DIVERSITY-AWARE PRETRAINING IN MATERIALS LEARNING VIA TASK SIMILARITY

Bhanu Teja Mamillapalli[1]
b.mamillapalli@mail.utoronto.ca

Mahyar Rajabi Kochi[1]
mahyar.rajabi@mail.utoronto.ca

Seyed Mohamad Moosavi[1,2]
mohamad.moosavi@utoronto.ca

[1] University of Toronto, Canada
[2] Vector Institute for Artificial Intelligence, Canada

ABSTRACT

Large-scale datasets underpin recent advances in machine learning; however, in materials science and chemistry, data acquisition remains expensive, making performance in low-data regimes a central challenge. Pretraining and transfer learning are effective strategies in this setting, yet their success critically depends on the choice of pretraining tasks. Poorly selected tasks can yield marginal gains or induce negative transfer, while principled criteria for assembling pretraining datasets remain underexplored. In this work, we leverage task similarity metrics to move beyond selecting a single source task and instead construct diverse, representative pretraining task subsets. Using similarity-derived structure among tasks, we show how pretraining datasets can be assembled to balance relevance and diversity, maximizing knowledge transfer under fixed data budgets. Experiments on the QM9 benchmark demonstrate that models pretrained on such diversity-aware task subsets achieve performance comparable to that of substantially larger pretraining datasets assembled without regard to task relationships. These results identify task diversity as a key factor governing transfer efficiency and provide a practical strategy for scaling general-purpose models in materials science under high labeling costs.

1 INTRODUCTION

Limited dataset size remains a fundamental challenge in materials science machine learning, where the cost, time, and expertise required to generate high-quality labeled data often restrict dataset scale (Xu et al., 2023; L'Heureux et al., 2017; Moosavi et al., 2020a). Transfer learning has therefore emerged as a promising approach for improving model performance in such data-scarce regimes by exploiting knowledge learned from related, data-rich tasks (Alzubaidi et al., 2023; Niu et al., 2020; Saha et al., 2016). In a typical transfer learning workflow, a model is first pretrained on a source task with abundant labeled data to learn generalizable representations, and is subsequently fine-tuned on a target task for which labeled data is limited to achieve improvement over scratch training.

Despite its demonstrated success across a range of applications, the effectiveness of transfer learning is sensitive to the choice of pretraining task (Azizian & Hasan, 2025; Peters et al., 2019). When the source and target tasks are well aligned, transfer learning can accelerate convergence and improve predictive accuracy (Yamada et al., 2019; Jha et al., 2019; Gupta et al., 2021). However, pretraining on tasks that are only weakly related to the target task may provide little benefit and can, in some cases, actively harm performance through negative transfer (Zhang et al., 2023; Wang et al., 2019; Achille et al., 2019). This sensitivity underscores the importance of understanding and quantifying task relatedness when designing transfer learning pipelines for materials science applications.

These limitations motivate the need for a more principled framework for selecting pretraining tasks in transfer learning (Zamir et al., 2018; Fifty et al., 2021; Shui et al., 2019). One of the earliest systematic efforts to formalize task relatedness was introduced by Zamir et al. (2018), who proposed the concept of task similarity as a means of characterizing relationships between learning objectives. Their work showed that transfer learning performance is strongly influenced by how similarly the two tasks organize inputs in a model's latent representation space (Liu et al., 2019; Lew & Buehler, 2021). This alignment enables effective feature reuse during transfer learning and, importantly, can be quantified rather than chosen heuristically.

Building on this foundation, Dwivedi & Roig (2019) introduced Representational Similarity Analysis (RSA) as a general methodology for comparing tasks by analyzing the latent embedding space learned by machine learning models. RSA quantifies task similarity by measuring the correspondence between learned feature representations, providing an architecture-agnostic means of assessing how similarly models trained on two different tasks structure input data (Zhong et al., 2016; Guo et al., 2019). The authors demonstrated that RSA-based similarity measures are predictive of transfer learning success in the computer vision domain, highlighting the value of representation-level analysis for task selection.

More recently, Li et al. (2022) adapted these ideas for application in materials science by combining representational similarity with additional task-level similarity metrics to form the Molecular Task Similarity Estimator (MoTSE). MoTSE was shown to effectively identify source tasks that yield improved transfer learning performance on downstream target tasks, providing empirical evidence that principled task similarity estimation can substantially enhance transfer learning outcomes in materials science. They demonstrate the efficacy of this workflow across multiple small molecule property prediction datasets, and demonstrate how it eliminates the issue of negative transfer.

In this work, we further investigate the concept of task similarity in order to utilize it as a mechanism for constructing pretraining datasets composed of multiple tasks, with the goal of improving transfer learning performance in the small-data regime. Rather than selecting a single source task, we explore how task similarity can be leveraged to assemble a representative and diverse set of pretraining tasks that collectively provide a more effective inductive bias for downstream learning (Moosavi et al., 2020b).

2 METHOD

In this work we use the QM9 dataset (Ramakrishnan et al., 2014) as a preliminary benchmark to investigate the benefits of creating a diverse pretraining set. The dataset is preprocessed by removing tasks whose labels have a Pearson correlation above 0.8 with any other task. The remaining tasks are split into training, validation, and test sets (80/10/10), with consistent molecule assignments across all tasks. To assess uncertainty, the training set is randomly permuted into five folds.

2.1 REPRESENTATIONAL SIMILARITY ANALYSIS (RSA) COMPUTATION

In this work, task similarity is quantified using the Representational Similarity Analysis (RSA) framework originally introduced by Dwivedi & Roig (2019). RSA measures the degree to which two learning tasks induce similar internal representations within a model, providing a task-agnostic metric of relatedness grounded in representation geometry. The computation of RSA-based task similarity proceeds in three stages, as described below.

Stage 1: Pretraining. For each task in the dataset, a separate model is pretrained using the full available labeled data for that task. In this study we employ Chemprop, a message-passing neural network (MPNN) architecture proposed by Graff et al. (2026), due to its strong performance in molecular property prediction. However, the RSA framework itself is model-agnostic and can be applied to any representation learning architecture that produces structured latent embeddings.

Stage 2: Representation Dissimilarity Matrix (RDM) A probe dataset $\mathcal{D}_p = \{x_1, x_2, \ldots, x_M\}$ is constructed, consisting of a small but representative set of input samples from the domain. Each input x_i is passed through a given task-specific model f_k to obtain its latent embedding, $z_i^{(k)}$.

For each model f_k, pairwise similarities between all embeddings in the probe dataset are computed using the Pearson correlation coefficient. These similarities are converted into a Representation

Dissimilarity Matrix (RDM), defined as

$$RDM_{i,j}^{(k)} = 1 - \rho\Big(\mathbf{z}_i^{(k)}, \mathbf{z}_j^{(k)}\Big), \tag{1}$$

The resulting RDM captures the relative geometric structure of the model's latent space based on the probe dataset. Each RDM therefore serves as a task-specific fingerprint, encoding how the model organizes inputs in representation space independently of task labels.

Stage 3: RSA Value To compute the similarity between tasks T_a and T_b, their RDMs, $RDM^{(a)}$ and $RDM^{(b)}$, are compared by vectorizing the upper triangular (non-redundant) elements and computing Spearman's rank correlation coefficient.

$$\mathrm{RSA}(T_a, T_b) = \rho_s\Big(\mathrm{vec}\Big(RDM^{(a)}\Big), \mathrm{vec}\Big(RDM^{(b)}\Big)\Big), \tag{2}$$

The resulting RSA value provides a measure of task similarity, with larger values indicating greater alignment between the latent representations learned for the two tasks. Because this metric compares representational structure rather than task outputs, it is well suited for guiding pretraining task selection in transfer learning workflows.

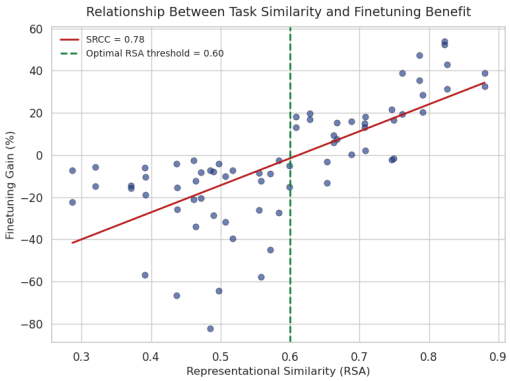
3 RESULTS

We began our analysis by computing pairwise similarity scores across all tasks. We assess the significance of this similarity measure by quantifying improvements from transfer learning in the small-data regime by fine-tuning a pretrained model on a target task. Performance gains are measured using the area under the curve (AUC) of the mean absolute error (MAE) as a function of training set size. The AUC improvement is expected to correlate positively with the similarity between the pretraining and target tasks. As shown in Figure 1a, we observe a weak but positive correlation between task similarity and fine-tuning gain.

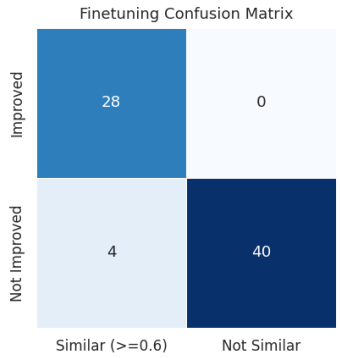
The relationship between fine-tuning gain and task similarity becomes clearer when both variables are treated as binary. Tasks are labeled as similar if their similarity score exceeds a threshold (chosen to maximize F1 score), and as improved if performance increases after fine-tuning. The resulting classification is shown in Figure 1b.

Figures 1a and1b show that the proposed similarity metric is indicative of improved transfer learning performance. Building on this, we propose a method for constructing a representative pretraining dataset by ensuring that, for any target task, at least one similar task is included in the pretraining set. The task subset $\{\text{A, HOMO, } C_v\}$ satisfies this coverage requirement, and we pretrain a model using only this task subset and evaluate its finetuned performance on the remaining target tasks, comparing the results to a model pretrained on all available tasks (excluding the target task). Results from this analysis shown in Figure 1c.

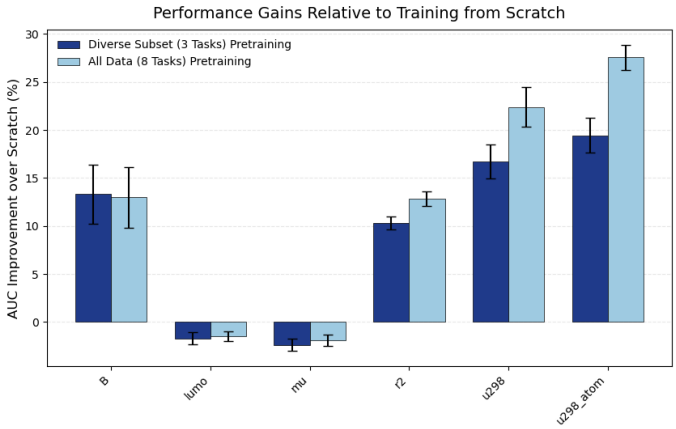
Across all five target tasks, the fine-tuned models pretrained on the diverse task subset achieve performance comparable to those pretrained on the full dataset. Overall, models pretrained on the diverse subset exhibit similar performance in the small data regime to those pretrained on the full dataset, despite being trained on approximately one third of the data. We propose that the increased diversity of the pretraining set allows for this improved data efficiency during learning.



(a) Correlation between similarity and finetuning gain.



(b) Confusion matrix from finetuning results.



(c) AUC metric for diverse pretrain set versus all task pretraining.

Figure 1: Relationship between task-pair similarity and transfer learning performance in small data conditions. (a) Continuous relationship between similarity and fine-tuning gain. (b) Confusion matrix using thresholds of 0.6 for similarity and 0% for fine-tuning gain; the dashed line indicates the similarity threshold that maximizes F1 score. (c) Transfer performance comparing a model pretrained on a diverse subset of three tasks vs. one pretrained on all eight non-target tasks. Performance is measured by AUC, computed as the integral of the MAE curve from 1,000 to 30,000 training samples, with error bars showing standard error across five training folds.

4 Conclusion

In this work, we show that task similarity, quantified through representational structure, provides a principled foundation for constructing compact yet effective pretraining datasets in materials learning. Rather than viewing task similarity solely as a tool for selecting individual source tasks, our results demonstrate its broader utility for designing diversity-aware pretraining strategies that preserve transfer performance while substantially reducing pretraining data labelling cost.

From a broader perspective, these findings suggest a shift in how datasets are assembled in data-scarce scientific domains. As materials datasets increasingly grow in breadth rather than depth, spanning heterogeneous properties, chemistries, and experimental conditions, task diversity, rather than dataset size alone, may emerge as a key driver of transferable modeling. Our results complement prior work on dataset diversity and data redundancy in single-task learning from Li et al. (2023); Moosavi et al. (2020b), together indicating that task-level and data-level diversity form a unified framework for guiding efficient materials data generation and discovery campaigns, and for advancing general-purpose materials models (Alampara et al., 2025) under realistic data and computational constraints.

REFERENCES

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6430–6439, 2019.

Nawaf Alampara, Anagha Aneesh, Martiño Ríos-García, Adrian Mirza, Mara Schilling-Wilhelmi, Ali Asghar Aghajani, Meiling Sun, Gordan Prastalo, and Kevin Maik Jablonka. General-purpose models for the chemical sciences: Llms and beyond. *arXiv preprint arXiv:2507.07456*, 2025.

Laith Alzubaidi, Jinshuai Bai, Aiman Al-Sabaawi, Jose Santamaría, A.s Albahri, Bashar Aldabbagh, Mohammed Fadhel, Mohamed Manoufali, Jinglan Zhang, Ali Al-Timemy, Ye Duan, Amjed Abdullah, Laith Farhan, Yi Lu, Ashish Gupta, Felix Albu, Amin Abbosh, and Yuantong Gu. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, 10, 04 2023. doi: 10.1186/s40537-023-00727-2.

Waïss Azizian and Ali Hasan. How does the pretraining distribution shape in-context learning? task selection, generalization, and robustness, 2025. URL https://arxiv.org/abs/2510.01163.

Kshitij Dwivedi and Gemma Roig. Representation similarity analysis for efficient task taxonomy transfer learning, 2019. URL https://arxiv.org/abs/1904.11740.

Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34:27503–27516, 2021.

David E. Graff, Nathan K. Morgan, Jackson W. Burns, Anna C. Doner, Brian Li, Shih-Cheng Li, Joel Manu, Angiras Menon, Hao-Wei Pang, Haoyang Wu, Akshat Shirish Zalte, Jonathan W. Zheng, Connor W. Coley, William H. Green, and Kevin P. Greenman. Chemprop v2: An efficient, modular machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 66(1):28–33, 01 2026. doi: 10.1021/acs.jcim.5c02332. URL https://doi.org/10.1021/acs.jcim.5c02332.

Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019. doi: 10.1109/ACCESS.2019.2916887.

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.

Dipendra Jha, Kamal Choudhary, Francesca Tavazza, Wei-keng Liao, Alok Choudhary, Carelyn Campbell, and Ankit Agrawal. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nature Communications*, 10(1):5316, 2019.

Andrew J. Lew and Markus J. Buehler. Encoding and exploring latent design space of optimal material structures via a vae-lstm model. *Forces in Mechanics*, 5:100054, 2021. ISSN 2666-3597. doi: https://doi.org/10.1016/j.finmec.2021.100054. URL https://www.sciencedirect.com/science/article/pii/S2666359721000457.

Han Li, Xinyi Zhao, Shuya Li, Fangping Wan, Dan Zhao, and Jianyang Zeng. Improving molecular property prediction through a task similarity enhanced transfer learning strategy. *iScience*, 25(10):105231, 2022. ISSN 2589-0042. doi: https://doi.org/10.1016/j.isci.2022.105231. URL https://www.sciencedirect.com/science/article/pii/S2589004222015036.

Kangming Li, Daniel Persaud, Kamal Choudhary, Brian DeCost, Michael Greenwood, and Jason Hattrick-Simpers. Exploiting redundancy in large materials datasets for efficient machine learning with less data. *Nature Communications*, 14(1):7283, 2023.

Yang Liu, Eunice Jun, Qisheng Li, and Jeffrey Heer. Latent space cartography: Visual analysis of vector space embeddings. *Computer Graphics Forum*, 38(3):67–78, 2019. doi: https://doi.org/10.1111/cgf.13672. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13672.

Alexandra L'Heureux, Katarina Grolinger, Hany F. Elyamany, and Miriam A. M. Capretz. Machine learning with big data: Challenges and approaches. *IEEE Access*, 5:7776–7797, 2017. doi: 10.1109/ACCESS.2017.2696365.

Seyed Mohamad Moosavi, Kevin Maik Jablonka, and Berend Smit. The role of machine learning in the understanding and design of materials. *Journal of the American Chemical Society*, 142(48): 20273–20287, 2020a.

Seyed Mohamad Moosavi, Aditya Nandy, Kevin Maik Jablonka, Daniele Ongari, Jon Paul Janet, Peter G Boyd, Yongjin Lee, Berend Smit, and Heather Kulik. Understanding the diversity of the metal-organic framework ecosystem. *Nature Communications*, 11:4068, 2020b.

Shuteng Niu, Yongxin Liu, Jian Wang, and Houbing Song. A decade survey of transfer learning (2010–2020). *IEEE Transactions on Artificial Intelligence*, 1(2):151–166, 2020. doi: 10.1109/TAI.2021.3054609.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. To tune or not to tune? adapting pretrained representations to diverse tasks. *CoRR*, abs/1903.05987, 2019. URL http://arxiv.org/abs/1903.05987.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, Aug 2014. ISSN 2052-4463. doi: 10.1038/sdata.2014.22. URL https://doi.org/10.1038/sdata.2014.22.

Budhaditya Saha, Sunil Gupta, Dinh Phung, and Svetha Venkatesh. Multiple task transfer learning with small sample sizes. *Knowledge and Information Systems*, 46(2):315–342, 2016.

Changjian Shui, Mahdieh Abbasi, Louis-Émile Robitaille, Boyu Wang, and Christian Gagné. A principled approach for learning task similarity in multitask learning. *CoRR*, abs/1903.09109, 2019. URL http://arxiv.org/abs/1903.09109.

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.

Pengcheng Xu, Xiaobo Ji, Minjie Li, and Wencong Lu. Small data machine learning in materials science. *npj Computational Materials*, 9(1):42, 2023. doi: 10.1038/s41524-023-01000-z. URL https://doi.org/10.1038/s41524-023-01000-z.

Hironao Yamada, Chang Liu, Stephen Wu, Yukinori Koyama, Shenghong Ju, Junichiro Shiomi, Junko Morikawa, and Ryo Yoshida. Predicting materials properties with little data using shotgun transfer learning. *ACS Central Science*, 5(10):1717–1730, 10 2019.

Amir R. Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3712–3722, 2018. doi: 10.1109/CVPR.2018.00391.

Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10(2):305–329, February 2023. ISSN 2329-9274. doi: 10.1109/jas.2022.106004. URL http://dx.doi.org/10.1109/JAS.2022.106004.

Guoqiang Zhong, Li-Na Wang, Xiao Ling, and Junyu Dong. An overview on data representation learning: From traditional feature learning to recent deep learning. *The Journal of Finance and Data Science*, 2(4):265–278, 2016. ISSN 2405-9188. doi: https://doi.org/10.1016/j.jfds.2017.05.001. URL https://www.sciencedirect.com/science/article/pii/S2405918816300459.