SOFTWARE

Materials Property Predictor for Crystalline Materials Using Cross-Property Deep Transfer Learning Framework

Vishu Gupta[1], Kamal Choudhary[2,3,4], Yuwei Mao[1], Kewei Wang[1], Francesca Tavazza[2], Carelyn Campbell[2], Wei-keng Liao[1], Alok Choudhary[1] and Ankit Agrawal[1*]

[*]Correspondence:
ankitag@eecs.northwestern.edu
[1]Department of Electrical and Computer Engineering, Northwestern University, Evanston, USA
Full list of author information is available at the end of the article

Abstract

The application of machine learning and deep learning techniques in the field of materials science is becoming increasingly common due to their promising ability to extract and utilize data-driven information from available materials data and accelerate materials discovery and design for future applications. In an attempt to assist with the process, we develop and deploy predictive models for multiple materials properties, given the composition of the material. The deep learning models described here are built using cross-property deep transfer learning technique which leverages source models trained on large datasets to build target models on small datasets with different properties. These models are deployed in an online software tool that takes a number of material compositions as input, performs pre-processing to generate composition-based attributes for each material, and feeds them into the predictive models to obtain up to 41 different materials property values. The online materials property predictor is available at http://ai.eecs.northwestern.edu/MPpredictor

Keywords: Materials Informatics; Supervised Learning; Density Functional Theory; Experimental Dataset; Transfer Learning

Introduction

Background

Traditionally, the field of materials science and engineering has involved conducting simulations and hands-on experiments in the lab to understand and discover new materials with desired properties. Over time, the data generated by such simulations and experiments has accumulated to a point that it has become possible to apply data-driven methods to build predictive models for materials properties. This led to the emergence of a new sub-field in materials science called materials informatics [1, 2, 3, 4, 5] also known as the fourth paradigm of science [6] which tries to unify first three paradigms of experiment, theory, and simulation. Recent years have seen a surge in research works that utilize data-driven methodologies to predict and optimize materials properties [7, 8, 9, 10, 11, 12, 13, 14]. In order to realize the vision of development of advanced materials, the US government launched the Materials Genome Initiative (MGI) [15] in 2011. The initiative aimed to reduce the time lag between the discovery of materials and their deployment to half at a fraction of the cost. Materials Genome Initiative Strategic Plan [16, 17], released and expanded in 2014 and 2021 respectively, also recognizes data analytics

and artificial intelligence as one of the main tools to integrate advanced modeling, computational and experimental tools, and quantitative data, to realize the vision of MGI. It is in the spirit and pursuit of the vision and approach of MGI that we discuss and present in this software article, an online materials informatics tool to predict 41 materials properties, including many properties for which large datasets are not available, making it harder to generate a highly accurate model by traditional model training methods. The online materials informatics tool deployed in this work can play a crucial role in the process of materials discovery and design.

Materials Representation

In this study, we mainly focus on two types of the most commonly used composition-based inputs, raw elemental fractions (EF) and MAGPIE (Materials Agnostic Platform for Informatics and Exploration) based physical attributes (PA) due to their consistency in providing promising results when used with powerful deep learning (DL) techniques [11, 18, 19]. EF is composed of 86 composition based-attributes where each attribute represents an element in the periodic table. For example, RbCl is represented as a 1D vector with 86 entries, with the Rb and Cl columns containing 0.5 and other entries as 0. PA consists of elemental property statistics, stoichiometric attributes, electronic structure attributes, and ionic compound attributes which are from domain perspective more informative set of descriptors as compared to EF.

Density functional theory

Most of the datasets used in this study are obtained using Density functional theory (DFT), which is a quantum mechanical simulation technique based on the electron density within the crystal structure of the material and is one of the most commonly used computational tools for studying the electronic scale properties of a material. As these calculations require atomistic structure of the materials we want to study as input, it can be extremely computationally costly and time consuming in nature where a single DFT calculation can take long amount of time (on the order of hours, days, or even weeks) depending on the size and complexity of the material being studied on modern computing systems.

Implementation

As this work focuses on improving the predictive model for materials properties for which a large number of data points are not available, we create a robust workflow that is divided into several steps. First, we perform data acquisition where different DFT-based and experimental datasets with various materials properties are collected. We then generate compositional attributes that consist of elemental fractions (EF) and physical attributes (PA) for each of the datasets to construct the database for each of the material property. We then deploy cross-property transfer learning framework [20] comprised of supervised learning techniques to determine the predictive models for material properties, given the composition of the materials without the use of structure-related attributes. The models trained using the cross-property transfer learning framework consist of a base model, scratch (SC) models and transfer learning (TL) models. The base model is a baseline model that uses the

average property value of the training data as the predicted value. For SC models, the model training is performed directly on the target dataset without providing the model with any form of pre-trained knowledge. SC models consist of traditional machine learning (ML) and deep learning (DL) models with EF and PA as model inputs. For TL models, the model training is performed by providing the target dataset with a pre-trained model trained on source dataset using only EF as the model input. The TL techniques incorporated in this work include fine-tuning, and feature extraction methods. We do not use traditional ML models for the TL-based model training as it has been shown to not benefit the results much but takes a lot of computational time and resources [20]. Source models used in the framework are obtained by training six different source properties: formation enthalpy, band gap, magnetic moment, stability, total energy, volume. We evaluate the models using standard validation techniques and using the most accurate models for each of the materials properties for predictive analysis in an online user-friendly software tool that can take a list of compositions as input. The whole workflow is shown in Figure 1.

Result

Dataset

We used the same dataset and material properties as used in [20], which consists of two datasets of DFT-computed materials properties (Open Quantum Materials Database, OQMD [21] and Joint Automated Repository for Various Integrated Simulations, JARVIS [22]) and two datasets of experimentally-measured materials properties in this work. Detailed descriptions of the datasets used in this work are shown in Table 1. For a detailed explanation of all the pre-processing performed (which mainly includes duplicate entry removal, property value range selection and attributes generation) on the datasets, the reader is referred to [20].

Methods

A deep learning framework, ElemNet [18], was implemented using Python and TensorFlow 2 [23] and Keras [24] and has shown to excel in producing a strong predictive model on raw inputs similar to [19, 25]. The traditional ML models used in this work are Elastic Net, Ada Boost, Stochastic Gradient Descent (SGD) regression, K-Nearest Neighbors, Ridge, Support Vector Machine, Linear Regression, Lasso, Extra Tree, Bagging, Random Forest, and Decision Tree. For all ML models, we carry out an extensive hyperparameter grid search to find the best hyperparameters. Readers interested in details of hyperparameters are referred to [20]. We use Mean Absolute Error (MAE) as the evaluation metric for all the models. We divide the available data for each property into training and validation sets in 9:1 ratio without the holdout test set to maximize the data used for model training, since these models are developed for deployment. Note that results on holdout test set are provided in [20].

Comparison Against Scratch Models

Table 2 presents the prediction accuracy of the best scratch (SC) and best transfer learning (TL) model selected based on validation MAE values for each of the 41 target properties.

As we can observe from Table 2, the TL models perform better than SC models in 31/41 cases, i.e., in $\approx$ 76 % of the cases. Additionally, we calculate one-tailed p-value for comparing the validation MAEs obtained on 41 target datasets (39 DFT-computed and two experimental datasets) to see if the observed improvement in accuracy of TL models over SC models is significant. As we are dealing with different datasets with different properties, whose MAE differences may not be directly comparable [26], we use Signed Test [27, 28] to calculate the p-value. Here, the null hypothesis is that "TL model is not better than SC model", and the alternate hypothesis is "TL model is better than SC model". After performing the sign test using sign test calculator [29], we get the p-value = 0.00052, thus rejecting the null hypothesis at α=0.01, suggesting that the difference in validation MAE between SC and TL models is unlikely to have arisen by chance. We can thus infer that in general TL models perform significantly better than SC models. Given these results, we believe that the machine learning and deep learning models trained and deployed in this work could be very useful for quickly and accurately estimating the different properties of materials (note that for the 39 DFT target properties, the prediction is by definition at 0 K temperature), which can help scan a large number of compositions in a short time for new promising materials with desired properties without expensive DFT calculations or experiments.

Materials Property Predictor

We have created an online materials property predictor that takes a list of compositions (satisfying the charge balance condition and common oxidation states of individual elements) as input, and returns predictions of materials properties. The best model with the least error for each of the materials properties (among the SC and TL) is used for prediction. The results are presented in the form of a table with columns representing each of the compositions given as input and rows representing the selected materials properties from the best model. Note that these models are trained on the known most stable structure for a given composition, and thus only take the composition as input. In other words, no structure information is needed to get the property predictions from this software tool, which on the one hand is advantageous as structure information can be unavailable or difficult to obtain in many cases, while on the other hand, these models cannot distinguish between structure polymorphs of a given composition. We believe that given the huge chemical space, composition-based models such as those deployed in this software tool can be instrumental in first identifying promising composition systems of elemental combinations, which could subsequently be analyzed further with structure-prediction methods and structure-aware property prediction methods. The screenshot of the main and result page of the materials property predictor is depicted in Figures 2 and 3, and the tool is available online at: http://ai.eecs.northwestern.edu/MPpredictor.

Conclusion/Significance

From a research point of view, this work explores the applicability of predictive modeling techniques to predict material properties. Unlike many other existing works which primarily focus on predicting a few materials properties [12, 13], here we perform materials property prediction for a wide range of materials properties, by

comparing the best SC and TL models on DFT and experimental datasets. We used the entire dataset to train the model unlike [20] to get more accurate models for deployment. From a practical perspective, we have deployed the most accurate predictive models in a software tool with a user-friendly and easy-to-access design. To make the developed predictive models for various materials properties readily accessible for use by the materials science and engineering community, we have created an online materials property predictor that can take a list of materials compositions as input and predict the materials properties of choice. The main advantage of this software tool is its ability to accurately predict multiple materials properties in a matter of seconds by just using the chemical composition of the material without the need for any structure-related information, which is, in general, hard to obtain and is also required when performing expensive DFT simulations. The deployed software tool is expected to be a valuable resource to assist in the process of searching for better materials with improved properties for researchers and practitioners in the materials science and engineering community. The code used to develop the online software tool in this work is also publicly available at https://github.com/GuptaVishu2002/MPpredictor.

Funding
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 award CMMI-2053929, and DOE awards DE-SC0019358, DE-SC0021399, and Northwestern Center for Nanocombinatorics.

Availability of data and materials
The code used to develop the online software tool in this work is also publicly available at https://github.com/GuptaVishu2002/MPpredictor. The codes and data required to perform TL used in this study is available at https://doi.org/10.5281/zenodo.5533023.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
V.G. designed and carried out the implementation and experiments for the Materials Property Predictor with the help of Y.M. and K.W. under the guidance of A.A., A.C., and W.L.. K.C., F.T., and C.C. provided the necessary domain expertise for this work. V.G., A.A., K.C. and F.T. wrote the manuscript. All authors discussed the results and reviewed the manuscript.

Author details
[1]Department of Electrical and Computer Engineering, Northwestern University, Evanston, USA. [2]Materials Measurement Laboratory, National Institute of Standards and Technology, 20899 Gaithersburg, USA. [3]Theiss Research, 92037 La Jolla, USA. [4]DeepMaterials LLC, 20906 Silver Spring, USA.

References

1. Agrawal, A., Choudhary, A.: Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. Apl Materials **4**(5), 053208 (2016)
2. Kalidindi, S.R., De Graef, M.: Materials data science: current status and future outlook. Annual Review of Materials Research **45**, 171–193 (2015)
3. Rajan, K.: Materials informatics: The materials "gene" and big data. Annual Review of Materials Research **45**, 153–169 (2015)
4. Agrawal, A., Choudhary, A.: Deep materials informatics: Applications of deep learning in materials science. MRS Communications **9**(3), 779–792 (2019)
5. Choudhary, K., DeCost, B., Chen, C., Jain, A., Tavazza, F., Cohn, R., Park, C.W., Choudhary, A., Agrawal, A., Billinge, S.J., *et al.*: Recent advances and applications of deep learning methods in materials science. npj Computational Materials **8**(1), 1–26 (2022)
6. Hey, A.J., Tansley, S., Tolle, K.M., *et al.*: The Fourth Paradigm: Data-intensive Scientific Discovery vol. 1. Microsoft research Redmond, WA, ??? (2009)
7. Gopalakrishnan, K., Agrawal, A., Ceylan, H., Kim, S., Choudhary, A.: Knowledge discovery and data mining in pavement inverse analysis. Transport **28**(1), 1–10 (2013)
8. Agrawal, A., Deshpande, P.D., Cecen, A., Basavarsu, G.P., Choudhary, A.N., Kalidindi, S.R.: Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters. Integrating Materials and Manufacturing Innovation **3**(1), 90–108 (2014)

9. Liu, R., Kumar, A., Chen, Z., Agrawal, A., Sundararaghavan, V., Choudhary, A.: A predictive machine learning approach for microstructure optimization and materials design. Scientific reports **5**(1), 1–12 (2015)
10. Liu, R., Yabansu, Y.C., Agrawal, A., Kalidindi, S.R., Choudhary, A.N.: Machine learning approaches for elastic localization linkages in high-contrast composite materials. Integrating Materials and Manufacturing Innovation **4**(1), 192–208 (2015)
11. Ward, L., Agrawal, A., Choudhary, A., Wolverton, C.: A general-purpose machine learning framework for predicting properties of inorganic materials. npj Computational Materials **2**(1), 1–7 (2016)
12. Agrawal, A., Choudhary, A.: A fatigue strength predictor for steels using ensemble data mining: steel fatigue strength predictor. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 2497–2500 (2016)
13. Agrawal, A., Meredig, B., Wolverton, C., Choudhary, A.: A formation energy predictor for crystalline materials using ensemble data mining. In: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), pp. 1276–1279 (2016). IEEE
14. Jha, D., Gupta, V., Liao, W.-k., Choudhary, A., Agrawal, A.: Moving closer to experimental level materials property prediction using ai. Scientific reports **12** (2022)
15. Science, N., (US), T.C.: Materials Genome Initiative for Global Competitiveness. Executive Office of the President, National Science and Technology Council, ??? (2011)
16. Drosback, M.: Materials genome initiative: advances and initiatives. JOM **66**(3), 334 (2014)
17. Initiative, M.G.: Strategic Plan. https://www.mgi.gov. Accessed = 2022-06-14 (2021)
18. Jha, D., Ward, L., Paul, A., Liao, W.-k., Choudhary, A., Wolverton, C., Agrawal, A.: ElemNet: Deep learning the chemistry of materials from only elemental composition. Scientific reports **8**(1), 17593 (2018)
19. Jha, D., Gupta, V., Ward, L., Yang, Z., Wolverton, C., Foster, I., Liao, W.-k., Choudhary, A., Agrawal, A.: Enabling deeper learning on big data for materials informatics applications. Scientific reports **11**(1), 1–12 (2021)
20. Gupta, V., Choudhary, K., Tavazza, F., Campbell, C., Liao, W.-k., Choudhary, A., Agrawal, A.: Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data. Nature communications **12**(1), 1–10 (2021)
21. Kirklin, S., Saal, J.E., Meredig, B., Thompson, A., Doak, J.W., Aykol, M., Rühl, S., Wolverton, C.: The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. npj Computational Materials **1**, 15010 (2015)
22. Choudhary, K., Garrity, K.F., Reid, A.C.E., DeCost, B., Biacchi, A.J., Walker, A.R.H., Trautt, Z., Hattrick-Simpers, J., Kusne, A.G., Centrone, A., Davydov, A., Jiang, J., Pachter, R., Cheon, G., Reed, E., Agrawal, A., Qian, X., Sharma, V., Zhuang, H., Kalinin, S.V., Sumpter, B.G., Pilania, G., Acar, P., Mandal, S., Haule, K., Vanderbilt, D., Rabe, K., Tavazza, F.: JARVIS: An Integrated Infrastructure for Data-driven Materials Design (2020). 2007.01831
23. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016)
24. Chollet, F., et al.: Keras. GitHub (2015)
25. Gupta, V., Liao, W.-k., Choudhary, A., Agrawal, A.: Brnet: Branched residual network for fast and accurate predictive modeling of materials properties. In: Proceedings of the 2022 SIAM International Conference on Data Mining (SDM), pp. 343–351 (2022). SIAM
26. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research **7**, 1–30 (2006)
27. Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures. Chapman and Hall/CRC, . (2003)
28. Salzberg, S.L.: On comparing classifiers: Pitfalls to avoid and a recommended approach. Data mining and knowledge discovery **1**(3), 317–328 (1997)
29. Statistics, S.S.: Sign Test Calculator. https://www.socscistatistics.com/tests/signtest/default.aspx. Accessed = 2022-06-10 (2018)
30. Kim, G., Meschel, S.V., Nash, P., Chen, W.: Experimental formation enthalpies for intermetallic phases and other inorganic compounds. Scientific Data **4**, 170162 (2017). Data Descriptor
31. Zhuo, Y., Mansouri Tehrani, A., Brgoch, J.: Predicting the band gaps of inorganic solids by machine learning. The Journal of Physical Chemistry Letters **9**(7), 1668–1673 (2018). doi:10.1021/acs.jpclett.8b00124. PMID: 29532658. https://doi.org/10.1021/acs.jpclett.8b00124

Figure 1 Block diagram depicting the workflow used in this work.

Figures

Figure 2 Screenshot of the main page of the deployed materials property predictor.

Tables

Figure 3 Screenshot of the result page of the deployed materials property predictor.

Table 1 Description of the datasets used in this work.

Dataset	Data Size	# Properties
OQMD [21]	341,443	6
JARVIS [22]	28,171	39
Experimental Formation Enthalpy [30]	1643	1
Experimental Bandgap [31]	4920	1

Table 2 Prediction performance in terms of validation MAE of the best SC and TL model selected for each of the target materials properties.

Property	Data Size	Base	Best SC	Best TL
Kpoints Length Unit (Å)	28056	18.77	**10.69**	11.15
Kpoints Array Average (Å)	28171	5.232	**2.692**	2.779
Bandgap Optb88vdw (eV)	28163	0.987	0.253	**0.228**
Formation Energy (eV/atom)	28155	0.850	0.127	**0.115**
Encut (eV)	28108	246.2	**76.02**	80.04
Ehull (eV/atom)	27297	0.130	0.054	**0.0471**
Magmom Oszicar (μ_B)	25844	1.222	0.408	**0.377**
Magmom Outcar (μ_B)	25357	1.172	0.385	**0.343**
Eps Refractive Index (x)	25150	3.829	1.259	**1.204**
P Powerfact (Wm_1K_1)	16250	650.2	486.3	**477.9**
N Powerfact (Wm_1K_1)	16250	657.8	478.1	**477.1**
P Effective Masses 300K Average (kg)	16763	1.917	**1.081**	1.109
N Effective Masses 300K Average (kg)	16760	1.917	**1.081**	1.120
P Seebeck ($\mu V/K$)	14439	163.1	**56.14**	56.53
N Seebeck ($\mu V/K$)	14144	108.6	**48.47**	49.73
Meps Refractive Index (x)	11349	4.904	1.738	**1.726**
Max (Phonon) Mode (cm^{-1})	10963	284.6	**55.97**	62.12
Min (Phonon) Mode (cm^{-1})	10930	40.76	22.89	**21.57**
Elastic Tensor C11 (GPa)	10839	81.56	34.25	**32.19**
Elastic Tensor C12 (GPa)	10759	44.95	16.52	**16.49**
Elastic Tensor C13 (GPa)	10846	42.51	13.86	**13.46**
Elastic Tensor C22 (GPa)	10832	84.01	33.70	**31.26**
Elastic Tensor C33 (GPa)	10856	84.03	35.11	**33.16**
Elastic Tensor C44 (GPa)	9986	29.52	14.94	**14.51**
Elastic Tensor C55 (GPa)	9755	26.63	12.35	**11.25**
Elastic Tensor C66 (GPa)	9739	27.56	13.43	**12.68**
Bulk Modulus KV (GPa)	10743	49.07	11.36	**10.52**
Shear Modulus GV (GPa)	10209	24.21	11.05	**10.28**
Bandgap MBJ (eV)	7296	1.910	0.539	**0.500**
Spillage ($Å^{-1}$)	3866	0.499	0.349	**0.334**
SLME (%)	3006	9.442	6.636	**5.827**
Max Ir Mode (cm^{-1})	2302	422.5	**83.68**	95.71
Min Ir Mode (cm^{-1})	2268	66.13	**37.39**	40.72
Dfpt Piezo Max Dielectric Electronic (ε_{11})	2126	5.715	2.809	**2.520**
Dfpt Piezo Max Dielectric (ε_{11})	2126	6.953	3.437	**3.206**
Dfpt Piezo Max Dielectric Ioonic (ε_{11})	2126	2.559	0.783	**0.691**
Dfpt Piezo Max Eij (cm^{-2})	1123	0.515	0.373	**0.359**
Dfpt Piezo Max Dij (cm^{-2})	689	44.05	21.51	**19.95**
Exfoliation Energy (eV/atom)	557	61.29	37.72	**35.74**
Experimental Formation Energy (eV/atom)	1643	1.033	0.101	**0.072**
Experimental Bandgap (eV)	4920	1.205	0.423	**0.348**