Science Bulletin 64 (2019) 612–616

Contents lists available at ScienceDirect

Science Bulletin

journal homepage: www.elsevier.com/locate/scib

Article

Discovering unusual structures from exception using big data and machine learning techniques

Jianshu Jie [a], Zongxiang Hu [a], Guoyu Qian [a], Mouyi Weng [a], Shunning Li [a], Shucheng Li [a], Mingyu Hu [a], Dong Chen [a], Weiji Xiao [a], Jiaxin Zheng [a], Lin-Wang Wang [b,*], Feng Pan [a,*]

[a] School of Advanced Materials, Peking University, Shenzhen Graduate School, Shenzhen 518055, China
[b] Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley 94720, USA

ARTICLE INFO

Article history:
Received 26 February 2019
Received in revised form 22 March 2019
Accepted 1 April 2019
Available online 5 April 2019

Keywords:
Machine learning
Gradient boosting decision tree
Band gap
Unusual structures

ABSTRACT

Recently, machine learning (ML) has become a widely used technique in materials science study. Most work focuses on predicting the rule and overall trend by building a machine learning model. However, new insights are often learnt from exceptions against the overall trend. In this work, we demonstrate that how unusual structures are discovered from exceptions when machine learning is used to get the relationship between atomic and electronic structures based on big data from high-throughput calculation database. For example, after training an ML model for the relationship between atomic and electronic structures of crystals, we find AgO_2F, an unusual structure with both Ag^{3+} and O_2^{2-}, from structures whose band gap deviates much from the prediction made by our model. A further investigation on this structure might shed light into the research on anionic redox in transition metal oxides of Li-ion batteries.

1. Introduction

In recent years, a variety of machine learning (ML) methods have been adopted to analyze high-throughput experimental or computational data [1–6], which helps to find new insights and spark novel ideas for material design. ML is the center piece of the artificial intelligence. One of its applications in materials science is aimed at establishing the relationship between structure and property, the ultimate goal in materials science. More specifically, it tries to establish a predictive relationship between the material fingerprints (including the features of the constituent elements, the atomic structure information, and any combination of these features) and the target property which we are interested in. The predictive power of ML schemes has been demonstrated in previous literatures focusing on properties such as band gap [7–10], elastic moduli [11], phase stability [12], ionic conductivity [13], thermal conductivity [14], melting temperature [15], and onset temperature of glass transition [16]. In these studies, the goal is to establish an ML model training on existing database. This model is then used to predict the properties for compounds where prior experimental and theoretical data do not exist. While tremendously useful, such model nevertheless does not help us to understand the rules and physics underlying the relationship between structure and property. Traditional materials science research is needed for a comprehensive understanding of them [17].

The success of ML schemes in previous studies is based on the presumption that a model developed based on the common trend of the data inside the database can be applied to most of the compounds. This presumption is valid for the "normal" compounds which have regular structural units as in the majority cases of the material database. Yet, there are always exceptions, and very often it is these exceptions which shed some new insights about the underlying physics, and open up new frontiers in science. The study of these exceptional cases can help us to establish new categories and rules. For example, while there is good correlation between the cohesion energy and melting point in a metallic system, Sn shows great deviation from the trend. Shao et al. [18,19] have made a careful examination over this and compared Sn to a "typical" system of Ag, finding the reason to be the relatively low melting point comparing to the Debye temperature for Sn. It has also been shown by Paul et al. [20] that we can learn from failing data by gathering the "abnormities" and deriving their common trends. However, such exceptional, or say the unusual cases are often difficult to find, especially in the traditional rational scientific inquiry approach. Here, we show how the ML can be used as a tool to pick up such unusual cases, and how such unusual cases can then be studied with traditional rational analysis to broaden our scientific knowledge. More specifically, in this work, we have identified 34 unusual compounds out of about 4,000 compounds in the

* Corresponding authors.
E-mail addresses: lwwang@lbl.gov (L.-W. Wang), panfeng@pkusz.edu.cn (F. Pan).

https://doi.org/10.1016/j.scib.2019.04.015
2095-9273/

database by looking at the outliners in the ML trained model prediction for the band gap, many of which show unusual structural units or other abnormities as reflected by the relationship between atomic and electronic structures. Afterwards, we choose one of them for investigation, AgO_2F, which has an unusual structure with both Ag^{3+} and O_2^{2-}.

2. Materials and methods

The workflow of our work is shown in Fig. 1. Data from existing database were used for machine learning, and structures with large prediction error were examined. Detailed HSE analysis were performed for those structures that contain unusual structural units and show rather large errors between the calculated band gaps and predictions by ML model. For other structures, we tried to find whether they show some abnormity by comparing with similar structures. Some unusual trends were found in this way, and for the rest of the structures, we believe that either our model still needs improvement, or the underlying physics are yet to be explored.

The original data were derived from MG database, a recently built electronic structure database constructed with both PBE and HSE calculations. The details of the database can be seen on website http://www.pkusam.com/bdm/. Structures with HSE band gaps larger than 0.1 eV were used.

In ML, we have built a model to predict the relationship between atomic structures and band gaps that were calculated using HSE. The model was trained using Lightgbm [21] with atomic positions in crystal structures with their symmetries (e.g. space group numbers) treated as categorical features in the scheme. 5-fold cross validation was used.

Features were generated in a similar way with Curtarolo et al. [10] in their work. Firstly, we chose 18 physical properties for each element. We used row number in periodic table, group number, electronegativity, enthalpy of fusion, enthalpy of atomization, enthalpy of vaporization, atomic mass, thermal conductivity (https://www.webelements.com/), covalent radii [22], element heat capacity, polarizability, ionization potential, standard molar entropy [23], molar volume (http://periodictable.com/Properties/A/MolarVolume.html), and valence electron that is commonly used in DFT calculations (http://cms.mpi.univie.ac.at/vasp/vasp/Recommended_PAW_potentials_DFT_calculations_using_vasp_5_2.html). Secondly, we made more properties by multiplication and division of these physical properties. Seven features were generated from each of these properties: the minimum of the property in all atoms in a unit cell, the maximum, the average, the standard error, the sum, and two others calculated with galvez matrix. Galvez matrix was obtained by multiplying the adjacency matrix **A** by the reciprocal square distance matrix **D**, and the last two features for a property **q** was calculated as

$$T^E = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |q_i - q_j| M_{ij},$$

$$T^E_{\text{bond}} = \sum_{\{ij\} \in \text{bonds}} |q_i - q_j| M_{ij}.$$

Thirdly, we calculated the coordination number, average bond lengths and the ratio between average bond lengths and covalent radii for each atom in the unit cell. We not only calculated minimum, maximum, sum, average and standard error for all these properties among each kind of atom in the unit cell, but also calculated these features for all metal atoms and all nonmetal atoms, respectively. We also used a feature to denote whether there were bonds between nonmetal atoms in the structure.

Fourthly, we classified structures into molecular crystals, ionic crystals with only simple ions, ionic crystals with groups like PO_4^{3-}, and covalent crystals. Features denoting whether the structure is 2D- or 3D-structures and whether all atoms are in their common oxidation states, as well as the *d* electron configurations, were taken into account.

Fifthly, some overall features were added regarding the crystal system, including space group number, number of elements, atomic density, and cell parameters.

Lastly, while keeping all categorical features, we used SelectKBest method in scikit-learn [24] to choose numerical features that have the greatest correlation with gaps and the final number of features used is 500.

The DFT calculations were performed by using PWmat [25–27]. NCPP-SG15-PBE pseudopotential [28,29] was used with a 50 Ryd cutoff in plane wave basis set. The unit cell for AgO_2F contains 2 formula units, and that for $KAgO_2$ contains 1. Kpoints for AgO_2F were $5 \times 4 \times 4$ in self-consistent calculation, and $6 \times 5 \times 5$ in non-self-consistent calculation. Kpoints for $KAgO_2$ were $7 \times 6 \times 4$ in self-consistent calculation, and $9 \times 7 \times 5$ in non-self-consistent calculation.

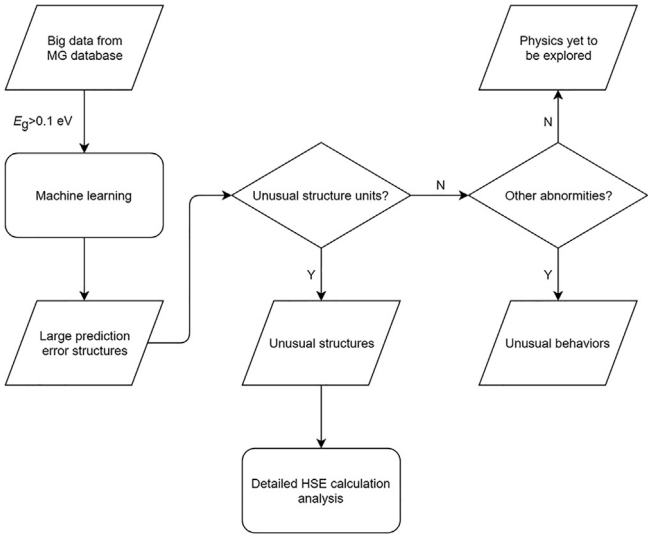


Fig. 1. The whole workflow.

3. Results and discussion

The overall performance of our model is comparable with existing works, despite the relatively small dataset we have used. R^2 score of our model is about 0.89. The scatter plots of results and distribution of prediction errors are shown in Fig. 2.

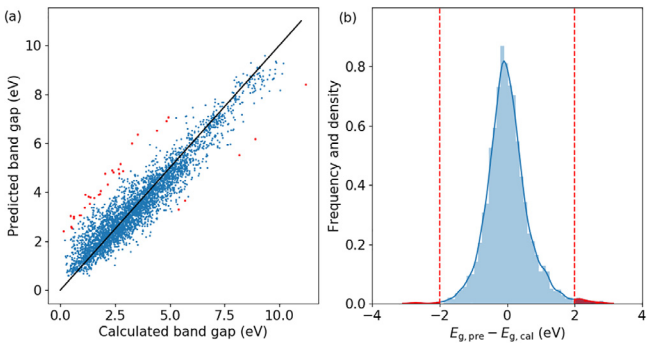


Fig. 2. (Color online) The machine learning results. (a) The scatter plot for predicted and calculated band gaps, and (b) the histogram of errors and the kernel density estimation of the probability density function. Errors > 2 eV are regarded as large and the corresponding structures are labeled as red points in (a) and in red region (outside of red dashed line) in (b).

There are many different kinds of structures with large prediction errors. Structures that have unusual structure units which affect the band gaps are listed in Table 1. Unusual structure units include unusual coordination circumstances, unusual oxidation states and ions that are not that common, like highly-distorted [MnO_6] octahedra in $Mn_9(PO_4)_8$, +4 formal charge of Bi atoms in $Y_2Bi_2O_7$, hybrid valence of transition metals in $Li_3Ni_3O_3F_5$, and metabisulfite ions in $K_2S_2O_5$. A more detailed analysis for these structures can be found in the Supplementary Data. Some other structures also have unusual structure units, but it needs more investigation to figure out how these units affect band gaps. Large-error structures may also help us to discover abnormities other than structures, like the abrupt increase in band gap of LiF comparing with other alkali halides, or the change in phase structure in IVA-VA covalent compounds. These will be listed and briefly discussed in the Supplementary Data, too.

There is one structure that raises our particular interest, AgO_2F, which is first found by global structure prediction with minima hopping method [30] and is included in ICSD database with collection code 670057. The formula implies some kind of interesting oxidation states in this structure, and this may have implications on anionic redox property, a recently hot topic on Li-ion batteries. Therefore, we take a deeper investigation in that structure. The first step will be the determination of oxidation states of atoms. The structure of AgO_2F is shown in Fig. 3a, and the calculated electronic structure is shown in Fig. 4a, b. The densities of states (DOS) plot is symmetric, and, as can be seen from Fig. S1 (online), the partial DOS of each O atom is symmetric as well. This means that the electrons in this system form pairs and there is no unbonded electron. Superoxide ions and Ag^{2+} have odd numbers of electrons, so they may not exist in this system. The O can then only exist as either peroxide ion or simple oxide ion [31]. The distance between O atoms is 1.30 Å, which indicates the existence of a chemical bond and precludes the existence of simple oxide ion. Besides, if O atoms exist as oxide ion, the silver will be in an irrational +5 oxidation states. As a result, O atoms can only be in the form of peroxide ions, and Ag should be in +3 oxidation state.

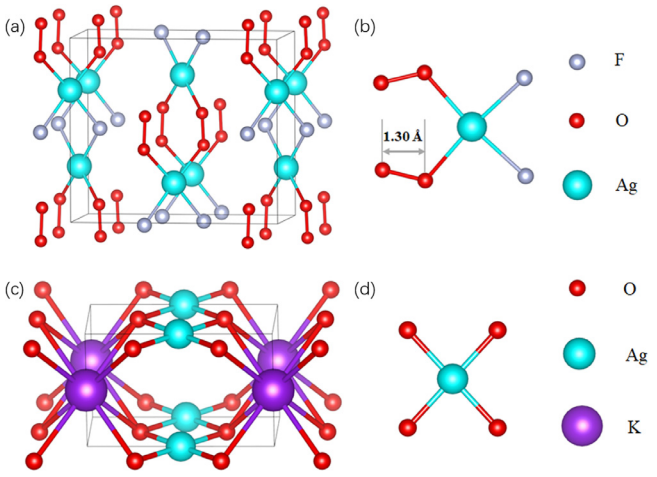
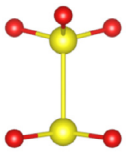
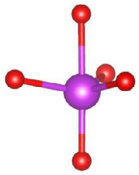
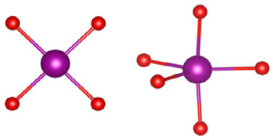
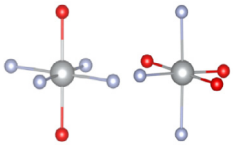


Fig. 3. (Color online) Comparison between atomic structure of AgO_2F and $KAgO_2$. (a) Crystal structure of AgO_2F, (b) coordination environment of Ag in AgO_2F, (c) crystal structure of $KAgO_2$, and (d) coordination environment of Ag in $KAgO_2$.

Then, we need to explain the failure of our model in this structure. It is found that this structure shows an unusually small band gap comparing with other Ag(III)-containing structures. As can be seen in Table 2, other such structures generally have a band gap near or larger than 1.8 eV, while the band gap of AgO_2F is only about 0.5 eV.

To better understand the small band gap, we choose $KAgO_2$, a structure in which Ag atoms are also in +3 oxidation state and planar square coordination, as a comparison. The atomic and electronic structure of $KAgO_2$ is shown in Figs. 3b and 4c, d, respectively. It can be seen that Ag and O have nearly equal contribution to the valence band maximum and conduction band minimum to

Table 1
(Color online) Structures with a prediction error >2 eV that have unusual structure units.

Formula	E_g (eV)		Structure units[a)]	Details in the Supplementary Data
	Calculated	Predicted		
$K_2S_2O_5$	4.407	6.530		Fig. S2
$Y_2Bi_2O_7$	0.517	2.529		Fig. S3
$Mn_9(PO_4)_8$	0.480	2.581		Fig. S4
$Li_3Ni_3O_3F_5$	0.594	2.938		Fig. S5

a) O, S, Mn, Bi, Ni and F.

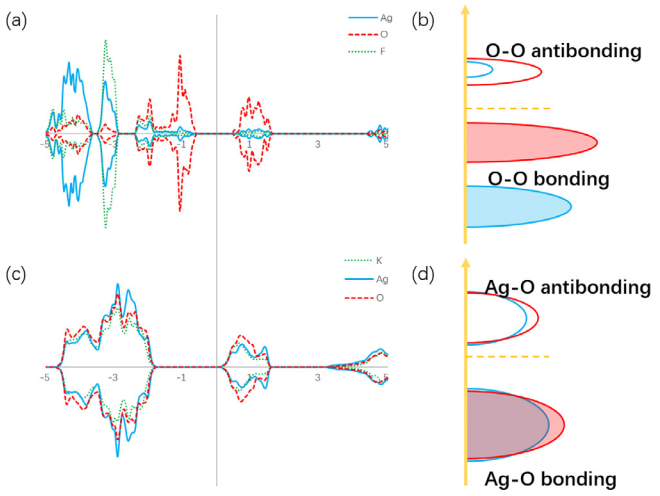


Fig. 4. (Color online) Comparison between electronic structure of AgO_2F and $KAgO_2$. (a) DOS plot of AgO_2F, (b) illustration for the band structure of AgO_2F, (c) DOS plot of $KAgO_2$ and (d) illustration for the band structure of $KAgO_2$.

Table 2
Calculated HSE band gaps for structures with Ag oxidation higher than +1.

Formula	Calculated band gap (eV)
$KAgO_2$[a)]	2.008
$KAgO_2$[a)]	2.616
$LiAgO_2$[a)]	1.829
$LiAgO_2$[a)]	1.853
$Ag_2P_2S_6$	2.462
$NaAgO_2$	1.997
$LiAgF_4$	2.632
Ta_2AgF_{12}	3.076
$ScAgO_3$	1.209
$SrAgO_2$	1.460
Cs_2KAgF_6	2.791
AgO_2F	0.461

[a)] Structures are shown in Fig. S9 (online).

$KAgO_2$, while in AgO_2F the contribution of Ag diminishes. Given that Ag ions are in the same oxidation state in both compounds and that the charge of K ions seldom differs from +1, the different scenario in electronic structures should be originated from the valence state of O ions. It is illustrated in Fig. 4 that the states near the band edges in AgO_2F are dominated by the non-bonding O-2*p* orbitals, which correspond to the unique features in peroxide dimer ions (O_2^{2-}) and exhibit similarity to typical peroxide materials such as Li_2O_2. The weak hybridization between these non-bonding orbitals and the orbitals of Ag eventually results in an unexpected decrease in band gap. Moreover, it should be emphasized that peroxide ions mostly exist in alkali compounds. The unstable +5 and metastable +3 states of Ag ion, as well as the unusual crystal structure of AgO_2F, could be the reason behind the stabilization of peroxide state of O ions in this little-explored material that has no alkaline ions. Similar phenomenon is also observed for CuO_2F, which has a crystal structure closely resembling that of AgO_2F and a large prediction error for band gap of 1.7 eV that, though smaller than the 2 eV threshold, can still distinguish the role of peroxide dimer ions on the electronic structure. The oxo-to-peroxo (O^{2-} to O_2^{2-}) transformation is generally recognized as anionic redox [32], a property that differs from the commonly-known cationic redox (e.g. the change in valence states of transition metal ions) and is currently a popular topic in materials science. Our result gives a new example of anionic redox, which offers a link to the phenomenon commonly found in Li-excess electrode materials. We believe that AgO_2F can be explored as a prototype for studying the anionic redox property in materials that contain ions (Ag^{3+} in this case) in uncommon valence states and manifest unusual crystal structures.

4. Conclusions

In this work, we perform machine learning for the prediction of band gap on a high throughput HSE calculation database to get the relationship between atomic and electronic structures. By analyzing the structures with remaining large errors, we find interesting physics, which might be due to the sudden change of a trend among common structures, the unusual valence states in the ions of ionic groups, or unusual coordination circumstances in the structures. Among them there is an interesting structure of AgO_2F. We have made a deep investigation in it and found that it contains both Ag in +3 oxidation states and peroxide ions. This offers a new example for anionic redox property, a hot topic in the investigation of Li-excess electrode materials. We demonstrate that combining the ML model to identify unusual structure, and traditional rational analysis to study these structures might provide an alternative use of machine learning in materials research, which can help us to uncover new physics and novel structural units from the existing materials science databases.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

Wang is supported by the Director, Office of Science (SC), Basic Energy Science (BES), Materials Science and Engineering Division (MSED), of the US Department of Energy (DOE) under Contract No. DE-AC02-05CH11231 through the Materials Theory program (KC2301) under Contract No. DE-AC02-05CH11231. This work was also financially supported by the National Key R&D Program of China (2016YFB0700600), Shenzhen Science and Technology Research Grant (ZDSYS201707281026184), and Guangdong Keylab Project (2017B0303010130).

Author contributions

Jiaxin Zheng, Lin-Wang Wang and Feng Pan designed the research. Mouyi Weng performed high-throughput calculation in MG database. Jianshu Jie performed the machine learning with the help form Weiji Xiao and Dong Chen. Jianshu Jie, Zongxiang Hu and Shunning Li analyzed the results and wrote the manuscript. Shucheng Li, Mingyu Hu and Dong Chen helped writing the manuscript. Shunning Li, Guoyu Qian and Feng Pan revised the manuscript.

Appendix A. Supplementary data

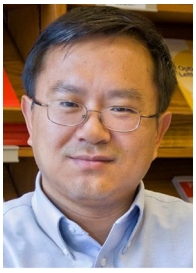
Supplementary data to this article can be found online at https://doi.org/10.1016/j.scib.2019.04.015.

References

[1] Pilania G, Gubernatis JE, Lookman T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. Comput Mater Sci 2017;129:156–63.
[2] Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys Rev Lett 2018;120:145301.

[3] Morales-García Á, Valero R, Illas F. An empirical, yet practical way to predict the band gap in solids by using density functional band structure calculations. J Phys Chem C 2017;121:18862–6.
[4] Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science 2015;349:255–60.
[5] Mueller T, Kusne AG, Ramprasad R. Machine learning in materials science: recent progress and emerging applications. Rev Comput Chem 2016;29:186–273.
[6] Liu Y, Zhao TL, Ju WW, et al. Materials discovery and design using machine learning. J Materiomics 2017;3:159–77.
[7] Partha D, Joe B, Somnath D, et al. Informatics-aided bandgap engineering for solar materials. Comput Mater Sci 2014;83:185–95.
[8] Grégoire M, Matthias R, Vivekanand G, et al. Machine learning of molecular electronic properties in chemical compound space. New J Phys 2013;15:095003.
[9] Ramakrishnan R, Hartmann M, Tapavicza E, et al. Electronic spectra from TDDFT and machine learning in chemical space. J Chem Phys 2015;143:084111.
[10] Isayev O, Oses C, Toher C, et al. Universal fragment descriptors for predicting properties of inorganic crystals. Nat Commun 2017;8:15679.
[11] de Jong M, Chen W, Notestine R, et al. A statistical learning framework for materials science: application to elastic moduli of k-nary inorganic polycrystalline compounds. Sci Rep 2016;6:34256.
[12] Ghiringhelli LM, Vybiral J, Levchenko SV, et al. Big data of materials science: critical role of the descriptor. Phys Rev Lett 2015;114:105503.
[13] Fujimura K, Seko A, Koyama Y, et al. Accelerated materials design of lithium superionic conductors based on first-principles calculations and machine learning algorithms. Adv Energy Mater 2013;3:980–5.
[14] Seko A, Togo A, Hayashi H, et al. Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and Bayesian optimization. Phys Rev Lett 2015;115:205901.
[15] Seko A, Maekawa T, Tsuda K, et al. Machine learning with systematic density-functional theory calculations: application to melting temperatures of single and binary component solids. Phys Rev B 2014;89:054303.
[16] Liu Y, Zhao TL, Yang G, et al. The onset temperature (T_g) of As_xSe_{1-x} glasses transition prediction: a comparison of topological and regression analysis methods. Comput Mater Sci 2017;140:315–21.
[17] Shi SQ, Gao J, Liu Y, et al. Multi-scale computation methods: their applications in lithium-ion battery research and development. Chin Phys B 2016;25:018212.
[18] Shao G. Melting of metallic and intermetallic solids: an energetic view from DFT calculated potential wells. Comput Mater Sci 2008;43:1141–6.
[19] Boer FR, Boom R, Matterns WCM, et al. Cohesion in metals: transition metal alloys. North-Holland; 1989.
[20] Paul R, Kathering CE, Philip DFA, et al. Machine-learning-assisted materials discovery using failed experiments. Nature 2016;533:73–7.
[21] Ke G, Meng Q, Finley T, et al. Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Proc Syst 2017;30:3146–54.
[22] Cordero B, Gómez V, Platero-Prats AE, et al. Covalent radii revisited. Dalton Trans 2008;21:2832.
[23] Haynes WM, Lide DR. CRC handbook of chemistry and physics. 92nd ed. CRC Press; 2011.
[24] Fabian P, Gael V, Alexandre G, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
[25] Jia W, Fu J, Cao Z, et al. Fast plane wave density functional theory molecular dynamics calculations on multi-GPU machines. J Comput Phys 2013;251:102–15.
[26] Jia W, Cao Z, Wang L, et al. The analysis of a plane wave pseudopotential density functional theory code on a GPU machine. Comput Phys Commun 2013;184:9–18.
[27] Lin L. Adaptively compressed exchange operator. J Chem Theory Comput 2016;12:2242–9.
[28] Hamann DR. Optimized norm-conserving vanderbilt pseudopotentials. Phys Rev B 2013;88:085117.
[29] Schlipf M, Gygi F. Optimization algorithm for the generation of ONCV pseudopotentials. Comput Phys Commun 2015;196:36–44.
[30] Tiago FTC, Lin S, Maximilian A, et al. Identification of novel Cu, Ag, and Au ternary oxides from global structure prediction. Chem Mater 2015;27:4562–73.
[31] Zhuo ZQ, Chaitanya DP, John V, et al. Spectroscopic signature of oxidized oxygen states in peroxides. J Phys Chem Lett 2018;9:6378–84.
[32] Zheng JX, Teng GF, Yang JL, et al. Mechanism of exact transition between cationic and anionic redox activities in cathode material Li_2FeSiO_4. J Phys Chem Lett 2018;9:6262–8.

Jianshu Jie received his B.S. degree from the College of Chemistry and Molecular Engineering, Peking University in 2016 and is now pursuing his M.S. degree under the supervision of Prof. Jiaxin Zheng at School of Advanced Materials, Peking University Shenzhen Graduate School. His current researches mainly relate to the application of machine learning techniques to electronic structures of crystals.



Lin-Wang Wang received his Ph.D. degree at Cornell University in 1991. He worked in Cornell University (1991–1992), National Renewable Energy Lab (1992–1995) as a postdoc, Biosym/Molecular Simulations Inc. (1995–1996) and National Renewable Energy Laboratory (1996–1999) as a staff scientist. From 1999, he has worked in Lawrence Berkeley National Laboratory and is now a senior staff scientist. His research interests mainly focus on the development of ab initio electronic structure calculation methods and the applications of these methods in materials design and discovery.

Feng Pan got his B.S. degree from Department of Chemistry, Peking University in1985 and Ph.D. degree from Department of P&A Chemistry, University of Strathclyde, UK, with "Patrick D. Ritchie Prize" for the best Ph.D. in 1994. Now he is a National 1000-plan Professor, Founding Dean of School of Advanced Materials, Peking University Shenzhen Graduate School, and Director of National Center of Electric Vehicle Power Battery and Materials for International Research. He is engaged in fundamental research and product development of novel energy conversion and storage materials & devices.