MULTI-FIDELITY TRANSFER LEARNING FOR QUANTUM CHEMICAL DATA USING A ROBUST DENSITY FUNCTIONAL TIGHT BINDING BASELINE

Mengnan Cui[1,2], **Karsten Reuter**[1], **Johannes T. Margraf**[1,2]
[1]Fritz Haber Institute of the Max Planck Society,
Berlin, Germany
[2]University of Bayreuth,
Bavarian Center for Battery Technology (BayBatt),
Bayreuth, Germany
johannes.margraf@uni-bayreuth.de

ABSTRACT

Machine learning has revolutionized the development of interatomic potentials over the past decade, offering unparalleled computational speed without compromising accuracy. However, the performance of these models is highly dependent on the quality and amount of training data. Consequently, the current scarcity of high-fidelity datasets (i.e. beyond semilocal density functional theory) represents a significant challenge for further improvement. To address this, this study investigates the performance of transfer learning (TL) across multiple fidelities for both molecules and materials. Crucially, we disentangle the effects of multiple fidelities and different configuration/chemical spaces for pre-training and fine-tuning, in order to gain a deeper understanding of TL for chemical applications. This reveals that negative transfer, driven by noise from low-fidelity methods such as a Density Functional Tight Binding (DFTB) baseline, can significantly impact fine-tuned models. Despite this, the multi-fidelity approach demonstrates superior performance compared to single-fidelity learning. Interestingly, it even outperforms TL based on foundation models in some cases, by leveraging an optimal overlap of pre-training and fine-tuning chemical spaces.

1 Introduction

Spurred by the high computational costs of first-principles electronic structure methods, the development of machine learning (ML) interatomic potentials has enabled accurate atomistic simulations for previously inaccessible systems.[1, 2, 3, 4] Early efforts are exemplified by the pioneering works of Behler and Parrinello[5], as well as the Gaussian Approximation Potentials of Csányi and co-workers[6]. These involve the construction of rotationally invariant representations of atomic environments, combined with shallow neural networks or kernel regression methods. Subsequently, graph neural networks were developed, which expand local environment representations via message passing, such as in the SchNet[7], PhysNet[8], DimeNet[9], HDNN[10], and GemNet[11] models. Most recently, equivariant networks such as NequIP[12], PaiNN[13], SpookyNet[14], NewtonNet[13], and MACE[15] have emerged that currently represent the state-of-the-art in atomistic ML.

These methodological developments have led to remarkable improvements in accuracy, but also increased the computational cost of training and inference.[15, 16, 17] Alternatively, instead of increasing model capacity by using deeper or more complex architectures, accuracy can usually also be improved by increasing the amount of training data[18]. For a given level of accuracy, the lower computational cost at inference time can thus be offset by increased computational cost for training data generation. Unfortunately, this can itself be prohibitively expensive, e.g., when highly accurate reference methods such as Coupled Cluster (CC) theory or Quantum Monte Carlo (QMC) are used. In such contexts, transfer learning (TL) is commonly used,[19, 20, 21] meaning that a model that is pre-trained on one large dataset (for example a general and/or low-fidelity one) is fine-tuned on another (for example a specialized

https://doi.org/10.26434/chemrxiv-2024-9734b **ORCID:** https://orcid.org/0000-0002-0862-5289 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

or high-fidelity one). In the best case, beneficial features of the pre-trained model can be maintained throughout the fine-tuning, leading to more accurate and robust models for a given training set size.

TL is an appealing idea and widely used in chemistry.[22, 23] It has for example been implemented by Hutchinson et al. to increase the accuracy of experimental band gap predictions using comparatively cheaper Density Functional Theory (DFT) band gaps for pre-training (transfer from low- to high-fidelity).[24] Similarly, Frey et al. found that a MEGNet model, pre-trained on tens of thousands of 3D bulk crystals, could be fine-tuned to predict the properties of 2D materials in a highly data-efficient manner (transfer from one chemical space to another)[25, 26]. With the recent advent of broadly applicable foundation models for materials and molecules, TL is becoming even more relevant.[27, 28, 25] However, there are two common issues that need to be avoided in this context. On one hand, negative transfer can occur, meaning that features learned during pre-training can in some cases be detrimental to the task in the fine-tuning step. On the other hand, catastrophic forgetting can occur, meaning that the fine-tuning essentially overwrites all pre-trained information, rendering the pre-training step irrelevant.[19, 29, 23] Both of these are especially pertinent when overlap between the pre-training and fine-tuning datasets is insufficient. In chemical applications, this idea of dataset overlap relates both to the types of structures included in each set (i.e. how similar are bulk crystals and 2D materials) and the fidelity of the reference data (i.e. how good is the agreement between low and high fidelity labels).

When foundation models are fine-tuned, the main focus is on structural overlap. Indeed, models like MACE-MP-0[27] (for materials) and MACE-OFF23[30] (for molecules) extrapolate remarkably well to highly diverse systems (including liquids, amorphous systems, and higher temperatures and pressures) despite being trained exclusively on near-ground state structures of inorganic crystals and isolated molecules and clusters, respectively. Nevertheless, the further the configurations of the intended application are from those in the pre-training dataset, the more additional data will be required to obtain an accurate fine-tuned model. This explains the appeal of TL in a multi-fidelity setting.[31, 32] Here, additional data can be generated cheaply, for exactly the kinds of structures that are of interest for a given application (i.e. with perfect structural overlap between pre-training and fine-tuning sets). For this reason, multi-fidelity approaches have been widely used in chemical applications, both in TL and other settings, such as Δ-learning[33, 34, 35, 36], multi-task learning[37, 38, 39, 40, 41], or meta-learning[42]. In principle, there is thus a trade-off between structural overlap and fidelity overlap. In practice, both aspects are usually confounded however, since pre-training is often performed with pre-existing databases (e.g. the Materials Project or SPICE datasets[43, 44]) with a predefined chemical space and level of theory. Meanwhile, the chemical space and level of theory for fine-tuning are defined by the target application.

The goal of this paper is to systematically explore the impact of structural and fidelity overlap in TL. To this end, we take advantage of the recently reported Periodic Table Baseline Parameters (PTBP)[45] for Density Functional Tight Binding (DFTB) calculations, which enable us to generate data for molecular and materials datasets with low computational cost. With this, we generate customized low-fidelity datasets for arbitrary configuration spaces, with perfect structural overlap. The corresponding multi-fidelity TL (MFTL) models are compared with TL models based on pre-trained foundation models, which by definition feature a lower degree of structural overlap, but are trained with higher fidelity reference data.

2 Results and Discussions

Figure 1a illustrates the general MFTL workflow used herein. In the initial stage, DFTB is employed as an efficient method for generating training labels for a large sample from the target configuration space. The interatomic potential (e.g. MACE in this case) is pre-trained using this low-fidelity data. As discussed in the introduction, the key advantage of the resulting models (compared to existing foundational models) is that they are trained on data with perfect structural overlap for the target application. The pre-trained low-fidelity model is subsequently fine-tuned using a smaller but more accurate high-fidelity dataset. This fine-tuning step enhances the model's predictive performance. In principle, this fine-tuning process can be iteratively extended across multiple fidelities until the desired level of accuracy for the final target is achieved, as demonstrated further below.[46] This MFTL approach is in contrast to configuration space transfer learning as illustrated in Fig. 1b.

To demonstrate the benefit of MFTL, we first consider the QM7x dataset[47], which represents an extensive configuration space of small organic molecules in equilibrium and non-equilibrium configurations. Specifically, QM7x comprises approximately 4.2 million configurations based on the enumeration of small organic molecules containing up to 7 heavy atoms (i.e. C, N, O, S, Cl). The molecular sizes span 4-23 atoms in total. For each configuration, total energies and forces were calculated at the hybrid PBE0 level[48] with the many-body dispersion correction[49], hereafter referred to as PBE0+MBD. To evaluate the impact of training set size, we randomly sampled training and validation sets with sizes of 0.5k, 1k, 3k, 10k, and 50k configurations, respectively. An additional 50k configurations were reserved as an independent test set for final evaluations.

https://doi.org/10.26434/chemrxiv-2024-9734b **ORCID:** https://orcid.org/0000-0002-0862-5289 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

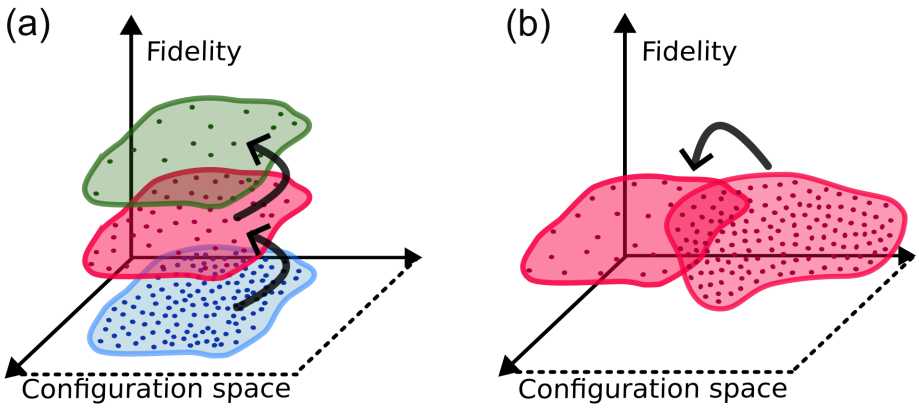

Figure 1: Schematic depiction of the MFTL approach, compared to TL in configuration space. In the former case (a), data is sampled with different levels of fidelity from the same region of configuration space. The low computational cost of lower fidelity methods allows more extensive sampling. In the latter case (b), the transfer occurs between one highly sampled region of configuration space to a less sampled region at the same (or a similar) fidelity.

The performance of the low-fidelity DFTB method (using the PTBP parameters) compared to the target PBE0+MBD method is shown in Fig. 2. For comparison, we also show the performance of the recent MACE-OFF23 foundation model. Perhaps surprisingly, the foundation model displays similar error statistics as the PTBP model (root mean squared errors (RMSEs) of 107.94 and 101.67 meV/atom for relative energies, respectively), despite it being trained on similar organic molecules, whereas PTBP is a simple DFTB model fitted on inorganic solids. These deviations can partially be attributed to the different levels of theory used for training MACE-OFF23 (ωB97M-D3) and for generating the QM7x data (PBE0+MBD). However, this does not explain the magnitude of the observed errors, since both methods are dispersion-corrected hybrid DFT functionals, which should perform similarly on this data. A more detailed investigation of the errors per molecule reveals that large deviations are exclusively observed for configurations with close interatomic contacts and/or broken covalent bonds. These occur in QM7x, because the non-equilibrium geometries are generated by normal mode sampling in rectilinear coordinates. In contrast, the SPICE set on which MACE-OFF23 is trained uses molecular dynamics (MD) to generate non-equilibrium structures, where close interatomic contacts or broken bonds are highly unlikely. As a consequence, the MACE-OFF23 RMSE is strongly impacted by a small number of outlier structures with unphysical bonding configurations. This becomes apparent when considering the histogram of force errors (Fig. 2c), which reveals that there is a lower density of errors in the intermediate range (around 1 eV/Å) for MACE-OFF23, but a tail of very large errors with low density. In contrast, the distribution of PTBP force errors lacks this tail, highlighting the robustness of this simple physics-based model. Representative configurations, for which MACE-OFF23 displays large errors are shown in Fig. 2d.

For initial MFTL tests on QM7x, new MACE models (using the MACE-OFF23 model architecture) were pre-trained on 10k and 50k DFTB(PTBP) datapoints, and subsequently fine-tuned on 0.5k PBE0+MBD datapoints. For robust statistics, each training was repeated three times with randomly initialized weights. In Fig. 3, the performance of these MFTL models on the PBE0+MBD test set is shown, as a function of the number of epochs used for pre-training. In all cases, errors initially decrease but quickly stagnate and even increase for longer training times. This trend is particularly clear for the larger pre-training set, and when training on forces. Overall, this figure shows that the MLTF concept is sound, since using more low-fidelity DFTB(PTBP) datapoints improves the performance for the high-fidelity test set, even though the size of the high-fidelity dataset is constant. On the other hand, this analysis also provides clear evidence of negative transfer, since training for more epochs increases the error on the high-fidelity data (even when it still decreases the error on the low-fidelity data, see Fig. S1).

These results may appear somewhat counter-intuitive at first glance, since MFTL appears to benefit from more data but not from more pre-training. However, they can be understood from the perspective of the widely used early-stopping approach for model regularization.[50] Neural networks tend to learn more general (and thus more transferable) concepts and features during early training epochs and more specific details in later epochs. In other words, fully converging the pre-training teaches the model irrelevant (and indeed detrimental) details about the PTBP potential energy surface, which cannot be corrected by the small fine-tuning dataset. Note that we use training epochs as a convenient measure for the length of the training of a given model here. However, this metric is not meaningful when comparing different training set sizes, since the number of weight updates per epoch increases with training set size.

https://doi.org/10.26434/chemrxiv-2024-9734b **ORCID:** https://orcid.org/0000-0002-0862-5289 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

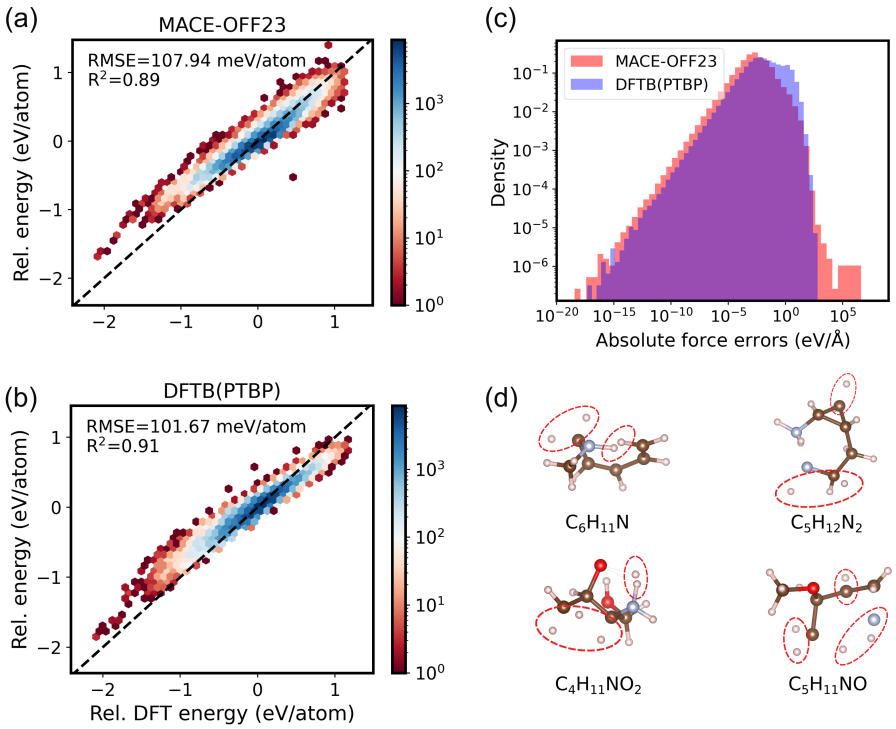

Figure 2: The performance of the MACE-OFF23 foundation model (a) and the semi-empirical DFTB(PTBP) model (b) for 150k randomly sampled configurations from the QM7x database. Histograms of force errors (c) and representative structures of MACE-OFF23 outliers (d).

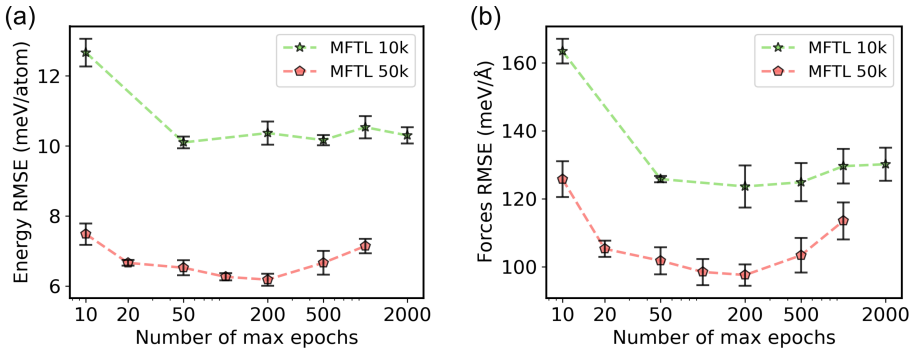

Figure 3: High-fidelity performance for energies (a) and forces (b) of MFTL models based on 10k and 50k low-fidelity training samples, respectively. The x-axis marks the maximum number of epochs allowed in pre-training, the y-axis shows the accuracy of the final fine-tuned model on the high-fidelity test set. All models are fine-tuned on 0.5k high-fidelity datapoints.

To investigate the influence of the size of the fine-tuning set, MFTL models trained on 50k DFTB(PTBP) samples over 200 epochs were fine-tuned on varying amounts of PBE0+MBD data (see Fig. 4). For comparison, we also trained single-fidelity models from scratch on the same data, as well as fine-tuning the MACE-OFF23 foundation model. We find that both TL approaches outperform the single-fidelity model, with the improvements being particularly pronounced for the smallest high-fidelity training set (0.5k configurations). Here, the MFTL model performs best with energy and force RMSEs of 6.1 meV/atom and 92.5 meV/Å, respectively, compared to 11.8 meV/atom and 195.0 meV/Å for single-fidelity learning.

https://doi.org/10.26434/chemrxiv-2024-9734b **ORCID:** https://orcid.org/0000-0002-0862-5289 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

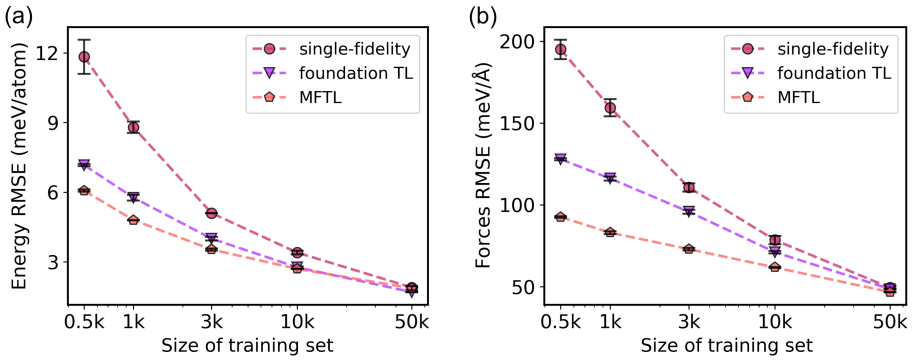

Figure 4: Learning curves for single-fidelity learning, foundation model TL, and MFTL on QM7x. The multi-fidelity models are pre-trained on 50k DFTB(PTBP) samples.

Overall, the differences between the MFTL and fine-tuned foundation models are small but still significant (RMSEs of 7.17 meV/atom and 127.87 meV/Å with 0.5k configurations). This shows the benefit of pre-training on configurations directly sampled from the target dataset. While MACE-OFF23 is certainly a better model than PTBP for describing organic molecules in general, the MLFT model benefits from a better description of short interatomic distances. These are important in QM7x but not included in the training of MACE-OFF23. Because of the efficiency of DFTB(PTBP) (and semi-empirical models in general), generating such custom pre-training data only leads to a small computational overhead relative to the generation of high-fidelity data.

Given the good performance of MLFT, it is also worth comparing with Δ-ML[33], which is perhaps the most straightforward multi-fidelity ML approach. Here, instead of targeting the full high-fidelity reference data, the difference between high- and low-fidelity targets is learned. This difference typically displays lower variance and is thus easier to learn than the full high-fidelity label. For QM7x, MLFT and Δ-ML display similar performance (see Fig. S2). However, Δ-ML has the downside that the low-fidelity method needs to be evaluated for each prediction. While semi-empirical methods are computationally efficient for small molecules, they display less favorable scaling with system size than atomistic ML models. This makes MLFT a more versatile approach than Δ-ML overall.

Although QM7x is a highly diverse dataset, small organic molecules are generally a manageable task for ML potentials. This is because of the highly systematic nature of organic chemistry, which can ultimately be reduced to a limited number of atomic environments (functional groups). In contrast, interatomic potentials for materials involving various surfaces, defects, and crystal structures can be more challenging. In particular, it was recently shown that transition metals display many-body interactions that are difficult to describe with interatomic potentials.[51] We therefore next investigate the performance of MFTL on a dataset containing 1.58k diverse configurations of elemental tungsten (i.e. vacancies, low-index surfaces, gamma-surfaces, and dislocation cores), previously reported in Ref[52]. Energies and forces for this dataset were computed with the PBE functional, which serves as the high-fidelity target.

Since the PTBP model was only fitted to simple crystals, its performance for the tungsten set is rather poor, with some large outliers (see Fig. 5 and Fig. S3), leading to energy and force RMSEs of 251.15 meV and 5.28 eV/Å, respectively. Nonetheless, it provides at least a qualitatively correct baseline in most cases. In contrast, the MACE-MP-0 foundation model performs much better. In terms of energies, non-equilibrium structures are systematically overestabilized, consistent with the previously reported mode-softening of MACE-MP-0 and other foundation models.[53] Nonetheless, the predicted energies and forces show excellent correlation with the reference DFT calculations and no significant outliers.

To develop MFTL models for this dataset, we isolated 1k structures each for validation and testing, and split the remaining dataset into training sets of 0.1k, 0.3k, 1k, 2k, and 7.6k configurations. As for the QM7x dataset, we observe negative transfer that can be mitigated by stopping the pre-training early. Indeed, we find that the best results are observed when stopping after just 6 epochs in this case (see Fig. S4). This is much earlier than for QM7x, likely due to the fact that the low-fidelity model is significantly noisier in this case. Learning curves for MFTL, single-fidelity models, and the fine-tuned foundation models can be found in Fig. 6. As above, we find that MFTL is highly beneficial for the smallest training set (100 configurations), with an almost four-fold improvement of the energy RMSE, compared to the single-fidelity model. For larger datasets, the performance of single- and multi-fidelity models is nearly indistinguishable, however.

https://doi.org/10.26434/chemrxiv-2024-9734b **ORCID:** https://orcid.org/0000-0002-0862-5289 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

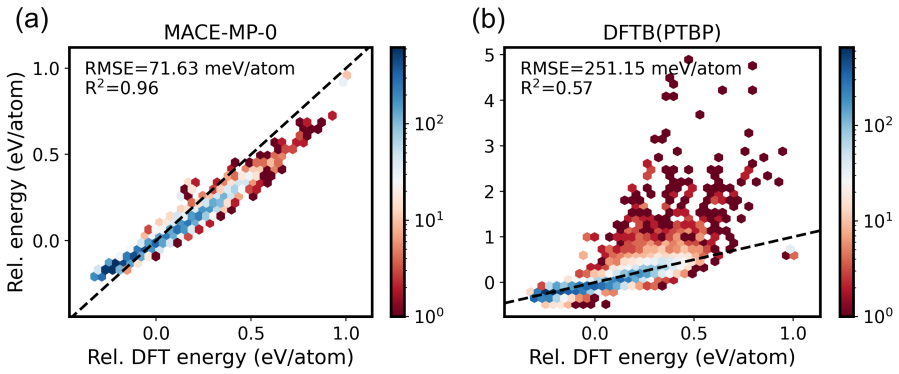

Figure 5: Performance of the foundation model MACE-MP-0 (a) and DFTB(PTBP) (b) for predicting relative energies on the tungsten dataset.

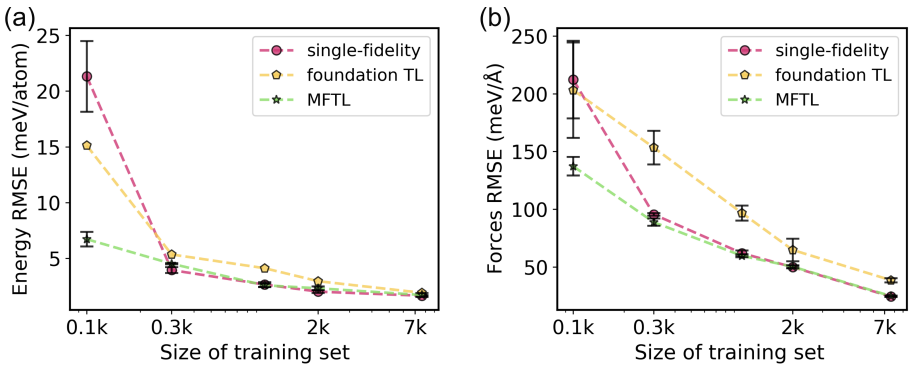

Figure 6: Learning curves for single-fidelity learning, foundation model TL, and MFTL on the tungsten dataset. The multi-fidelity models are pre-trained on 7.6k DFTB(PTBP) configurations.

Interestingly, the foundation model TL scheme is a much smaller improvement over single-fidelity learning for the smallest dataset. In fact, the fine-tuned foundation model performs somewhat worse than the single fidelity model for the larger training sets. This indicates that significant negative transfer is occurring here, despite the good performance of the foundation model as is. This can be attributed to the fact that the MPTraj dataset used to train MACE-MP-0 only contains very few pure tungsten configurations. Specifically, the Materials Project only contains eight different Tungsten samples, all of which are simple crystals. In contrast, the MFTL model is pre-trained on the full range of atomic environments included in the dataset.

It should be emphasized that all examples discussed up to this point use the simplest TL strategy of retraining all pre-trained weights on the new dataset. For MACE-MP-0, a multi-head TL approach was recently developed, which uses separate read-out heads for the pre-training and fine-tuning data. Additionally, this approach retains a subsample of the pre-training data during the fine-tuning step. This can mitigate both catastrophic forgetting and negative transfer. We also applied this multi-head strategy to the tungsten set, finding much improved results, almost en par with MLFT (see Fig. S5). This indicates that negative transfer is indeed the likely cause of the discrepancy between MLFT and foundation model fine-tuning. For comparison, Δ-ML models were also developed for this dataset. As shown in Fig. 7, exceptionally high errors (larger than for single-fidelity models) were observed for these models, however. This can be attributed to the high level of noise in the DFTB(PTBP) baseline data. With early stopping during the pre-training phase, MFTL is nevertheless highly robust, even under these circumstances.

So far, all MFTL examples we discussed used a single low fidelity level (DFTB) and a target high fidelity level (DFT). As demonstrated by von Lilienfeld and co-workers in the Δ-ML context, quantum chemical data is also well suited for developing models with more levels of fidelity, e.g. including highly accurate wavefunction methods like Coupled Cluster (CC) theory.[54] To explore this idea further, we used the DFTB(PTBP) pre-trained and PBE0+MBD fine-tuned models developed on the QM7x dataset (see above) as baselines for further TL on CCSD(T) targets from the MD22 dataset[55].

https://doi.org/10.26434/chemrxiv-2024-9734b **ORCID:** https://orcid.org/0000-0002-0862-5289 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

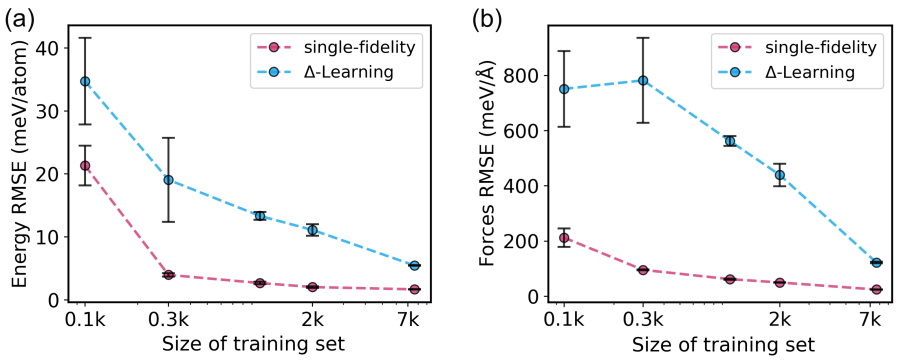

Figure 7: Learning curves for Δ-ML and single-fidelity learning on the tungsten dataset.

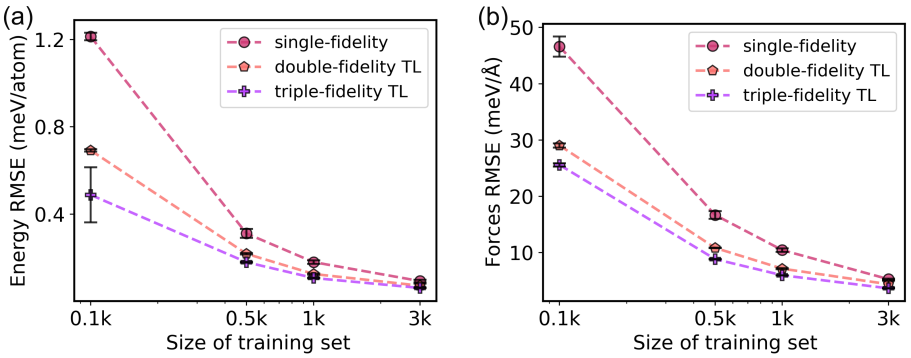

Figure 8: Learning curves for single-fidelity, double-fidelity, and triple-fidelity models on CCSD(T) data from the MD22 set. The multi-fidelity models use 50k DFTB(PTBP) and 10k DFT datapoints for pre-training, respectively.

Specifically, we used a set of 4500 non-equilibrium configurations of Benzene, Malonaldehyde, and Toluene (1500 structures for each molecule), for which CCSD(T)/cc-pVDZ energies and forces are available. From this combined dataset, we uniformly sampled training sets with 0.1k, 0.5k, 1k, and 3k configurations. Validation and test sets of 0.75k structures each were also generated. We then trained single-fidelity models (trained from scratch on CCSD(T) data), double-fidelity models (TL from DFT to CCSD(T)), and triple-fidelity models (TL from DFTB to DFT to CCSD(T)). Here, 50k DFTB and 10k DFT datapoints from the QM7x dataset were used as the pre-training samples. The corresponding results are shown in Fig. 8.

This shows that using multiple levels of fidelity is indeed beneficial, as the triple-fidelity model performs best across all training set sizes. Compared to the single-fidelity model, it achieves up to 59.75% and 45.09% improvement in energy and forces error, respectively, yielding RMSEs of 0.49 meV/atom and 25.59 meV/Å with only 100 CCSD(T) datapoints. Importantly, it even outperforms the double-fidelity model pre-trained on 10k DFT configurations. This confirms that the extensive amount of DFTB data contains information relevant to the CCSD(T) learning task, beyond what is provided by the DFT pretraining.

3 Conclusion

In this study, we have investigated the properties of MFTL models based on a robust DFTB(PTBP) baseline. By drawing low- and high-fidelity configurations from the same datasets, the effects of structural and fidelity overlap in quantum chemical TL could be disentangled. We find that noise in the low-fidelity labels can be detrimental in both TL and Δ-ML settings. Early-stopping of the pre-training proved to be an efficient way to mitigate this issue, however. With this approach, MFTL even outperforms the straightforward fine-tuning of high quality foundation models, both for molecular and materials datasets. This is somewhat surprising, since the foundation models are generally more accurate than the DFTB(PTBP) baseline. However, the lower structural overlap between pre-training and fine-tuning datasets causes some negative transfer in the fine-tuning of the foundation models. For the challenging

https://doi.org/10.26434/chemrxiv-2024-9734b **ORCID:** https://orcid.org/0000-0002-0862-5289 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

tungsten dataset, we found that this can be mitigated with a more sophisticated multi-head fine-tuning strategy for the foundation model.

More broadly, our results indicate that the current approach of training foundation models to achieve the highest possible accuracy on large single-fidelity databases is non-optimal from the perspective of fine-tuning for specific applications. In future work, multi-fidelity approaches and early-stopping should be investigated for foundation models as well. Inexpensive electronic structure models like DFTB(PTBP) or small basis DFT would allow a massive exploration of materials configuration space[45, 56]. This could increase the applicability and robustness of the next generation of foundation models.

Code and Data Availability: The dataset sets used in this work and training scripts can be found at `https://gitlab.com/mncui/ptbplus.git`.

Acknowledgements: The authors gratefully acknowledge the Max Planck Computing and Data Facility (MPCDF) for providing computing time.

https://doi.org/10.26434/chemrxiv-2024-9734b **ORCID:** https://orcid.org/0000-0002-0862-5289 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

References

[1] Y. Shi, Z. Yang, S. Ma, P. Kang, C. Shang, P. Hu, and Z. Liu. Machine learning for chemistry: Basics and applications. *Engnieering*, 27:70–83, 2023.

[2] N. Fedik, R. Zubatyuk, M. Kulichenko, N. Lubbers, J. Smith, B. Nebgen, R. Messerly, Y. Li, A. Boldyrev, K. Barros, O. Isayev, and S. Tretiak. Extending machine learning beyond interatomic potentials for predicting molecular properties. *Nat. Rev. Chem.*, 6(9):653–672, 2022.

[3] J. Margraf, H. Jung, C. Scheurer, and K. Reuter. Exploring catalytic reaction networks with machine learning. *J. Chem. Theory Comput.*, 6(2):112–121, 2023.

[4] J. Margraf. Science-driven atomistic machine learning. *Angew. Chem. Int. Ed.*, 62(26):e202219170, 2023.

[5] J. Behler and M. Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98(14):1–4, 2007.

[6] A. Bartók, M. Payne, R. Kondor, and G. Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.*, 104(13):136403, 2010.

[7] K. Schütt, P. Kindermans, H. Sauceda, S. Chmiela, A. Tkatchenko, and K. Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *Adv. Neural Inf. Process. Syst.*, 2017-Decem(1):992–1002, 2017.

[8] O. Unke and M. Meuwly. PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges. *J. Chem. Theory Comput.*, 15(6):3678–3693, 2019.

[9] J. Gasteiger, J. Groß, and S. Günnemann. Directional message passing for molecular graphs, 2022.

[10] K. Schütt, F. Arbabzadah, S. Chmiela, K. Müller, and A. Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.*, 8(1):13890, 2017.

[11] J. Gasteiger, F. Becker, and S. Günnemann. GemNet: Universal directional graph neural networks for molecules. *Adv. Neural Inf. Process. Syst.*, 34:6790–6802, 2021.

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

[13] M. Haghighatlari, J. Li, X. Guan, O. Zhang, A. Das, C. J.ăStein, F. Heidar-Zadeh, M. Liu, M. Head-Gordon, L. Bertels, H. Hao, I. Leven, and T. Head-Gordon. NewtonNet: a newtonian message passing network for deep learning of interatomic potentials and forces. *Digit. Discov.*, 1(3):333–343, 2022.

[14] O. Unke, S. Chmiela, M. Gastegger, K. Schütt, H. Sauceda, and K. Müller. SpookyNet: Learning force fields with electronic degrees of freedom and nonlocal effects. *Nat. Commun.*, 12(1):7273, 2021.

[15] I. Batatia, D. 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.*, 35:11423–11436, 2022.

[16] D. Zhang, H. Bi, F. Dai, W. Jiang, X. Liu, L. Zhang, and H. Wang. Pretraining of attention-based deep learning potential model for molecular simulation. *npj Comput. Mater.*, 10(1):1–8, 2024.

[17] Z. Zhouyin, Z. Gan, S. Pandey, L. Zhang, and Q. Gu. Learning local equivariant representations for quantum operators, 2024.

[18] S. Stocker, J. Gasteiger, F. Becker, S. Günnemann, and J. Margraf. How robust are modern graph neural network potentials in long and hot molecular dynamics simulations? *Mach. Learn.: Sci. Technol.*, 3(4):045010, 2022.

[19] S. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359, 2010.

[20] S. Käser, L. ItzaăVazquez-Salazar, M. Meuwly, and K. Töpfer. Neural network potentials for chemistry: concepts, applications and prospects. *Digit. Discov.*, 2(1):28–58, 2023.

[21] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proc. IEEE*, 109(1):43–76, 2021.

[22] P. Rowe, V. Deringer, P. Gasparotto, G. Csányi, and A. Michaelides. Erratum: An accurate and transferable machine learning potential for carbon (j. chem. phys. (2020) 153 (034702) DOI: 10.1063/5.0005084). *J. Chem. Phys.*, 156(15):2020–2022, 2022.

[23] C. Chen and S. Ong. AtomSets as a hierarchical transfer learning framework for small and large materials datasets. *npj Comput. Mater.*, 7(1):1–9, 2021.

https://doi.org/10.26434/chemrxiv-2024-9734b **ORCID:** https://orcid.org/0000-0002-0862-5289 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

[24] M. Hutchinson, E. Antono, B. Gibbons, S. Paradiso, J. Ling, and B. Meredig. Overcoming data scarcity with transfer learning, 2017.

[25] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.*, 31(9):3564–3572, 2019.

[26] N. Frey, D. Akinwande, D. Jariwala, and V. Shenoy. Machine learning-enabled design of point defects in 2d materials for quantum and neuromorphic information processing. *ACS Nano*, 14(10):13406–13417, 2020.

[27] I. Batatia, P. Benner, Y. Chiang, A. Elena, D. Kovács, J. Riebesell, X. Advincula, M. Asta, M. Avaylon, W. Baldwin, F. Berger, N. Bernstein, A. Bhowmik, S. Blau, V. Cărare, J. Darby, S. De, F. Della Pia, V. Deringer, R. Elijošius, Z. El-Machachi, F. Falcioni, E. Fako, A. Ferrari, A. Genreith-Schriever, J. George, R. Goodall, C. Grey, P. Grigorev, S. Han, W. Handley, H. Heenen, K. Hermansson, C. Holm, J. Jaafar, S. Hofmann, K. Jakob, H. Jung, V. Kapil, A. Kaplan, N. Karimitari, J. Kermode, N. Kroupa, J. Kullgren, M. Kuner, D. Kuryla, G. Liepuoniute, J. Margraf, I. Magdău, A. Michaelides, J. Moore, A. Naik, S. Niblett, S. Norwood, N. O' Neill, C. Ortner, K. Persson, K. Reuter, A. Rosen, L. Schaaf, C. Schran, B. Shi, E. Sivonxay, T. Stenczel, V. Svahn, C. Sutton, T. Swinburne, J. Tilly, C. van der Oord, E. Varga-Umbrich, T. Vegge, M. Vondrák, Y. Wang, W. Witt, F. Zills, and G. Csányi. A foundation model for atomistic materials chemistry, 2024.

[28] C. Devereux, J. Smith, K. Huddleston, K. Barros, R. Zubatyuk, O. Isayev, and A. Roitberg. Extending the applicability of the ANI deep learning molecular potential to sulfur and halogens. *J. Chem. Theory Comput.*, 16(7):4192–4202, 2020.

[29] T. Fitzgerald and A. Thomaz. Skill demonstration transfer for learning from demonstration. *Sci. Data*, pages 187–188, 2015.

[30] D. Kovács, J. Moore, N. Browning, I. Batatia, J. Horton, V. Kapil, W. Witt, I. Magdău, D. Cole, and G. Csányi. MACE-OFF23: Transferable machine learning force fields for organic molecules, 2023.

[31] R. Batra, G. Pilania, B. Uberuaga, and R. Ramprasad. Multifidelity information fusion with machine learning: A case study of dopant formation energies in hafnia. *ACS Appl. Mater. Interfaces*, 11(28):24906–24918, 2019.

[32] S. Goodlett, J. Turney, and H. Schaefer, I. Comparison of multifidelity machine learning models for potential energy surfaces. *Int. J. Chem. Phys.*, 159(4):044111, 2023.

[33] R. Ramakrishnan, P. Dral, M. Rupp, and O. von Lilienfeld. Big data meets quantum chemistry approximations: The delta-machine learning approach. *J. Chem. Theory Comput.*, 11(5):2087–2096, 2015.

[34] S. Wengert, G. Csányi, K. Reuter, and J. T.ăMargraf. Data-efficient machine learning for molecular crystal structure prediction. *Chem. Sci.*, 12(12):4536–4546, 2021.

[35] C. Staacke, H. Heenen, C. Scheurer, G. Csányi, K. Reuter, and J. Margraf. On the role of long-range electrostatics in machine-learned interatomic potentials for complex battery materials. *Chem. Sci.*, 4(11):12562–12569, 2021.

[36] S. Wengert, G. Csányi, K. Reuter, and J. Margraf. A hybrid machine learning approach for structure stability prediction in molecular co-crystal screenings. *J. Chem. Theory Comput.*, 18(7):4586–4593, 2022.

[37] A. Rosen, V. Fung, P. Huck, C. ODonnell, M. Horton, D. Truhlar, K. Persson, J. Notestein, and R. Snurr. High-throughput predictions of metal–organic framework electronic properties: theoretical challenges, graph neural networks, and data exploration. *npj Comput. Mater.*, 8(1):1–10, 2022.

[38] C. Fare, P. Fenner, M. Benatan, A. Varsi, and E. Pyzer-Knapp. A multi-fidelity machine learning approach to high throughput materials screening. *npj Comput. Mater.*, 8(1):1–9, 2022.

[39] D. Buterez, J. Janet, S. Kiddle, D. Oglic, and P. Lió. Transfer learning with graph neural networks for improved molecular property prediction in the multi-fidelity setting. *Nat. Commun.*, 15(1):1517, 2024.

[40] R. Zubatyuk, J. Smith, J. Leszczynski, and O. Isayev. Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Sci. Adv.*, 5(8):eaav6490, 2019.

[41] K. Chen, C. Kunkel, B. Cheng, K. Reuter, and J. T.ăMargraf. Physics-inspired machine learning of localized intensive properties. *Chem. Sci.*, 14(18):4913–4922, 2023.

[42] A. Allen, N. Lubbers, S. Matin, J. Smith, R. Messerly, S. Tretiak, and K. Barros. Learning together: Towards foundational models for machine learning interatomic potentials with meta-learning, 2023.

[43] A. Jain, S. Ong, G. Hautier, W. Chen, W. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.*, 1(1):011002, 2013.

[44] P. Eastman, P. Behara, D. Dotson, R. Galvelis, J. Herr, J. Horton, Y. Mao, J. Chodera, B. Pritchard, Y. Wang, G. De Fabritiis, and T. Markland. SPICE, a dataset of drug-like molecules and peptides for training machine learning potentials. *Sci. Data*, 10(1):11, 2023.

https://doi.org/10.26434/chemrxiv-2024-9734b **ORCID:** https://orcid.org/0000-0002-0862-5289 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

[45] M. Cui, K. Reuter, and J. Margraf. Obtaining robust density functional tight-binding parameters for solids across the periodic table. *J. Chem. Theory Comput.*, 20(12):5276–5290, 2024.

[46] P. Zaspel, B. Huang, H. Harbrecht, and O. von Lilienfeld. Boosting quantum machine learning models with a multilevel combination technique: Pople diagrams revisited. *J. Chem. Theory Comput.*, 15(3):1546–1559, 2019.

[47] J. Hoja, L. Medrano Sandonas, B. Ernst, A. Vazquez-Mayagoitia, R. DiStasio Jr., and A. Tkatchenko. QM7-x, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. *Sci. Data*, 8(1):43, 2021.

[48] C. Adamo and V. Barone. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *Int. J. Chem. Phys.*, 110(13):6158–6170, 1999.

[49] A. Tkatchenko, R. DiStasio, R. Car, and M. Scheffler. Accurate and efficient method for many-body van der waals interactions. *Phys. Rev. Lett.*, 108(23):236402, 2012.

[50] R. Caruana, S. Lawrence, and C. Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

[51] C. Owen, S. Torrisi, Y. Xie, S. Batzner, K. Bystrom, J. Coulter, A. Musaelian, L. Sun, and B. Kozinsky. Complexity of many-body interactions in transition metals via machine-learned force fields from the TM23 data set. *npj Comput. Mater.*, 10(1):1–16, 2024.

[52] W. Szlachta, A. Bartók, and G. Csányi. Accuracy and transferability of gaussian approximation potential models for tungsten. *Phys. Rev. B*, 90(10):104108, 2014.

[53] B. Deng, Y. Choi, P. Zhong, J. Riebesell, S. Anand, Z. Li, K. Jun, K. Persson, and G. Ceder. Overcoming systematic softening in universal machine learning interatomic potentials by fine-tuning, 2024.

[54] Peter Zaspel, Bing Huang, Helmut Harbrecht, and O. Anatole von Lilienfeld. Boosting quantum machine learning models with a multilevel combination technique: Pople diagrams revisited. *J. Chem. Theo. Comput.*, 15(3):1546–1559, 2019. PMID: 30516999.

[55] S. Chmiela, V. Vassilev-Galindo, O. Unke, A. Kabylda, H. Sauceda, A. Tkatchenko, and K. Müller. Accurate global machine learning force fields for molecules with hundreds of atoms, 2022.

[56] E. Keller, J. Morgenstein, K. Reuter, and J. Margraf. Small basis set density functional theory method for cost-efficient, large-scale condensed matter simulations. *Int. J. Chem. Phys.*, 161(7):074104, 2024.

https://doi.org/10.26434/chemrxiv-2024-9734b **ORCID:** https://orcid.org/0000-0002-0862-5289 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0