Combining Transfer Learning and Representation Learning to Improve Predictive Analytics on Small Materials Data

Vishu Gupta
ECE Department
Northwestern University
Evanston, Illinois, USA
vishugupta2020@
u.northwestern.edu

Wei-keng Liao
ECE Department
Northwestern University
Evanston, Illinois, USA
wkliao@
eecs.northwestern.edu

Alok Choudhary
ECE Department
Northwestern University
Evanston, Illinois, USA
choudhar@
eecs.northwestern.edu

Ankit Agrawal
ECE Department
Northwestern University
Evanston, Illinois, USA
ankitag@
eecs.northwestern.edu

***Abstract*—Modern data mining methods have seen a widespread and growing application in the field of materials science for regression-based predictive modeling due to their effectiveness in extracting and utilizing the hidden information from the materials datasets. However, due to the costly and time-consuming nature of the methods involved in obtaining the experimental and computational data, the majority of the materials datasets are small in size. Moreover, limited hand-engineered representations available from the raw materials data make it harder to improve the accuracy of predictive models on such small and specialized training datasets. In this paper, we introduce a novel technique that combines transfer learning (TL) and representation learning (RL) using a pre-trained deep neural network to maximize accuracy without additional computational costs on inorganic material properties. The performance of the proposed method is compared against traditional machine learning (ML), and deep neural network models trained from scratch (SC) with elemental fraction (EF) as input, more informative physical attributes (PA) as input (for a stringent comparison), as well as conventional TL and RL techniques using deep neural networks. The results demonstrate that the proposed method can improve the accuracy as compared to SC models and conventional TL and RL techniques.**

***Index Terms*—deep learning, transfer learning, representation learning, deep regression, predictive modeling**

I. Introduction

Modern data mining methods have seen a widespread and growing application in the field of materials science for regression-based predictive modeling due to their effectiveness in extracting and utilizing the hidden information from the materials datasets and aid in the process of materials discovery [1]–[7]. This has been made possible due to the availability of computationally calculated large materials databases [8], [9] as well as easy-to-use data mining tools and advancement in the machine learning (ML) and deep learning (DL) algorithms to extract hidden information from raw inputs and build accurate and robust models for various material properties [10]–[15]. Since materials property prediction is a regression-based task and the representation used as model input to train various ML/DL methods usually comprises of a one-dimensional numerical vector obtained by pre-processing raw materials input, traditional ML algorithms [10], [11] and DL models composed of fully connected layers [16]–[21] are extensively used. However, due to the costly and time-consuming nature of the methods involved in obtaining the experimental and, in some cases even computational data, the majority of the materials datasets are small in size, limiting the highly accurate models to a selected few materials properties with a large amount of data [22], [23]. Moreover, limited generalized hand-engineered representations available from the raw materials data [24], [25] make it harder to improve the accuracy of predictive models built on such small and specialized training datasets. Therefore, various advanced data mining techniques such as transfer learning (TL) [26]–[29] and representation learning (RL) [30]–[35] are often applied to tackle the bottleneck of small data size by reusing the existing knowledge in a bid to boost the predictive performance of the model.

Our goal in this work is to design a data mining method that can improve the predictive ability of the model by using a raw materials representation as input. We introduce the idea of combining TL and RL to perform model training for deep regression networks on materials datasets where the input data used to train the model is different from what the pre-trained model was trained on. Although both the TL and RL methods [29], [32]–[35] have been widely used separately to handle small training datasets, to the best of our knowledge, no previous work investigates the effectiveness of the combination of TL and RL techniques for building deep regression networks composed of fully connected layers for numerical vector inputs, especially for materials science applications.

In this paper, we propose a novel TL+RL method where we combine the weights and activations extracted from a pre-trained model trained on a large dataset to train small datasets to maximize accuracy without additional computational costs. We compare the proposed TL+RL method against state-of-the-art ML/DL models trained directly on the small target dataset from scratch (SC) and conventional TL (fine-tuning)

and DL-based RL methods using deep neural networks. We perform predictive analysis of formation energy of inorganic materials from a numerical input vector composed of 86 features representing composition-based elemental fractions with OQMD [8] as the source dataset used to train the source model, and JARVIS [9] as the target dataset used to train the target model. Our proposed TL+RL method achieves better results in terms of the 10-fold cross-validation error and test error than other approaches for the regression-based prediction tasks. Overall, the proposed TL+RL method provided a more accurate model as compared to other models used for the comparison and is expected to be widely helpful for efficient and accurate predictive modeling without additional computational costs on not only materials datasets but for datasets from other domains as well.

II. Design

Here we describe the proposed TL+RL method with the novel combination of data mining techniques and the primary DL architecture deployed in this work. For the primary architecture, we use a 17-layered neural network (NN-17) comprising of stacks of fully connected layers and ReLU as the activation function as used in [19], [22], [23]. For the proposed TL+RL mechanism, we use activations of the pre-trained model from a given layer as the materials representation for each compound of the small target dataset and the weights of the pre-trained model as the initial set of weights of the deep neural network. The activations of the pre-trained model from a given layer can have varying sizes as it depends on the number of neurons in a given layer. For example, if we represent the materials representation using the first layer of the DL model, each compound will be represented as a 1024-dimensional feature vector. As our primary architecture consists of 16 layers (except the output layer), we can have 15 different types of representations (excluding the layer before the output layer, which is used for freezing). Moreover, as each layer in the network comprises of a dense block and ReLU activation, we call the representation obtained from the dense block and ReLU activation as DL(Dense) and DL(ReLU), respectively. The main difference between DL(Dense) and DL(ReLU) is that the DL(Dense) features comprise of both the positive and negative activations, whereas DL(ReLU) representation truncates the negative values to zero, causing a potential loss of information during the process. The weights of the pre-trained model are obtained by training the primary architecture on a large source dataset (OQMD) using elemental fraction (EF) composed of 86 composition-based attributes only. Hence, although in our proposed TL+RL method, the input representation used to train the model on the target dataset is obtained from the activations of the pre-trained model, the input representation used to train the target model is different from the input representation used to train the source model. We also initialize the target model's weights using those from the pre-trained model before starting the training process (except for the first layer, which is randomly initialized due to the differing inputs between the source and target models). A block diagram depicting the workflow of our proposed method is shown in Figure 1. After training the models on representations obtained from different layers as model input, we report the result of the best-performing layer on the validation set (which are usually the features extracted from the first four layers [32]). The notations for specific methods used in this work are stated in Table I.

TABLE I
NOTATIONS FOR THE DIFFERENT SCRATCH (SC) AND TRANSFER LEARNING/REPRESENTATION LEARNING (TL/RL/TL+RL) MODELING CONFIGURATIONS USED IN THIS WORK.

Notation	Description
SC:ML (EF)	AutoML model trained from scratch (SC) using Elemental Fractions (EF)
SC:ML (PA)	AutoML model trained from SC using Physical Attributes (PA)
SC:DL(EF)	NN-17 model trained from SC using EF
SC:DL(PA)	NN-17 model trained from SC using PA
TL: Fine-Tune	Fine-tuning on the same NN-17 framework using the pre-trained weights of source model
RL: Freezing	NN-17 model trained from the representation obtained from the last layer of the source model
RL:DL (ReLU)	NN-17 model trained using the representation obtained after the ReLU activation function of the source model
RL:DL (Dense)	NN-17 model trained using the representation obtained after the dense layer of the source model
TL+RL:DL (ReLU)	NN-17 model trained from the representation obtained after the dense layer of the source model and using the pre-trained weights of source model
TL+RL:DL (Dense)	NN-17 model trained from the representation obtained after the dense layer of the source model and using the pre-trained weights of source model

III. Empirical Evaluation

In this section, we present a detailed analysis and assessment of the proposed combination of transfer learning and representation learning (TL+RL) method. Before presenting the results, we discuss the experimental settings and datasets used in this work.

1) Experimental Settings: For the DL model, we use a 17-layered neural network comprising of stacks of fully connected layers and ReLU as the activation function as used in [19], [22], [23]. For ML, we used an AutoML library called hyperopt sklearn [36] to find the best-performing ML model. We used mean absolute error (MAE) as the loss function and the error metric for all the results. In general, we follow a two-step comparison for the result analyses performed in our work. For step one, we perform 10-fold cross-validation (CV) for all the methods used in this work, and the mean validation MAE from 10-fold CV is used to select the best-performing method among two categories SC and TL/RL/TL+RL (as all the methods discussed in TL, RL, and TL+RL use some form of pre-obtained knowledge to assist the model training). For step two, we compare the test MAEs among SC and TL/RL/TL+RL by performing model testing on the holdout test set using the selected methods in step one. Since the selected method from each of the categories would have 10 models (from the 10-fold CV), the one with the least validation

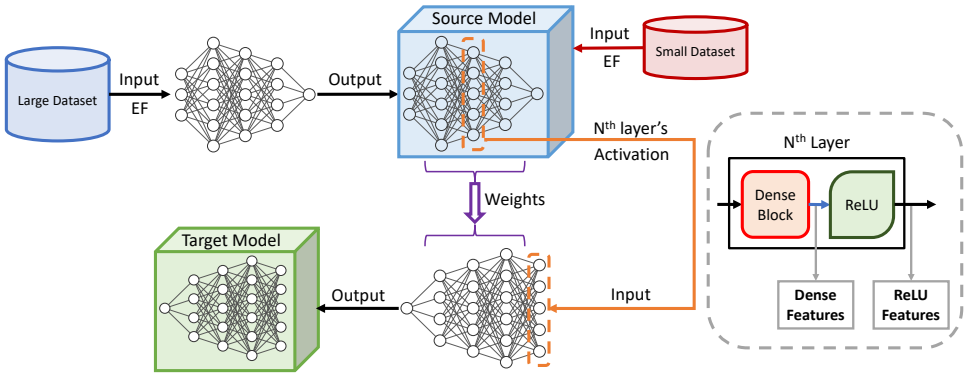

Fig. 1. Block diagram depicting the workflow of our proposed method.

TABLE II
PREDICTION PERFORMANCE BENCHMARKING. THE TABLE SHOWS THE MEAN ± STANDARD DEVIATION OF MAE FROM 10-FOLD CROSS-VALIDATION FOR ALL METHODS AND TEST MAE OF THE BEST SC, TL/RL, AND TL+RL MODEL.

Model	Validation MAE	Test MAE
SC:DL(EF)	**0.135±0.007**	0.142
SC:DL(PA)	0.167±0.005	
SC:ML(EF)	0.206±0.011	
SC:ML(PA)	0.141±0.005	
TL:Fine-Tune	0.121±0.007	
RL:Freezing	0.230±0.008	
RL:DL(ReLU)	0.128±0.005	
RL:DL(Dense)	**0.118±0.006**	0.123
TL+RL:DL(ReLU)	0.124±0.006	
TL+RL:DL(Dense)	**0.117±0.005**	**0.121**

MAE is used to perform model testing on the holdout test set, thereby ensuring that the test set is seen only once.

2) Datasets: The two datasets and their respective data sizes used to evaluate our methods are as follows: OQMD (345,220), and JARVIS (28,171). OQMD and JARVIS databases were downloaded from their respective websites [8], [9]. The formation energy from JARVIS as the target dataset and OQMD as the source dataset are used for model training. The datasets are randomly split with a fixed random seed of 12345 into training, validation, and test sets in the ratio of 81:9:10.

A. Prediction performance benchmarking

We conduct predictive analysis for formation energy, using EF as input across all methods. Additionally, we introduce a more stringent comparison by employing a composition-based input physical attributes (PA), for the baseline SC models. This allows us to assess the potential of the proposed TL+RL:DL method, which relies solely on EF-based inputs.

In Table II, we show the 10-fold cross-validation MAE for all methods and test MAE of the best SC, TL/RL, and TL+RL models. For the SC methods, SC:DL(EF) achieved the best 10-fold cross-validation MAE of 0.135±0.007 eV/atom as compared to other SC models. Interestingly, for SC models, the PA attributes benefited only the ML models and not the DL models. We believe this might be because, compared to ML, the neural-network-based DL is much better equipped to work well on raw inputs due to its hierarchical feature learning capability [22]. For TL/RL methods, RL:DL(Dense) achieved the best 10-fold cross-validation MAE of 0.118±0.006 eV/atom, whereas RL:Freezing performed the worst with an MAE of 0.230±0.008 eV/atom. Among our proposed TL+RL methods, TL+RL:DL(Dense) method performs the best with a 10-fold cross-validation MAE of 0.117±0.005 eV/atom.

We then use the best model with the least validation (out of 10 models) for SC:DL(EF), RL:DL(Dense), and TL+RL:DL(Dense) to perform model testing on the holdout test set and obtain test MAE of 0.142 eV/atom, 0.123 eV/atom, and 0.121 eV/atom respectively. Even though the SC models were allowed potentially more informative PA-based attributes for their input, the results show that the best TL+RL method outperforms the SC and TL/RL methods in terms of the 10-fold cross-validation MAE and test MAE.

IV. CONCLUSION AND FUTURE WORK

In this paper, we analyzed and proposed a combination of data mining methods for deep regression networks composed of fully connected layers with numerical vectors as inputs. We introduced a novel combination of transfer learning and representation learning method (TL+RL) in a deep regression network, which leverages the both ”representation” and ”weights” learned from the pre-trained source model for building the target model on the small target dataset. The proposed TL+RL method not only outperformed ML/DL models trained from scratch for predictive modeling using the same input but also performed better than conventional TL and DL-based RL methods where SC methods were even allowed to use more informative PA-based model input. The insights obtained from the proposed combination of transfer learning and representation learning methods can help build predictive models for other applications with numerical vector inputs without additional computational costs. The code, data, and models used in this work are publicly available at https://github.com/GuptaVishu2002/TLRL to facilitate reproducibility and further building upon this work.

ACKNOWLEDGMENT

This work was performed under the following financial assistance award 70NANB19H005 from U.S. Department of Commerce, National Institute of Standards and Technology as part of the Center for Hierarchical Materials Design (CHiMaD). Partial support is also acknowledged from NSF awards CMMI-2053929, OAC-2331329, DOE award DE-SC0021399, and Northwestern Center for Nanocombinatorics.

REFERENCES

[1] A. Agrawal and A. Choudhary, "Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science," *APL Materials*, vol. 4, no. 5, p. 053208, 2016.

[2] L. Ward and C. Wolverton, "Atomistic calculations and materials informatics: A review," *Current Opinion in Solid State and Materials Science*, vol. 21, no. 3, pp. 167–176, 2017.

[3] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, "Machine learning in materials informatics: recent applications and prospects," *npj Computational Materials*, vol. 3, no. 1, p. 54, dec 2017.

[4] A. Agrawal and A. Choudhary, "Deep materials informatics: Applications of deep learning in materials science," *MRS Communications*, vol. 9, no. 3, pp. 779–792, 2019.

[5] K. Choudhary, B. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C. W. Park, A. Choudhary, A. Agrawal, S. J. Billinge *et al.*, "Recent advances and applications of deep learning methods in materials science," *npj Computational Materials*, vol. 8, no. 1, pp. 1–26, 2022.

[6] V. Gupta, W.-k. Liao, A. Choudhary, and A. Agrawal, "Evolution of artificial intelligence for application in contemporary materials science," *MRS communications*, vol. 13, no. 5, pp. 754–763, 2023.

[7] K. Choudhary, D. Wines, K. Li, K. F. Garrity, V. Gupta, A. H. Romero, J. T. Krogel, K. Saritas, A. Fuhr, P. Ganesh *et al.*, "Jarvis-leaderboard: a large scale benchmark of materials design methods," *npj Computational Materials*, vol. 10, no. 1, p. 93, 2024.

[8] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, "The open quantum materials database (oqmd): assessing the accuracy of dft formation energies," *npj Computational Materials*, vol. 1, p. 15010, 2015.

[9] K. Choudhary, K. F. Garrity, A. C. Reid, B. DeCost, A. J. Biacchi, A. R. Hight Walker, Z. Trautt, J. Hattrick-Simpers, A. G. Kusne, A. Centrone *et al.*, "The joint automated repository for various integrated simulations (jarvis) for data-driven materials design," *npj Computational Materials*, vol. 6, no. 1, pp. 1–13, 2020.

[10] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials," *npj Computational Materials*, vol. 2, no. August, p. 16028, 2016.

[11] A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, and I. Tanaka, "Representation of compounds for machine-learning prediction of physical properties," *Physical Review B*, vol. 95, no. 14, p. 144110, 2017.

[12] G. Pilania, "Machine learning in materials science: From explainable predictions to autonomous design," *Computational Materials Science*, vol. 193, p. 110360, 2021.

[13] D. Morgan and R. Jacobs, "Opportunities and challenges for machine learning in materials science," *Annual Review of Materials Research*, vol. 50, pp. 71–103, 2020.

[14] J. Westermayr, M. Gastegger, K. T. Schütt, and R. J. Maurer, "Perspective on integrating machine learning into computational chemistry and materials science," *The Journal of Chemical Physics*, vol. 154, no. 23, p. 230903, 2021.

[15] V. Gupta, K. Choudhary, Y. Mao, K. Wang, F. Tavazza, C. Campbell, W.-k. Liao, A. Choudhary, and A. Agrawal, "Mppredictor: An artificial intelligence-driven web tool for composition-based material property prediction," *Journal of Chemical Information and Modeling*, 2023.

[16] Q. Zhou, P. Tang, S. Liu, J. Pan, Q. Yan, and S.-C. Zhang, "Learning atoms for materials discovery," *Proceedings of the National Academy of Sciences*, vol. 115, no. 28, pp. E6411–E6417, 2018.

[17] D. Jha, V. Gupta, W.-k. Liao, A. Choudhary, and A. Agrawal, "Moving closer to experimental level materials property prediction using ai," *Scientific reports*, vol. 12, 2022.

[18] E. O. Pyzer-Knapp, K. Li, and A. Aspuru-Guzik, "Learning from the harvard clean energy project: The use of neural networks to accelerate materials discovery," *Advanced Functional Materials*, vol. 25, no. 41, pp. 6495–6502, 2015.

[19] V. Gupta, A. Peltekian, W.-k. Liao, A. Choudhary, and A. Agrawal, "Improving deep learning model performance under parametric constraints for materials informatics applications," *Scientific reports*, vol. 13, no. 1, p. 9128, 2023.

[20] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Machine learning of molecular electronic properties in chemical compound space," *New Journal of Physics, Focus Issue, Novel Materials Discovery*, 2013, to appear.

[21] V. Gupta, Y. Li, A. Peltekian, M. N. T. Kilic, W.-k. Liao, A. Choudhary, and A. Agrawal, "Simultaneously improving accuracy and computational cost under parametric constraints in materials property prediction tasks," *Journal of Cheminformatics*, vol. 16, no. 1, p. 17, 2024.

[22] D. Jha, L. Ward, A. Paul, W.-k. Liao, A. Choudhary, C. Wolverton, and A. Agrawal, "ElemNet: Deep learning the chemistry of materials from only elemental composition," *Scientific reports*, vol. 8, no. 1, p. 17593, 2018.

[23] D. Jha, V. Gupta, L. Ward, Z. Yang, C. Wolverton, I. Foster, W.-k. Liao, A. Choudhary, and A. Agrawal, "Enabling deeper learning on big data for materials informatics applications," *Scientific reports*, vol. 11, no. 1, pp. 1–12, 2021.

[24] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, "Big data of materials science: Critical role of the descriptor," *Physical Review Letters*, vol. 114, no. 10, p. 105503, 2015.

[25] A. D. Sendek, Q. Yang, E. D. Cubuk, K.-A. N. Duerloo, Y. Cui, and E. J. Reed, "Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials," *Energy & Environmental Science*, vol. 10, no. 1, pp. 306–320, 2017.

[26] X. Li, Y. Zhang, H. Zhao, C. Burkhart, L. C. Brinson, and W. Chen, "A transfer learning approach for microstructure reconstruction and structure-property predictions," *Scientific reports*, vol. 8, no. 1, pp. 1–13, 2018.

[27] M. H. Segler, T. Kogej, C. Tyrchan, and M. P. Waller, "Generating focused molecule libraries for drug discovery with recurrent neural networks," *ACS central science*, vol. 4, no. 1, pp. 120–131, 2018.

[28] D. Jha, K. Choudhary, F. Tavazza, W.-k. Liao, A. Choudhary, C. Campbell, and A. Agrawal, "Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning," *Nature communications*, vol. 10, no. 1, pp. 1–12, 2019.

[29] B. Kailkhura, B. Gallagher, S. Kim, A. Hiszpanski, and T. Y.-J. Han, "Reliable and explainable machine-learning methods for accelerated material discovery," *npj Computational Materials*, vol. 5, no. 1, pp. 1–9, 2019.

[30] R. E. Goodall and A. A. Lee, "Predicting materials properties without crystal structure: Deep representation learning from stoichiometry," *Nature communications*, vol. 11, no. 1, pp. 1–9, 2020.

[31] P. K. Routh, Y. Liu, N. Marcella, B. Kozinsky, and A. I. Frenkel, "Latent representation learning for structural characterization of catalysts," *The Journal of Physical Chemistry Letters*, vol. 12, no. 8, pp. 2086–2094, 2021.

[32] V. Gupta, K. Choudhary, F. Tavazza, C. Campbell, W.-k. Liao, A. Choudhary, and A. Agrawal, "Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data," *Nature communications*, vol. 12, no. 1, pp. 1–10, 2021.

[33] C. Shu, J. He, G. Xue, and C. Xie, "Grain knowledge graph representation learning: A new paradigm for microstructure-property prediction," *Crystals*, vol. 12, no. 2, p. 280, 2022.

[34] V. Gupta, W.-k. Liao, A. Choudhary, and A. Agrawal, "Pre-activation based representation learning to enhance predictive analytics on small materials data," in *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2023, pp. 1–8.

[35] V. Gupta, K. Choudhary, B. DeCost, F. Tavazza, C. Campbell, W.-k. Liao, A. Choudhary, and A. Agrawal, "Structure-aware graph neural network based deep transfer learning framework for enhanced predictive analytics on diverse materials datasets," *npj Computational Materials*, vol. 10, no. 1, p. 1, 2024.

[36] B. Komer, J. Bergstra, and C. Eliasmith, "Hyperopt-sklearn: automatic hyperparameter configuration for scikit-learn," in *ICML workshop on AutoML*, vol. 9. Citeseer, 2014, p. 50.