Data-centric approach to improve machine learning models for inorganic materials

Christopher J. Bartel1,*
1Department of Materials Science and Engineering, University of California, Berkeley, Berkeley, CA, USA
*Correspondence: cbartel@berkeley.edu
https://doi.org/10.1016/j.patter.2021.100382

Pandey et al. (2021) demonstrate the importance of diversifying training data to make balanced predictions of thermodynamic properties for inorganic crystals.

The big-data revolution has arrived for materials science. Open materials databases, such as the Materials Project1 and NRELMatDB,2 house the results of tens to hundreds of thousands of density-functional theory (DFT) calculations for inorganic materials. These databases have transformed the way computational and experimental materials scientists perform research, allowing users to query the properties of materials spanning the periodic table with a few lines of code or a few clicks of the mouse.3 This access to materials properties—until recently—was unprecedented and is ripe for the picking with machine learning (ML) models, and indeed materials databases are the typical starting point for any foray in the materials informatics space. In the recent work by Pandey et al.,4 the authors ask the question, “do these databases have the best set of materials for training generalizable models”?

There has been a recent push in the ML community toward data-centric instead of model-centric development. For years in ML, researchers have focused on improving the code, the data, and algorithms applied to some training tasks for benchmark datasets. This embodies the model-centric approach to ML. However, when embarking on a new problem where a carefully curated dataset has not yet been developed, a substantial fraction of researcher effort often goes into data generation, cleaning, processing, etc. A data-centric approach tries to systematically improve the quality of training (and validation) data to enhance the accuracy and generalizability of a more-or-less fixed model architecture. This push toward data engineering is embodied in the recently introduced Data-Centric AI Competition,5 where DeepLearning.AI and LandingAI are providing competitors an initial dataset and fixed ML model and judging who can best engineer the data to maximize the model’s performance.

For open materials databases, the initial dataset typically begins with the Inorganic Crystal Structure Database (ICSD),6 which houses crystallographic data (compositions and site occupations) for >100,000 inorganic materials, most of which have been synthesized. This is a great starting point for ML, as it accounts for essentially all the solid-state inorganic materials known to date, but if we want to develop models that can predict new materials, the ICSD may not provide a sufficient training dataset. Indeed, Pandey et al. cleverly show that by augmenting their training dataset to include DFT calculations of hypothetical non-ground-state structures, in addition to known structures from the ICSD, they generate models that perform well for both low-energy (mostly known) materials and high-energy (mostly unknown) materials.

Starting with a typical crystal graph convolutional neural network (CGCNN) representation, introduced in 2018,7 Pandey et al. first train their representation on a set of ~15,000 structures taken from the ICSD. They show that their CGCNN performs quite well for a left-out set of ~500 ICSD structures, achieving a mean absolute error (MAE) of ~40 meV/atom on the DFT total energy, approaching the resolution of DFT with respect to experiment8 and comparable to previously reported models. However, problems arise when this ICSD-trained model is asked to predict the energy of hypothetical crystal structures (i.e., structures that have not been synthesized and logged in the ICSD). For ~6,000 of these materials, the model achieves an MAE of ~240 meV/atom, ~6x worse than on ICSD structures, and predominantly predicts energies that are far too negative (i.e., it suggests that these hypothetical structures should be more stable than they are). This is especially problematic when we consider how ML models might be applied to materials discovery in practice. Because materials in the ICSD have already been discovered, we are not as interested in predictions of their energies except as potential competing phases for newly proposed hypothetical materials. Instead, it is important that we accurately capture the relative energetics of hypothetical candidate materials compared to these known compounds in the ICSD. Models that overestimate the stability of hypothetical candidates will inevitably have high false positive rates and lead to wasted resources investigating materials predicted to be stable that will not be upon further investigation.9

Pandey et al. show this problem can be remedied by simply including hypothetical materials during training (Figure 1). Adding ~10,000 hypothetical crystal structures, predominantly comprised of thermodynamically unstable materials, to the training database results in a model that achieves ~40 meV/atom MAEs on an excluded test set of ICSD structures (on par with the previous model’s performance) and also an excluded test set of high-energy hypothetical structures for which the previous model (trained only on the ICSD) failed. This beneficial effect of diversifying the training dataset is further demonstrated by improving the prediction of polymorph energy orderings for a diverse set of compounds—i.e., the model trained on both low- and high-energy structures does a better job of...
determining which structures are potentially accessible for a given chemical composition. The advantage of including high-energy structures can most succinctly be summed up from the authors’ thermodynamic stability analysis, where, for a set of ~2,000 materials, the false positive rate (materials predicted to be stable by the models that are not calculated to be stable with DFT) is significantly decreased from ~50% to < 2% by including hypothetical structures during training.

The work by Pandey et al. presents compelling evidence that training datasets for ML models should include materials outside those that have already been discovered and justifies the inclusion of hypothetical compounds in open materials databases. As pointed out by the authors, there remains a major caveat when considering the performance of these models for genuine materials discovery problems. To date, models are almost exclusively trained and tested on DFT-relaxed crystal structures, but if we want to chart the space of unexplored materials without first-principles calculations, we will need ML models that generate reasonable crystal structures for candidate compositions. This is the next major challenge that must be addressed for ML to enable the rapid discovery of novel and synthesizable materials with exciting properties.

REFERENCES

1. Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., et al. (2013). Commentary: The Materials Genome approach to accelerating materials innovation. APL Mater. 1, 011002.
2. NREL MatDB. NREL Materials Database. https://materials.nrel.gov/.
3. Horton, M.K., Dwaraknath, S., and Persson, K.A. (2021). Promises and perils of computational materials databases. Nature Computational Science 1, 3–6.
4. Pandey, S., Qu, J., Stevanović, V., St. John, P., and Gori, P. (2021). Predicting energy and stability of known and hypothetical crystals using graph neural network. Patterns 2. https://doi.org/10.1016/j patter.2021.100361.
5. DeepLearningAI. (2021). Data-Centric A.I. Competition. https://https-deeplearning-ai.github.io/data-centric-mp/.
6. Hellenbrandt, M. (2004). The Inorganic Crystal Structure Database (ICSD)—Present and Future. Crystallogr. Rev. 10, 17–22.
7. Xie, T., and Grossman, J.C. (2018). Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. Phys. Rev. Lett. 120, 145301.
8. Zhang, Y., Kitchaev, D.A., Yang, J., Chen, T., Dacek, S.T., Sarmentio-Pérez, R.A., Marques, M.A.L., Peng, H., Ceder, G., Perdew, J.P., et al. (2018). Efficient first-principles prediction of solid stability: Towards chemical accuracy. npj Computational Materials 4, 9.
9. Bartel, C.J., Trewhatha, A., Wang, Q., Dunn, A., Jain, A., and Ceder, G. (2020). A critical examination of compound stability predictions from machine-learned formation energies. npj Computational Materials 6, 97.

Figure 1. Improving generalizability to high- and low-energy structures
Training a model only on known (ICSD) structures leads to poor predictions for high-energy structures (left). Adding hypothetical structures to the training data dramatically improves performance on high-energy structures without losing accuracy on ground-state structures (right). Note that the axes in the right panel have been expanded to show the full range of energies whereas the axes in the left panel highlight “high-energy” structures.