MATBENCH DISCOVERY
CAN MACHINE LEARNING IDENTIFY STABLE CRYSTALS?

Janosh Riebesell
University of Cambridge, Lawrence Berkeley National Laboratory
janosh@lbl.gov

Rhys Goodall
University of Cambridge

Anubhav Jain
Lawrence Berkeley National Laboratory

Kristin Persson
University of California - Berkeley, Lawrence Berkeley National Laboratory

Alpha Lee
University of Cambridge

ABSTRACT

We present a new machine learning (ML) benchmark for thermodynamic materials stability predictions named `Matbench Discovery`. A goal of this benchmark is to highlight the need to focus on metrics that directly measure their utility in prospective discovery campaigns as opposed to analyzing models based on predictive accuracy alone. Our benchmark consists of a task designed to closely simulate the deployment of ML energy models in a high-throughput search for stable inorganic crystals. We explore a wide variety of models covering multiple methodologies ranging from random forests to GNNs, and from one-shot predictors to iterative Bayesian optimizers and interatomic potential-based relaxers. We find M3GNet to achieve the highest F1 score of 0.58 and R^2 of 0.59 while MEGNet wins on discovery acceleration factor (DAF) with 2.70. Our results provide valuable insights for maintainers of high throughput materials databases to start using these models as triaging steps to more effectively allocate compute for DFT relaxations.

1 INTRODUCTION

For nearly two decades, ever since the work of Behler and Parrinello (Behler & Parrinello, 2007) who introduced a custom neural network for learning the density-functional theory (DFT) potential energy surface (PES), material scientists have devoted significant effort to developing custom model architectures for tackling the problem of learning the PES. Initially, most of these models were trained and deployed as interatomic potentials to study known materials of interest which required curating custom training data for each application (Bartók et al., 2018; Deringer et al., 2020). As larger and more diverse datasets emerged from initiatives like the Materials Project (MP) (Jain et al., 2013) or the Open Quantum Materials Database (OQMD) (Saal et al., 2013), researchers have begun to train models that cover the full periodic table opening up the prospect of ML-guided materials discovery.

Yet despite many advances in ML for materials discovery, it is unclear which methodology performs best at predicting material stability. Recent areas of progress include one-shot predictors like Wren (Goodall et al., 2022), universal force predictors such as M3GNet (Chen & Ong, 2022) that emulate density functional theory to relax crystal structures according to Newton's laws, and Bayesian optimizers like BOWSR that paired with any energy model treat structure relaxation as a black-box optimization problem (Zuo et al., 2021). In this work, we aim to answer which of these is the winning methodology in a future-proof benchmark that closely simulates using ML to guide a real-world discovery campaign.

2 Related Work

2.1 Using Stability rather than Formation Energies

In 2020, Chris Bartel et al. (Bartel et al., 2020) benchmarked 7 models, finding all of them able to predict DFT formation energies with useful accuracy. However, when asked to predict stability (specifically decomposition enthalpy), the performance of all models deteriorated sharply. This insight meant that ML models are much less useful than DFT for discovering new solids than prior studies had suggested. The paper identified two main reasons for the sharp decline in predictive power: 1. Stability is a property not only of the material itself but also the chemical space of competing phases it occupies. Current ML algorithms are given an input that only describes the single material they are asked to predict, leaving them clueless of competing phases. 2. Unlike DFT, ML models appear to benefit less from systematic error cancellation across similar chemistries.

Bartel et al. showed that to demonstrate the utility of ML for materials discovery, the vanity metric of formation energy accuracy must be replaced with stability predictions. Moreover, the qualitative leap in performance from Roost (Goodall & Lee, 2020), the best compositional model benchmarked, to CGCNN (Xie & Grossman, 2018), the single structural model they tested, shows structure plays a crucial role in determining the stability of materials. However, using the DFT-relaxed structure as input to CGCNN renders the discovery pipeline circular as the input becomes the very thing we aim to find. A true test of prospective utility requires using unrelaxed structures as the next most information-rich input.

2.2 Matbench

As the name suggests, this work seeks to expand upon the original Matbench suite of property prediction tasks (Dunn et al., 2020). By providing a standardized collection of datasets along with canonical cross-validation splits for model evaluation, Matbench helped focus the field of ML for materials, increase comparability across papers and attempt to accelerate the field similar to what ImageNet did for computer vision.

Matbench released a test suite of 13 supervised tasks for different material properties ranging from thermal (formation energy, phonon frequency peak), electronic (band gap), optical (refractive index) to tensile and elastic (bulk and shear moduli). They range in size from ~300 to ~132,000 samples and include both DFT and experimental data sources. 4 tasks are composition-only while 9 provide the relaxed crystal structure as input. Importantly, all tasks were exclusively concerned with the properties of known materials. We believe a task that simulates a materials discovery campaign by requiring materials stability predictions from unrelaxed structures to be a missing piece here.

2.3 The Open Catalyst Project

The Open Catalyst Project (OCP) is a large-scale initiative to discover substrate-absorbate combinations that catalyze key industrial reactions processing said absorbates into more useful products. The OCP has released two data sets thus far, OCP20 (Chanussot et al., 2021) and OCP22 (Tran et al., 2022), that can be used for training and benchmarking ML models.

However, the ambition and scale of OCP comes with limitations for its use as a benchmark. OCP20 is already 10x larger than the largest crystal structure data sets available imposing a high barrier to entry for researchers without access to cloud-scale computing. In contrast, we believe the discovery of stable materials is a problem where ML methods have matured enough to be usefully deployed at scale after training for only $\mathcal{O}(10^2)$ GPU hours.

3 Data Sets

The choice of data for the train and test sets of this benchmark fell on the latest Materials Project (MP) (Jain et al., 2013) database release (2021.05.13 at time of writing) and the WBM dataset (Wang et al., 2021).

3.1 The Materials Project - Training Set

The Materials Project is a well-known effort to calculate the properties of all inorganic materials using high-throughput ab-initio methods. At the time of access, the Materials Project database contains approximately 154k crystals (providing relaxed+initial structure and the relaxation trajectory for each of them) covering a diverse range of chemistries. For our benchmark, the training set is all data available from the 2021.05.13 MP release. Models are free to train on relaxed and/or unrelaxed structures or the full DFT relaxation trajectory. This flexibility is intended to allow authors to experiment and exploit the large variety of data available.

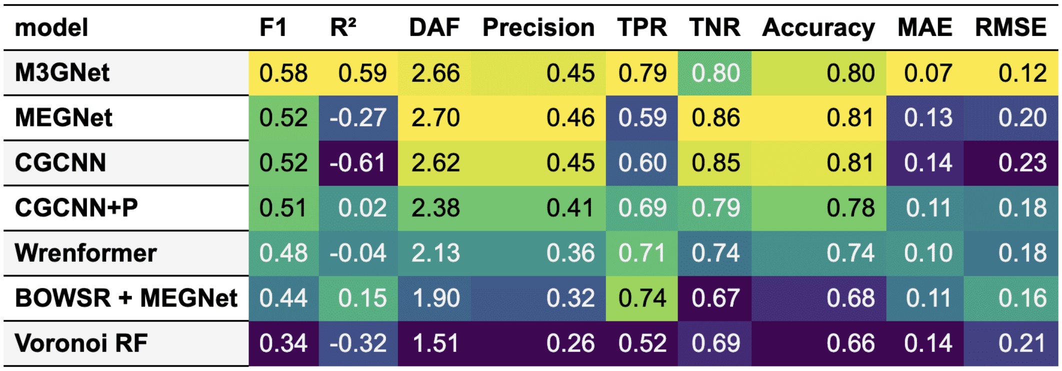
3.2 WBM - Test Set

The WBM data set (Wang et al., 2021) consists of ~257k structures generated via chemical similarity-based elemental substitution of MP source structures followed by DFT relaxation and convex hull distance calculation. Throughout this work, we define stability in terms of being on or below the convex hull of the MP training set. ~42k out of ~257k materials in WBM satisfy this criterion. As WBM explores regions of materials space not well sampled by MP, many of these materials discovered that are stable w.r.t. MP's convex hull are not stable with respect to each other. Only around ~20k were found to remain on the convex hull when merging the MP and WBM hulls. This observation highlights a critical aspect of this benchmark in that we purposely operate with an incomplete convex hull. Only current knowledge is accessible to a real discovery campaign. Hence our metrics are designed to reflect this.

Moreover, to simulate a discovery campaign our test set inputs are unrelaxed structures obtained from element substitution of MP source structures but our target labels are the relaxed PBE formation energies. This opens up the opportunity to explore how different approaches (one-shot, force-based pseudo-relaxation, black-box pseudo-relaxation, etc.) compare for materials discovery.

4 Models

Our initial benchmark release includes 8 models. We present metrics for all in table 1 but focus on the 6 best performers in fig. 1 for visual clarity.



model	F1	R²	DAF	Precision	TPR	TNR	Accuracy	MAE	RMSE
M3GNet	0.58	0.59	2.66	0.45	0.79	0.80	0.80	0.07	0.12
MEGNet	0.52	-0.27	2.70	0.46	0.59	0.86	0.81	0.13	0.20
CGCNN	0.52	-0.61	2.62	0.45	0.60	0.85	0.81	0.14	0.23
CGCNN+P	0.51	0.02	2.38	0.41	0.69	0.79	0.78	0.11	0.18
Wrenformer	0.48	-0.04	2.13	0.36	0.71	0.74	0.74	0.10	0.18
BOWSR + MEGNet	0.44	0.15	1.90	0.32	0.74	0.67	0.68	0.11	0.16
Voronoi RF	0.34	-0.32	1.51	0.26	0.52	0.69	0.66	0.14	0.21

Table 1: Regression and classification metrics for all models tested on our benchmark. The heat map ranges from yellow (best) to blue (worst) performance. DAF = discovery acceleration factor (see text), TPR = true positive rate, TNR = true negative rate, MAE = mean absolute error, RMSE = root mean squared error

1. **Voronoi+RF** (Ward et al., 2017) - A random forest trained to map a combination of composition-based Magpie features and structure-based relaxation-invariant Voronoi tessellation features (effective coordination numbers, structural heterogeneity, local environment properties, ...) to DFT formation energies.
2. **Wrenformer** (Goodall et al., 2022) - For this benchmark, we introduce Wrenformer which is a variation on Wren (Goodall et al., 2022) constructed using standard QKV-Transformer blocks to reduce memory usage, allowing it to scale to structures with >16 Wyckoff positions.
3. **CGCNN** (Xie & Grossman, 2018) - The Crystal Graph Convolutional Neural Network (CGCNN) was the first neural network model to directly learn 8 different DFT-computed material properties from a graph representing the atoms and bonds in a periodic crystal.
4. **CGCNN+P** (Gibson et al., 2022) - This work proposes a simple, physically motivated structure perturbations to augment stock CGCNN's training data of relaxed structures with structures resembling unrelaxed ones but mapped to the same DFT final energy. Here we chose $P = 5$, meaning the training set was augmented with 5 random perturbations of each relaxed MP structure mapped to the same target energy.
5. **MEGNet** (Chen et al., 2019) - MatErials Graph Network is another GNN similar to CGCNN for material properties of relaxed structures that also updates the edge and global features (like pressure, temperature, entropy) in its message passing operation.
6. **M3GNet** (Chen & Ong, 2022) - M3GNet is a GNN-based universal (as in full periodic table) interatomic potential (IAP) for materials trained on up to 3-body interactions in the initial, middle and final frame of MP DFT relaxations. The model takes the unrelaxed input and emulates structure relaxation before predicting energy for the pseudo-relaxed structure.
7. **BOSWR + MEGNet** (Zuo et al., 2021) - BOWSR uses a symmetry-constrained Bayesian optimizer (BO) with a surrogate energy model (here MEGNet) to perform an iterative exploration-exploitation-based search of the potential energy landscape. The high sample count needed to explore the PES with BO makes this by far the most expensive model.

5 Results

Table 1 shows performance metrics for all models considered in v1 of our benchmark. M3GNet takes the top spot on most metrics and emerges as current SOTA for ML-guided materials discovery. The discovery acceleration factor (DAF) measures how many more stable structures a model found among the ones it predicted stable compared to the dummy discovery rate of 43k / 257k $\approx$ 16.7% achieved by randomly selecting test set crystals, consequently the maximum possible DAF is ~ 6. This highlights the fact that our benchmark is made more challenging by deploying models on an already enriched space with a much higher fraction of stable structures over randomly exploring materials space. As the convex hull becomes more thoroughly sampled by future discovery, the fraction of unknown stable structures decreases, naturally leading to less enriched future test sets which will allow for higher maximum DAFs. The reason MEGNet outperforms M3GNet on DAF becomes clear from fig. 1 *right* by noting that MEGNet's line ends closest to the total number of stable materials. The other models overpredict this number, resulting in large numbers of false positive predictions that drag down their DAFs.

Figure 1 *left* visualizes a model's reliability as a function of a material's hull distance. The lower its rolling MAE exits the shaded triangle, the better. Inside this area, the model's mean error is larger than the distance to the convex hull, making misclassifications likely. Outside the triangle even if the model's error points toward the stability threshold at 0 eV from the hull (the plot's center), the mean error is too small to move a material over the stability threshold which would cause a false stability classification. M3GNet achieves the lowest overall MAE and exits the peril zone much sooner than other models on the right half of the plot. This means it rarely misclassifies unstable materials that lie more than 40 meV above the hull. On the plot's left half, CGCNN+P exits the peril zone first, albeit much further from the hull at more than 100 meV below. Essentially, all models are prone to false negative predictions even for materials far below the known hull.

Despite their low accuracy in one-shot predicting relaxed energies from unrelaxed structures, CGCNN and MEGNet achieve high F1 scores and DAFs. This can be explained by unrelaxed in-

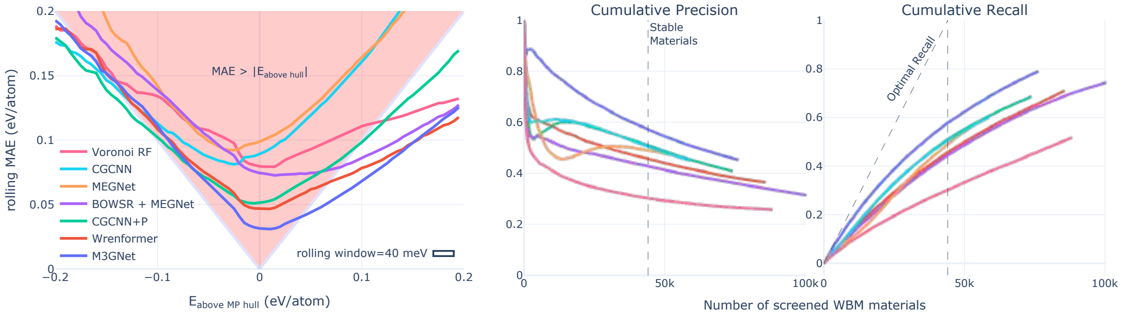


Figure 1: *Left*: Rolling MAE on the WBM test set as the energy to the convex hull of the MP training set is varied. The white box in the bottom left indicates the size of the rolling window. The highlighted 'triangle of peril' shows where the models are most likely to misclassify structures. *Middle*: Cumulative precision over the course of a simulated discovery campaign. *Right*: Cumulative recall over the course of a simulated discovery campaign.

puts being in higher energy configurations than their relaxed counterparts which makes them more likely to be unstable with respect to the training set convex hull, even in cases where the relaxed structure is stable. This biases one-shot GNNs towards predicting unrelaxed input structures as unstable, resulting in higher true negative rates that offset the lower true positive rates in the F1 and DAF metrics. This phenomenon is expected due to the training and testing configuration mismatch and explains the success of previous screening attempts that used such GNN models for screening unrelaxed inputs Park & Wolverton (2020).

6 Discussion

From table 1 we see several models achieve a DAF greater than 2 in this realistic benchmark scenario. Consequently, the benefits of deploying ML-based triage in high-throughput computational materials discovery applications likely warrant the time and setup required. The results obtained from version 1 of our benchmark show that ML universal interatomic potentials like M3GNet are the most promising methodology to pursue going forward, being both 20x cheaper to run than black box optimizers like BOWSR and having access to more training structures than coordinate-free approaches like Wrenformer.

Although the task of discovery will necessarily become more challenging over time as currently undersampled regions of materials space are explored, the path to making ML a ubiquitous discovery tool appears straightforward and is one the field is already pursuing: training foundational IAPs on significantly more data may get us there even without further algorithmic or model improvements.

We welcome further model submissions as well as data contributions for version 2 of this benchmark to the GitHub repo at https://github.com/janosh/matbench-discovery.

Author Contributions

Janosh Riebesell: Methodology, Software, Data Curation, Formal Analysis. Rhys Goodall: Conceptualization, Software, Formal Analysis. Anubhav Jain: Supervision. Kristin Persson: Supervision. Alpha Lee: Supervision.

Acknowledgments

Janosh Riebesell acknowledges support from the German Academic Scholarship Foundation (Studienstiftung) and gracious hosting as a visiting affiliate in the groups of Kristin Persson and Anubhav Jain.

References

Christopher J. Bartel, Amalie Trewartha, Qi Wang, Alexander Dunn, Anubhav Jain, and Gerbrand Ceder. A critical examination of compound stability predictions from machine-learned formation energies. *npj Computational Materials*, 6(1):1–11, July 2020. ISSN 2057-3960. doi: 10.1038/s41524-020-00362-y. URL https://www.nature.com/articles/s41524-020-00362-y.

Albert P. Bartók, James Kermode, Noam Bernstein, and Gábor Csányi. Machine Learning a General-Purpose Interatomic Potential for Silicon. *Physical Review X*, 8(4):041048, December 2018. doi: 10.1103/PhysRevX.8.041048. URL https://link.aps.org/doi/10.1103/PhysRevX.8.041048.

Jörg Behler and Michele Parrinello. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Physical Review Letters*, 98(14):146401, April 2007. doi: 10.1103/PhysRevLett.98.146401. URL https://link.aps.org/doi/10.1103/PhysRevLett.98.146401.

Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi. Open Catalyst 2020 (OC20) Dataset and Community Challenges. *ACS Catalysis*, 11(10):6059–6072, May 2021. doi: 10.1021/acscatal.0c04525. URL https://doi.org/10.1021/acscatal.0c04525.

Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, November 2022. ISSN 2662-8457. doi: 10.1038/s43588-022-00349-3. URL https://www.nature.com/articles/s43588-022-00349-3.

Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chemistry of Materials*, 31(9):3564–3572, May 2019. ISSN 0897-4756. doi: 10.1021/acs.chemmater.9b01294. URL https://doi.org/10.1021/acs.chemmater.9b01294.

Volker L. Deringer, Miguel A. Caro, and Gábor Csányi. A general-purpose machine-learning force field for bulk and nanostructured phosphorus. *Nature Communications*, 11(1):5461, October 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-19168-z. URL https://www.nature.com/articles/s41467-020-19168-z.

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):1–10, September 2020. ISSN 2057-3960. doi: 10.1038/s41524-020-00406-3. URL https://www.nature.com/articles/s41524-020-00406-3.

Jason Gibson, Ajinkya Hire, and Richard G. Hennig. Data-augmentation for graph neural network learning of the relaxed energies of unrelaxed structures. *npj Computational Materials*, 8(1):1–7, September 2022. ISSN 2057-3960. doi: 10.1038/s41524-022-00891-8. URL https://www.nature.com/articles/s41524-022-00891-8.

Rhys E. A. Goodall and Alpha A. Lee. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *Nature Communications*, 11(1):6280, December 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-19964-7. URL https://www.nature.com/articles/s41467-020-19964-7.

Rhys E. A. Goodall, Abhijith S. Parackal, Felix A. Faber, Rickard Armiento, and Alpha A. Lee. Rapid discovery of stable materials by coordinate-free coarse graining. *Science Advances*, 8(30): eabn4117, July 2022. doi: 10.1126/sciadv.abn4117. URL https://www.science.org/doi/10.1126/sciadv.abn4117.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, July 2013. doi: 10.1063/1.4812323. URL https://aip.scitation.org/doi/10.1063%2F1.4812323.

Cheol Woo Park and Chris Wolverton. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Physical Review Materials*, 4(6): 063801, June 2020. doi: 10.1103/PhysRevMaterials.4.063801. URL https://link.aps.org/doi/10.1103/PhysRevMaterials.4.063801.

James E. Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and C. Wolverton. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM*, 65(11):1501–1509, November 2013. ISSN 1543-1851. doi: 10.1007/s11837-013-0755-4. URL https://doi.org/10.1007/s11837-013-0755-4.

Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon M. Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, Anuroop Sriram, Felix Therrien, Jehad Abed, Oleksandr Voznyy, Edward H. Sargent, Zachary Ulissi, and C. Lawrence Zitnick. The Open Catalyst 2022 (OC22) Dataset and Challenges for Oxide Electrocatalysts, November 2022. URL http://arxiv.org/abs/2206.08917.

Hai-Chen Wang, Silvana Botti, and Miguel A. L. Marques. Predicting stable crystalline compounds using chemical similarity. *npj Computational Materials*, 7(1):1–9, January 2021. ISSN 2057-3960. doi: 10.1038/s41524-020-00481-6. URL https://www.nature.com/articles/s41524-020-00481-6.

Logan Ward, Ruoqian Liu, Amar Krishna, Vinay I. Hegde, Ankit Agrawal, Alok Choudhary, and Chris Wolverton. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Physical Review B*, 96(2):024104, July 2017. doi: 10.1103/PhysRevB.96.024104. URL https://link.aps.org/doi/10.1103/PhysRevB.96.024104.

Tian Xie and Jeffrey C. Grossman. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters*, 120(14):145301, April 2018. doi: 10.1103/PhysRevLett.120.145301. URL https://link.aps.org/doi/10.1103/PhysRevLett.120.145301.

Yunxing Zuo, Mingde Qin, Chi Chen, Weike Ye, Xiangguo Li, Jian Luo, and Shyue Ping Ong. Accelerating materials discovery with Bayesian optimization and graph deep learning. *Materials Today*, October 2021. ISSN 1369-7021. doi: 10.1016/j.mattod.2021.08.012. URL https://www.sciencedirect.com/science/article/pii/S1369702121002984.