Perovs-Dopants: Machine Learning Potentials for Doped Bulk Structures

Xiaoxiao Wang
Intel Labs
xiaoxiao.wang@intel.com

Suehyun Park
Intel Corporation
suehyun.park@intel.com

Santiago Miret
Intel Labs
santiago.miret@intel.com

Abstract

Exploring new dopant materials is crucial for enhancing the performance, efficiency, and versatility of semiconductors. Perovskites, with their diverse structures and tunability, have emerged as promising candidates for the next generation of semiconductors. Machine learning potentials (MLPs) have shown great promise in efficiently predicting material properties for bulk materials. However, the lack of comprehensive dopant datasets for perovskites has hindered the application of data driven techniques for high-throughput screening and material discovery in this domain. In this work, we propose a dopant dataset "Perovs-Dopants" comprising over 20,000 density functional theory (DFT) data points from 438 different doped perovskite material relaxation trajectories. Using Perovs-Dopants, we evaluate MACE-MP, a foundation model pretrained on bulk material trajectories, to benchmark the performance of state-of-the-art MLPs. Our results show that despite MACE-MP's robust performance on bulk crystals, Perovs-Dopants represents an out-of-distribution challenge with significant prediction errors. We redeem these errors by finetuning MACE-MP to achieve comparative modeling of Perovs-Dopants and pristine bulk crystals.

1 Introduction

The development of new doped materials is pivotal for advancing semiconductor technologies. The introduction of dopants into semiconductor materials can tailor their electronic and optical properties, enabling the design of materials with specific functionalities for complex logic devices. Among the various classes of semiconductors, perovskites have emerged as promising candidates due to their versatile crystal structures and rich chemical compositions [1–3]. Perovskites are the class of material with a general chemical formula of $\mathrm{ABX_3}$ which share the same crystal structure as $\mathrm{CaTiO_3}$. These materials exhibit remarkable properties, such as high carrier mobility, tunable band gaps, and strong light absorption, making them ideal for a wide range of applications, including photovoltaics, light-emitting diodes, and sensors [4–6].

In recent years, there has been a surge of research focusing on the use of machine learning (ML) to accelerate the discovery of new materials, chemicals, and drugs [7–12]. The scientific community has made significant efforts to construct computational datasets that facilitate the development of machine learning potential (MLP) for diverse types of materials to broader and more efficient materials modeling [13–21]. Prior work also proposed MLP foundation models, such as MACE [22], CHGNet [23], and M3GNet [24], that claim fast and accurate predictions of material properties for a wide range of chemical systems and applications. Despite the growing interest in perovskites, the exploration of

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

ML models for dopant-infused perovskites remains underdeveloped. While the perovskite family offers a vast compositional space, the lack of dopant datasets poses a significant challenge to the systematic design and optimization of doped perovskite materials. Existing computational datasets mainly focus on pristine perovskite structures [25, 26], leaving a gap for doped Perovskite systems.

To address this gap, we propose a dopant dataset specifically tailored for perovskites, "Perovs-Dopants", containing relaxation trajectories of various dopant materials within oxide and halide perovskites. This dataset is composed of over 20,000 data points on atomic configuration and Density Functional Theory (DFT) calculated energy and forces. While we have already collected a significant amount of DFT data, we note that we are actively working on populating the chemical space with additional perovskite and dopant combinations with a greater set of data available at the time of the workshop. Nevertheless, this dataset serves as a benchmark for evaluating the performance of universal ML potentials in predicting relevant properties of doped perovskite systems, and can subsequently be used to facilitate the identification of promising dopant candidates.

2 Method

2.1 Perovs-Dopants

To construct a broad dopant dataset, we begin by selecting a diverse set of perovskite materials and dopant elements to ensure extensive coverage of different chemical environments and structural variations. The base perovskite materials were selected from the Materials Project database [27], a cubic perovskite dataset [28], and a halide perovskite dataset [29]. Based on the structures contained in these datasets, we selected the perovskites that are stable, and have a band gap ranging from 1 to 3 eV. For the dopants, we focused on the d-block transition metals from group 3 to 12 of the periodic table based on prior work related to doped semiconductor materials. Concretely, these elements were chosen due to their diverse electronic configurations and their potential to introduce significant property modifications to the host perovskite material. The doped perovskite structures were then generated by substituting one atom in either the A or B site of the perovskite base structure with the selected dopant. Additionally, we included scenarios where vacancies were introduced at the A or B site to capture both substitutional and vacancy doping effects.

The workflow for constructing the dopant dataset was developed using Atomate2 [30]. CP2K was employed as the DFT code for all calculations [31], and the Perdew-Burke-Ernzerhof (PBE) was applied as the exchange-correlation functional [32]. The calculations were conducted using the Orbital Transformations method from the Quickstep code in CP2K [33]. TZVP basis set and GTH Pseudopotential was used. Geometric optimization simulations were performed to generate the relaxation trajectories for the doped structures. During these simulations, the atomic positions were optimized with Broyden–Fletcher–Goldfarb–Shanno algorithm (BFGS) to minimize the forces acting on the atoms to ensure the structure reached a stable state. The relaxation process was iterated until the forces on all atoms were reduced to below 0.02 eV/Å.

2.2 Machine Learning Model Analysis

For this study, we analyzed the MACE-MP model, which was pretrained on the MPtrj dataset that contains 1.5M atomic configuration and DFT calculated properties [23, 34]. For testing the model performance on the Perovs-Dopants dataset, we split it into an 8:1:1 ratio for training, validation and testing. The training set also included 89 isolated atoms to adjust the atomic energies when training the MACE models and account for the variation in the atomic energy between CP2K and VASP as DFT codes. The pretrained model serves as a starting point for testing on the Perovs-Dopants dataset. Given the differences between the Perovs-Dopants dataset and the majority of chemical systems in the MPtrj dataset, we expect some level of fine-tuning to be necessary to adapt the pretrained model to the new data. We fine-tuned the pretrained MACE-MP-0 model [1] using two different learning rates: 0.0005 and 0.00001 [2]. The fine-tuning process was run for 20 epochs. Additionally, we randomly selected 2000 DFT data points from the MPtrj test set to create a quick evaluation dataset to measure

[1]The pretrained model (2024-01-07-mace-128-L2_epoch-199.model) was downloaded from `https://github.com/ACEsuit/mace-mp/releases/tag/mace_mp_0`

[2]the MACE-MP-0 model was pretrained with a learning rate of 0.005

the finetuned model's performance on its original dataset. Additionally, we train the MACE model from scratch on Perovs-Dopants and analyze its performance.

The pretrained (MACE-MP), finetuned MACE-MP (MACE-MP-ft), and trained from scratch MACE (MACE-init) were evaluated on both the MPtrj subset and the Perovs-Dopants test set, and the experiments are referred to as *MACE-MP MPtrj*, *MACE-MP dopant*, *MACE-MP-ft MPtrj*, *MACE-MP-ft dopant*, *MACE-init MPtrj* and *MACE-init dopant*, respectively.

3 Results

The dopant dataset contains 438 doped perovskite materials, and the element distribution is shown in Figure 1a. We performed a t-distributed stochastic neighbor embedding (t-SNE) [35] analysis to help qualitatively analyze the difference in the chemical space coverage between the MPtrj dataset and the dopant dataset from the model's perspective. The node feature for 10000 randomly selected systems from MPtrj training dataset and the entire Perovs-Dopants test set were extracted from the pretrained MACE-MP model. These 256-dimensional vector features represent the atomic neighborhood of each atom in a chemical system. We averaged the per-atom vectors within each system to obtain a system-level descriptor. t-SNE was then applied to reduce the dimensionality and visualize the distribution of the systems. As shown in Figure 1b, while there is some overlap suggesting shared features between the datasets, a significant portion of the dopant dataset lies in the areas that are not covered by MPtrj. This observation confirms that the Perovs-Dopants dataset explores new chemical spaces, emphasizing the need for fine-tuning the pretrained MACE model to better adapt to these out-of-distribution data points.

Figure 2 summarizes the results of the proposed experiments. The pretrained MACE-MP model performance on the MPtrj subset aligns with the previously published results with an energy mean absolute error (MAE) of 0.02 eV/atom and force MAE of 0.04 eV/Å. When the pretrained model was directly applied to the Perovs-Dopants dataset, we observed a decline in performance as the model tends to overestimate the atomic forces. This result is expected and consistent with our earlier analysis: as indicated by Figure 1b, the Perovs-Dopants dataset represents an out-of-domain challenge.

The comparison between the two learning rate in the fine-tuning process can be found in Figure A1 and Table A1. The fine-tuning result from the lower learning rate (0.00001) is included in Figure 2. Finetuning on the Perovs-Dopants dataset significantly improved the model performance and showed a better alignment with the behavior of the foundation model. However, when we tested the fine-tuned model on the original MPtrj dataset, the force MAE increased to 0.14 eV/Å, indicating the occurrence of catastrophic forgetting where the finetuned model loses learnt information from its original training domain. The higher learning rate results in larger energy and forces errors. More investigation into robust fine-tuning strategies to preserve generalization across different dataset is necessary.

The MACE model trained from scratch on the Perovs-Dopants dataset performed comparably well. However, it comes at the cost of not leveraging the advantages of pretraining on other chemical interactions. Most likely this model will fail to accurately predict new perovskite dopant materials if they fall outside of the scope of the current training data.

4 Conclusion

In this study, we demonstrate our work towards the development of a dopant dataset for perovskite materials. The current benchmark dataset consists of over 20000 DFT data points from the relaxation trajectories of 438 doped perovskite systems. We demonstrated the potential of ML models, specifically the MACE Universal model, to predict the properties of these doped systems efficiently. The t-SNE analysis with the MACE-MP embeddings illustrates that the perovskite dopant materials are outside of the distribution of its training set. This emphasizes the importance to build new computational datasets for advancing the development of foundation models and accelerating material discovery. Our results highlight that while the pretrained MACE-MP model shows extraordinary performance on its original MPtrj dataset, it struggles with the dopants in perovskites. Finetuning the model improved its accuracy on the Perovs-Dopants dataset, but also resulted in catastrophic forgetting on the MPtrj dataset.

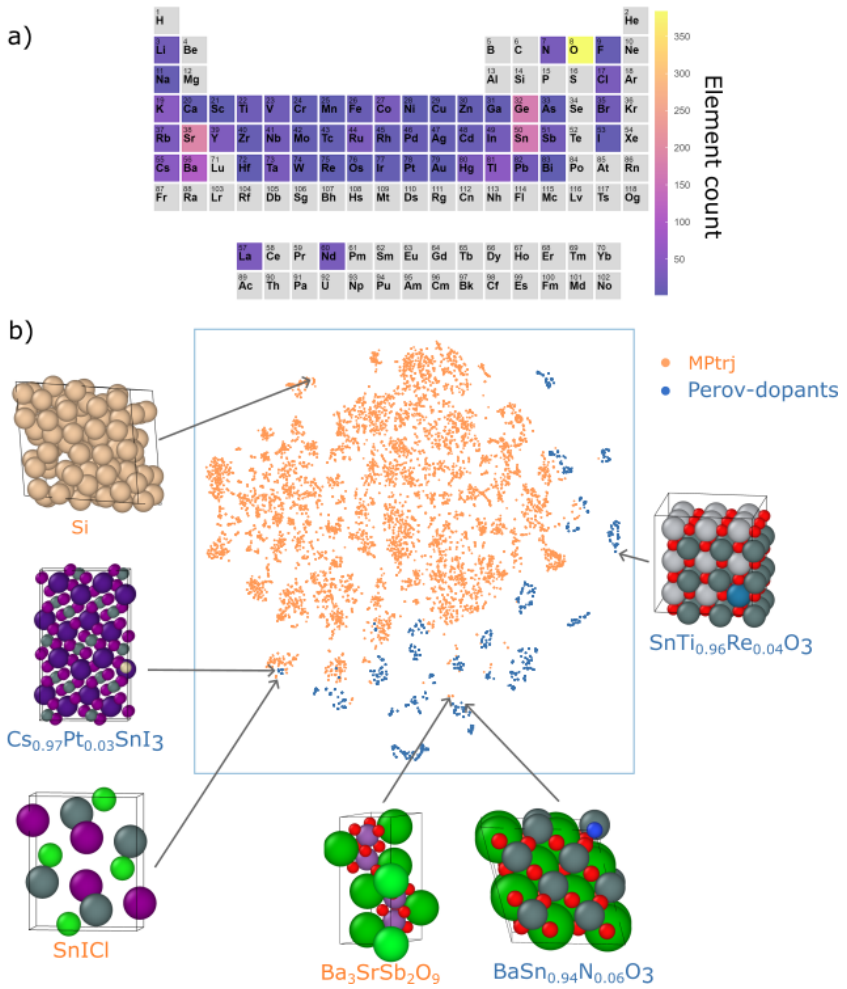
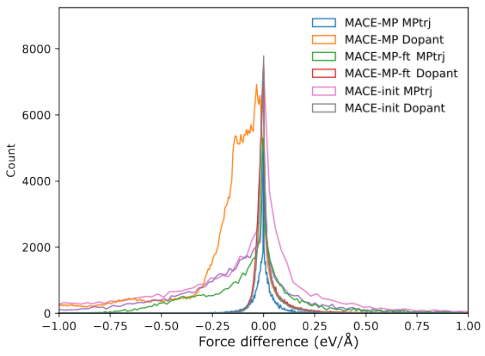


Figure 1: (a) Distribution of elements in the dopant dataset. (b) t-SNE plot comparing the chemical space covered by the MPtrj and Perovs-Dopants datasets.



(a) Distribution of the difference in atomic forces between DFT and MACE. $\mathrm{DFT_F}$ and $\mathrm{MACE_F}$ stand for atomic forces calculated from DFT and MACE model respectively.

Model	Test set	Energy MAE (eV/atom)	Force MAE (eV/Å)
MACE-MP	MPtrj	0.02	0.04
	Dopant	0.18	0.12
MACE-MP-ft	MPtrj	0.16	0.14
	Dopant	0.02	0.03
MACE-init	MPtrj	0.59	0.44
	Dopant	0.04	0.03

(b) Summary of MAE for energy and force prediction across different models.

Figure 2: MACE performance on Perovs-Dopants

This study aims for providing a valuable perovskite dopant dataset for the material science community to fill a critical gap in the field of semiconductor research. Future efforts will focus on expanding the dataset to cover more chemical spaces, and exploring other pretrained models', i.e.: ChGNet and M3GNet, performance on the Perovs-Dopants dataset with the Open MatSci ML Toolkit by Miret et al. [21]. Moving forward, there is a need to improve fine-tuning strategies to maintain the generalization ability of MLP foundation models. We plan to integrate multi-head architectures to address the heterogeneity of DFT settings across different datasets.

References

[1] Xu Zhang, Xiaodong Ren, Bin Liu, Rahim Munir, Xuejie Zhu, Dong Yang, Jianbo Li, Yucheng Liu, Detlef M. Smilgies, Ruipeng Li, Zhou Yang, Tianqi Niu, Xiuli Wang, Aram Amassian, Kui Zhao, and Shengzhong Liu. Stable high efficiency two-dimensional perovskite solar cells via cesium doping. *Energy and Environmental Science*, 10:2095–2102, 10 2017. ISSN 17545706. doi: 10.1039/c7ee01145h.

[2] Martin A. Green, Anita Ho-Baillie, and Henry J. Snaith. The emergence of perovskite solar cells, 2014. ISSN 17494893.

[3] Jeffrey W. Fergus. Perovskite oxides for semiconductor-based gas sensors, 5 2007. ISSN 09254005.

[4] Ahmed L. Abdelhady, Makhsud I. Saidaminov, Banavoth Murali, Valerio Adinolfi, Oleksandr Voznyy, Khabiboulakh Katsiev, Erkki Alarousu, Riccardo Comin, Ibrahim Dursun, Lutfan Sinatra, Edward H. Sargent, Omar F. Mohammed, and Osman M. Bakr. Heterovalent dopant incorporation for bandgap and type engineering of perovskite crystals. *Journal of Physical Chemistry Letters*, 7:295–301, 1 2016. ISSN 19487185. doi: 10.1021/acs.jpclett.5b02681.

[5] Silvia Colella, Edoardo Mosconi, Paolo Fedeli, Andrea Listorti, Francesco Gazza, Fabio Orlandi, Patrizia Ferro, Tullo Besagni, Aurora Rizzo, Gianluca Calestani, Giuseppe Gigli, Filippo De Angelis, and Roberto Mosca. Mapbi3-xclx mixed halide perovskite for hybrid solar cells: The role of chloride as dopant on the transport and structural properties. *Chemistry of Materials*, 25: 4613–4618, 11 2013. ISSN 0897-4756. doi: 10.1021/cm402919x.

[6] Gencai Pan, Xue Bai, Dongwen Yang, Xu Chen, Pengtao Jing, Songnan Qu, Lijun Zhang, Donglei Zhou, Jinyang Zhu, Wen Xu, Biao Dong, and Hongwei Song. Doping lanthanide into perovskite nanocrystals: Highly improved and expanded optical properties. *Nano Letters*, 17: 8005–8011, 12 2017. ISSN 1530-6984. doi: 10.1021/acs.nanolett.7b04575.

[7] Santiago Miret, NM Anoop Krishnan, Benjamin Sanchez-Lengeling, Marta Skreta, Vineeth Venugopal, and Jennifer N Wei. Perspective on ai for accelerated materials design at the ai4mat-2023 workshop at neurips 2023. *Digital Discovery*, 2024.

[8] Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169, 2024.

[9] Gary Tom, Stefan P Schmid, Sterling G Baird, Yang Cao, Kourosh Darvish, Han Hao, Stanley Lo, Sergio Pablo-García, Ella M Rajaonson, Marta Skreta, et al. Self-driving laboratories for chemistry and materials science. *Chemical Reviews*, 2024.

[10] Alexandre Duval, Simon V Mathis, Chaitanya K Joshi, Victor Schmidt, Santiago Miret, Fragkiskos D Malliaros, Taco Cohen, Pietro Lio, Yoshua Bengio, and Michael Bronstein. A hitchhiker's guide to geometric gnns for 3d atomic systems. *arXiv preprint arXiv:2312.07511*, 2023.

[11] Santiago Miret and NM Krishnan. Are llms ready for real-world materials discovery? *arXiv preprint arXiv:2402.05200*, 2024.

[12] Mara Schilling-Wilhelmi, Martiño Ríos-García, Sherjeel Shabih, María Victoria Gil, Santiago Miret, Christoph T Koch, José A Márquez, and Kevin Maik Jablonka. From text to insight: Large language models for materials science data extraction. *arXiv preprint arXiv:2407.16867*, 2024.

[13] Kin Long Kelvin Lee, Carmelo Gonzales, Marcel Nassar, Matthew Spellings, Mikhail Galkin, and Santiago Miret. Matsciml: A broad, multi-task benchmark for solid-state materials modeling. *arXiv preprint arXiv:2309.05934*, 2023.

[14] Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon M. Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, Anuroop Sriram, Félix Therrien, Jehad Abed, Oleksandr Voznyy, Edward H. Sargent, Zachary Ulissi, and C. Lawrence Zitnick. The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts. *ACS Catalysis*, 13:3066–3084, 2023. ISSN 21555435. doi: 10.1021/acscatal.2c05426. URL http://arxiv.org/abs/2206.08917.

[15] Anuroop Sriram, Sihoon Choi, Xiaohan Yu, Logan M. Brabson, Abhishek Das, Zachary Ulissi, Matt Uyttendaele, Andrew J. Medford, and David S. Sholl. The open dac 2023 dataset and challenges for sorbent discovery in direct air capture. 2023. URL http://arxiv.org/abs/2311.00341.

[16] Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. 2 2021. URL http://arxiv.org/abs/2102.09548.

[17] Peter Eastman, Pavan Kumar Behara, David L. Dotson, Raimondas Galvelis, John E. Herr, Josh T. Horton, Yuezhi Mao, John D. Chodera, Benjamin P. Pritchard, Yuanqing Wang, Gianni De Fabritiis, and Thomas E. Markland. Spice, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data*, 10, 12 2023. ISSN 20524463. doi: 10.1038/s41597-022-01882-6.

[18] Till Siebenmorgen, Filipe Menezes, Sabrina Benassou, Erinc Merdivan, Kieran Didi, André Santos Dias Mourão, Radosław Kitel, Pietro Liò, Stefan Kesselheim, Marie Piraud, Fabian J. Theis, Michael Sattler, and Grzegorz M. Popowicz. Misato: machine learning dataset of protein–ligand complexes for structure-based drug discovery. *Nature Computational Science*, 4: 367–378, 5 2024. ISSN 26628457. doi: 10.1038/s43588-024-00627-2.

[19] Scott Kirklin, James E. Saal, Bryce Meredig, Alex Thompson, Jeff W. Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The open quantum materials database (oqmd): Assessing the accuracy of dft formation energies. *npj Computational Materials*, 1, 12 2015. ISSN 20573960. doi: 10.1038/npjcompumats.2015.10.

[20] Peder Lyngby and Kristian Sommer Thygesen. Data-driven discovery of 2d materials by deep generative models. *npj Computational Materials*, 8, 12 2022. ISSN 20573960. doi: 10.1038/s41524-022-00923-3.

[21] Santiago Miret, Kin Long Kelvin Lee, Carmelo Gonzales, Marcel Nassar, and Matthew Spellings. The open matsci ML toolkit: A flexible framework for machine learning in materials science. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=QBMyDZsPMd.

[22] Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. 6 2022. URL http://arxiv.org/abs/2206.07697.

[23] Bowen Deng, Peichen Zhong, Kyu Jung Jun, Janosh Riebesell, Kevin Han, Christopher J. Bartel, and Gerbrand Ceder. Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5:1031–1041, 9 2023. ISSN 25225839. doi: 10.1038/s42256-023-00716-3.

[24] Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2:718–728, 11 2022. ISSN 26628457. doi: 10.1038/s43588-022-00349-3.

[25] Chiho Kim, Tran Doan Huan, Sridevi Krishnan, and Rampi Ramprasad. A hybrid organic-inorganic perovskite dataset. *Scientific Data*, 4, 5 2017. ISSN 20524463. doi: 10.1038/sdata.2017.57.

[26] Jiaqi Yang, Panayotis Manganaris, and Arun Mannodi-Kanakkithodi. A high-throughput computational dataset of halide perovskite alloys. *Digital Discovery*, 2:856–870, 2023. ISSN 2635-098X. doi: 10.1039/D3DD00015J.

[27] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation, 2013. ISSN 2166532X.

[28] Ivano E. Castelli, David D. Landis, Kristian S. Thygesen, Soren Dahl, Ib Chorkendorff, Thomas F. Jaramillo, and Karsten W. Jacobsen. New cubic perovskites for one- and two-photon water splitting using the computational materials repository. *Energy and Environmental Science*, 5:9034–9043, 10 2012. ISSN 17545692. doi: 10.1039/c2ee22341d.

[29] Christopher P. Muzzillo, Cristian V. Ciobanu, and David T. Moore. High-entropy alloy screening for halide perovskites. *Materials Horizons*, 5 2024. ISSN 20516355. doi: 10.1039/d4mh00464g.

[30] Alex Ganose, Janosh Riebesell, J. George, Jimmy Shen, Andrew S. Rosen, Aakash Ashok Naik, nwinner, Mingjian Wen, rdguha1995, Matthew Kuner, Guido Petretto, Zhuoying Zhu, Matthew Horton, Hrushikesh Sahasrabuddhe, Aaron Kaplan, Jonathan Schmidt, Christina Ertural, Ryan Kingsbury, Matt McDermott, Rhys Goodall, Alexander Bonkowski, Thomas Purcell, Daniel Zügner, and Ji Qi. atomate2, January 2024. URL `https://github.com/materialsproject/atomate2`.

[31] Thomas D. Kühne, Marcella Iannuzzi, Mauro Del Ben, Vladimir V. Rybkin, Patrick Seewald, Frederick Stein, Teodoro Laino, Rustam Z. Khaliullin, Ole Schütt, Florian Schiffmann, Dorothea Golze, Jan Wilhelm, Sergey Chulkov, Mohammad Hossein Bani-Hashemian, Valéry Weber, Urban Borštnik, Mathieu Taillefumier, Alice Shoshana Jakobovits, Alfio Lazzaro, Hans Pabst, Tiziano Müller, Robert Schade, Manuel Guidon, Samuel Andermatt, Nico Holmberg, Gregory K. Schenter, Anna Hehn, Augustin Bussy, Fabian Belleflamme, Gloria Tabacchi, Andreas Glöß, Michael Lass, Iain Bethune, Christopher J. Mundy, Christian Plessl, Matt Watkins, Joost VandeVondele, Matthias Krack, and Jürg Hutter. Cp2k: An electronic structure and molecular dynamics software package -quickstep: Efficient and accurate electronic structure calculations. *Journal of Chemical Physics*, 152, 5 2020. ISSN 10897690. doi: 10.1063/5.0007045.

[32] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple, 1996.

[33] Joost VandeVondele and Jürg Hutter. An efficient orbital transformation method for electronic structure calculations. *The Journal of Chemical Physics*, 118(10):4365–4369, 03 2003. ISSN 0021-9606. doi: 10.1063/1.1543154. URL `https://doi.org/10.1063/1.1543154`.

[34] Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Samuel M. Blau, Vlad Cărare, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Fabio Falcioni, Edvin Fako, Andrea C. Ferrari, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Christian Holm, Jad Jaafar, Stephan Hofmann, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Viktor Svahn, Christopher Sutton, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry. 12 2023. URL `http://arxiv.org/abs/2401.00096`.

[35] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

A Appendix

A.1 Fine-tune analysis

We fine-tuned the pretrained MACE-MP-0 model using two learning rates: 0.005 and 0.0001. For reference, the pretrained model was originally trained with a learning rate of 0.0005. Both fine-tuning processes were run for 20 epoches with other training parameters same as the pretraining process. Both learning rates exhibit catastrophic forgetting, but the effect is more pronounced with the larger learning rate of 0.0005. The high energy and forces error on the MPtrj dataset indicates that the fine-tuned model is overfitted to the dopant dataset.

Model	Test set	Energy MAE (eV/atom)	Force MAE (eV/A)	Force RMSE (eV/A)
MACE-MP	MPtrj	0.02	0.04	0.08
	Dopant	0.18	0.12	0.23
MACE-MP-ft lr = 0.00001	MPtrj	0.16	0.14	0.27
	Dopant	0.02	0.03	0.08
MACE-MP-ft lr = 0.0005	MPtrj	0.62	0.23	0.55
	Dopant	0.01	0.02	0.07
MACE-init	MPtrj	0.59	0.44	4.03
	Dopant	0.04	0.03	0.08

Table A1: Summary of energy and forces MAE, and force root mean squared error (RMSE) prediction across different models.

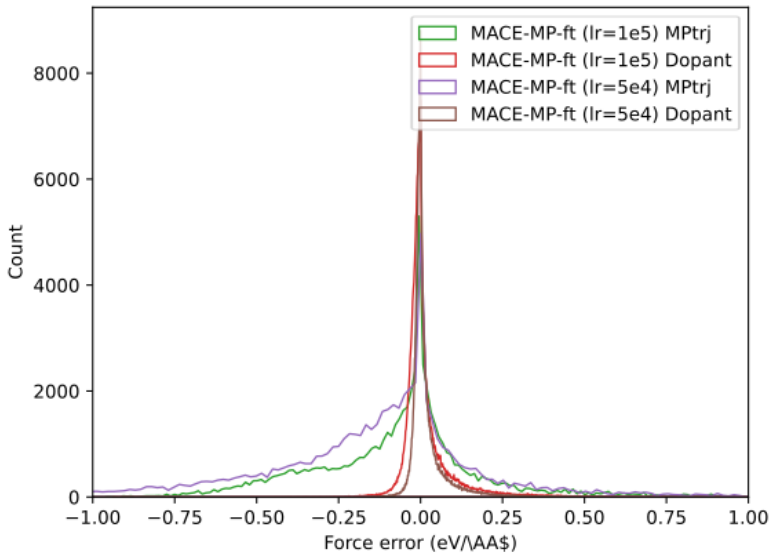


Figure A1: Distribution of the differences in the atomic forces predicted by fine-tuned and pretrained MACE models.