Noname manuscript No.
(will be inserted by the editor)

Extracting knowledge from DFT: Experimental band gap predictions through ensemble learning

Steven K. Kauwe · Taylor M. Welker · Taylor D. Sparks

Received: date / Accepted: date

Abstract The field of materials science has seen an explosion in the amount of accessible high quality data. With this sudden surge of data, the application of machine learning (ML) onto materials data has led to great results. Particular success has been found in training models based on chemical formulae. Such models have traditionally focused on learning from density functional theory (DFT) or experimental data. Though some researchers have explored the use of DFT calculated properties as features for learning, this has not gained much traction since the machine learning predictions would be limited by the DFT computation time and accuracy. In this work, we explore the use of a stacked ensemble learning system that combines machine learning from DFT calculations to improve learning on experimental data. This is accomplished by handling the DFT and experimental data separately, training distinct models for each. The DFT models are used to generate a "predicted DFT" value for the formulae in the experimental data. A meta-learner—trained using predictions generated by the experimental models combined with predictions from the DFT models—is shown to improve root-mean-squared-error by over 9% in the test data, when compared to a baseline model that only learns from the training data.

Keywords First keyword · Second keyword · More

Taylor D. Sparks
Department of Materials Science and Engineering, University of Utah, Salt Lake City
Tel.: 1+801-581-8632
E-mail: sparks@eng.utah.edu

1 Introduction

The discovery of new materials is critical for achieving the grand engineering goals of the future [1–3]. Despite the importance of materials discovery, revolutionary discoveries are a rarity and often serendipitous in nature. The resources needed for the synthesis and characterization of new materials frequently limits the materials discovery process [2]. Because of the often prohibitive costs of synthesis, researchers in the field of materials science and engineering routinely use first-principles calculations such as density functional theory (DFT) to aid in the materials discovery process [1, 2,4,5]. These computational techniques are used to estimate materials properties and effectively suggest unexpected candidates [6–9], providing insight to promising synthesis actions. Despite the success of first-principles approaches, these physics-based calculations can take days or weeks and consequently are not well suited for exhaustively screening chemical composition space [10, 11]. Moreover, these approaches require crystal structure to be known prior to calculations which precludes unknown structures from most screening methods.

The Materials Genome Initiative in 2011 is evidence that the current rate of discovery will not address the technological and scientific needs we face. The opportunity for machine learning (ML) to help accelerate materials discovery has already been made evident by companies and research programs including: Citrine Informatics [12], the Materials Project (MP) [1], the AFLOW distributed materials property repository (AFLOW) [13], as well as many individual research groups. These organizations have made considerable progress towards providing simple access to computational data, machine learning techniques, and experimental materials information.

A simple approach for using such data relies on creating a composition-based feature vector (CBFV) [14, 15]. Models using this approach have successfully predicted materials properties on both experimental and first-principles data [15–20].

Successful machine learning has also taken advantage of DFT data as features. In particular, the work of Lee *et al.* [21], and Seko *et al.* [22] have shown that DFT can effectively be used to improve predictions on calculated band gap ($G_\circ W_\circ$) and experimental melting temperatures, respectively. Despite their successes, few other researchers have adopted the use of DFT-based features as screening time is a primary advantage of a machine learning approach. These works are great example of combining composition and physics-based features together, however they are still costly in terms of time due to the direct use of DFT calculations.

In this work, we evaluate whether first-principles-based features can be incorporated into a machine learning approach without the computational limits associated with first-principles calculations. This is done by using deep learning networks to learn trends in the DFT data. These trends are then transferred to an ensemble model to aid in learning on experimental data. An excellent property to test this hypothesis on is band gap, as both DFT calculated [5, 13], and experimental [23] band gap are readily available. Recently, Zhou *et al.* [23] showed that band gap can be predicted with good accuracy using machine learning and a general CBFV. However, this work was based only on experimental values and ignored the more than 40,000 DFT computed band gap values accessible through AFLOW and the MP. While it is well known that DFT computed band gaps are not perfect, it is also known that they systematically underestimate the correct value and therefore this large data set should be a useful source of knowledge.

A commonly cited challenge in materials informatics is the lack of data. Even as data repositories come online, it has been observed that they are extremely heterogeneous in nature. Therefore, developing a model which relies on multiple distinct data sets will inherently face challenges of mismatched data and missing features. Moreover, some data sets have the size and characteristics best suited for a certain type of algorithm: take neural networks as an example, which perform best with data on the order of 10^4. Previous approaches have tried to pick the one algorithm and the one data set that gives the best predictive model. However, there is the possibility that disparate data sets can be trained with individual algorithms and later unified to capture information that would otherwise be unavailable. In much the same way, the political environment relies on a poll of polls in order to overcome the biases and shortcomings associated with any individual poll. We hypothesis that a model constructed of different algorithms trained on different data sets will yield a more accurate prediction.

2 Methodology

2.1 Data Acquisition and Feature Development

The idea of integrating DFT-learned predictions with experimental data via model ensembling was tested using band gap. We base our approach on the work of Zhou *et al.* who recently predicted band gap by way of support vector regression (SVR) trained on experimental compounds only. They provided band gap data for both metal and non-metal entries, as well as optimized model parameters, which we adapted for use when generating and testing our models.

Using the experimental data for non-metals only and removing duplicates, there were a total of 2,483 unique entries. This set of compounds is referred to as the "Experimental Database". The experimental database was then randomly split into a training set of 1,986 (75%) and testing set of 497 (25%), which was withheld until all models were optimized to ensure proper validation of model performance. DFT band gaps were obtained from the AFLOW and MP databases. The MP data were retrieved using Citrine's online platform Citrination [12]; AFLOW data was readily available at AFLOW.org. In total, 44,568 unique band gap predictions were accessed: 30,784 from the materials project and 13,784 band gap predictions from AFLOW project. The aggregated DFT data contained a few compounds which already matched those listed in the experimental data. These were systematically removed in order to ensure that the model made a bias-free prediction on experimental compounds. The resulting data sizes are shown in Figure 1.

After retrieving the necessary data, chemical compositions were vectorized into features based off of elemental properties. The CBFV employed in this work uses the weighted average, the sum, the variance, and the biggest range of the elemental properties. When features cannot be generated due to missing elemental properties the missing value is replaced with the mean value for the property across all compositions [14]. When learning with non-tree-based algorithms, the CBFV's undergo scaling and normalization using Scikit-learn's StandardScalar and Normalizer functions [24].

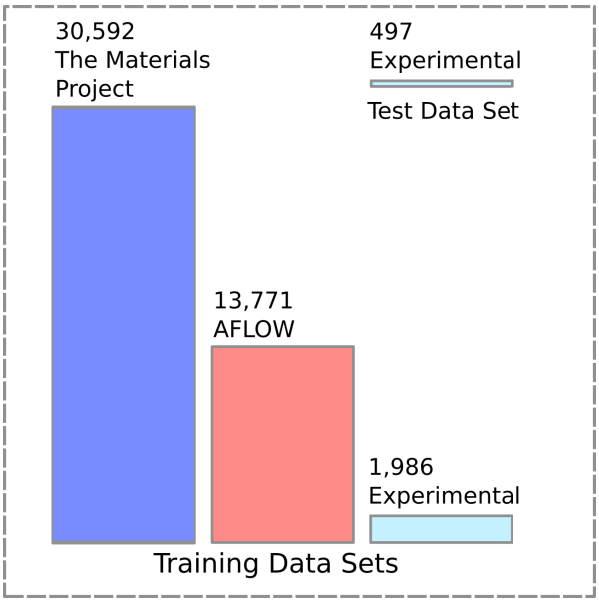

Fig. 1 DFT data is combined and used for learning with a neural network. The experimental data is separated into a training set and a test set.

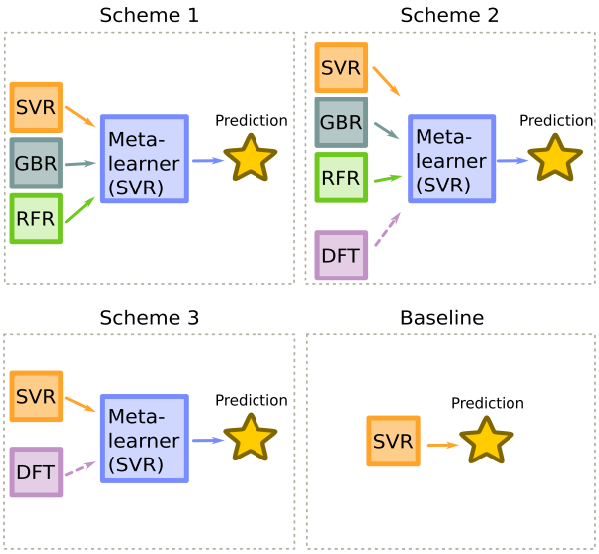

Fig. 2 The three ensemble schemes and the baseline are shown above. Each model sends interim-predictions to the meta-learner which then generates the final prediction

2.2 Ensemble Models

Model ensembling works by combining the predictions of multiple model types. For this work we employed stacked ensembling. Stacked ensembling is a technique that uses two layers of machine learning. The first layer contains multiple models which generate predictions on the data. The second layer then uses the first layer's predictions as features for a meta-learner which generates predictions on the true value.

We created three separate stacked ensembles which we compared against a baseline (see "Schemes" in Figure 2) to investigate whether the supposed wealth of information captured in the DFT data has an impact on our ability to learn from the experimental band gap data.

1. Models trained on experimental data using three dissimilar learning methods: support vector regression (SVR), gradient boosting regression (GBR), and random forest regression (RFR).
2. Addition of DFT prediction to the three experimental-trained models.
3. Addition of DFT to SVR only.
4. Baseline: A single non-ensemble model trained off experimental data using an SVR algorithm (reproducing Zhou *et al.*)

Scheme 1: Cross-validated predictions from the three experimental models were used as features for the meta-learner (Layer 2). This allowed us to optimize parameters for the meta-learner before predicting on the test set. Once optimized, the first layer models were retrained on the full experimental training data. These models then generated predictions for compositions in the test set, which the meta-learner used to generate final predictions.

Scheme 2: The same experimental models were utilized, but the first layer also incorporated a DFT model trained on the aggregated DFT data. After retraining, first layer predictions were again used by the meta-learner to generate predictions on the test data.

Scheme 3: In order to explicitly examine the role of DFT-based predictions, we included a scheme which learns off of SVR and DFT predictions alone. By excluding GBR and RFR, we are able to make a comparison which illustrates the role of DFT-base predictions by itself. Using the same approach as schemes 1 and 2, the DFT model was ensembled with the first-layer SVR which had the best cross-validation scores.

2.3 DFT-based Model

The DFT data was modeled with a Sequential Neural Network. The python implementation of Keras (2.2.2) was used with a Tensor Flow (1.10.0) back-end running CUDA (9.2) on the Ubuntu 18.04.1 LTS operating system using an Intel 8600K CPU and NVIDIA GTX 1080 Ti graphics card. The network structure was obtained through optimization of dropout rate and node count using a genetic algorithm followed by trial and error optimization using the AFLOW DFT data (Figure 3). Each node in the network used a ReLU activation function, and dropout was implemented between layers

to prevent over-fitting. The model used for DFT predictions was built using the aggregate DFT data after applying scaling and normalization to the CBFV.

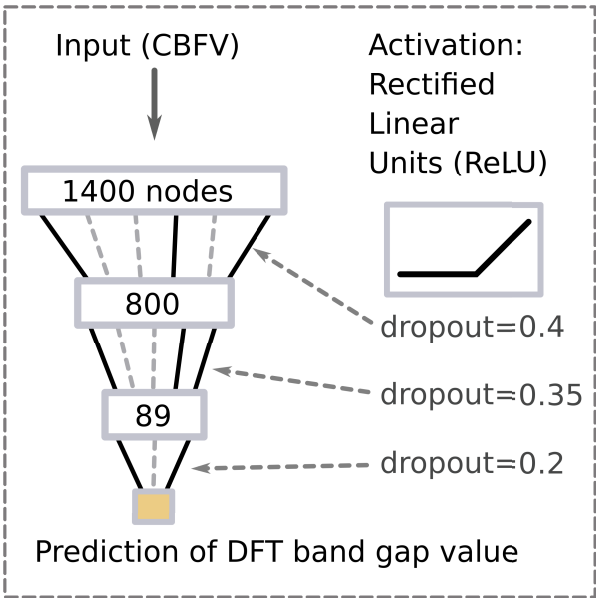

Fig. 3 The network structure used is generally effective for learning problems using a CBFV

3 Results and Discussion

3.1 Analysis of First Layer Models

3.1.1 Experimental-based Machine Learning Models

It is well known in the machine learning community that many learning algorithms provide unique advantages and disadvantages within different data sets. For example SVRs capture non-linearity but lack scalability, where as faster learners, such as a Linear or Random Forest Regression, may have difficulty capturing the same complexity in the data. The ability to capture and express different patterns in data is the heart of ensembling; the combined outputs from each model can contribute to the overall learning process to better represent the data as a whole.

Parameters for the RFR and GBR were selected with the help of a grid search, which iterates over a selection of predetermined parameter values to find the optimal set. For the SVR we used the model parameters found by Zhou and coworkers. Five-fold cross-validation predictions were used during optimization and when comparing model performance. Table 1 shows the coefficient of determination R^2, root-mean-square error (RMSE), and mean absolute percentage error (MAPE) for the first layer models. Cross-validation performance indicated that the SVR was the best individual learning method for this data. This aligns with the work of Zhou *et al.*, and justified the use of SVR performance as a baseline when evaluating the ensemble performance.

Table 1 Cross-validation Metrics for Experimental Models

First Layer Model	R^2	RMSE	MAPE
SVR	0.808	0.627 eV	37.6
GBR	0.782	0.669 eV	43.0
RFR	0.772	0.688 eV	44.6

3.1.2 DFT-trained Neural Network

Figure 4 shows the deep-network cross-validation performance. Overlaid on top of this data are neural network predictions of compounds which were removed from training because they coincided with experimental data. The neural network-based predictions of the experimental compounds are a necessary part of the first layer in both schemes 2 and 3.

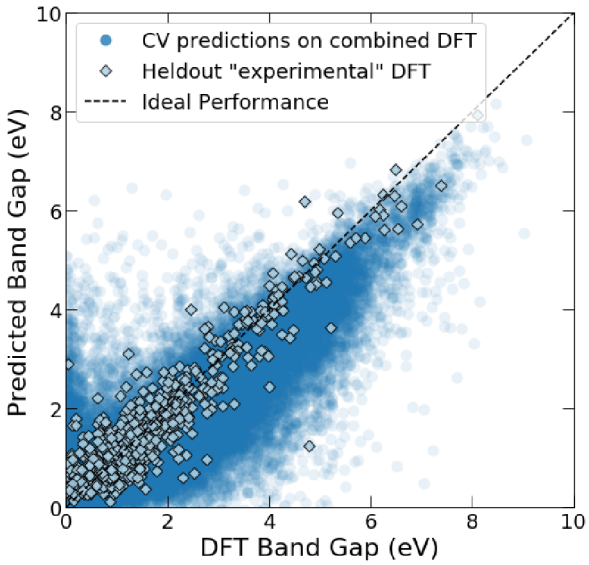

Fig. 4 Performance was estimated using cross-validation on the combined DFT data set. Predictions of the DFT band gap value for the compounds that were removed due to overlap with the experimental data are quite accurate.

At first glance there is significantly more error in the cross-validated data compared to the data points associated with the experimental compounds. One possible reason for this is that the DFT data includes band gap predictions for many exotic hypothetical structures, whereas the experimental data only contains "real" compounds with a stable crystal structure.

It is possible to compare the accuracy of DFT calculations vs neural network predictions for the 763 compounds where we have experimental data. In either of these cases, a residual can be determined by subtracting the DFT or neural network prediction from the experimental value. An initial comparison of the residuals (Figure 5) might suggest that the neural network is slightly more accurate than the DFT predictions. This could technically be possible due to the difference in distribution between the training data and the experimental data from which the residuals were calculated. However, it is more likely due to chance, as a Student's t-test shows that there is no statistically significant difference in the mean of the absolute error for the two residuals (p-value = 0.50).

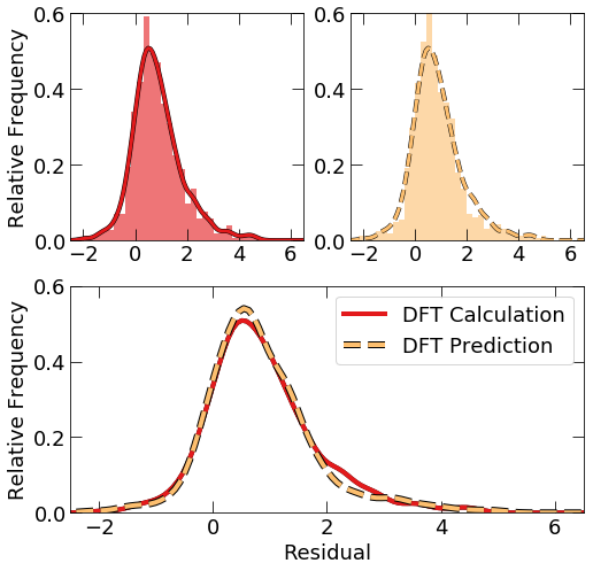

Fig. 5 The calculated residual indicates that the error associated with predicted DFT values is effectively indistinguishable from the calculated DFT values.

The similarity in neural network predictions and DFT calculations is a welcome finding for those seeking to use machine learning on stacked ensembles that are built on heterogeneous data sets. The ability to extract knowledge from distinct but related data, and use that information as features while learning, represents an important step beyond previous approaches. Until the time comes that a unified data infrastructure exists, future work will continue to rely on individual repositories of data. With this approach, we can capture the knowledge in these different data sets without performing computationally expensive calculations.

3.2 Analysis of Ensemble Schemes

With an understanding of the performance of the first layer models (refer to Table 1), we are now able to evaluate how these compare both to the baseline (SVR model) as well as the stacked ensembles defined as Schemes 1, 2, and 3.

The most straight forward ensembling is built on the original data using the outputs from various machine learning algorithms; this is common practice and is routinely used to improve model performance. This first approach was used for the stacked ensemble from Scheme 1 which provides improvement in band gap predictions with a 5.1% reduction in RMSE from the baseline SVR (see Table 2 for performance of all schemes). Note, this is more powerful than a poll-of-polls approach in that this is not an averaging of predictions (3.3% reduction in RMSE), but rather each model is treated as an independent feature and therefore the meta-learner can capture the non-linear improvements associated with each method.

Table 2 Performance on Test Data

	R^2	RMSE	MAPE
Baseline	0.857	0.606 eV	0.471
Scheme 1	0.872	0.575 eV	46.9
Improvement to baseline	1.6%	5.1%	0.3%
Scheme 2	0.883	0.548 eV	46.0
Improvement to baseline	3.0%	9.5%	2.5%
Scheme 3	0.880	0.556 eV	47.2
Improvement to baseline	2.7%	8.3%	0.3%
Poll of polls	0.867	0.586 eV	50.0
Improvement to baseline	1.1%	3.3%	-6.1%

The ideal scenario for learning band gap would include DFT computed band gap as an additional feature to scheme 1; however, computation time is a major problem when relying on DFT as a feature. To overcome this barrier, a neural network was used to generate high quality predictions after learning from the vast amount of existing DFT values. The use of this neural network, in proxy for actual DFT calculations, combined with the experimental models leads to Scheme 2. Analysis of Scheme 2 shows 9.5% improvement over the baseline. Notably, this is better than the 5.1% improvement Scheme 1 offers over the baseline (Figure 6). This demonstrates an ability to extract knowledge from the independent DFT database.

It might be surprising that we can combine DFT data in this way. A more obvious approach would be to simply treat the DFT and experimental values as indistinguishable in a combined data set. A domain expert is unlikely to advise this approach, as adding the DFT data would only water down the high quality ex-

perimental data. Rather than treating them as indistinguishable, we keep them separate in heterogeneous data sets using a stacked ensemble to learn from both distinct data sets, then build a meta-learner which efficiently captures knowledge from each.

Scheme 3 resembles Scheme 2 in that it learns off the DFT predictions and experimental data. However, instead of incorporating multiple experimental models, it only uses the SVR. Surprisingly, we observe an 8.3% improvement in predictions. Which is only 1 percent worse than Scheme 2. This suggests that there is an overlap in the knowledge which is gained from including the DFT model in addition to the ensembling of multiple experimental models.

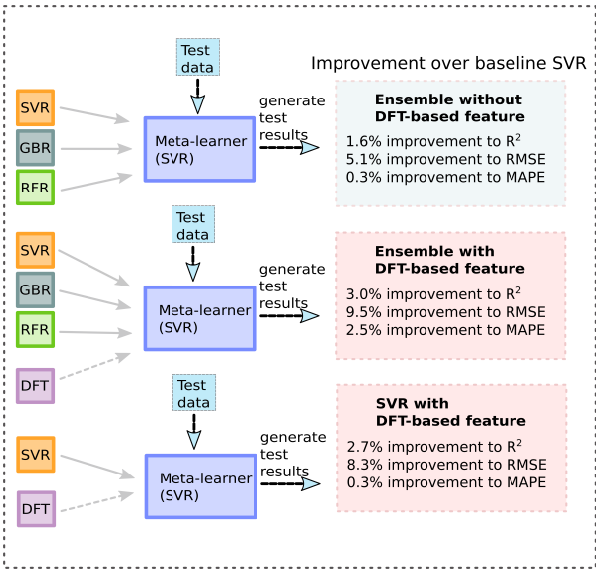

Fig. 6 Performance on the test set shows significant improvements when incorporating DFT-based predictions in the learning process.

4 Conclusion

In this work we were able to build a machine learning-based band gap prediction model that is better than any single prediction model. The stacked ensemble utilized neural network predicted DFT values rather than the time intensive DFT calculations. Three different schemes were used to probe the added value associated with each first layer model. The most accurate ensemble included the experimental and DFT models, with resulting RMSE of 0.55 eV which constitutes a 9.5% improvement over the baseline. This effectively took advantage of the 2,483 experimental as well as the 44,568 DFT compounds.

The findings from this work suggest that researchers could benefit by departing from the traditional approaches that are being used in materials informatics. Specifically, we find the following:

1. Learning on a single data set can be improved by considering an ensemble approach. Resembling a poll-of-polls, multiple algorithms are used to inform learning but can do so in a non-linear fashion. An additional benefit from considering multiple algorithms is that it provides a first measure of variability in predictions as a function of algorithm selection.
2. Distinct heterogeneous data sets do not need to be unified prior to learning. In fact, unifying data by excluding entries with significant missing features is harmful in that you end up excluding a great deal of information. Instead, it is possible to model data from distinct sources and use the models in a stacked ensemble. This has important implications in materials science. Relative to some other fields, such as computer science, materials science is plagued by mismatched, non-standardized, heterogeneous and widely distributed data sets. A technique such as one described in this work, which allows learning without unification is a powerful new idea. By retaining heterogeneous data, we do not throw out the baby with the bath water.
3. Enormous effort and computational resources are being used to generate materials structure and properties from first-principles calculations. Forward thinking paradigms such as the Materials Genome Initiative have cited the necessity of experimental and computational research going hand in hand to minimize the cost and risk of new materials development. However, despite the effort going into DFT this is not sufficient to bridge the knowledge gap. The machine learning approach described here allows us to generate high quality predictions by extracting knowledge from DFT—leveraging the wealth of information stored within these first principle calculations even when the data lack parity.

Acknowledgements We would like to thank the National Science Foundation for their support of this research under NSF CAREER Award 1651668. We would also like to thank the Brgoch group at the University of Houston for inspiring this research and for readily supplying data in a way which adheres to FAIR Data Principles.

References

1. A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, et al., Commentary: The materials project: A materials

genome approach to accelerating materials innovation, Apl Materials **1**(1), 011002 (2013)
2. A. Jain, K.A. Persson, G. Ceder, Research update: The materials genome initiative: Data sharing and the impact of collaborative ab initio databases, APL Materials **4**(5), 053102 (2016)
3. R. Seshadri, T.D. Sparks, Perspective: Interactive material property databases through aggregation of literature data, APL Materials **4**(5), 053206 (2016)
4. G. Hautier, C.C. Fischer, A. Jain, T. Mueller, G. Ceder, Finding natures missing ternary oxide compounds using machine learning and density functional theory, Chemistry of Materials **22**(12), 3762 (2010)
5. A. Jain, G. Hautier, S.P. Ong, K. Persson, New opportunities for materials informatics: Resources and data mining techniques for uncovering hidden relationships, Journal of Materials Research **31**(8), 977 (2016)
6. W. Chen, J.H. Pöhls, G. Hautier, D. Broberg, S. Bajaj, U. Aydemir, Z.M. Gibbs, H. Zhu, M. Asta, G.J. Snyder, et al., Understanding thermoelectric properties from high-throughput calculations: trends, insights, and comparisons with experiment, Journal of Materials Chemistry C **4**(20), 4414 (2016)
7. G. Hautier, A. Jain, S.P. Ong, From the computer to the laboratory: materials discovery and design using first-principles calculations, Journal of Materials Science **47**(21), 7317 (2012)
8. S. Curtarolo, G.L. Hart, M.B. Nardelli, N. Mingo, S. Sanvito, O. Levy, The high-throughput highway to computational materials design, Nature materials **12**(3), 191 (2013)
9. E.B. Isaacs, C. Wolverton, Inverse band structure design via materials database screening: Application to square planar thermoelectrics, Chemistry of Materials **30**(5), 1540 (2018)
10. B. Meredig, A. Agrawal, S. Kirklin, J.E. Saal, J. Doak, A. Thompson, K. Zhang, A. Choudhary, C. Wolverton, Combinatorial screening for new materials in unconstrained composition space with machine learning, Physical Review B **89**(9), 094104 (2014)
11. K. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. Müller, E. Gross, How to represent crystal structures for machine learning: Towards fast prediction of electronic properties, Physical Review B **89**(20), 205118 (2014)
12. Citrine. https://citrination.com (2018). URL https://citrination.com. [Online; accessed 1-March-2018]
13. Aflow. Aflow - automatic - flow for materials discovery (2018). URL http://aflowlib.org/. [Online; accessed 14-July-2018]
14. S.K. Kauwe, J. Graser, A. Vazquez, T.D. Sparks, Machine learning prediction of heat capacity for solid inorganics, Integrating Materials and Manufacturing Innovation pp. 1–9
15. L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, npj Computational Materials **2**, 16028 (2016)
16. A.O. Oliynyk, A. Mar, Discovery of intermetallic compounds from traditional to machine-learning approaches, Accounts of chemical research **51**(1), 59 (2017)
17. A.O. Oliynyk, L.A. Adutwum, B.W. Rudyk, H. Pisavadia, S. Lotfi, V. Hlukhyy, J.J. Harynuk, A. Mar, J. Brgoch, Disentangling structural confusion through machine learning: Structure prediction and polymorphism of equiatomic ternary phases abc, Journal of the American Chemical Society **139**(49), 17870 (2017)
18. J. Carrete, W. Li, N. Mingo, S. Wang, S. Curtarolo, Finding unprecedentedly low-thermal-conductivity half-heusler semiconductors via high-throughput materials modeling, Physical Review X **4**(1), 011019 (2014)
19. A. Mansouri Tehrani, A.O. Oliynyk, M. Parry, Z. Rizvi, S. Couper, F. Lin, L. Miyagi, T.D. Sparks, J. Brgoch, Machine learning directed search for ultraincompressible, superhard materials, Journal of the American Chemical Society **140**(31), 9844 (2018)
20. J. Graser, S.K. Kauwe, T.D. Sparks, Machine learning and energy minimization approaches for crystal structure predictions: A review and new horizons, Chemistry of Materials **30**(11), 3601 (2018)
21. J. Lee, A. Seko, K. Shitara, K. Nakayama, I. Tanaka, Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques, Physical Review B **93**(11), 115104 (2016)
22. A. Seko, T. Maekawa, K. Tsuda, I. Tanaka, Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single- and binary-component solids, Physical Review B **89**(5), 054303 (2014)
23. Y. Zhuo, A. Mansouri Tehrani, J. Brgoch, Predicting the band gaps of inorganic solids by machine learning, The journal of physical chemistry letters **9**(7), 1668 (2018)
24. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research **12**, 2825 (2011)