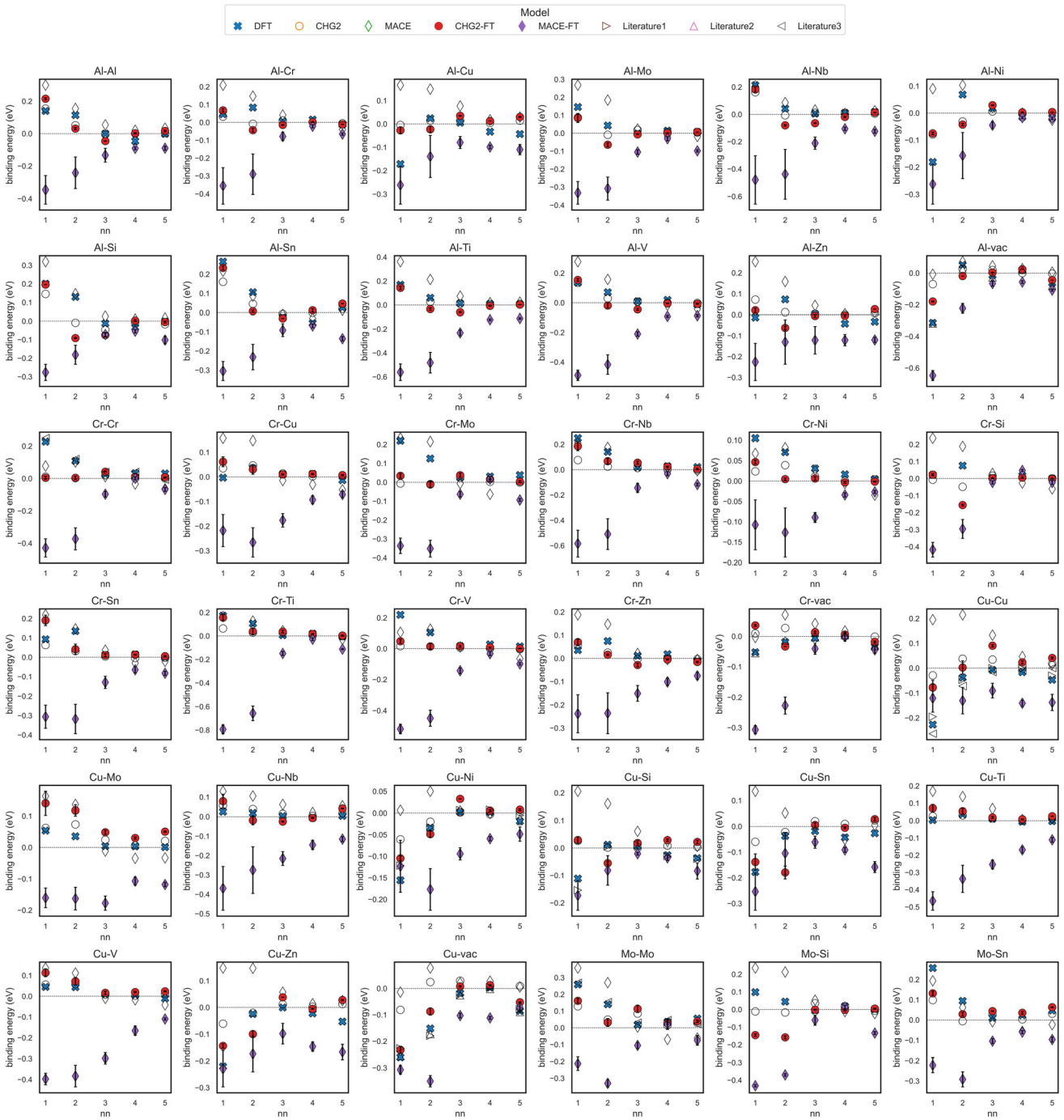

Fig. 18 Binding energy of first five nearest neighbours for various elements in bcc Fe matrix predicted by DFT, CHG2, MACE, CHG2-FT, MACE-FT and Literature. Literature1, literature2 and literature3 correspond to[73–75] respectively

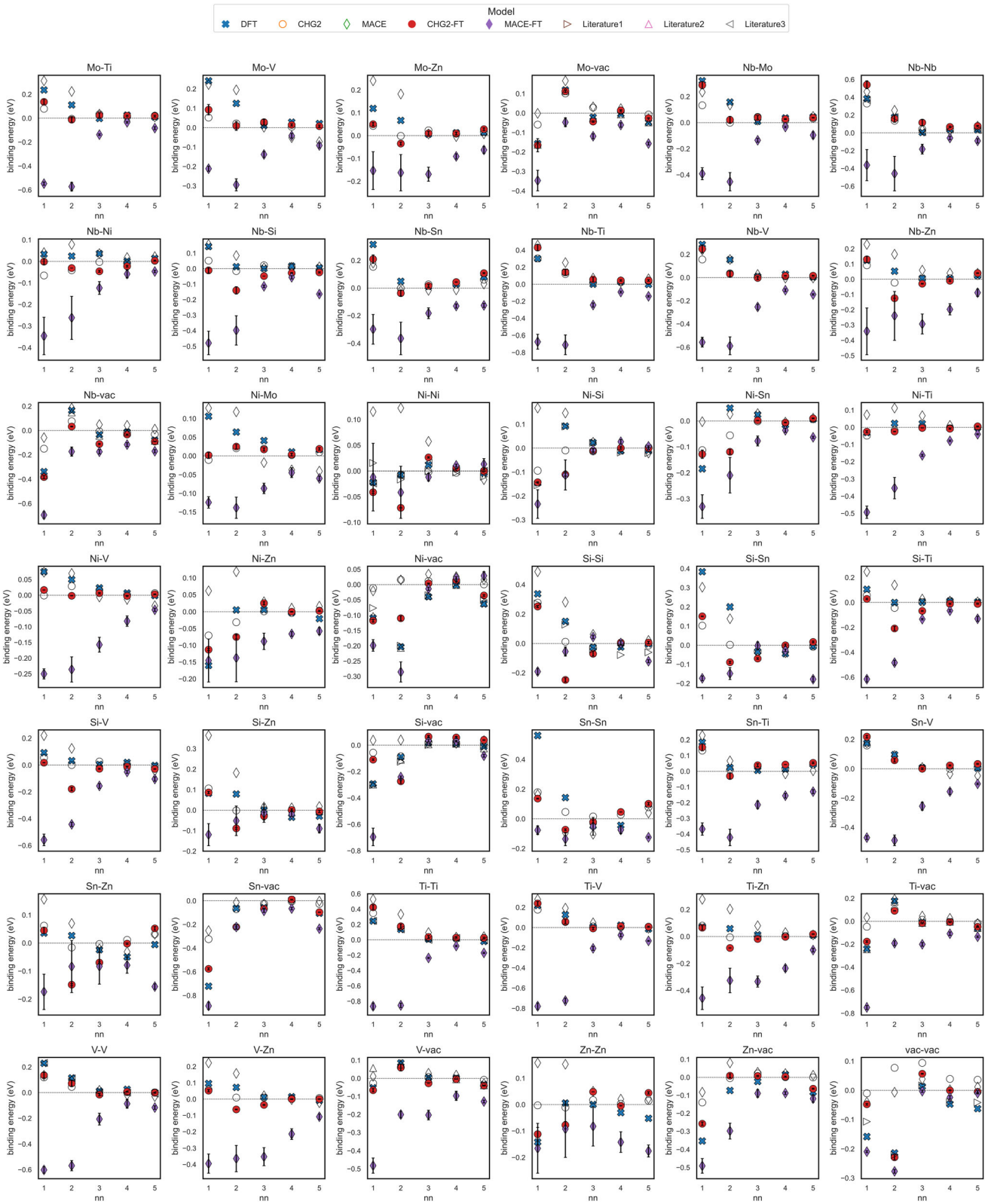

Fig. 19 (Contd.) Binding energy of first five nearest neighbours for various elements in bcc Fe matrix predicted by DFT, CHG2, MACE, CHG2-FT, MACE-FT and Literature. Literature1, litearture2 and literature3 correspond to[73–75]

literature[73–75] Despite employing a replay strategy, MACE-FT consistently yields large errors in binding energy predictions, indicating significant catastrophic forgetting.

Acknowledgements The authors acknowledge the use of computational resources of the DelftBlue supercomputer, provided by Delft High Performance Computing Centre (https://www.tudelft.nl/dhpc).

Funding This research is part of the DEPMAT project (with project number P20-22 / N21022) of the research programme Perspectief which is partly financed by the Dutch Research Council (NWO).It is also part of the Partnership Program of the Materials innovation institute M2i (www.m2i.nl).

Data Availability The Fe dataset, alongside the fine-tuned models, will be made available in a dedicated repository with DOI: 10.4121/d17c84d4-4a17-418b-911b-5495ad7f61cb.

Code Availability The code used to evaluate the MLIPs will be made available under URL https://github.com/naveenmohandas/eval_mlip_on_fe

References

1. S. Bell, B. Davis, A. Javaid, E. Essadiqi, Final Report on Effect of Impurities in Steel, Government of Canada (2006) https://doi.org/10.13140/RG.2.2.33946.85440
2. W. Kohn, Nobel Lecture: Electronic Structure of Matter–Wave Functions and Density Functionals, *Rev. Mod. Phys.*, 1999, **71**(5), p 1253–1266. https://doi.org/10.1103/RevModPhys.71.1253
3. K. Burke, Perspective on Density Functional Theory, *J. Chem. Phys.*, 2012, **136**(15), p 150901. https://doi.org/10.1063/1.4704546
4. A.J. Cohen, P. Mori-Sánchez, and W. Yang, Challenges for Density Functional Theory, *Chem. Rev.*, 2012, **112**(1), p 289–320. https://doi.org/10.1021/cr200107z
5. M.H. Müser, S.V. Sukhomlinov, and L. Pastewka, Interatomic Potentials: Achievements and Challenges, *Adv. Phys. X*, 2023. https://doi.org/10.1080/23746149.2022.2093129
6. J. Behler, and M. Parrinello, Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces, *Phys. Rev. Lett.*, 2007, **98**(14), p 146401. https://doi.org/10.1103/PhysRevLett.98.146401
7. A.V. Shapeev, Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials, *Multiscale Model. Sim.*, 2016, **14**(3), p 1153–1173. https://doi.org/10.1137/15M1054183
8. R. Drautz, Atomic Cluster Expansion for Accurate and Transferable Interatomic Potentials, *Phys. Rev. B*, 2019, **99**(1), p 014104. https://doi.org/10.1103/PhysRevB.99.014104
9. A.P. Bartók, R. Kondor, and G. Csányi, On Representing Chemical Environments, *Phys. Rev. B*, 2013, **87**(18), p 184115. https://doi.org/10.1103/PhysRevB.87.184115
10. R. Jana, and M.A. Caro, Searching for Iron Nanoparticles with a General-Purpose Gaussian Approximation Potential, *Phys. Rev. B*, 2023, **107**(24), p 245421. https://doi.org/10.1103/PhysRevB.107.245421
11. F.-S. Meng, J.-P. Du, S. Shinzato, H. Mori, P. Yu, K. Matsubara, N. Ishikawa, and S. Ogata, General-Purpose Neural Network Interatomic Potential for the α-Iron and Hydrogen Binary System: Toward Atomic-Scale Understanding of Hydrogen Embrittlement, *Phys. Rev. Mater.*, 2021, **5**(11), p 113606. https://doi.org/10.1103/PhysRevMaterials.5.113606
12. F.-S. Meng, S. Shinzato, S. Zhang, K. Matsubara, J.-P. Du, P. Yu, W.-T. Geng, and S. Ogata, A Highly Transferable and Efficient Machine Learning Interatomic Potentials Study of α-Fe-C Binary System, *Acta Mater.*, 2024, **281**, p 11360120408. https://doi.org/10.1016/j.actamat.2024.120408
13. I. Novikov, B. Grabowski, F. Körmann, and A. Shapeev, Magnetic Moment Tensor Potentials for Collinear Spin-Polarized Materials Reproduce Different Magnetic States of BCC Fe, *npj Comput. Mater. Sci.*, 2022, **8**(1), p 1–6. https://doi.org/10.1038/s41524-022-00696-9
14. X.-G. Li, C. Hu, C. Chen, Z. Deng, J. Luo, and S.P. Ong, Quantum-Accurate Spectral Neighbor Analysis Potential Models for Ni-Mo Binary Alloys and fcc Metals, *Phys. Rev. B*, 2018, **98**(9), p 094104. https://doi.org/10.1103/PhysRevB.98.094104
15. A.S. Kotykhov, K. Gubaev, M. Hodapp, C. Tantardini, A.V. Shapeev, and I.S. Novikov, Constrained DFT-Based Magnetic Machine-Learning Potentials for Magnetic Alloys: A Case Study of Fe–Al, *Sci. Rep.*, 2023, **13**(1), p 19728. https://doi.org/10.1038/s41598-023-46951-x
16. A.S. Kotykhov, K. Gubaev, V. Sotskov, C. Tantardini, M. Hodapp, A.V. Shapeev, and I.S. Novikov, Fitting to Magnetic Forces Improves the Reliability of Magnetic Moment Tensor Potentials, *Comput. Mater. Sci.*, 2024, **245**, p 113331.
17. K.C. Pitike, and W. Setyawan, Accurate Fe-He Machine Learning Potential for Studying He Effects in BCC-Fe, *J. Nucl. Mater.*, 2023, **574**, p 154183. https://doi.org/10.1016/j.jnucmat.2022.154183
18. V.L. Deringer, A.P. Bartók, N. Bernstein, D.M. Wilkins, M. Ceriotti, and G. Csányi, Gaussian Process Regression for Materials and Molecules, *Chem. Rev.*, 2021, **121**(16), p 10073–10141. https://doi.org/10.1021/acs.chemrev.1c00022
19. R. Jacobs, D. Morgan, S. Attarian, J. Meng, C. Shen, Z. Wu, C.Y. Xie, J.H. Yang, N. Artrith, B. Blaiszik, G. Ceder, K. Choudhary, G. Csanyi, E.D. Cubuk, B. Deng, R. Drautz, X. Fu, J. Godwin, V. Honavar, O. Isayev, A. Johansson, B. Kozinsky, S. Martiniani, S.P. Ong, I. Poltavsky, K. Schmidt, S. Takamoto, A.P. Thompson, J. Westermayr, and B.M. Wood, A Practical Guide to Machine Learning Interatomic Potentials—Status and Future, *Curr. Opin. Solid State Mater. Sci.*, 2025, **35**, p 101214. https://doi.org/10.1016/j.cossms.2025.101214
20. R. Drautz, Atomic Cluster Expansion of Scalar, Vectorial, and Tensorial Properties Including Magnetism and Charge Transfer, *Phys. Rev. B*, 2020, **102**(2), p 024104. https://doi.org/10.1103/PhysRevB.102.024104
21. A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K.A. Persson, Commentary: The Materials Project: A Materials Genome

Approach to Accelerating Materials Innovation, *APL Mater.*, 2013, **1**(1), p 011002. https://doi.org/10.1063/1.4812323
22. J. Schmidt, T.F.T. Cerqueira, A.H. Romero, A. Loew, F. Jäger, H.-C. Wang, S. Botti, and M.A.L. Marques, Improving Machine-Learning Models in Materials Science Through Large Datasets, *Mater. Today Phys.*, 2024, **48**, p 101560. https://doi.org/10.1016/j.mtphys.2024.101560
23. L. Barroso-Luque, M. Shuaibi, X. Fu, B.M. Wood, M. Dzamba, M. Gao, A. Rizvi, C.L. Zitnick, and Z.W. Ulissi, Open Materials 2024 (OMat24) Inorganic Materials Dataset and Models. arXiv:2410.12771 [cond-mat] 2024, https://doi.org/10.48550/arXiv.2410.12771.
24. C. Chen, and S.P. Ong, A Universal Graph Deep Learning Interatomic Potential for the Periodic Table, *Nat. Comput. Sci.*, 2022, **2**(11), p 718–728. https://doi.org/10.1038/s43588-022-00349-3
25. B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C.J. Bartel, and G. Ceder, CHGNet as a Pretrained Universal Neural Network Potential for Charge-Informed Atomistic Modelling, *Nat. Mach. Intell.*, 2023, **5**(9), p 1031–1041. https://doi.org/10.1038/s42256-023-00716-3
26. I. Batatia, D.P. Kovacs, G. Simm, C. Ortner, and G. Csanyi, MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields, *Adv. Neural Inf. Process Syst.*, 2022, **35**, p 11423–11436.
27. Y. Park, J. Kim, S. Hwang, and S. Han, Scalable Parallel Algorithm for Graph Neural Network Interatomic Potentials in Molecular Dynamics Simulations, *J. Chem. Theory Comput.*, 2024, **20**(11), p 4857–4868. https://doi.org/10.1021/acs.jctc.4c00190
28. A. Bochkarev, Y. Lysogorskiy, and R. Drautz, Graph Atomic Cluster Expansion for Semilocal Interactions Beyond Equivariant Message Passing, *Phys. Rev. X*, 2024, **14**(2), p 021036. https://doi.org/10.1103/PhysRevX.14.021036
29. H. Yang,C. Hu, Y. Zhou, X. Liu, Y. Shi, J. Li, G. Li, Z. Chen, S.Chen, C. Zeni, M. Horton, R. Pinsler, A. Fowler, D. Zügner, T. Xie, J. Smith, L. Sun, Q. Wang, L. Kong, C. Liu, H. Hao, and Z. Lu, MatterSim: a deep learning atomistic model across elements, temperatures and pressures. arXiv:2405.04967 [cond-mat] 2024, https://doi.org/10.48550/arXiv.2405.04967.
30. Y.-L. Liao, B. Wood, A. Das, and T. Smidt, EquiformerV2: improved equivariant transformer for scaling to higher-degree representations. arXiv:2306.12059 [cs] 2024, https://doi.org/10.48550/arXiv.2306.12059 .
31. Y. Liu, X. He, and Y. Mo, Discrepancies and Error Evaluation Metrics for Machine Learning Interatomic Potentials, *npj Comput. Mater.*, 2023, **9**(1), p 1–13. https://doi.org/10.1038/s41524-023-01123-3
32. Y. Liu, and Y. Mo, Learning from Models: High-Dimensional Analyses on the Performance of Machine Learning Interatomic Potentials, *npj Comput. Mater.*, 2024, **10**(1), p 1–14. https://doi.org/10.1038/s41524-024-01333-3
33. B. Focassio, L.P.M. Freitas, and G.R. Schleder, Performance Assessment of Universal Machine Learning Interatomic Potentials: Challenges and Directions for Materials' Surfaces, *ACS Appl. Mater. Interfaces*, 2024. https://doi.org/10.1021/acsami.4c03815
34. B. Deng, Y. Choi, P. Zhong, J. Riebesell, S. Anand, Z. Li, K. Jun, K.A. Persson, and G. Ceder, Systematic Softening in Universal Machine Learning Interatomic Potentials, *npj Comput. Mater.*, 2025, **11**(1), p 1–9. https://doi.org/10.1038/s41524-024-01500-6
35. S. Echeverri Restrepo, N.K. Mohandas, M. Sluiter, and A.T. Paxton, Applicability of Universal Machine Learning Interatomic Potentials to the Simulation of Steels, *Model. Simul. Mater. Sci. Eng.*, 2025. https://doi.org/10.1088/1361-651X/adb483
36. M. Iman, H.R. Arabnia, and K. Rasheed, A Review of Deep Transfer Learning and Recent Advancements, *Technologies*, 2023, **11**(2), p 40. https://doi.org/10.3390/technologies11020040
37. A. Hosna, E. Merry, J. Gyalmo, Z. Alom, Z. Aung, and M.A. Azim, Transfer Learning: A Friendly Introduction, *J. Big Data*, 2022, **9**(1), p 102. https://doi.org/10.1186/s40537-022-00652-w
38. H. Kaur, F..D. Pia, I. Batatia, R. Advincula, X. X, Shi B, J. Lan, G. Csányi, A. Michaelides, and V. Kapil, Data-Efficient Fine-Tuning of Foundational Models for First-Principles Quality Sublimation Enthalpies, *Faraday Discuss*, 2025, **256**(0), p 120–138. https://doi.org/10.1039/D4FD00107A
39. M. Radova, W.G. Stark, C.S. Allen, R.J. Maurer, and A.P. Bartók, Fine-tuning foundation models of materials interatomic potentials with frozen transfer learning. arXiv:2502.15582 [cond-mat] 2025, https://doi.org/10.48550/arXiv.2502.15582 .
40. S. Ju, J. You, G. Kim, Y. Park, H. An, and S. Han, Application of pretrained universal machine-learning interatomic potential for physicochemical simulation of liquid electrolytes in Li-ion battery. arXiv:2501.05211 [cond-mat] 2025, https://doi.org/10.48550/arXiv.2501.05211 .
41. E.L. Aleixo, J.G. Colonna, M. Cristo, and E. Fernandes, Catastrophic Forgetting in Deep Learning: A Comprehensive Taxonomy. arXiv:2312.10549 [cs] 2023, https://doi.org/10.48550/arXiv.2312.10549 .
42. P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. Hoesel, H. Schopmans, T. Sommer, and P. Friederich, Graph Neural Networks for Materials Science and Chemistry, *Commun. Mater.*, 2022, **3**(1), p 1–18. https://doi.org/10.1038/s43246-022-00315-6
43. T.B. Massalski, and D.E. Laughlin, The Surprising Role of Magnetism on the Phase Stability of Fe (Ferro), *Calphad*, 2009, **33**(1), p 3–7. https://doi.org/10.1016/j.calphad.2008.07.009
44. S. Batzner, A. Musaelian, L. Sun, M. Geiger, J.P. Mailoa, M. Kornbluth, N. Molinari, T.E. Smidt, and B. Kozinsky, E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials, *Nat. Commun.*, 2022, **13**(1), p 2453. https://doi.org/10.1038/s41467-022-29939-5
45. V. Vovk, A. Gammerman, and G. Shafer, Algorithmic Learning in a Random World. Springer, New York (2005) https://doi.org/10.1007/b106715 .
46. J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A.A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, Overcoming Catastrophic Forgetting in Neural Networks, *Proc. Natl. Acad. Sci.*, 2017, **114**(13), p 3521–3526. https://doi.org/10.1073/pnas.1611835114
47. M. Masana, X. Liu, B. Twardowski, M. Menta, A.D. Bagdanov, and J. Weijer, Class-Incremental Learning: Survey and Performance Evaluation on Image Classification, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023, **45**(5), p 5513–5533. https://doi.org/10.1109/TPAMI.2022.3213473
48. Fine-tuning Foundation Models – mace 0.3.13 documentation. https://mace-docs.readthedocs.io/en/latest/guide/finetuning.html. Accessed 22 Aug 2025
49. mace-torch: None. https://github.com/ACEsuit/mace Accessed 04 Sept 2025
50. E.V. Podryabinkin, and A.V. Shapeev, Active Learning of Linearly Parametrized Interatomic Potentials, *Comput. Mater. Sci.*, 2017, **140**, p 171–180. https://doi.org/10.1016/j.commatsci.2017.08.031
51. K. Gubaev, E.V. Podryabinkin, G.L.W. Hart, and A.V. Shapeev, Accelerating High-Throughput Searches for New Alloys with Active Learning of Interatomic Potentials, *Comput. Mater. Sci.*, 2019, **156**, p 148–156. https://doi.org/10.1016/j.commatsci.2018.09.031

52. A.P. Bartók, M.C. Payne, R. Kondor, and G. Csányi, Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, Without the Electrons, *Phys. Rev. Lett.*, 2010, **104**(13), p 136403. https://doi.org/10.1103/PhysRevLett.104.136403
53. A. A. Peterson, R. Christensen, and A. Khorshidi, Addressing Uncertainty in Atomistic Machine Learning, *Phys. Chem. Chem. Phys.*, 2017, **19**(18), p 10978–10985. https://doi.org/10.1039/C7CP00375G
54. G. Kresse, and J. Furthmüller, Efficiency of Ab-Initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set, *Comput. Mater. Sci.*, 1996, **6**(1), p 15–50. https://doi.org/10.1016/0927-0256(96)00008-0
55. S.P. Ong, W.D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V.L. Chevrier, K.A. Persson, and G. Ceder, Python Materials Genomics (Pymatgen): A Robust, Open-Source Python Library for Materials Analysis, *Comput. Mater. Sci.*, 2013, **68**, p 314–319. https://doi.org/10.1016/j.commatsci.2012.10.028
56. J.P. Perdew, M. Ernzerhof, and K. Burke, Rationale for Mixing Exact Exchange with Density Functional Approximations, *J. Chem. Phys.*, 1996, **105**(22), p 9982–9985. https://doi.org/10.1063/1.472933
57. B. Deng, Materials Project Trajectory (MPtrj) Dataset (2023) https://doi.org/10.6084/m9.figshare.23713842.v2 .
58. A. Hjorth Larsen, J. Jørgen 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. Bjerre Jensen, J. Kermode, J.R. Kitchin, E. Leonhard Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. Bergmann 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, and K.W. Jacobsen, The Atomic Simulation Environment–A Python Library for Working with Atoms, *J. Phys. Condens. Matter*, 2017, **29**(27), p 273002. https://doi.org/10.1088/1361-648X/aa680e
59. M. Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C. Krishna Ande, S. Zwaag, J.J. Plata, C. Toher, S. Curtarolo, G. Ceder, K.A. Persson, and M. Asta, Charting the Complete Elastic Properties of Inorganic Crystalline Compounds, *Sci. Data*, 2015, **2**(1), p 150009. https://doi.org/10.1038/sdata.2015.9
60. S. Echeverri Restrepo, Density Functional Theory Characterisation of Cementite (Fe3C) with Substitutional Molybdenum (Mo) Atoms, *Rev. B Condens. Matter*, 2022, **631**, p 413669. https://doi.org/10.1016/j.physb.2022.413669
61. G. Gottstein, and L.S. Shvindlerman, *Grain Boundary Migration in Metals: Thermodynamics, Kinetics, Applications*, 2nd edn. CRC Press, Boca Raton, 2009. https://doi.org/10.1201/9781420054361
62. H. Zheng, X.-G. Li, R. Tran, C. Chen, M. Horton, D. Winston, K.A. Persson, and S.P. Ong, Grain Boundary Properties of Elemental Metals, *Acta Mater.*, 2020, **186**, p 40–49. https://doi.org/10.1016/j.actamat.2019.12.030
63. H. Bhadeshia, and R. Honeycombe, Chapter 2 - strengthening of iron and its alloys, in *Steels: Microstructure and Properties*, 4th edn. H. Bhadeshia and R. Honeycombe, Eds., Butterworth-Heinemann, Oxford, 2017, p23–57. https://doi.org/10.1016/B978-0-08-100270-4.00002-0
64. C. Domain, C.S. Becquart, and J. Foct, Ab Initio Study of Foreign Interstitial Atom (C, N) Interactions with Intrinsic Point Defects in α-Fe, *Phys. Rev. B*, 2004, **69**(14), p 144112. https://doi.org/10.1103/PhysRevB.69.144112
65. T.Q. Nguyen, K. Sato, and Y. Shibutani, Development of Fe-C Interatomic Potential for Carbon Impurities in α-Iron, *Comput. Mater. Sci.*, 2018, **150**, p 510–516. https://doi.org/10.1016/j.commatsci.2018.04.047
66. H.C. Herper, E. Hoffmann, and P. Entel, Ab Initio Full-Potential Study of the Structural and Magnetic Phase Stability of Iron, *Phys. Rev. B*, 1999, **60**(6), p 3839–3848. https://doi.org/10.1103/PhysRevB.60.3839
67. R. Tran, Z. Xu, B. Radhakrishnan, D. Winston, W. Sun, K.A. Persson, and S.P. Ong, Surface Energies of Elemental Crystals, *Sci. Data*, 2016, **3**(1), p 160080. https://doi.org/10.1038/sdata.2016.80
68. D. Dragoni, D. Ceresoli, and N. Marzari, Vibrational and Thermoelastic Properties of BCC Iron from Selected EAM Potentials, *Comput. Mater. Sci.*, 2018, **152**, p 99–106. https://doi.org/10.1016/j.commatsci.2018.05.038
69. C. Domain, and C.S. Becquart, Ab Initio Calculations of Defects in Fe and Dilute Fe-Cu Alloys, *Phys. Rev. B*, 2001, **65**(2), p 024103. https://doi.org/10.1103/PhysRevB.65.024103
70. A.F. Bialon, T. Hammerschmidt, and R. Drautz, Ab Initio Study of Boron in α -Iron: Migration Barriers and Interaction with Point Defects, *Phys. Rev. B*, 2013, **87**(10), p 104109. https://doi.org/10.1103/PhysRevB.87.104109
71. M. Souissi, Y. Chen, M.H.F. Sluiter, and H. Numakura, Ab Initio Characterization of B, C, N, and O in BCC Iron: Solution and Migration Energies and Elastic Strain Fields, *Comput. Mater. Sci.*, 2016, **124**, p 249–258. https://doi.org/10.1016/j.commatsci.2016.07.037. **Accessed 2025-03-10**
72. P.P. Ippolito, Hyperparameter tuning: the art of fine-tuning machine and deep learning models to improve metric results. In: Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications, pp. 231–251. Springer (2022), https://doi.org/10.1007/978-3-030-88389-8 .
73. P. Olsson, T.P.C. Klaver, and C. Domain, Ab Initio Study of Solute Transition-Metal Interactions with Point Defects in BCC Fe, *Phys. Rev. B*, 2010, **81**(5), p 054102. https://doi.org/10.1103/PhysRevB.81.054102
74. L. Messina, M. Nastar, N. Sandberg, and P. Olsson, Systematic Electronic-Structure Investigation of Substitutional Impurity Diffusion and Flux Coupling in BCC Iron, *Phys. Rev. B*, 2016, **93**(18), p 184302. https://doi.org/10.1103/PhysRevB.93.184302
75. T.M. Whiting, P.A. Burr, D.J.M. King, and M.R. Wenman, Understanding the Importance of the Energetics of Mn, Ni, Cu, Si and Vacancy Triplet Clusters in BCC Fe, *J. Appl. Phys.*, 2019, **126**(11), p 115901. https://doi.org/10.1063/1.5109483