EXPLORING TRANSFER LEARNING FOR MATERIALS PROPERTY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Transfer learning based on foundation models has shown strong gains in low-data regimes, and similar benefits are emerging for materials property prediction where labeled data are scarce due to the cost of DFT calculations and experimental measurements. We study transfer learning for materials properties by finetuning MACE-MP-0, a relatively small foundation model pretrained on large-scale structure–energy DFT data, on different tasks from Matbench spanning mechanical and electronic targets. Energy-based pretraining yields consistent improvements over random initialization. We further show that negative transfer can limit gains and that regularization can improve results. Additionally, we provide some interpretations on what is transferred via layer-wise mutual information and weight-space drift.

1 INTRODUCTION

Transfer learning based on foundation models have enabled strong gains in low-data regimes across domains such as computer vision and language modeling (Bommasani et al., 2021), and have recently shown similar benefits for materials property prediction (Shoghi et al., 2024). In materials science, labeled data are often scarce due to the cost of DFT calculations and experimental measurements, making transfer learning particularly attractive. However, many foundation models are large and resource-intensive. In this work, we study transfer learning for materials properties by finetuning MACE-MP-0 (Batatia et al., 2025) on Matbench (Dunn et al., 2020) for inorganic bulk materials. MACE-MP-0 is a relatively small foundation model pretrained on large-scale structure-energy DFT data. We show that MACE-MP-0 with only 10% of the parameter count of JMP-S can achieve comparable performance for the bulk modulus prediction task.(Table 1). We fine-tune MACE-MP-0 on four Matbench tasks spanning mechanical and electronic targets (bulk modulus, shear modulus, band gap, and metal classification). Energy-based pretraining yields consistent improvements over random initialization.

Beyond performance, we show that negative transfer can limit gains and that accounting for negative transfer with regularization can improve results. Finally, we begin to interpret what is transferred by analyzing layer-wise mutual information and weight-space drift (Fig. 2). We notice that even for tasks with low mutual information between the features and the labels as well as for tasks where the physical relations between the target and the source label are small, the model still benefits from the pretrained weights. Even in these cases initializing the training with the weights of the foundational model leads to better models than random initialization.

2 RELATED WORK

2.1 TRANSFER LEARNING AND NEGATIVE TRANSFER

It has been observed by Yosinski et al. (2014), that neural networks tend to learn general features in the early layers. These features that are essentially re-learned for various tasks, and it is hence said that they generalize to various target domains.
This insight formulates the foundation for recent advances in transfer learning in fields such as computer vision (Yosinski et al., 2014; Sharif Razavian et al., 2014) and language modeling (Devlin et al., 2019), where given a pretrained model on task A, the performance of the model is improved

when finetuned on task B. This has been shown to be successful especially in the case of working with low-data regimes (Hernandez et al., 2021). However, in the case where the target task domain is not the same as the source task domain, negative transfer (Wang et al., 2019; Rosenstein et al., 2005) can occur, where the model pretrained on the source domain has a poorer performance compared to the same model trained from scratch on the target task.
Additionally, Neyshabur et al. (2020) conclude that a successful transfer depends much on the ability of the model to reuse features, which can be determined by measuring the similarities in representations in the pretrained model and finetuned model. The authors also show that even though the features correlation is distorted on the input space the model still benefits from the pretrained weights at initialization even when low level statistics are not changing (eg. pixel distribution).
Prior work (Wang et al., 2019) suggests that large divergence between the joint distribution $p(X_s, Y_s)$ of features X_s and labels Y_s in the source domain and the joint distribution of features and labels in the target domain $p(X_t, Y_t)$ may cause negative transfer. On the other hand, Chen et al. (2019) shows that general features are extracted via the large singular values of the weight matrix. Based on this observation, the authors propose to mitigate negative transfer by encouraging the model to keep the small singular values (obtained with Singular Value Decomposition SVD) of the weight matrix close to zero during the fine tuning process.

Additionally, different methods (Mehra et al., 2024; Huang et al., 2022) have been developed to estimate the effectiveness of transfer by approximating the mutual information between source domain and the target domain.
Not limited to computer vision tasks and language modeling, transfer learning has shown success in computational chemistry and materials informatics (Chithrananda et al., 2020; Magar et al., 2022; King-Smith, 2024; Gupta et al., 2021), with various pretraining approaches either with specifying pretraining auxiliary tasks or in a self supervised manner.

2.2 MATERIALS PROPERTY PREDICTION

For evaluating machine learning models for materials-related tasks, Matbench (Dunn et al., 2020) provides 13 different tasks and the corresponding datasets, with a standardized benchmarking schema. An advantage of Matbench is that the dataset provided for different properties prediction tasks is curated from Materials Project database, where most of foundation models, based on graph neural networks (Deng et al., 2023; Batatia et al., 2025; Chen & Ong, 2022), have been trained on for energy prediction. This makes Matbench tasks attractive for transfer learning since the source domain task (energy) and downstream tasks from Matbench share the same input space.

3 METHODS

3.1 EXPERIMENTAL SETUP

To study the transferability among different materials properties we choose Matbench, which provides data for 13 different tasks, where nine of them contain the structure information (described in Table2 in Appendix A.1). We use MACE-MP-0 (Batatia et al., 2025) as a foundation model, which is based on the architecture of (Batatia et al., 2022) and pretrained on a large atomistic structure-energy DFT dataset. Due to computational limitations, we pick four tasks (highlighted in grey in Table1) for finetuning, which (1) have different data sizes (2) different physical relatedness to structure energy (3) include both regression and classification objectives.
To study the benefits of transfer learning for the four tasks, we asses the performance of two models, the pretrained model and the randomly initialized version of it.

3.2 NEGATIVE TRANSFER AND CATASTROPHIC FORGETTING

During training we apply full transfer, this means all the network parameters are trainable.
As mentioned before, to mitigate for negative transfer during the finetuning, we use the work of Chen et al. (2019), where we minimize the smallest singular values of the pooled representations of the last interaction layer. We also require the model to stay close to the original weights w_0 with an L2-SP regularization term $\lambda \|w - w_0\|_2^2$ introduced by Xuhong et al. (2018) with the penalty parameter λ which is determined with hyperparameter search.

Table 1: Matbench tasks and model performance finetuned on the tasks highlighted tasks with grey. we finetune the randomly initialized **MACE-RI** as well as the pretrained model **MACE-MP-0**. Relative improvement between the two models computed as $\frac{\text{RI}-\text{MP-0}}{\text{RI}}$ for regression and $\frac{\text{MP-0}-\text{RI}}{\text{RI}}$ for classification (higher is better). Although JMP-S (Shoghi et al., 2024) is larger in size compared to MACE, it performs slighlty better for the physically related tasks `log_kvrh` and `log_gvrh` as shown in the relative improvement between JMP-S and MACE-MP-0 calculated as $\frac{\text{JMP-S}-\text{MP-0}}{\text{JMP-S}}$.

TASK	MB	MACE-RI	MACE-MP-0	JMP-S	REL. IMPROV. (RI vs MP-0)	REL. IMPROV. (JMP-S vs MP-0)
jdft2d	33.1918	-	-	30.16	-	-
phonons	28.7606	-	-	22.77	-	-
dielectric	0.2711	-	-	0.252	-	-
log_gvrh	0.0670	0.158	0.069	0.062	56.3%	10.14%
log_kvrh	0.0491	0.112	0.048	0.046	57.1%	4.17%
perovskites	0.0269	-	-	0.028	-	-
mp_gap	0.1559	0.883	0.325	0.121	63.2%	62.77%
mp_is_metal	0.9520	0.844	0.921	-	9.12%	-
mp_e_form	0.0170	-	-	13.3	-	-

4 Results and Discussion

Matbench provides a 5-fold train-test splits, which we use to investigate the transferability and evaluate our results. In Figure 1, we show the final results of training on the four mentioned tasks. The pretrained model is superior to its randomly initialized counterpart on the four tasks. We also report the regularization losses (penalized) in 1.c, where the SVD loss is consistently minimized in conjunction with the primary loss function, in contrast to the L2-SP loss. However, in Figure 2.b we show the weights drift with respect to the pretrained model at each layer, where larger drifts appear more on the final layers, which supports the idea that preserving general features at the early layers may assist learning new tasks. The weights drift also peaks for the last interaction layer for the metal classification task, which is expected due to scale difference and the sigmoid activation on the output.

To further understand why the pretrained model is outperforming the randomly initialized model, we measure the mutual information of the pretrained representations (at each layer l) and the four tasks formalized as $I({Z^l}_t; Y_t)$. We approximate the mutual information as proposed by Huang et al. (2022). Figure 2a shows the normalized mutual information (MI) for each of the nine task representations across layers of the network. We observe that the mutual information decays with depth, consistent with the data (information) processing inequality. Interestingly, tasks that are physically related to structure–energy prediction have higher mutual information (e.g., `jdft2d`, `gvrh`, and `kvrh`), whereas tasks involving electronic properties tend to have much lower mutual information with the representations learned during pretraining. However, we expect to have higher mutual between the representations of the pretrained network and the formation energy task compared to eg. the dielectric task. It is important to mention that the dataset size is crucial for estimating MI, for example, the dielectric dataset (around 4k) is much smaller than the formation energy dataset (around 100k), which can lead to higher entropy estimation on the representations. In addition, we estimate MI on graph-level embeddings obtained by mean pooling node embeddings. Since mean pooling (a non-invertible operation) causes information loss, it discards node-level information and can only reduce the information content of the representation, making the values in the Figure 2.a partially informative. Despite this limitation, the results are in general consistent with our expectations. Moreover, as shown in Figure 1.b, the test loss for the `mp_gap` task in the fine-tuned model is much lower than in the model trained from scratch. This aligns with the conclusion of Neyshabur et al. (2020), where the model has better weight initialization with the pretrained weights compared to random initialization, given that low level statistics such as (eg. atom distribution and Bbnd-length) are not changing much.

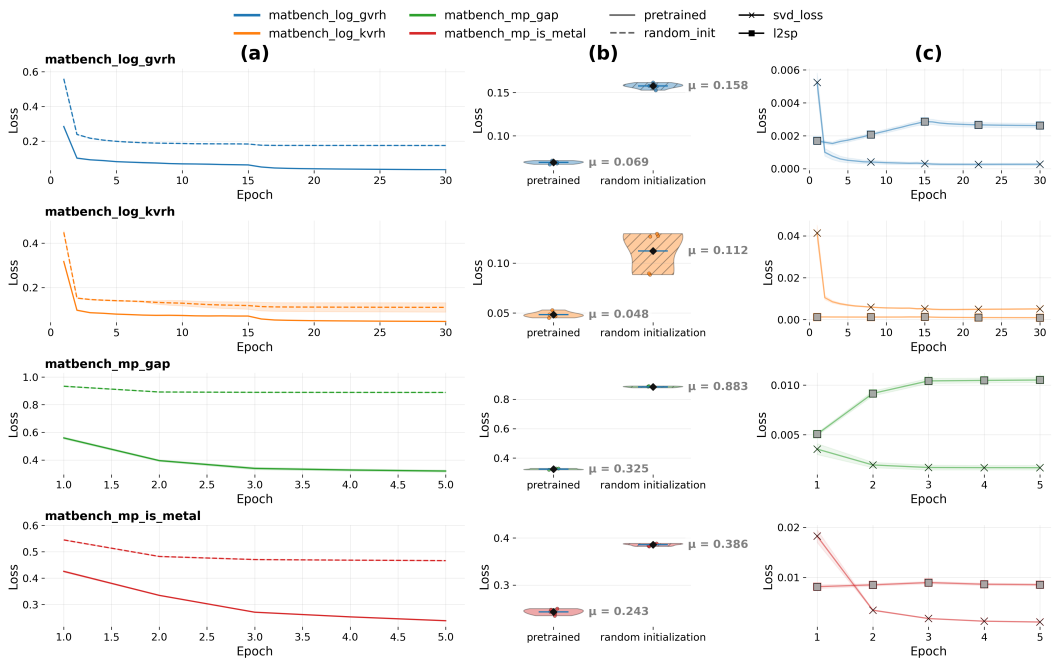

Figure 1: (a) The train loss on the four tasks for the pretrained model and randomly initialized model. (b) The test loss on the 5-fold splits provided from Matbench, where we show that the pretrained model consistently outperforms the randomly initialized model. (c) The weighted loss of the SVD term and the L2-SP term. We see that the SVD term is consistently minimized along the primary loss, where L2-SP is mostly staying constant. This indicates that depending on the strength of SVD and the primary objective loss, more deviation from the pretrained model is required.

5 Conclusion

MACE-MP-0 is a suitable foundation model for predicting material properties. Table 1 shows that transfer learning methods based on this foundation yield higher performance, than random initialization. We show that MACE-MP-0 is able to transfer general feature maps for the downstream tasks pretrained on structure energy similar to JMP model (Shoghi et al., 2024). We also show that when accounting for negative transfer we achieve good results with MACE-MP-0 compared to JMP-S, where MACE has only 10% of JMP-S parameters. We believe that further improvements on avoiding negative transfer could potentially make transferable models more economic by improving the performance-to-parameter quotient. Figure 2.a provides evidence that we can expect that MACE-MP-0 learned, analog to image detection models, to extract general features in the lower layers. Additionally, we see in Figure 2 that even for tasks, which have low mutual information with the original task such as the band gap task, still benefit from the pretraining. We believe this is due to alignment in the low level statistics as observed by Neyshabur et al. (2020) (e.g. distribution of atoms) which provides better weight initialization than the random initialization.

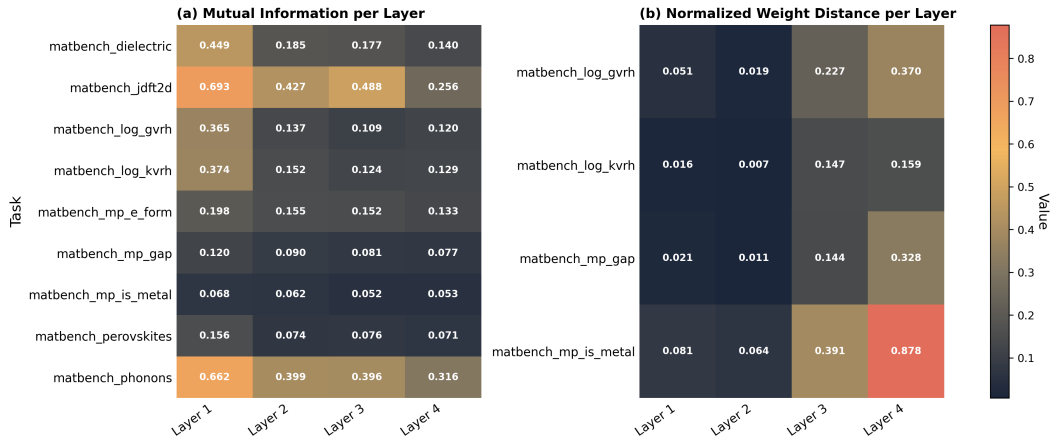

Figure 2: (a) The approximated (normalized) mutual information $\frac{I(Z^l{}_t;Y_t)}{H(Z^l{}_t)}$ of the representations of each task at each layer of the pretrained network before finetuning (normalized). Consistent with prior studies, where we see from many tasks that the neural network shares general features at early layers. We also see the mutual information varies depending on the task, where physically related tasks have high mutual information. (b) The normalized model weights distance to the pretrained model, where we see minimal changes in the first two layers.

REFERENCES

Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in neural information processing systems*, 35:11423–11436, 2022.

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, Filippo Bigi, Samuel M. Blau, Vlad Cărare, Michele Ceriotti, Sanggyu Chong, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Edvin Fako, Fabio Falcioni, Andrea C. Ferrari, John L. A. Gardner, Mikołaj J. Gawkowski, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Jonas Grandel, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Cheuk Hin Ho, Stephan Hofmann, Christian Holm, Jad Jaafar, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Panagiotis Kourtis, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Chen Lin, 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, Louise A. M. Rosset, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Christopher Sutton, Viktor Svahn, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Santiago Vargas, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Thomas Wolf, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry. *The Journal of Chemical Physics*, 163(18):184110, 11 2025. ISSN 0021-9606. doi: 10.1063/5.0297006. URL https://doi.org/10.1063/5.0297006.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Ku-

ditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022.

Xinyang Chen, Sinan Wang, Bo Fu, Mingsheng Long, and Jianmin Wang. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. *Advances in neural information processing systems*, 32, 2019.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *CoRR*, abs/2010.09885, 2020. URL https://arxiv.org/abs/2010.09885.

Bowen Deng, Peichen Zhong, KyuJung 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(9):1031–1041, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):138, 2020.

Vishu Gupta, Kamal Choudhary, Francesca Tavazza, Carelyn Campbell, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data. *Nature communications*, 12(1):6595, 2021.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *CoRR*, abs/2102.01293, 2021. URL https://arxiv.org/abs/2102.01293.

Long-Kai Huang, Junzhou Huang, Yu Rong, Qiang Yang, and Ying Wei. Frustratingly easy transferability estimation. In *International conference on machine learning*, pp. 9201–9225. PMLR, 2022.

Emma King-Smith. Transfer learning for a foundational chemistry model. *Chemical Science*, 15 (14):5143–5151, 2024.

Rishikesh Magar, Yuyang Wang, and Amir Barati Farimani. Crystal twins: self-supervised learning for crystalline material property prediction. *npj Computational Materials*, 8(1):231, 2022.

Akshay Mehra, Yunbei Zhang, and Jihun Hamm. Understanding the transferability of representations via task-relatedness. *Advances in Neural Information Processing Systems*, 37:116513–116546, 2024.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *Advances in neural information processing systems*, 33:512–523, 2020.

Michael T. Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G. Dietterich. To transfer or not to transfer. In *NIPS 2005 Workshop on Transfer Learning*, volume 898, pp. 1–4, Vancouver, BC, Canada, December 2005.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813, 2014.

Nima Shoghi, Adeesh Kolluru, John R. Kitchin, Zachary W. Ulissi, C. Lawrence Zitnick, and Brandon M. Wood. From molecules to materials: Pre-training large generalizable models for atomic property prediction. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=PfPnugdxup.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11293–11302, 2019.

LI Xuhong, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *International conference on machine learning*, pp. 2825–2834. PMLR, 2018.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

A Appendix

A.1 A.1

Table 2: Matbench tasks with dataset size, learning type, and brief descriptions.

TASK NAME	SAMPLES	TYPE	SHORT DESCRIPTION
matbench_jdft2d	636	Reg.	Exfoliation energies from crystal structure (meV/atom).
matbench_phonons	1,265	Reg.	Vibration properties from crystal structure (cm^{-1}) .
matbench_dielectric	4,764	Reg.	Dielectric constant (unitless).
matbench_log_gvrh	10,987	Reg.	$\log_{10}$ shear modulus G_{VRH} (GPa) from structure.
matbench_log_kvrh	10,987	Reg.	$\log_{10}$ bulk modulus K_{VRH} (GPa) from structure.
matbench_perovskites	18,928	Reg.	Perovskite formation energy (eV/unit cell) from structure.
matbench_mp_gap	106,113	Reg.	DFT band gap from Materials Project (eV) from structure.
matbench_mp_is_metal	106,113	Cls.	Metal vs non-metal (DFT-derived label) from structure.
matbench_mp_e_form	132,752	Reg.	Formation energy from Materials Project (eV/atom) from structure.