npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-025-01727-x

Fine-tuning foundation models of materials interatomic potentials with frozen transfer learning

Check for updates

Mariia Radova[1,2], Wojciech G. Stark[1], Connor S. Allen[2,3], Reinhard J. Maurer[1,2] ✉ & Albert P. Bartók[2,3] ✉

Machine-learned interatomic potentials are revolutionising atomistic materials simulations by providing accurate and scalable predictions within the scope covered by the training data. However, generation of an accurate and robust training data set remains a challenge, often requiring thousands of first-principles calculations to achieve high accuracy. Foundation models have started to emerge with the ambition to create universally applicable potentials across a wide range of materials. While foundation models can be robust and transferable, they do not yet achieve the accuracy required to predict reaction barriers, phase transitions, and material stability. This work demonstrates that foundation model potentials can reach chemical accuracy when fine-tuned using transfer learning with partially frozen weights and biases. For two challenging datasets on reactive chemistry at surfaces and stability and elastic properties of tertiary alloys, we show that frozen transfer learning with 10–20% of the data (hundreds of datapoints) achieves similar accuracies to models trained from scratch (on thousands of datapoints). Moreover, we show that an equally accurate, but significantly more efficient surrogate model can be built using the transfer learned potential as the ground truth. In combination, we present a simulation workflow for machine learning potentials that improves data efficiency and computational efficiency.

Foundation models are large-scale machine learning models pre-trained on vast and diverse datasets that have revolutionised many domains in machine learning, enabling remarkable transferability and adaptability across various tasks. They were first popularised in Natural Language Processing (NLP) in the 2010s through models such as BERT[1] and GPT[2], then became used in computer vision, for example, in Vision Transformers[3]), and image recognition applications such as ResNet[4] and DenseNet[5] Recently, foundation models were generated in atomistic modelling and materials science, with a particular focus on materials discovery applications. Examples of foundation models trained on diverse databases of chemical structures are MEGNet[6], GemNet[7], CHGNet[8], MACE-MP[9], ALIGNN[10], GNoME[11], and more recent models: GRACE[12], EquiformerV2[13], MatterSim[14] and Orb[15]. Foundation models represent a shift towards well-generalised models as opposed to problem-specific potentials, although often at the price of reduced accuracy in predictions. Fine-tuning foundation models for a specific task presents a data-efficient compromise, as data generation can be exceedingly costly, especially in atomistic modelling, where it relies on first principles electronic structure methods.

In the domain of atomistic modelling, foundation models typically use Graph Neural Network (GNN) architectures that aim to capture atomic interactions via message passing. Message passing allows learning atomic representation through the exchange of messages between atoms, represented by nodes in the graph. The MACE[16] architecture incorporates many-body messages and equivariant features which facilitate capturing symmetry properties of the atomic structures. Recently, large foundation models based on MACE were trained using the Materials Project dataset (MPtrj)[8] that have shown impressive performance on a wide variety of benchmark systems[9].

However, when the dynamics of complex systems are modelled, such as gas-surface dynamics or studying phase transitions, the training database of foundation models will inevitably under-represent atomic environments relevant to achieving quantitative predictions. In such cases, fine-tuning exploits the capability of the model to generalise, achieving excellent accuracy using a limited amount of data. Naive fine-tuning of MACE-MP foundation models – starting from the final checkpoint of the pre-trained model – has been shown to be data-efficient[17]. However, this approach may lead to catastrophic forgetting and can pose risks of training instability due

[1]Department of Chemistry, University of Warwick, Coventry, UK. [2]Department of Physics, University of Warwick, Coventry, UK. [3]Warwick Centre for Predictive Modelling, School of Engineering, University of Warwick, Coventry, UK. ✉e-mail: r.maurer@warwick.ac.uk; apbartok@gmail.com

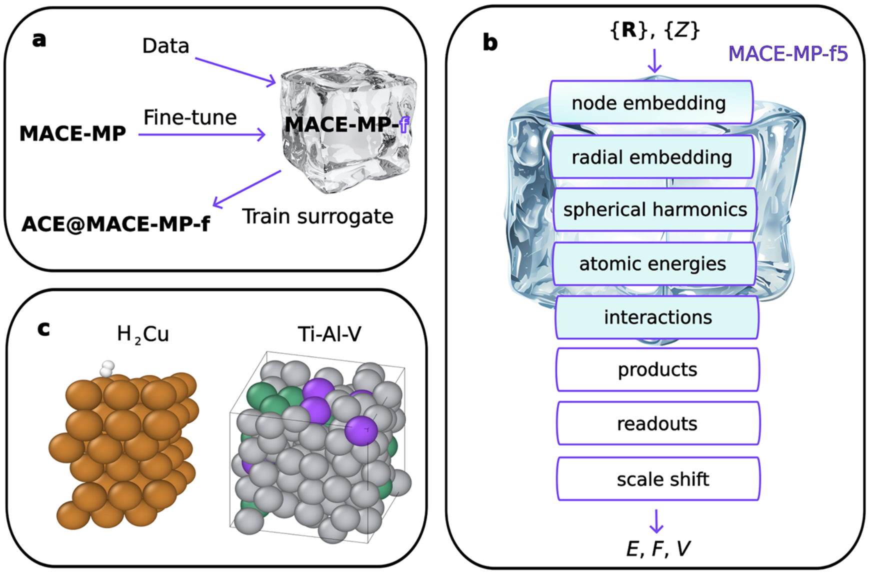

Fig. 1 | A schematic representation of the key concepts of the MACE-freeze method. a The workflow for training a model based on MACE-MP using transfer learning and using it to generate accurate data to fit a fast ACE surrogate model. **b** Transfer learning by freezing parameters in the model layers. **c** The atomic systems used in our benchmarks.

to continued updating of deeper network layers, which are particularly susceptible to divergence[18,19]. To address catastrophic forgetting, multi-head fine-tuning has recently been introduced for MACE-MP models[9]. It focuses on maintaining transferability across the systems represented in the MPtrj dataset and allows training on data obtained from multiple levels of electronic structure theory. Another fine-tuning method has been suggested that transforms the MACE-MP descriptors into random-feature (RF) maps, focussing on data efficiency[20]. Frozen transfer learning has been implemented for CHGNet[8], demonstrating its robustness and accuracy, however, the aspect of data efficiency of their approach remained unexplored. Notably, the fine-tuning of CHGNet required a substantial dataset, with more than 196,000 structures used in the fine-tuning database, demonstrating a similar performance and database requirement as those of a from-scratch CHGNet model.

In this work, we apply the transfer learning method with partially frozen weights and biases for fine-tuning Machine Learning Interatomic Potential (MLIP) foundation models (Fig. 1a). Our aim is to use the foundation model as a stepping stone to create a tailor-made model that can describe the dynamics of a specific system as accurately as possible with as little data as possible. We discuss two challenging systems to show how transfer learning on the atomistic foundation models can be a more data-efficient way of generating highly accurate models than training them from scratch with only task-specific data: dissociative adsorption of molecular hydrogen on copper surfaces, and a ternary alloy (Fig. 1c). Furthermore, we show that the fine-tuned foundation model can be used to generate ground truth labels for a more efficient surrogate model based on the Atomic Cluster Expansion (ACE)[21]. In doing so, we benefit from the data efficiency of the fine-tuning process and can still produce a model capable of rapid inference to tackle large-scale or massively parallel simulations.

Results and Discussion

Transfer learning

The transfer learning technique implemented in this work for fine-tuning MACE-MP foundation models involves controlled freezing of neural network model layers. In this process, the model parameters corresponding to a particular model layer that are kept fixed during training. In other words, backpropagation is only carried out on the active neural network layers (Fig. 1b). This technique has proved to be efficient in the fine-tuning of convolutional neural network models trained for image recognition[22–24]. The data efficiency of transfer learned models is justified by the fact that common features or patterns deduced in the original training phase are retained. These low-level components of the model are expected to be general, and the remaining adjustable parameters can be reliably fitted using the scarce training data provided during the transfer learning procedure. Fixing parameters in some of the layers also reduces the training time for frozen transfer learned models compared to back-propagating the information through the whole model. In cases when there is only little data available for training a from-scratch-trained neural network, transfer learning represents an efficient alternative, especially if generating new data is computationally expensive, for example in atomistic simulations.

We created the `mace-freeze` patch[25] to the MACE software suite that allows to freeze layers or parameter tensors in any MACE model to fine-tune them using a particular dataset of interest. This approach allows for retaining the features learned from the MPtrj dataset of the MACE-MP foundational model and adapting the later layers to the new task. The models, which we refer to as MACE-freeze, trained using `mace-freeze` patch retain the same architecture as the original MACE model used for fine-tuning and only differ in which layers or parameter groups are frozen.

Transfer learned model data efficiency and performance: H_2/Cu

To show the efficiency of the MACE-freeze models, we compare their accuracy to from-scratch-trained MACE models trained on the same dataset. The hyperparameters of the from-scratch-trained MACE models are fully optimised on the dataset. We aim to demonstrate that a “universal” model such as MACE-MP can be fine-tuned using transfer learning to perform a specific task at least as well as a from-scratch MACE model that was trained specifically for this task only. Moreover, we show that the transfer learned model can outperform, at least in the low-data regime, the from-scratch model in the high-data regime. As a consequence, a much smaller number of training points are required to achieve similar accuracy of predictions with the MACE-freeze model, than the bespoke MACE model.

The “small”, “medium” and “large” MACE-MP foundation models were used as the basis of our transfer learned models using a dataset on reactive hydrogen chemistry on various facets of copper surfaces[26]. The database contains 4230 structures and was obtained using reactive gas-surface scattering molecular dynamics simulations and a committee uncertainty-driven active learning algorithm. A from-scratch MACE model was trained using this dataset, with hyperparameters optimised previously, which were validated with k-point cross-validation and Molecular Dynamics (MD)[27].

The MACE-freeze models retain the architecture of the foundation models and use varying numbers of frozen layers of the original MACE-MP “small”, “medium” and “large” foundation models. We show the learning curves of different “small” freeze models, as well as the from-scratch-trained

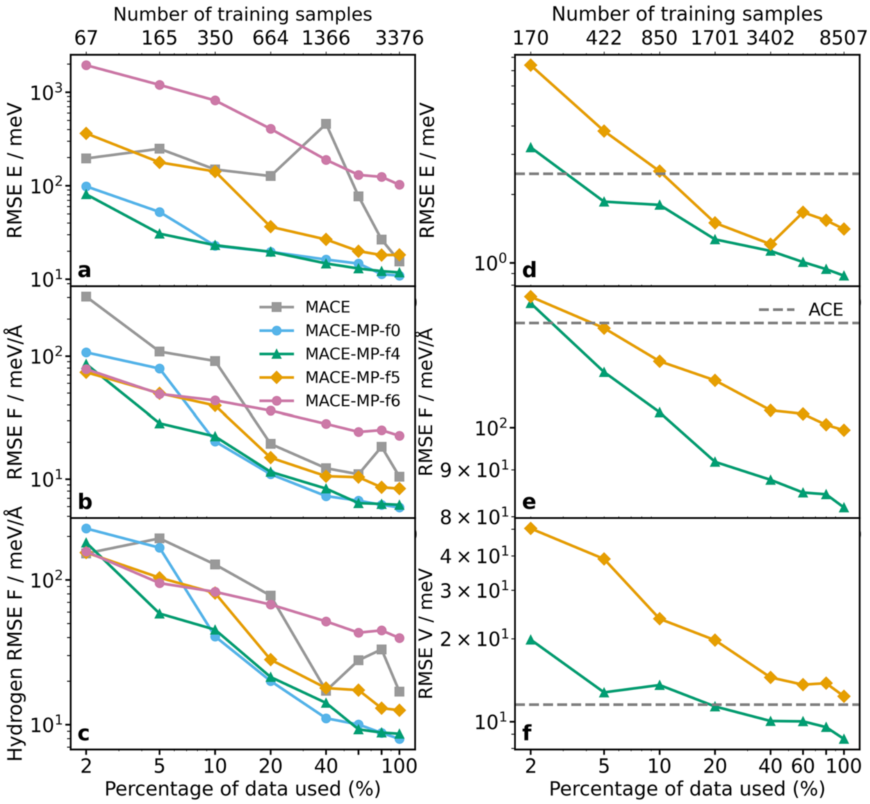

Fig. 2 | Transfer learning curves for Hydrogen on Copper surface and for the Ti-Al-V alloy systems, trained based on the "small" foundation model. For the Cu-H_2 system, root mean squared errors (RMSEs) of energies, force components and force components of the H atoms only are shown in panels (**a**–**c**), respectively. For the Ti-Al-V alloy system, root mean squared errors (RMSEs) of energies, force components and virial stress components are shown in panels (**d**–**f**), respectively. The points correspond to the percentages of the respective datasets, namely 2, 5, 10, 20, 40, 60, 80 and 100% for both systems. The layers were frozen correspondingly to f6 (pink circles), f5 (yellow diamonds), f4 (green triangles), and f0 (blue circles). Grey squares mark the learning curve of the from-scratch-trained MACE model in case of the Cu-H_2 system, and the grey dashed line marks the errors of the custom ACE model in case of the Ti-Al-V alloy system.

MACE model in Fig. 2a–c. We trained models where parameters are fixed in all layers except the readouts (MACE-MP-f6) and then incrementally we allow parameters to vary in the product layer (MACE-MP-f5), and the interaction parameters (MACE-MP-f4), while in MACE-MP-f0 all layers are active.

In the low-data regime, where the models are trained on a small percentage of the original training datapoints, all the represented freeze models with the exception of the least flexible model, f6, perform better than the from-scratch-trained MACE model. As the number of frozen layers decreases, the predictive performance on energies and forces improves, peaking at f4. Allowing more flexibility by further reducing the number of frozen layers does not improve the predictive performance: the MACE-MP-f0 model, in which all parameters were allowed to update, has similar validation errors to that of MACE-MP-f4. The reduced number of trainable parameters in MACE-MP-f4, however, have the minor added benefit of reducing the computational cost of training (Supplementary Figures S1 and S2). Figure 2 also demonstrates that the lower layers of the network, as fitted in the original MACE-MP model do not benefit from further fine-tuning and may be reused. The benefit of the reduced cost of training can be of particular importance for users with limited computational resources. The superior performance of the transfer-learned models suggests that the models benefit from the pre-training due to other structures present in the MACE-MP training set, resulting in transferable and robust descriptor embeddings.

Having found the optimum number of frozen layers to be four in our applications, we used this setting for further benchmarks. Supplementary Figure S3 shows that the size of the foundation model does not significantly contribute to the accuracy of the transfer-learned models, but "medium" and "large" models are computationally more demanding. For this reason, in the subsequent models, we use the "small" MACE-MP foundation model for our further investigations.

At 20% of the training set (664 configurations), the MACE-MP-f4 model shows a similar level of accuracy (measured by root mean squared error, RMSE) on energies and total forces as the from-scratch-trained MACE model trained using all configurations in the training set (3376 points). Notably, MACE-MP-f4 predicts forces on hydrogen atoms that are significantly more accurate than the from-scratch-trained model hydrogen-only forces (Fig. 2c). Stark et al. previously found that the force errors on hydrogen atoms predicted by a from-scratch-trained MACE model are considerably higher, which is the key limiting factor in determining the accuracy of dynamic reaction probabilities such as sticking probabilities[27]. In contrast, all the transfer-learned models have resulted in better force error measures and more balanced force RMSEs across copper and hydrogen atoms.

Model validation and ACE-surrogate for H_2/Cu dynamics

To independently validate the transfer-learned MACE-MP models, we assess the prediction accuracy of the structure, reaction barriers and dynamic scattering probabilities. Supplementary Figure S4 reports the energy-volume curve of bulk Cu as predicted by Density Functional Theory (DFT), MACE-MP, MACE-MP-f4 10%, MACE-MP-f4 20%, and the previously published MACE model[27]. While the results obtained with the MACE-MP models deviate from DFT methods, we may attribute the differences to the underlying methods and employed electronic structure packages: the PBE functional[28] and VASP[29–31] were used to evaluate the MP database while the SRP functional[32] and FHI-aims[33] were used to generate the Cu-H database. Using only 10% of data from our previous database for transfer learning, the agreement of our MACE-MP-f4 models is in excellent agreement with DFT, outperforming the from-scratch-trained MACE model which was fitted using the entire database.

An accurate description of lattice expansion can play a significant role in predicting reaction probabilities at metal surfaces[34]. Thus, we explore the ability of the various approaches to capture the surface temperature-mediated lattice expansion by running NPT simulations at 9 different temperatures, between 200 and 1000 K. We observed that the foundation MACE-MP models compared to our other models underestimate the lattice constants by 0.035 Å across all the considered temperatures (Supplementary Figure S5), which can be attributed to differences in DFT functionals and

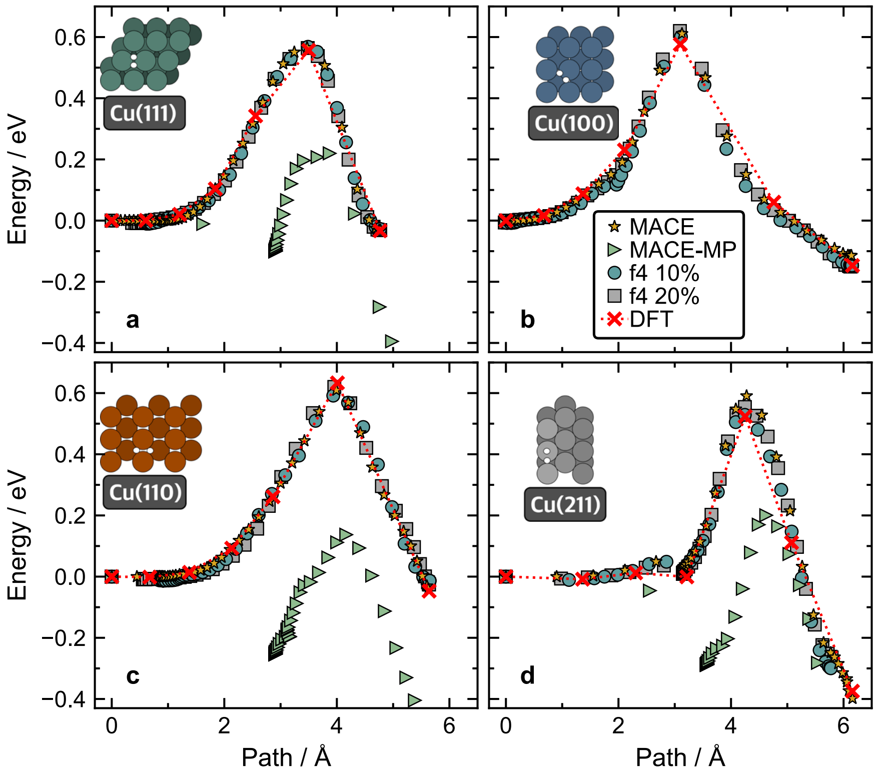

Fig. 3 | Minimum energy paths obtained using CI-NEB method for H_2 dissociative adsorption on different copper surfaces. Potential energy values are shown along the reaction path (Å), evaluated using DFT (red ×) and the MLIPs included in our study: MACE (yellow stars), MACE-MP (green triangles), MACE-MP-f4 10% (blue circles), and MACE-MP-f4 20% (grey squares) for H_2 dissociation at Cu(111) (**a**), Cu(100) (**b**), Cu(110) (**c**), and Cu(211) (**d**) surfaces. The top-down views of the configurations at the transition states are included on the respective plots. All DFT data points included in the plot were taken from ref. 26. For Cu(100), MACE-MP predictions are beyond the shown energy scale in panel b.

codes used in generating the respective training databases. Given that the MACE-MP foundation models were originally trained on DFT data obtained using the PBE functional, this behaviour is expected as with PBE it was reportedly not possible to reproduce adsorption-related experimental observables for H_2 at Cu surfaces. To remedy this failure, the SRP48 functional was constructed as a combination of PBE and RPBE functionals (52% and 48%, respectively), to fit the experimental data, giving the most reliable prediction capabilities for this system[35,36]. Notably, already after including 10% of the structures from our Cu-H database the results obtained with a from-scratch-trained MACE model, trained on 100% of the data, and MACE-MP-f4 transfer models are in very close agreement.

To assess the ability of the models to predict the reaction barriers of H_2 dissociation at different Cu surfaces we evaluated the minimum energy paths (MEPs) using climbing image nudged elastic band (CI-NEB) method (Fig. 3). MACE-MP models were not able to predict the MEPs well for any of the surfaces we considered. They also predicted spurious local minima that were not found with DFT or MACE models, e.g. at Cu(111) surface around 3 Å or at Cu(211) at 3.5 Å. The spurious minima disappear in MEPs generated with both the 10%, and 20% transfer models, and the predictions of barriers match the reference results well.

One of the most important dynamical observables for investigating adsorption processes at metal surfaces is the sticking probability (P_{stick}), which is the probability of an H_2 molecule with a given vibrational (v) and rotational (J) initial state to dissociatively adsorb at the surface, as opposed to scattering from it. The sticking probability is calculated as the ratio between reactive events and the total number of simulated scattering events ($P_{\mathrm{stick}} = n_{\mathrm{traj}}^{\mathrm{dissoc}}/n_{\mathrm{traj}}^{\mathrm{all}}$). Here, we evaluate the sticking probability for the adsorption of H_2 in the ground (v=0) and first excited (v=1) vibrational states and the first excited rotational state (J = 1) at Cu(111) (925 K) and H_2(v=1, J = 1) at Cu(211) (925 K) with MACE-MP-f4 10%, and compare the results with the models trained on the entire database (Fig. 4). As previously discussed, reactive scattering at 925 K surface temperature is a challenging validation of the models as the training dataset was generated from data samples drawn from low-temperature scattering, which does not guarantee that the trained models generalise to high-temperature scattering[27]. The dynamics simulated with MACE-MP in all cases led to an overestimation of sticking probabilities. This overestimation is unsurprisingly largest at lower collision energies, where the prediction of the reaction barrier is the greatest determining factor in obtaining accurate sticking probabilities. As shown before in Fig. 3, MACE-MP models severely underestimate the MEPs, leading to increased sticking probabilities. Transfer learning on just 10% of our database leads to a significant improvement, resulting in sticking probabilities very close to the experimental results depicted by the red line. The agreement is comparatively good with our previous from-scratch-trained MACE model trained exclusively on our DFT-based database. Here, the only visible discrepancy in the predictions of the sticking probability is at higher collision energies (above 0.5 eV for H_2(v=0), and 0.3 eV for H_2(v=1)). The deviation is less than 10%, which is at a similar level to the difference between the experiment and the MACE predictions. Notably, as the experimental reference is based on permeation experiments[37], the results have to be scaled to be comparable to theoretical predictions[26]. As described in detail in previous work, we have scaled the experiment to match the previous theoretical predictions of the from-scratch-trained MACE model[26]. The simulation results suggest that we have reached the level of accuracy of the full-DFT MACE model with only 10% of structures (DFT evaluation), thereby reducing the cost of model training by approximately a factor of ten. The generation of training data for the Ti-Al-V system required approximately 1 million CPU hours, while the Cu-H dataset required approximately 500,000 CPU hours.

While the transfer-learned MACE-MP-f4 model provides highly accurate predictions, the time-to-solution for energy and force predictions is significantly slower than for the from-scratch-trained MACE model. The latter is a highly optimised and small model tailored for the given dataset (5 Å cutoff, correlation order 2, 16 × 0e model size), whereas MACE-MP "small" has larger cutoffs and more parameters (6 Å cutoff, correlation order 3, 128 × 0e model size. The hyperparameters of the reference from-scratch model were justified by Stark et al. (ref. 27), who reported that increasing the

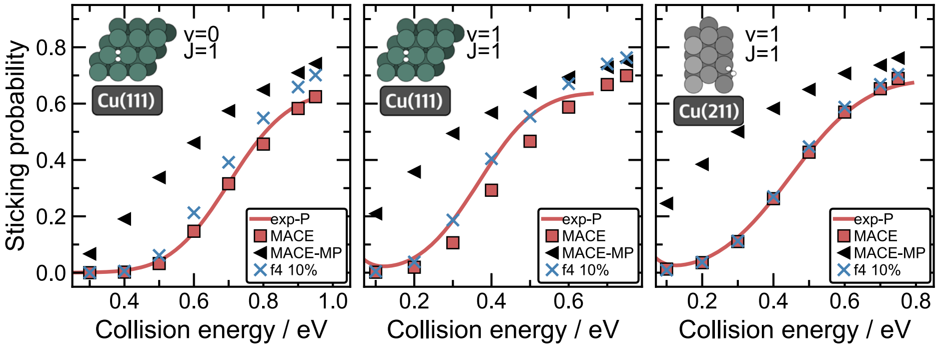

Fig. 4 | Sticking probabilities for H_2 scattering on Cu(111) and Cu(211) at 925 K. Probabilities were calculated at different collision energies using MACE (red squares), MACE-MP (black triangles), MACE-MP-f4 10% (f4 10%) (blue ×) models for the ground (v=0) (left) and excited (v=1) (middle) H_2 vibrational states at Cu(111) surface and excited (v=1) state at Cu(211) (J = 1 in every case). MACE refers to the model based on the DFT-based database from ref. 26. The red line represents a sticking probability obtained from the experimental results of Kaufmann et al.[37] (exp-P) at 923 ± 3 K, scaled to match theoretical probabilities from ref. 26 at the highest incidence energy (saturation parameter A=0.64 for both Cu(111) sticking functions, and A=0.66 for Cu(211)).

Table 1 | Performance of MACE-MP-f4 10%, ACE-f 10%, and from-scratch (f-s) ACE and MACE models

Property	MACE-MP-f 10%	ACE-f 10%	ACE (f-s)	MACE (f-s)
MAE (E)	6.7	4.8	8.6	11.3
RMSE (E)	13.5	6.4	13.7	15.8
MAE (F)	3.1	7.2	10.4	8.1
RMSE (F)	10.1	10.9	18.2	12.9
MAE (H-atom F)	4.3	9.0	28.5	17.7
RMSE (H-atom F)	17.1	17.7	47.6	32.4
t_{eval} (E)	372.6	9.6	9.1	60.0
t_{eval} (F)	390.5	22.6	20.7	60.0

Energy (E) and force (F) errors are included in meV and meV/Å, respectively. Evaluation times (t_{eval}) are in ms and were calculated on a single AMD EPYC 7742 (Rome) 2.25 GHz CPU processor core.

model size and complexity yielded minimal improvement in the predictive performance of the from-scratch model. As the dynamics simulations require tens of thousands of trajectories at various incidence energies to provide converged sticking probabilities, the transfer-learned MACE-MP model, despite its accuracy, comes at an almost prohibitive computational cost (Table 1). Whereas the from-scratch-trained model evaluates energies and forces per geometry in 60 ms, MACE-MP-f requires 390.5 ms on an AMD EPYC 7742 (Rome) CPU processor core.

To address this, we take MACE-MP-f4 predictions as ground truth labels and construct a new model using the ACEpotentials.jl[38] package, which fits linear Atomic Cluster Expansion (ACE) potentials[39], enabling significantly faster force evaluations[27] than the more complex MACE-MP models. We will refer to this model as ACE-f 10%. Previously, optimised ACE models trained on the full database were not able to provide sufficiently low force errors and accurate sticking probability predictions[27]. We utilise data generated with MACE-freeze 10% potential exclusively to construct a new training set for ACE-f 10%. This was carried out by generating a total of 600 trajectories, comprised of 5 trajectories each for H_2 scattering at 4 different Cu surfaces, namely, (111), (110), (100), and (211), at 3 surface temperatures (300, 600, and 900 K), at two rovibrational states (H_2(v=0, J = 0) and H_2(v=1, J = 1)), at 5 different collision energies. Such comprehensive sampling of training data would not be possible with on-the-fly ab-initio dynamics while active learning requires many more simulations and training/learning loops[26]. For each trajectory, structures were saved at the interval of 1 fs. Then the k-means clustering method was used to choose 2000 of the most diverse configurations. Furthermore, we included the MEPs obtained from nudged elastic band (NEB) calculations, 50 structures along reaction paths at each studied surface. The final database contained 2200 structures. ACE (ACE-f 10%) models were trained based on this database employing the hyperparameter settings used for the same system in our previous study[27]. The final ACE-f 10% model achieves excellent accuracy, which even combined with the MACE-MP-f4 10% evaluation errors reaches chemical accuracy (Table 1). This is a significant improvement compared to the previous ACE model trained directly on the full DFT database. The previous ACE model was unable to reach the accuracy of the neural network-based methods[27], especially for forces on hydrogen atoms, where the best performance was limited to an MAE (mean absolute error) 28.5 meV/Å, and RMSE of 47.6 meV/Å). These errors are approximately 3 times larger than for ACE-f 10% (Tab. 1). We attribute this discrepancy to the presence of outliers in our previous database, which was adaptively sampled using less accurate and unstable SchNet models[26]. Thus, as previously found by others[40], constructing a large synthetic database directly using a model of high accuracy and smoothness is an efficient approach to training accurate linear models.

To examine the ability of the ACE-f 10% model to predict the actual dynamical observables, we evaluated sticking probabilities for the same systems and settings as for MACE-freeze 10% model (Fig. 5). In all cases, the agreement between ACE-f 10% and MACE-freeze model is good. Additionally, we evaluated sticking probabilities with the ACE model trained, using the same settings, on our previous database (ACE-S). The probabilities obtained with ACE-S model match the probabilities obtained with the ACE-f 10% model well for both rovibrational states of Cu(111), however, this is not the case for the dynamics at Cu(211), where the ACE-S model significantly underestimates, by more than 10%, the sticking probabilities for collision energies above 0.3 eV.

Evaluating probabilistically determined dynamical observables, such as sticking probability requires simulating hundreds of thousands of MD trajectories, which leads to high computational expenses. Employing MLIPs instead of traditional methods, such as DFT, reduces that time significantly. Due to the high generalizability of MACE-MP models, their complexity needs to be higher than MLIPs generated for a specific system, meaning the evaluation times are significantly higher. Training fast, small models based on specialized MACE-MP models, such as MACE-MP-f models, can help mitigate this issue. Here, a single force evaluation on an AMD EPYC7742 2.25 GHz CPU processor with our MACE-MP-f4 10% model takes approximately 390 ms, however, by training an ACE-f 10% model we reduce this time by more than 17 times (22.6 ms, Table 1), while preserving the prediction quality, as shown in Fig. 5.

Ti-Al-V data and benchmarks

A dataset was constructed for the purpose of building MLIPs to accurately model the crystalline and liquid phases of Ti-6Al-4V (Ti 90 wt%, Al 6 wt%, V 4 wt%) alloy up to 30 GPa. To model crystalline Ti-6Al-4V, the three physically observable phases below 30 GPa were considered: α (hcp, P6_3/mmc), β (bcc, Im-3m) and ω-Ti (hexagonal, P6/mmm).

To benchmark this dataset, we considered a validation set, elastic properties and vibrational properties. The validation set consisted of a series

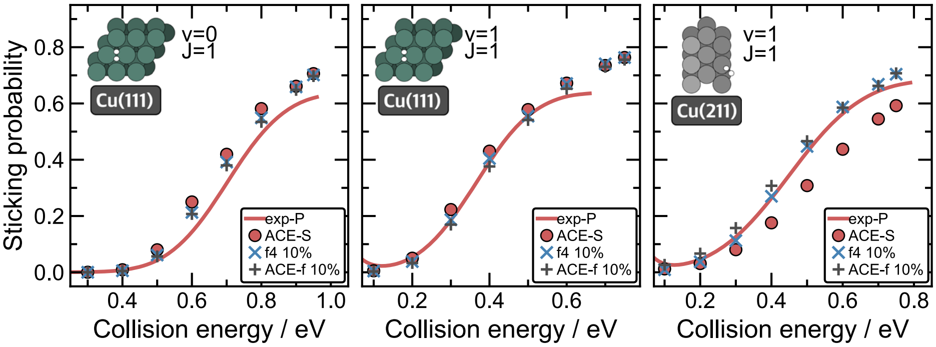

Fig. 5 | Sticking probabilities for H_2 scattering on Cu(111) and Cu(211) at 925 K. Probabilities were calculated at different collision energies using ACE-S (red circles), MACE-MP-f4 10% (f4 10%) (blue ×), and ACE-f 10% (grey +) models for the ground (ν=0) (left) and excited (ν=1) (middle) H_2 vibrational states at Cu(111) surface and excited (ν=1) state at Cu(211) (J = 1 in every case). ACE-S refers to the model based on the DFT-based database from ref. 26. The red line represents a sticking function obtained from the experimental results of Kaufmann et al.[37] (exp-P) at 923 ± 3 K, scaled to match theoretical probabilities from ref. 26 at the highest incidence energy (saturation parameter A=0.64 for both Cu(111) sticking functions, and A=0.66 for Cu(211)).

of large simulation cells that resemble the Ti-6Al-4V stoichiometry, where developed models can be compared against the configurational energy, forces and virial stresses. We also calculated the elastic constants for simulation cells representative of Ti-6Al-4V in each crystalline phase using `matscipy`[41]. Due to the dilute amount of Al and V in Ti-6Al-4V, it is not tractable to compute phonon properties in a cell representative of this stoichiometry. Therefore, to characterise vibrational properties we consider phonon dispersion, density of states and quasi-harmonic free energy calculations of simulation cells that represent minor alloying component nearest neighbour interactions, even though the compositions correspond to significantly higher Al and V concentrations than that of Ti-6Al-4V. Calculations were performed in 8 (α and β) and 12 (ω) atom simulation cells, with a vibrational Brillouin Zone grid sampling of 2 × 2 × 2.

The reader is referred to ref. 42 where the full details of the database construction, benchmarks, and alternative MLIP developments are presented. Also discussed in ref. 42 is the development of the ACE model for Ti-6Al-4V which is presented here as a baseline model alongside the MACE models developed in this work. In this work, we show that frozen transfer learning, using as little as 10% of the database, is able to achieve superior predictive performance on our benchmarks compared to ACE potentials fitted using the entire database.

The "small" foundation model was fine-tuned using the `mace-freeze` patch on different percentages of the training database, amounting to 8507 structures in total. We show in Fig. 2, that the validation error metrics of the MACE-MP-f4 model start to surpass the ACE baseline figures at less than 5% of training data when considering energy and force values, and at 20% of training data for virial stress components. Compared to ACE, MACE-MP-f5 becomes more accurate at around 10% data for energies, and 5% for forces. We note that the validation accuracy of MACE-MP-f5 on the virial stress components approaches that of the ACE model at 40% training data, but never surpasses it.

We have applied transfer learning to all the "small", "medium" and "large" tiers of the MACE-MP foundation models. We observed that, similarly to the copper-hydrogen system, the validation accuracy figures are very similar (Supplementary Figure S3) across the differently sized models. We note that with the validation accuracy of the "large" model being the worst, it may indicate a small degree of overfitting. For computational efficiency, we have used the transfer learned model based on the "small" MACE-MP model for our further benchmarking experiments.

The transfer models along with the ACE model fitted on the entire database were evaluated using a set of benchmark calculations, designed to explore the accuracy of the potential energy surface (PES) of the Ti-Al-V system at various compositions in the relevant crystalline phases. Using DFT calculations as our reference, elastic and vibrational properties, as well as root mean squared error (RMSE) values of energies, forces and virial stress components in a set of independent configurations were predicted with our models. To compare our models, we define a figure of merit (FOM) to compress the error metrics into a single scale. For each benchmark property p that was calculated using model m we first calculated the absolute error $\Delta_{p,m}$ which was transformed to

$$\tilde{\Delta}_{p,m} = \frac{\Delta_{p,m} - \Delta_{p,\mathrm{worst}}}{\Delta_{p,\mathrm{best}} - \Delta_{p,\mathrm{worst}}}$$

and used to determine the average FOM as

$$\mathrm{FOM} = \frac{1}{N_p}\sum_p \tilde{\Delta}_{p,m}$$

We note that the FOM value of a model may vary depending on the other models included in the comparison.

We present all the metrics considered for benchmarking the Ti-6Al-4V models in Supplementary Figures S6–S9 in the SI. Various benchmarks grouped together into four distinct categories ("validation", "training", "elastics" and "vibrational") are shown in Fig. 6, showing the performance of transfer learned models as a function of increasing dataset size. We collated individual benchmarks into their respective categories to demonstrate how balanced the performance of the various models is, although note that the uniform weighting of the $\tilde{\Delta}_{p,m}$ values is an arbitrary choice. Within the "training" and "validation" categories, we consider the RMSEs of the predicted energy, forces and virial stress quantities, compared to DFT reference values evaluated on crystalline and liquid configurations. Within the "elastics" category, we consider the absolute difference in elastic constants between our models and DFT across the physically relevant crystalline phases, characterising second derivatives of the PES with respect to deformations of the lattice. Finally, we characterise vibrational properties by calculating the RMSE in phonon dispersions, as well as considering the absolute error differences in the quasi-harmonic free energy values evaluated at 0 and 2000 K, thus providing insight into the reproduction of the force constant matrices of our models and their finite-temperature behaviour.

We compare the performance of the transfer learned models to the baseline ACE model trained on 100% of the data, with the aim of achieving the same or better accuracy with as little data as possible. We have demonstrated that transfer learned potentials based on the "small" MACE-MP model at the f4 (Fig. 6) and f5 level (Supplementary Figure S10) achieved comparable performance to the ACE model in the "validation", "elastics" and "vibrational" categories using only 10% or 20% of the database, respectively. It is important to note that even though the Ti-6Al-4V database contains 8507 individual configurations, the diversity of the data means that some compositions in the ordered phases are only represented by a few structures. In those cases, we ensured that the structures are included in the reduced database. To assess the uncertainty due to downsampling the original database, we trained a committee of 5 f4 models using 5 different random samples containing 10% of the original configurations. Using our FOM metrics, we present the performance of the committee members and

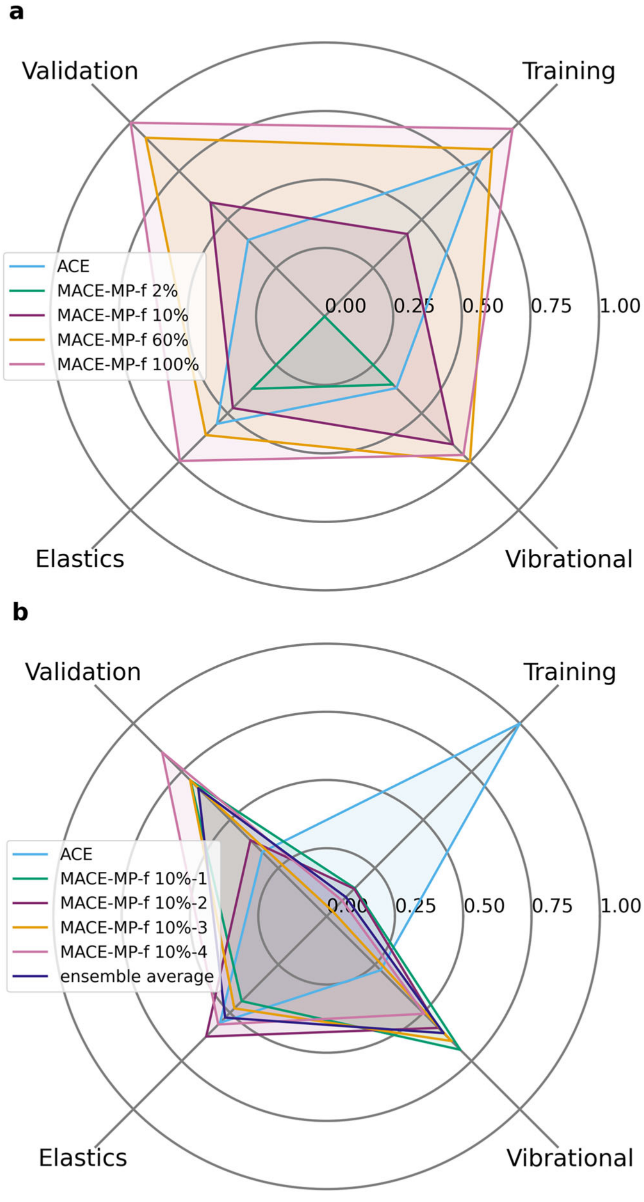

Fig. 6 | Figures of merit (FOM) for f4 models relative to the custom ACE trained on the Ti-Al-V data. Panel (**a**) shows how the FOMs of the transfer models (f4) improve as the size of the database increases. Panel (**b**) displays the FOMs of an ensemble of f4 10% models, quantifying the uncertainty obtained via random sampling of the training set.

that of their ensemble average compared to the ACE model trained on the entire database. While there is considerable variation between the committee members, their performance on the "validation", "elastic" and "vibrational" benchmark groups are similar to or better than that of the ACE model. The fact that these committee members are better than ACE indicates that transfer learning fills the gaps in the Ti-Al-V database which the database leaves out.

To demonstrate the performance of our transfer learned models, we present phonon dispersions along high symmetry lines and the phonon density of states for the relevant α, β and ω phases of pure titanium (Fig. 7). Excellent qualitative agreement with the DFT reference can be observed at using 5% of the database, which improves consistently as we increase the size of the database.

We also compared the f4 models to from-scratch MACE models (Supplementary Figure S11), using the improvement on the overall FOM relative to the ACE model. To facilitate a practical comparison, our from-scratch MACE models were trained with the same hyperparameters as the "small" MACE-MP foundation model[9]. The results show that the accuracy of the from-scratch Ti-Al-V MACE model is poorer in the low-data regime than that of the MACE-f4 Ti-Al-V model, while they perform similarly if more than 40% of the data is used in training. Even in the higher training data regime the transfer models have the advantage that fitting them is computationally less expensive, as they have fewer adjustable parameters and require fewer epochs in training. This reaffirms that the transfer models are able to provide significant benefits over from-scratch-trained models where little data is available.

Comparison to Δ-learning

Finally, Δ machine learning represents a commonly employed alternative to transfer learning to exploit correlations between distinct sources of data and to utilise multi-fidelity datasets simultaneously. We compare Δ learning to frozen transfer learning by training MACE models to fit the difference between the MACE-MP predictions, regarded as baseline, and the ab initio target using the Cu-H dataset. To evaluate the performance of the Δ-learning approach, we compared learning curves of the Δ-models to that of a MACE model trained from scratch (Supplementary Figure S12). We find that the from-scratch model remains superior for all but the smallest training set sizes, where the achieved prediction errors are insufficient for production simulations. While it may be possible to improve the performance of the Δ-model by using a more complex MACE architecture, it would result in a more computationally expensive potential. Overall, transfer learning in our tests achieved a significantly superior accuracy across low and high data regimes.

Discussion

Herein, we find that transfer learning of foundation models is a viable strategy to train MLIPs with a small amount of training data, while achieving the accuracy of from-scratch-trained potentials trained with a significantly larger number of atomic configurations. By fixing, or freezing, parameter groups of existing foundation models the complexity of the fitting procedure is reduced. Benchmarking our MACE-freeze approach demonstrated that the transfer learned models benefit from retaining the low-level layers in the MACE-MP neural network, which are often interpreted as descriptors[43,44], and generalised from the extensive MP database. We have shown that transfer learned models can not only achieve the accuracy of the from-scratch-trained models, but can surpass them in the accuracy of the fit and, subsequently, the predictions, sometimes able to reach the DFT noise level of RMSEs. It is also worth pointing out that the need for hyperparameter optimisation is reduced to the weights of the observables (energies, forces, stresses) only. Other hyperparameters are inherent to the foundation model, and do not need optimising. Transfer learned models based on MACE foundation models can be used in place of the first-principle methods such as DFT for data generation in active learning algorithms. The challenge of the accurate, but large models leading to slow, inefficient MD simulations can be solved by training an efficient surrogate model, based on the potentials of the large model, as demonstrated in this work.

Our work provides a workflow that allows rapid development of fast and highly accurate tailormade MLIPs using a minimal amount of ab initio reference data, by leveraging the information in foundation models and is, in principle, independent of the specific architecture of the foundation model. Having said that, further data efficiency and accuracy improvements may be possible if only specific types of parameter groups are frozen rather than whole layers of parameters. This will be a topic of future investigations.

We note that there are other viable transfer learning approaches to exploit scarce amounts of ab initio data to train accurate MLIPs. For example, Gardner et al. suggested training, from scratch, NequIP models based on atomic configurations obtained from simulations using MLIPs and using transfer learning to refine the resulting models with a small amount of DFT labels[45]. Such a targeted approach could be very ab initio data-efficient, as the original model is specifically adapted to the configurational space of interest. However, as discussed by Gardner et al. "the synthetic source

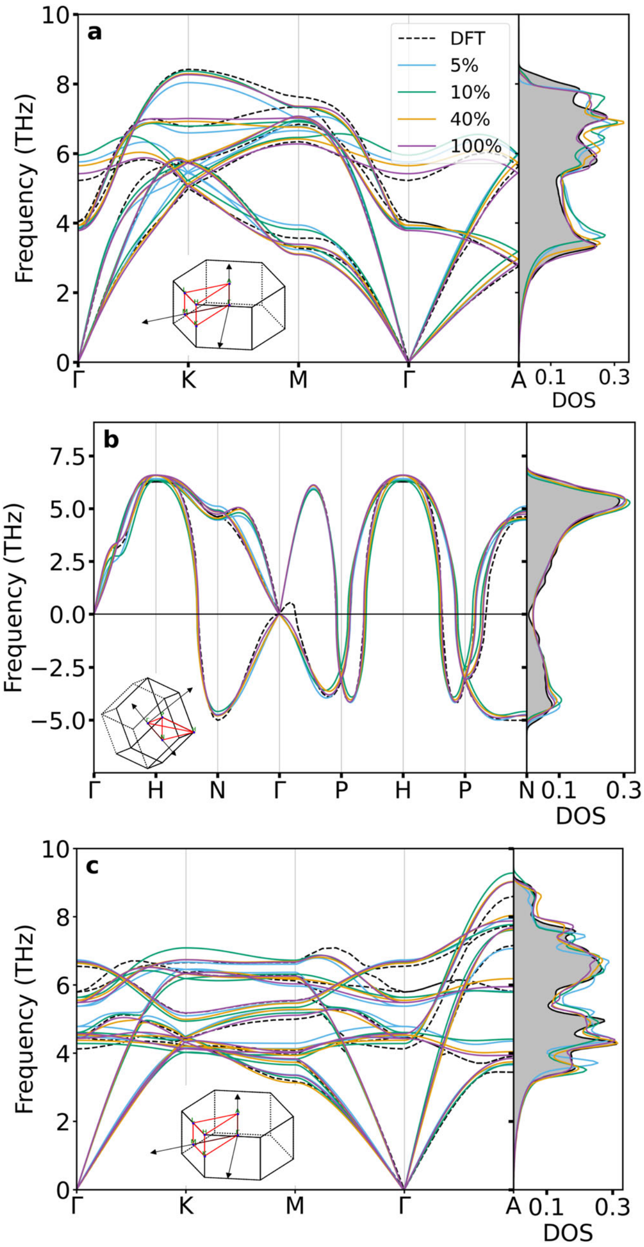

Fig. 7 | Calculated phonon dispersions for (**a**) *α*-Ti, (**b**) *β*-Ti, and (**c**) *ω*-Ti. Each subplot compares the predicted phonon dispersions from the f4 "small" models (5% - blue, 10% - green, 40% - yellow, 100% - purple) against DFT (dashed lines and gray density of states (DOS) projection).

matters", which in most cases means a good quality MLIP is a prerequisite to generating relevant atomic configurations. Our frozen learning relies on a much more general foundational model and achieves specificity by fixing parameter groups in the refining step, thereby avoiding the need to generate synthetic data and train a from-scratch model to be transfer-learned.

Frozen transfer learning inevitably reduces the transferability that foundation models provide in favour of creating a model suitable for a narrow application domain. The extent of this scope-narrowing requires further investigations. Other techniques, such as multi-head learning, have been proposed to retain transferability by replaying the original data on which the foundation model has been trained. Such approaches could be combined with frozen transfer learning in the future to retain a higher degree of transferability.

Negative transfer, where knowledge from a source domain impedes rather than enhances performance in the target domain, can pose a challenge in transfer learning[46]. This phenomenon often arises when the source and target distributions exhibit large discrepancies, causing the transferred features to introduce biases that hinder model convergence. However, when the pre-trained model has been trained on a sufficiently diverse and representative dataset that captures the chemical and structural diversity relevant to the target task, the learned embeddings and interaction patterns are more likely to generalise effectively[47]. Therefore, selecting a relevant foundation model for the target task is important to minimise the risk of negative transfer. For instance, if one is interested in inorganic materials domain, using a foundation model trained on organic molecules such as MACE-OFF may cause negative transfer effects. Supplementary Figure S13 illustrates the spatial relationships between two specific training datasets and a sampled subset of MPtrj data, as interpreted through the learned features. It is apparent that both datasets exhibit overlap with the MPtrj dataset; however, the Ti-Al-V dataset demonstrates a notably smaller degree of overlap with the foundation model data compared to the H_2/Cu dataset. Furthermore, the Ti-Al-V dataset occupies a substantially larger feature space than the H_2/Cu dataset. We observed that the Ti-Al-V MACE-MP-f4 model converged more rapidly than the MACE-MP-f5 model. This may indicate that as the representation of a given system diverges further from that captured by the foundation model, more flexibility is required to achieve an accurate fit. In our context, a greater number of parameters may need to remain unfrozen. We have shown that negative transfer did not occur even in this case. However, fully leveraging the strengths of the frozen transfer learning method relies on a key assumption: that the data of interest is well-represented at the lower-level features learned by the foundation model. In other words, that the data of interest is not too far from the foundation model training set.

Methods

Model hyperparameters

Hydrogen on Copper surface transfer models. The following hyperparameters were used for training on this dataset for all frozen transfer learning models based on the "small", "medium", and "large" MACE-MP foundation models: a learning rate of 0.01, a cutoff distance of 6 Å, and a batch size of 16. The force and energy weights were set to 100 and 1, respectively, for the first 1200 epochs, and 1000 and 100, respectively, for the last 300 epochs. The models trained for a total of 1500 epochs. The freeze parameter was set to 0, 3, 4, 5, or 6. We note that training for such a large number of epochs was not necessary, but here and for other models, it was done to ensure convergence and stability across the models trained on different subset sizes.

Hydrogen on Copper surface delta models. The delta learning models used for refining the "small", "medium" and "large" MACE-MP foundation models had the following hyperparameters: a correlation order of 2, a cutoff distance of 4 Å, a model size of 16 × 0e, and 2 interaction layers. Notably, increasing the cutoff distance to 6 Å has shown no improvement in the on the test set RMSE. The initial loss function had energy and force weightings of 1 and 100, respectively, with the energy weighting increasing to 1000 after 1200 epochs. The models were trained for a total of 1500 epochs.

Ti-Al-V transfer models. The "small" MACE-MP foundation model was fine-tuned for this task using a learning rate of 0.01, a cutoff distance of 6 Å, and a batch size of 32. The force and energy weightings were both set to 100, with stochastic weight averaging (SWA) starting at 1200 epochs, using SWA force and energy weightings of 10 and 1000, respectively. The freeze parameter was set to 3, 4, or 5. The maximum number of epochs was 2000, with SWA starting at 1500 epochs, for consistency across models trained on different subset sizes.

ACE surrogate model. The models were trained with ACEpotentials (https://github.com/wgst/ACEpotentials.jl)[38], version 0.8.2. The cutoff distance was set to 6 Å. A correlation order of 4 was used, with polynomial degrees of 18, 12, 10, and 8, respectively. The energy and force loss function weightings were set to 0.54 and 0.01, respectively.

Ti-Al-V dataset

The underlying Ti-Al-V data was obtained from DFT calculations with the plane wave package CASTEP(v24.1)[48]. On-the-fly ultrasoft pseudopotentials were generated for Al, V and Ti with respective valence electronic structures: $3s^2 3p^1$, $3s^2 3p^6 3d^2 4s^2$, and $3s^2 3p^6 3d^2 3s^2$. The PBE[28] level of theory was used to model exchange-correlation. The parameters for our DFT calculations are configured to ensure convergence to within sub-meV per atom, compared to an excessive basis set and k-point sampling. This convergence criterion was met by using a plane wave energy cutoff of 800 eV and by sampling the electronic Brillouin Zone with a Monkhorst-Pack grid spacing of 0.02 $Å^{-1}$.

To characterise potential local ordering in each crystalline phase (α (hcp, $P6_3$/mmc), β (bcc, Im-3m) and ω-Ti (hexagonal, P6/mmm)), the Non-Diagonal Supercell (NDSC) method[49,50] was extended as a data reduction tool to efficiently sample atomic species ordering. In this method, a series of NDSCs are generated for pure Ti in each crystalline symmetry, from which atomic species are randomly swapped from Ti to Al and V. The simulation cells are then volumetrically scaled and randomly deformed, with atomic positions being perturbed according to a normal distribution with a standard deviation of 0.10 Å.

To characterise the liquid phase of Ti-6Al-4V, we utilise the machine-learning accelerated ab-initio molecular dynamics (MLMD) feature in CASTEP by Stenczel et al.[51]. In MLMD, molecular dynamics simulations are performed using a combination of DFT and on-the-fly generated MLIPs to propagate the dynamics, whereby a decision making algorithm is used to switch between the DFT and MLIP calculator, which is constantly updated with DFT datapoints. The ultimate result of switching between DFT and the surrogate MLIP is that in a given simulation one may consider a much larger number of timesteps in a set amount of computation time. MLMD was performed on simulation cells containing 54 and 128 atoms, where the stoichiometry resembled Ti-6Al-4V as close as possible. For full details, the reader is again referred to ref. 42.

Simulation details

Molecular dynamics simulations (excluding NVT and NPT simulations) for H_2/Cu were run using NQCDynamics[52] package (https://github.com/NQCD/NQCDynamics.jl, version 0.14.0). NPT and NVT (Langevin MD) simulations, as well as NEBs, were evaluated using the Atomic Simulation Environment[53] (https://gitlab.com/ase/ase, version 3.23.0).

Initial structures for simulations included structures in which the hydrogen molecule is situated 7 Å above the copper surface. The initial vibrational and rotational states were established using the Einstein-Brillouin-Keller (EBK) method[54], implemented within NQCDynamics. Polar and azimuthal angles were chosen randomly. Initial positions and velocities of surface atoms were established by running Langevin molecular dynamics at set temperatures using adequate lattice constants, based on NPT simulations, as detailed in ref. 27. Sticking probabilities were evaluated using data from 10,000 H_2/Cu trajectories (at every collision energy, surface facet and rovibrational state). The maximum simulation time was set to 3 ps with a time step of 0.1 fs. However, the trajectories were stopped when the following conditions were met: the distance between adsorbed hydrogen atoms was larger than 2.25 Å (dissociation), or the distance between hydrogen molecule and surface exceeded 7.1 Å (scattering). Sticking probabilities (P_{sticking}) can be defined as $P_{\text{sticking}} = n_{\text{dissociation}}/n_{\text{all}}$, where $n_{\text{dissociation}}$ is the number of trajectories that ended with dissociation, and n_{all} is the number of all the trajectories.

Data availability

The H2 on Cu(111) dataset has previously been published and is publicly available. [Stark, W. G. et al. Machine learning interatomic potentials for reactive hydrogen dynamics at metal surfaces based on iterative refinement of reaction probabilities. J. Phys. Chem. C 127, 24168 (2023).] The Ti-Al-V dataset, alongside the baseline Ti-Al-V ACE model we compare against, is made available in the dedicated repository: https://zenodo.org/records/15114121.The MACE-freeze patch to the MACE software suite is available under URL https://github.com/7radians/mace-freeze/tree/mace-freeze.

Received: 20 February 2025; Accepted: 3 July 2025;
Published online: 18 July 2025

References

1. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. Conf. North Am. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol*. 4171–4186 (2019).
2. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. OpenAI (2018). Available at: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
3. Dosovitskiy, A. et al. An image is worth 16×16 words: Transformers for image recognition at scale. *Int. Conf. Learn. Represent*. *arXiv preprint arXiv:2010.11929* (2021).
4. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*. 770–778 (2016).
5. Huang, G., Liu, Z., Maaten, L. V. D. & Weinberger, K. Q. Densely connected convolutional networks. In *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2261–2269 (2017).
6. Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **31**, 3564–3572 (2018).
7. Gasteiger, J., Becker, F. & Günnemann, S. GemNet: Universal directional graph neural networks for molecules. *Adv. Neural Inf. Process. Syst*. **34**, 6790–6802 (2021).
8. Deng, B. et al. CHGNet: Pretrained universal neural network potential for charge-informed atomistic modeling. *Nat. Mach. Intell.* **5**, 1031–1041 (2023).
9. Batatia, I. et al. A foundation model for atomistic materials chemistry. Preprint at https://arxiv.org/abs/2401.00096v2 (2023).
10. Choudhary, K. & DeCost, B. Atomistic line graph neural network for improved materials property predictions. *npj Comput. Mater.* **7**, 185 (2021).
11. Merchant, A. et al. Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023).
12. Bochkarev, A., Lysogorskiy, Y. & Drautz, R. Graph atomic cluster expansion for semilocal interactions beyond equivariant message passing. *Phys. Rev. X* **14**, 021036 (2024).
13. Barroso-Luque, L. et al. Open materials 2024 (OMAT24) inorganic materials dataset and models. Preprint at https://arxiv.org/abs/2410.12771 (2024).
14. Yang, H. et al. MatterSim: A deep learning atomistic model across elements, temperatures and pressures. Preprint at https://arxiv.org/abs/2405.04967 (2024).
15. Neumann, M. et al. Orb: A fast, scalable neural network potential. Preprint at https://arxiv.org/abs/2410.22570 (2024).
16. Batatia, I., Kovacs, D. P., Simm, G., Ortner, C. & Csányi, G. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. *Adv. Neural Inf. Process. Syst*. **35** (2022).
17. Kaur, H. et al. Data-efficient fine-tuning of foundational models for first-principles quality sublimation enthalpies. *Faraday Discuss.* **256**, 120–138 (2025).
18. Kumar, A., Raghunathan, A., Jones, R. M., Ma, T. & Liang, P. Fine-tuning can distort pretrained features and underperform out-of-distribution. *Int. Conf. Learn. Represent*. *arXiv preprint arXiv:2202.10054* (2022).
19. Ramasesh, V. V., Dyer, E. & Raghu, M. Anatomy of catastrophic forgetting: Hidden representations and task semantics. Preprint at https://arxiv.org/abs/2007.07400 (2020).

20. Novelli, P. et al. Fine-tuning foundation models for molecular dynamics: A data-efficient approach with random features. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)* (2024).
21. Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B.* **99**, 014104 (2019).
22. Gao, Y. & Mosalam, K. M. Deep transfer learning for image-based structural damage recognition. *Comput.-Aided Civ. Infrastruct. Eng.* **33**, 748–768 (2018).
23. Zhou, Z. et al. Fine-tuning convolutional neural networks for biomedical image analysis: Actively and incrementally. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*. 4761–4772 (2017).
24. Li, H. et al. Parkinson's image detection and classification based on deep learning. *BMC Med. Imaging* **24**, 187 (2024).
25. *MACE-freeze* https://github.com/7radians/mace-freeze/tree/mace-freeze (2024).
26. Stark, W. G. et al. Machine learning interatomic potentials for reactive hydrogen dynamics at metal surfaces based on iterative refinement of reaction probabilities. *J. Phys. Chem. C.* **127**, 24168–24182 (2023).
27. Stark, W. G. et al. Benchmarking of machine learning interatomic potentials for reactive hydrogen dynamics at metal surfaces. *Mach. Learn.: Sci. Technol.* **5**, 030501 (2024).
28. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
29. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals.*Phys. Rev. B.* **47**, 558–561 (1993).
30. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
31. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50 (1996).
32. Smeets, E. W., Voss, J. & Kroes, G. J. Specific reaction parameter density functional based on the meta-generalized gradient approximation: application to $H_2 + Cu(111)$ and $H_2 + Ag(111)$. *J. Phys. Chem. A.* **123**, 5395–5406 (2019).
33. Blum, V. et al. Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **180**, 2175–2196 (2009).
34. Mondal, A., Wijzenbroek, M., Bonfanti, M., Díaz, C. & Kroes, G.-J. Thermal lattice expansion effect on reactive scattering of H_2 from Cu(111) at T_s=925 K. *J. Phys. Chem. A.* **117**, 8770–8781 (2013).
35. Díaz, C. et al. Chemically accurate simulation of a prototypical surface reaction: H_2 dissociation on Cu(111). *Sci.* **326**, 832–834 (2009).
36. Nattino, F., Díaz, C., Jackson, B. & Kroes, G.-J. Effect of surface motion on the rotational quadrupole alignment parameter of D_2 reacting on Cu(111). *Phys. Rev. Lett.* **108**, 236104 (2012).
37. Kaufmann, S., Shuai, Q., Auerbach, D. J., Schwarzer, D. & Wodtke, A. M. Associative desorption of hydrogen isotopologues from copper surfaces: characterization of two reaction mechanisms. *J. Chem. Phys.* **148**, 194703 (2018).
38. Witt, W. C. et al. ACEpotentials.jl: A Julia implementation of the atomic cluster expansion. *J. Chem. Phys.* **159**, 164101 (2023).
39. Kovács, D. P. et al. Linear atomic cluster expansion force fields for organic molecules: Beyond RMSE. *J. Chem. Theory Comput.* **17**, 7696–7711 (2021).
40. Morrow, J. D. & Deringer, V. L. Indirect learning and physically guided validation of interatomic potential models. *J. Chem. Phys.* **157**, 104105 (2022).
41. Grigorev, P. et al. matscipy: materials science at the atomic scale with Python. *J. Open Source Softw.* **9**, 5668 (2024).
42. Allen, C. S. & Bartók, A. P. Multi-phase dataset for Ti and Ti-6Al-4V. Preprint at https://arxiv.org/abs/2501.06116 (2025).
43. Nigam, J., Pozdnyakov, S., Fraux, G. & Ceriotti, M. Unified theory of atom-centered representations and message-passing machine-learning schemes. *J. Chem. Phys.* **156**, 204115 (2022).
44. Batatia, I. et al. The design space of E(3)-equivariant atom-centred interatomic potentials. *Nature Mach. Intell.* **7**, 56–67 (2025).
45. Gardner, J. L. A., Baker, K. T. & Deringer, V. L. Synthetic pre-training for neural-network interatomic potentials. *Mach. Learn.: Sci. Technol.* **5**, 015003 (2024).
46. Zhang, W., Deng, L., Zhang, L. & Wu, D. A survey on negative transfer. *IEEE CAA J. Autom. Sin.* **10**, 305–329 (2023).
47. Hu, W. et al. Strategies for pre-training graph neural networks. Preprint at https://arxiv.org/abs/1905.12265 (2020).
48. Clark, S. J. et al. First principles methods using CASTEP. *Z. Kristallogr. Cryst. Mater.* **220**, 567–570 (2005).
49. Lloyd-Williams, J. H. & Monserrat, B. Lattice dynamics and electron-phonon coupling calculations using nondiagonal supercells. *Phys. Rev. B.* **92**, 184301 (2015).
50. Allen, C. & Bartók, A. P. Optimal data generation for machine learned interatomic potentials. *Mach. Learn.: Sci. Technol* **3**, 045031 (2022).
51. Stenczel, T. K. et al. Machine-learned acceleration for molecular dynamics in CASTEP. *J. Chem. Phys.* **159**, 044803 (2023).
52. Gardner, J. et al. NQCDynamics.jl: A Julia package for nonadiabatic quantum classical molecular dynamics in the condensed phase. *J. Chem. Phys.* **156**, 174801 (2022).
53. Larsen, A. H. et al. The atomic simulation environment – a Python library for working with atoms. *J. Phys.: Condens. Matter* **29**, 273002 (2017).
54. Larkoski, A. J., Ellis, D. G. & Curtis, L. J. Numerical implementation of Einstein-Brillouin-Keller quantization for arbitrary potentials. *Am. J. Phys.* **74**, 7 (2006).

Acknowledgements

The authors acknowledge financial support from the UKRI Future Leaders Fellowship program (MR/S016023/1, MR/X023109/1), a UKRI frontier research grant (ERC StG, EP/X014088/1), the CASTEP-USER grant funded by UK Research and Innovation (EP/W030438/1) and the EPSRC-funded Centre for Doctoral Training in Modelling of Heterogeneous Systems (HetSys CDT, EP/S022848/1). High-performance computing resources were provided via the Scientific Computing Research Technology Platform of the University of Warwick.

Author contributions

R.J.M. and A.B.P. designed the research and supervised the work. M.R. wrote the software patch. M.R., W.J.S. and C.S.A. carried out the research: M.R. trained and evaluated the MACE models, WJS carried out the simulations of the Cu-H system, and CSA carried or the Ti-Al-V benchmark calculations. M.R. prepared figures 1, 2 and 6–8. W.J.S. prepared figures 3–5. All authors contributed to the manuscript text and all authors reviewed the manuscript.

Competing interests

R.J.M. is an associate editor for npj Computational Materials. All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-025-01727-x.

Correspondence and requests for materials should be addressed to Reinhard J. Maurer or Albert P. Bartók.

Reprints and permissions information is available at http://www.nature.com/reprints