

Deep Learning for Chemical Compound Stability Prediction

Ruoqian Liu
Electrical Engineering and
Computer Science
Northwestern University
Evanston, IL 60208
rosanne@northwestern.edu

Logan Ward
Materials Science and
Engineering
Northwestern University
Evanston, IL 60208
loganward2012@u.northwestern.edu

Ankit Agrawal
Electrical Engineering and
Computer Science
Northwestern University
Evanston, IL 60208
ankitag@northwestern.edu

Wei-keng Liao
Electrical Engineering and
Computer Science
Northwestern University
Evanston, IL 60208
wkliao@eecs.northwestern.edu

Chris Wolverton
Materials Science and
Engineering
Northwestern University
Evanston, IL 60208
c-wolverton@northwestern.edu

Alok Choudhary
Electrical Engineering and
Computer Science
Northwestern University
Evanston, IL 60208
choudhar@northwestern.edu

ABSTRACT

This paper explores the idea of using deep neural networks with various architectures and a novel initialization method, to solve a critical topic in the field of materials science. Understanding the relationship between the composition and the property of materials is essential for accelerating the course of materials discovery. Data driven approaches using advanced machine learning to derive knowledge from that of existing compounds, and/or from simulations of non-existing ones, have only started to play a crucial role. We demonstrate an application with a large-scale data set containing 300K organic and inorganic compounds. Deep multi-layer perceptrons are used to capture nonlinear mappings between chemical composition and compound stability characterized by a continuous value, known as the formation energy. It is surprising to see that input features as raw and sparse as the compositional fractions of elements can lead to a remarkably accurate modeling of a far-fetched regression prediction. The performance is shown to be outperforming state-of-the-art predictions by as much as 54%.

Keywords

Deep learning; multi-layer perceptrons; materials design; materials discovery; chemical comp

1. INTRODUCTION

The practice of data mining, in light of the availability of large-scale data, has been causing a paradigm change in sci-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD Workshop on Large-Scale Deep Learning for Data Mining August 13–17, San Francisco, CA USA

© 2016 ACM. ISBN 978-1-4503-2138-9.

DOI: 10.1145/1235

entific discovery, from empirical science to theoretical, computational science, and now to data science [2, 12]. Specifically, there are spurring interests in applying advanced machine learning techniques to assist materials discovery, advocated by the Materials Genome Initiative (MGI) [1], which envisions computationally assisted discovery, development, manufacturing, and deployment of advanced materials at a much accelerated speed and reduced cost. The initiative of Integrated Computational Materials Engineering (ICME) [4, 15] also encourages the development of methods that help objectively create new materials with desired properties for specific applications, by linking what we see (structure of a crystal, composition of a compound) to what we want (a certain strength demand of an alloy). Interdisciplinary fields like materials informatics, materials data science have been formed to address such need.

In materials science, the primary goal concerns the discovery of new, alternative materials (metals, crystals, ceramics, polymers, etc.), from the prior knowledge of past experiments and simulations. A recent work by Raccuglia et.al [23] has accomplished crystal-structure prediction by learning from failed experiment observations. The work uses support vector machine (SVM) as the predictor. Kernel methods like SVM has limitations due to its shallow structure, and cannot handle large-scale data effectively. Deep neural networks (DNN), with its extensible and flexible structure and huge number of parameters, are powerful for capturing highly nonlinear mappings between inputs and outputs. They are able to automatically extract high level features from low level one through multi-layer abstraction. Their capability of utilizing large data has demonstrated great success with various data types like image [18, 26], speech [13], video [16], text [31] and more.

However, the application of DNN to modeling materials phenomena with large data is rarely seen. The main reason, arguably, is that materials data are never large enough. Materials data come from two sources: empirical experiments, and computational simulations. For a long time trial-

and-error experiments are the only way of observation and hence data generation, and the collection, standardization and maintenance of such data is expensive and laborious. Until fairly recently, progresses have been made in first principles calculations [9, 10], such as Density functional theory (DFT), as a theoretical and computational technique for materials property calculation. Although the DFT simulation for compounds is rather slow (approximately 10 hours per compound for formation energy calculation), it has offered opportunities for large scale data collection. Several public databases containing computed materials properties are hence built. Examples include the Materials Project [14]¹, and Open Quantum Materials Database (OQMD) [17]². The latter stores over 285,780 compounds.

In this work, we attempt to utilize the opportunity provided by the large-scale available data, to address an intricate compound stability prediction problem by use of deep learning. We intend to fully utilize the capability of deep neural networks in learning hierarchy of features from raw inputs, for that purpose we only draft the input feature to the network in a raw, sparse, unprocessed form. It is also our intention to develop general purpose learning models without much guidance from domain experts, and therefore the input and output variables are designed to be easily obtained and interpretable.

The rest of paper is organized as follows. In Section 2, we review the related work on the use of neural networks in many materials problems. Section 3 introduces the problem of chemical compound stability prediction. In Section 4 we discuss the proposed deep learning solution framework, and Section 5 the experiments and results. Further implementation details and discussions of deep net training are provided in Section 6. We finally conclude the paper and discuss future work in Section 7.

2. RELATED WORK

The past few years have witnessed a renewed interest in neural networks. With the availability of large scale data and fast computation tools, the practice of training deep networks has been favored. We have seen its application to not only traditional computer science problems such as image, speech, languages, but also other areas of science. A recent work for drug discovery making use of a multitask deep learning framework [24] is an example.

Similarly to drug discovery, the problem of materials discovery is about deciding on certain composition, formulation, and processing steps of materials with a desired target property [3, 19–21]. Neural networks as a tool have been used in materials science applications such as spectroscopy classification and structural identification [27], characterizing constitutive relationship for alloys [28], the evaluation of descriptors [7], etc. However, neither the size of data or complexity of networks in these works have gone large enough.

Montavon et al. [22] have used about 7000 compounds to train a 4 layer MLP to predict molecular electronic properties. In [8], Support Vector Machines (SVMs) are built to predict which atoms in a liquid are most mobile, a key question in understanding the structure of liquids. The data size was up to 10,000 particles. In the work by Meredig

et.al [21], around 90,000 compound entries in the Inorganic Crystal Structure Database (ICSD) are used to construct an ensemble learning model for the prediction of formation energy. We tackle a similar problem as [21] in this paper but are using a much larger dataset. The OQMD data contains the entirety of ICSD.

To this date, the actual combination of deep learning and materials discovery has been scarce. Moreover, the size of data in these applications are not large enough to qualify the notion of big data. To the best of our knowledge, this work is the first work utilizing a dataset as large as 300K compounds.

3. COMPOUND STABILITY PREDICTION

3.1 Problem Definition

The problem deals with the prediction of chemical properties from the compound compositions. It is an application strongly related to materials discovery, as it helps explore the chemical compositional space to find compounds that are stable and thus are possible to form.

It is known, from the basic theory of matter, that atoms of different elements combine in small, generally whole-number ratios to form compounds. For example, hydrogen and oxygen atoms combine in a ratio of 2:1 to form the compound water, H₂O. Carbon and oxygen atoms combine in a ratio of 1:2 to form the compound carbon dioxide, CO₂. Iron and oxygen atoms combine in a ratio of 2:3 to form the familiar substance rust, Fe₂O₃. The set of element types and its corresponding ratio is the simplest, most straightforward form of representation of a compound’s composition.

For chemical compounds, composition is what we observe, and property is what we desire to know. The property of interest to this work is the the formation energy of a compound. It is the energy released when forming that material from its constituent elements, and is often the most important design criterion when searching for new materials. By comparing the formation energy of a material to those of known compounds, it is possible to determine if a material is likely to be actually made in a lab. Consequently, the ability to actually and quickly obtain formation energies and, thereby the stability of materials is crucial for the discovery of new materials. Commonly, the formation energy is predicted using Density Functional Theory (DFT) which, while accurate, requires significant computational resources. The estimated time to simulate one compound given composition is 10 hours.

In the machine learning setup, each distinct compound is a data entry. Suppose the only available information about a given compound is its composition, i.e. the set of ions and the associated ratios in which they appear. The objective is to determine chemical properties of it, after studying a set of known composition-property pairs. And the property to be predicted here is the formation energy. We take DFT simulation results as the ground truth, and build models that predict the energy as close to the simulated value as possible. However, we have to note that DFT is not absolutely accurate. The formation energy simulation is known to be within an error bound of 0.1 electronvolt (eV) per atom. Past works have shown performances similar magnitude.

3.2 Existing Work and Our Distinction

Many existing work from the materials science domain

¹<https://www.materialsproject.org>

²<http://www.oqmd.org>

have studied the relationship between compound composition and chemical property. In [30], a composition similarity function is defined to quantify how similar two compositions are, and in turn how possible they are to take the same crystal structure prototype. The function basically computes the probability that two ions will substitute for each other within the same prototype. Meredig et al. [21] developed heuristic and machine learning models on a similar DFT database for the prediction of formation energy. A total of 129 attributes are used to represent each compound entry. Among those, 112 are the elemental composition percentage, or atomic ratios, which we will use in our deep learning model evaluation. The other 17 attributes are heuristic quantities developed using chemical intuition, such as the average atomic mass, average column and row numbers on the periodic table, average and maximum difference of atomic numbers, average and maximum difference of electronegativity, etc. The learning system used is Rotation Forest [25] and the mean absolute error (MAE) achieved for prediction is 0.12 eV.

Our proposed work is a lot similar to this work in [21] but does offer some key differences and improvements:

1. The data size is over three times bigger. In [21] around 90,000 compound entries in ICSD are used to construct an ensemble learning model, while we have accessed the whole OQMD, a collection of over 285,780 entries.
2. We further eliminated intuitive based attributes in compound representation as inputs, and therefore restricted the representation to the most straightforward set - atomic fractions in compound composition.
3. A deep learning architecture is used as a replacement to Rotation Forest.

4. PROPOSED DEEP LEARNING SOLUTION

Our goal is to build a deep learning regression system, which each time takes a fixed-length vector of atom fraction values (0 when the element does not appear in the compound) and produces a real-numbered value that stands for the formation energy. A snapshot of the input and output data is shown in Fig. 1(a). The deep neural network architecture is schematically illustrated in Fig. 1(b). We propose to use an MLP architecture, with a number of fully connected (FC) layers. We also propose a novel treatment on the weight initialization.

4.1 Architectures

We train several MLPs with 5 to 9 hidden layers and varying numbers of hidden units. Mostly the number of hidden units per layer would start with a largest number and decreases towards the output layer. Since it is a regression problem, there is one output node, producing the formation energy with values roughly between $[-20, 5]$. The unit is electronvolt (eV) per atom. For each architecture we train a total epoch of 1000, minibatch of 128, with SGD as the learning algorithm, and learning rate 0.001 which is halved every 100 epochs. The loss function used is the mean absolute error between predicted output and target. We also tried mean squared error and the difference is negligible.

4.2 Initialization method

In this solution we make an innovative design in the training method of MLP. The idea comes from observing the

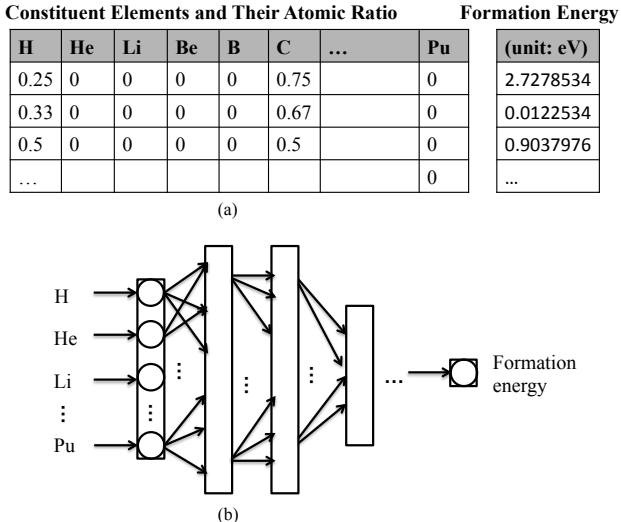


Figure 1: (a) Snapshot of data used in the composition based compound energy prediction problem. The input data is for each compound the ratio of its constituent elements. For a binary compound it is a row of two nonzeros and others all zero. The nonzeros add up to be 1. (b) A schematic of the proposed deep neural network architecture.

input data. It is noticeable from the data (and valid in theory) that not all combinations of elements are possible, given by the relatively small number of compound collection compared to the whole combinatorial space of all elements. Bearing such restrictions, we can guide the learning so that the network starts with a structure that lies in a more probable parameter space.

Even without domain knowledge of compound formation principles, insights can be drawn from data that some combinations of elements happen more frequently and some never do. This inspires us to model the a priori distribution of elements, from the input data in an unsupervised fashion. The probability of each element is then used as the variance of a Gaussian distribution with which the initial weights of the first layer is sampled from. For example, if we have element C present in the dataset at a probability of 0.05, and the first layer has 1024 nodes, weights between the input node for C and every one of the 1024 first-layer nodes would be initialized with a Gaussian distribution with 0 mean and standard deviation of 0.05. In this way, more frequently present elements would have a larger influence factor at the very starting stage of the training process. This gives the network a more promising area to start the searching of all parameters. We demonstrate in the following result section that the proposed initialization method improves the prediction performance by a large margin.

5. EXPERIMENT RESULTS

5.1 Data and network configuration

The public materials database of OQMD contains DFT calculations of more than 285,780 compounds as of February 2016. Out of all 118 known elements in the periodic table as of 2016, OQMD contains compounds composed by

Table 1: Configuration of MLPs designed for compound prediction.

MLP	Configuration	Init.	MAE
mlp5r	1024x2-512-128-1	random	0.115
mlp5p	1024x2-512-128-1	prob.	0.096
mlp7r	1024x2-512x2-128x2-1	random	0.092
mlp7p	1024x2-512x2-128x2-1	prob.	0.072
mlp9r	1024x2-512x2-256x2-128x2-1	random	0.112
mlp9p	1024x2-512x2-256x2-128x2-1	prob.	0.090

89 of them (whose pseudopotentials are available for DFT calculations). And the number of compositionally unique compounds is 275,759. In the case of more than one compounds with the same composition (but with different microstructures), the one with the lowest formation energy is chosen, for that a low formation energy indicates a more chemically stable compound and is of more interest in this study.

Three rounds of outlier removal is conducted before training. Firstly, all non-compound, single-element materials are removed, as their formation energy is considered zero. Secondly, data entries with out-of-bound energy values are discarded, with a rational bounds set to be $[-20, 5]$. Lastly, we further remove entries whose energy value is outside of the $\pm 5\sigma$ bound, σ being the standard deviation in the training data. The reason for removing outliers outside of several σ s is to prevent calculation errors undetected by strict value bounds. To validate we tried experimenting without removing σ -based outliers and ended up having performances that are in average the 16% worse.

There remains 256,673 compounds after outlier removal. This set contains 16,354 (6.37%) binary compounds, 208,859 (81.37%) ternaries, and 31,460 (12.26%) quaternaries and above. The highest order composition contains 7 elements. The data is randomly split into training and test with a 9:1 ratio. All results reported in this section are on the test set.

Three MLP architectures are designed, as shown in Table 1. The configuration 1024x2-512-128-1 means that there are 4 FC layers with 1024, 1024, 512, 128 nodes and an output layer with 1 node. We insert Dropout modules wherever the number of node changes from the previous layer to the current layer. That is, for mlp7*, we have a Dropout between layer 2 and 3, between 4 and 5, and between 7 and the output. In this way, the mlp5* networks have 3 Dropout modules, and mlp9* have 4 Dropout modules.

Two types of initialization methods are experimented: random and our proposed probability based, denoted "prob." in Table 1. The corresponding network is therefore named with the number of FC layers, and either a "r" or "p" affixed. In random initialization, weights are initialized with a Gaussian random distribution with zero mean and standard deviation of 0.02.

5.2 Prediction results

The MAE results shown on the last column in Table 1 is the final error on test data at the end of a training of around 1000 epochs. Figure 2 shows the change of test MAE along training epochs, for all 6 networks. If we compare every pair of networks under the same architecture, it is consistent that the when we initialize the network with probability based random weights, the result outperforms pure random

initialization. A zoomed in view of the beginning 50 epochs, shown in the bottom figure in Fig. 2, indicates clearly that the network mlp7p is able to search from a much more optimistic area at the very beginning of training, and could have finished the training much earlier.

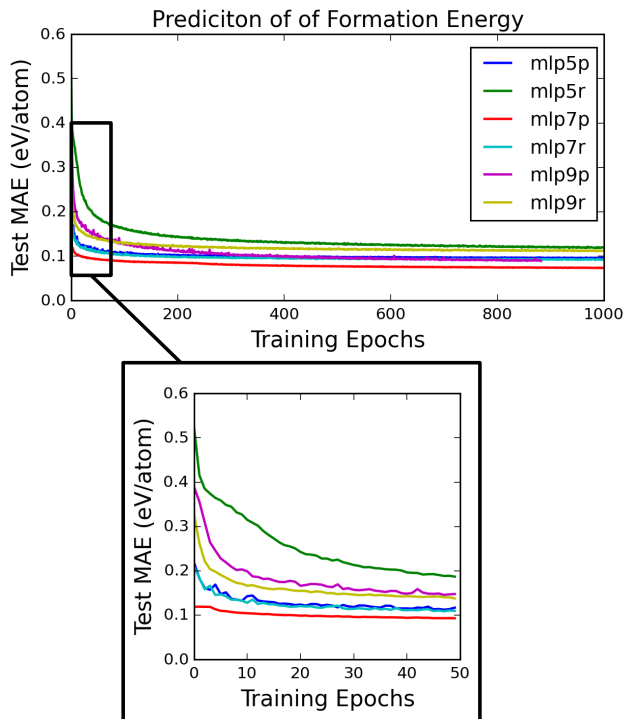


Figure 2: Test MAE (unit: eV/atom) over training epochs in chemical compound application. All 6 networks with 3 different architectures and 2 initialization methods. The probability based method (network with name ending in "p") consistently outperforms random method. The influence is clearer at the beginning of training, shown in the bottom zoomed-in figure.

The benchmark we chose to compare our result with include 1-Nearest Neighbor (1-NN), and Random Forest (RF) regressors. Table 2 shows three benchmark systems results. Different number of member trees in RF from 50 to 150 is used, resulting in systems RF50, 100 and 150. The run time is based on the scikit-learn³ implementation. The distance function used in 1-NN is selected to be cosine. The run time of 1-NN is based on an implementation with Theano [5, 6], taking advantage of batch processing with GPU when calculating matrix-matrix multiplications.

We can see that RF gives relatively consistent performance despite the number of trees. Our best MLP outperforms RF by a margin of 54%. Cosine distance based 1-NN is the worst performing method, indicating that comparing the proximity of the composition ratios across every element dimension is not reliable in predicting a compound's formation stability. We outperform 1-NN prediction by as much as 94%.

6. IMPLEMENTATION AND DISCUSSION

³<http://scikit-learn.org/stable/modules/ensemble.html#random-forests>

Table 2: Benchmark results for compound prediction, using 1-NN and RF with different number of trees.

Benchmark System	MAE (eV/atom)	Run time (s)
RF-50	0.155	203.78
RF-100	0.154	423.05
RF-150	0.154	797.79
1-NN	1.299	2062.74

All experiments were carried on an NVIDIA DIGITS DevBox with a Core i7-5930K 6 Core 3.5GHz desktop processor, 64GB DDR4 RAM, and 4 TITAN X GPUs with 12GB of memory per GPU. Although only one GPU is used during each deep net training. The implementation of deep neural networks is done using Theano [5, 6]. The remainder of this section discusses several observations in the training process.

Training time. The numbers of data samples is around 300,000. There are over 2 million free parameters, or weights, in the largest MLP we designed. The MLP architecture we proposed are not considered typically deep in recent applications, when successful designs are seen with 150 layers and more [11] in the case of convolutional networks. When we use one TITAN X GPU through Theano implementation, the MLP training takes about 2 hours.

Preprocessing. We do not particularly preprocess the input features as atomic fractions of elements are already between 0 and 1. We tried center each input column by subtracting the mean and dividing it by its standard deviation, and we do not observe particular improvement of results.

Dropouts. We find that adding dropouts to some extent were helpful. We find that as we kept adding dropout layers, starting from no dropout layer to 3 in MLP, the performance steadily increased. However, trying to add more than that ends up hurting the performance.

Hyperparameter tuning. For every network structure, we tried a number of different combinations of hyperparameter settings before fixing on one, through a grid search. Hyperparameters include the learning rate, minibatch size, learning rate drop rate (how many epochs to reduce learning rate to half), among others. For each setting, we train for a couple of iterations and then observe the dropping rate of validation error. The set with the steepest drop is chosen.

7. CONCLUSION

In this paper we demonstrated the application of deep learning as a step towards automated and accelerated materials discovery. We make use of fully-connected multi-layer perceptrons for chemical property prediction of compounds, by directly looking at their atomic composition ratios. To the best of our knowledge, this study is the first application of deep learning strategy towards large-scale materials discovery, utilizing a dataset as large as 300,000 and generating a higher prediction accuracy than present state-of-the-art. The best network achieves a prediction mean absolute error of 0.072 eV, within the bound of DFT simulation error which is known to be around 0.1eV.

As popular as deep learning has become these days, there is no free lunch. A single universal architecture, parameter set, loss function, training method or initialization cannot work for all kinds of problems. It can hardly be used as an

off-the-shelf classifier. We managed to treat the application problems from an agnostic point of view, but there is still extensive data-driven exploration conducted to adapt the learning model to each given problem.

A key challenge in applying deep learning to materials science problems is that the available labeled training samples are very limited. Simulated and experimental data are the two most common types of data in materials science. Experimental data refer to the trial-and-error iterations of experimental observations, examples being microscopic images taken directly from a materials sample. Such data are often very limited in the size and may contain noise of non-stationary variance affecting the data quality. Simulation data, on the other hand, are more abundant, available, and stable - less prone to human operation errors. This work makes use of large scale simulation data from DFT calculations.

However, even with the help of computational apparatus like DFT, extensive calibration of parameters is required and the running time is unbearable. Therefore the search of new materials is still limited. Besides, while DFT results are regarded reliable for organic materials, for inorganic materials the structure-property relationships are often more complicated and DFT less reliable. This is ascribed to the diversity of chemical elements, crystal structures and target properties [29]. In our experiments, bot organic and inorganic materials are included and randomly distributed between training and test set. A future work could be looking into how performances differ between these two types of compounds.

A more fundamental challenge persists, in the next step after an accurate model construction: how to distill information from learned model and extract crucial insights that relate structure to property; how to not only build robust predictive, quantitative models, but also interpret them with domain knowledge, and systematically integrate them in the discovery, engineering process of materials.

As databases containing various type of materials data are growing, being refined and becoming available everyday, the application of data mining techniques begins to gain large expectation in the business of faster and smarter materials discovery. Similar development has been successful in biological science, drug discovery and healthcare, and is yet to be seen in the materials domain. Our demonstration in this work could play a role in encouraging the use of large datasets, efficient analytics, and advanced computational models for all kinds of applications in this field.

Acknowledgment

This work is supported in part by the following grants: NIST award 70NANB14H012; AFOSR award FA9550-12-1-0458; DARPA award N66001-15-C-4036; NSF awards CCF-1029166, IIS-1343639, CCF-1409601; DOE awards DE-SC0007456, DE-SC0014330.

8. REFERENCES

- [1] Materials genome initiative for global competitiveness. www.whitehouse.gov/mgi, 2011. [Online; accessed 12-February-2016].
- [2] A. Agrawal and A. Choudhary. Perspective: Materials informatics and big data: Realization of the Fourth

- paradigm of science in materials science. *APL Materials*, 4(5):053208, 2016.
- [3] A. Agrawal, P. D. Deshpande, A. Cecen, G. P. Basavarsu, A. N. Choudhary, and S. R. Kalidindi. Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters. *Integrating Materials and Manufacturing Innovation*, 3(1):1–19, 2014.
- [4] J. Allison, D. Backman, and L. Christodoulou. Integrated computational materials engineering: A new paradigm for the global materials profession. *JOM: the journal of the Minerals, Metals & Materials Society*, 58(11):25–27, 2006.
- [5] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.
- [6] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010. Oral Presentation.
- [7] H. Bhadeshia, R. Dimitriu, S. Forsik, J. Pak, and J. Ryu. Performance of neural networks in materials science. *Materials Science and Technology*, 25(4):504–510, 2009.
- [8] E. D. Cubuk, S. S. Schoenholz, J. M. Rieser, B. D. Malone, J. Rottler, D. J. Durian, E. Kaxiras, and A. J. Liu. Identifying Structural Flow Defects in Disordered Solids Using Machine-Learning Methods. *Physical Review Letters*, 114(March):108001, 2015.
- [9] S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy. The high-throughput highway to computational materials design. *Nature materials*, 12(3):191–201, 2013.
- [10] J. Hafner, C. Wolverton, and G. Ceder. Toward computational materials design: the impact of density functional theory on materials research. *MRS bulletin*, 31(09):659–668, 2006.
- [11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.
- [12] A. J. Hey, S. Tansley, K. M. Tolle, et al. *The fourth paradigm: data-intensive scientific discovery*, volume 1. Microsoft research Redmond, WA, 2009.
- [13] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.
- [14] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. a. Persson. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013.
- [15] S. R. Kalidindi and M. D. Graef. Materials data science: Current status and future outlook. *Annual Review of Materials Research*, 45(1):171–193, 2015.
- [16] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.
- [17] 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*, 1:15010, 2015.
- [18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [19] R. Liu, A. Kumar, Z. Chen, A. Agrawal, V. Sundararaghavan, and A. Choudhary. A predictive machine learning approach for microstructure optimization and materials design. *Scientific reports*, 5, 2015.
- [20] R. Liu, Y. C. Yabansu, A. Agrawal, S. R. Kalidindi, and A. N. Choudhary. Machine learning approaches for elastic localization linkages in high-contrast composite materials. *Integrating Materials and Manufacturing Innovation*, 4(1):1, 2015.
- [21] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Physical Review B*, 89(9):094104, 2014.
- [22] 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*, 15(9):095003, 2013.
- [23] P. Raccuglia, K. C. Elbert, P. D. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, and A. J. Norquist. Machine-learning-assisted materials discovery using failed experiments. *Nature*, 533(7601):73–76, 2016.
- [24] B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, and V. Pande. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*, 2015.
- [25] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso. Rotation forest: A new classifier ensemble method. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(10):1619–1630, 2006.
- [26] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.
- [27] B. G. Sumpter and D. W. Noid. On the design, analysis, and characterization of materials using computational neural networks. *Annual Review of Materials Science*, 26(1):223–277, 1996.
- [28] Y. Sun, W. Zeng, Y. Zhao, Y. Qi, X. Ma, and Y. Han. Development of constitutive relationship model of ti600 alloy using artificial neural network. *Computational Materials Science*, 48(3):686–691, 2010.

- [29] I. Tanaka and A. Seko. Toward materials discovery with first-principles datasets and learning methods. In *Information Science for Materials Discovery and Design*, pages 173–186. Springer, 2016.
- [30] L. Yang and G. Ceder. Data-mined similarity function between material compositions. *Physical Review B*, 88(22):224107, 2013.
- [31] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, pages 649–657, 2015.