PREDICTION OF THE FORMATION ENERGY OF LI-BASED COMPOUNDS USING MACHINE LEARNING METHODS

Abduqahhorov Foziljon Ismoiljon o‘g‘li
Center for Advanced Technologies
E-mail: aryzen179@gmail.com
Tursunkulov Oybek Muydinovich
Center for Advanced Technologies
E-mail: oybtmuz@gmail.com
Toxtabayev Ernazar Rustembay uli
Center for Advanced Technologies
E-mail: tokhtabaevernazar@gmail.com
https://doi.org/10.5281/zenodo.23086641

Annotatsiya

Ushbu tadqiqotda 21 764 ta birikmani o'z ichiga olgan Materials Project tizimidagi zichlik funksionali nazariyasi hisoblari asosida litiyli birikmalarning hosil bo'lish energiyasini bashorat qiluvchi mashinali o'rganish modeli yaratildi.

Vazifani bajarish uchun algoritmi qo'llanildi. 20 ta turli xususiyatlar asosida o'qitilgan model R^2 = 0.8932 determinatsiya koeffitsiyentiga va 0.1584 eV/atom o'rtacha mutlaq xatolikka erishdi. Ushbu natijalar materialshunoslikda yangi birikmalar xususiyatlarini tez va aniq baholashda mazkur usullarning samaradorligini tasdiqlaydi.

Abstract

In this study, a machine learning model was developed to predict the formation energy of lithium-based compounds using density functional theory (DFT) calculation results obtained from the Materials Project database, comprising 21,764 compounds. The XGBoost algorithm was employed for the prediction task. Trained on twenty distinct features, the model achieved a coefficient of determination of R^2 = 0.8932, with a mean absolute error of 0.1584 eV/atom. These results confirm the effectiveness of machine learning methods in materials science for the rapid and accurate evaluation of the properties of novel compounds.

Аннотация

В данном исследовании была разработана модель машинного обучения для прогнозирования энергии образования соединений на основе лития с использованием результатов расчетов теории функционала плотности (DFT), полученных из базы данных Materials Project, включающей 21 764 соединения. Для выполнения задачи прогнозирования применялся алгоритм XGBoost. Обученная на двадцати различных признаках, модель достигла коэффициента детерминации R^2 = 0.8932 со средней абсолютной ошибкой 0.1584 эВ/атом. Эти результаты подтверждают эффективность методов машинного обучения в материаловедении для быстрой и точной оценки свойств новых соединений.

Keywords: machine learning, XGBoost, Materials Project, formation energy, lithium compounds, DFT, regression.

Introduction

In modern materials science, the preliminary assessment of the physicochemical properties of new compounds is of great importance. Although conventional DFT (Density Functional Theory) calculations provide high accuracy, they require substantial computational resources and time. Computing several thousand compounds by the DFT method may take weeks or even months. For this reason, machine learning methods have been widely applied in materials science in recent years [5, 8].

Li-based compounds are used as key materials in the modern energy sector, particularly in lithium-ion batteries. The formation energy of such compounds is an important indicator that determines their thermodynamic stability. The aim of this paper is to develop a model that predicts the formation energy of lithium-based compounds with high accuracy using machine learning methods.

Methodology and Dataset

The open Materials Project database was used for this study [1]. Using the MP-API and pymatgen libraries, the DFT calculation results for a total of 21,764 lithium-containing compounds were retrieved [7]. For each compound, the following key properties were obtained: energy per atom, formation energy, band gap, density, unit cell volume, number of atoms, number of chemical elements, and thermodynamic stability status.

All results in the database were obtained through a unified computational workflow: the DFT calculations were performed in the VASP software package using plane-wave basis sets and PAW (Projector Augmented Wave) pseudopotentials, with the PBE functional within the generalized gradient approximation employed as the exchange–correlation functional [1]. This ensures the mutual comparability of the values obtained for different compounds.

In addition to the original properties, 13 derived features were engineered to improve the accuracy of the model: atomic density, volume per atom, logarithmic transformations of volume and density, the square root of the number of atoms, the product of density and the number of elements, and others. As a result, a total of 20 parameters were prepared for the model.

XGBoost was selected as the machine learning algorithm [4]. This algorithm sequentially combines a large number of weak decision trees to produce a powerful, highly accurate regression model. For model training, the dataset was split into two parts in an 80:20 ratio using the Scikit-learn library [6]: 17,411 compounds were used for training, and 3,050 compounds were used for testing the accuracy of the model. The model employed 3,000 decision trees, a tree depth of 8, and a learning rate of 0.02. An early stopping mechanism was applied to prevent overfitting of the model.

Results and Discussion

The trained XGBoost model demonstrated the following results on the test dataset:

Table 1. Main results obtained in the study

Indicator	Value
R^2 — coefficient of determination	0.8932
MAE — mean absolute error	0.1584 eV/atom
RMSE — root mean square error	0.2989 eV/atom
Train size — training set size	17,411 compounds
Test size — test set size	3,050 compounds
n_features — number of features	20

The coefficient of determination $R^2 = 0.8932$ indicates that our model predicts the formation energy with an accuracy of 89.32 %. The mean absolute error amounted to 0.16 eV/atom, which lies within the accuracy limits of DFT calculations themselves. This result is comparable to the results ($R^2 = 0.85–0.92$) obtained for similar tasks in the scientific papers published by Faber et al. (2016) [2] and Ward et al. (2018) [3].

Feature importance analysis showed that the most significant indicators for predicting the formation energy were ranked as follows: the product of the number of atoms and the number of elements 37.4 %, energy per atom 7.8 %, number of chemical elements 7.2 %, square root of the number of atoms 6.3 %, and volume per atom 6.0 %. This result confirms that the compositional complexity of the crystal structure and the atomic density are the main factors determining the formation energy.

Physicochemical Interpretation of the Results

Formation energy is the amount of energy released or absorbed when a compound is formed from its constituent elements (in their most stable, standard states). If this value is negative, the compound is thermodynamically stable relative to its constituent elements; the more negative the value, the more robust and resistant to decomposition the compound is considered. Therefore, the ability to predict this quantity accurately makes it possible to assess, prior to experiment, whether a new compound can be synthesized at all and whether it will remain stable under laboratory conditions.

The model's $R^2 = 0.8932$ means that it can explain nearly 89 percent of the variability in formation energy based solely on the chemical composition and simple structural features of a compound. The mean absolute error (MAE) is 0.16 eV/atom, which is close to the typical deviation of DFT calculations from experimental results. From a practical standpoint, this means that a demanding DFT calculation lasting several days can be replaced by a prediction of nearly the same accuracy performed within seconds.

The feature importance analysis also yields meaningful conclusions from a physical point of view. The considerable importance of the "energy per atom" feature (7.8 %) is natural, since the formation energy is calculated precisely through the difference of total energies. The dominance of the features representing compositional complexity — the numbers of atoms and elements, as well as the number of chemical elements — reveals an important trend: in multicomponent compounds with complex compositions, the distribution of chemical bonding energy changes, and this directly affects thermodynamic stability. At the same time, the high contribution of the engineered product-type feature (37.4 %) should be interpreted with caution, as a purely statistical correlation: it does not represent a direct physical cause but serves as a generalized measure of compositional complexity.

Overall, the results show that stability can be estimated with sufficient accuracy based only on the composition and a few simple features of a compound, without knowledge of its full crystal structure. This enables a transition in materials science from the traditional "trial-and-error" approach to the stage of rational design and opens the way for the rapid screening of thousands of candidate compounds.

Application of the Results in Nanotechnology

Lithium-based compounds are key materials for a number of areas of modern nanotechnology; therefore, a model capable of rapidly assessing their stability is of practical importance. First, in the design of nanostructured electrode materials for lithium-ion batteries (nanoparticles, nanotubes, thin films), the model rapidly identifies, among thousands of candidate compositions, those that are stable, i.e., those with a negative formation energy. Second, in the search for solid electrolytes for solid-state batteries, it helps to identify stable phases in advance.

In addition, the model can be useful in the design of nanocatalysts and two-dimensional (2D) materials for the pre-experimental screening of the synthesizability of new compositions, as well as in assessing the stability of nanoscale interfaces and protective coatings inside batteries. Such rapid assessment significantly accelerates the discovery of new nanomaterials and saves research resources.

Conclusion

In this study, a machine learning model for predicting the formation energy was developed using the XGBoost algorithm based on the DFT calculation results of 21,764 lithium-based compounds obtained from the Materials Project database. The model achieved $R^2 = 0.89$ and MAE = 0.16 eV/atom. This result confirms that machine learning methods can be applied with high efficiency in materials science. The developed model makes it possible to assess the stability of new lithium-based compounds within seconds rather than hours or days. In future research, it would be reasonable to improve the model using neural networks and to extend it to the prediction of other physical properties.

References

1. Jain A., Ong S. P., Hautier G., Chen W., Richards W. D., Dacek S., et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation // APL Materials. — 2013. — Vol. 1, № 1. — P. 011002.
2. Faber F. A., Lindmaa A., von Lilienfeld O. A., Armiento R. Machine Learning Energies of 2 Million Elpasolite (ABC_2D_6) Crystals // Physical Review Letters. — 2016. — Vol. 117, № 13. — P. 135502.
3. Ward L., Dunn A., Faghaninia A., Zimmermann N. E. R., Bajaj S., Wang Q., et al. Matminer: An open source toolkit for materials data mining // Computational Materials Science. — 2018. — Vol. 152. — P. 60–69.
4. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — San Francisco: ACM, 2016. — P. 785–794.
5. Schmidt J., Marques M. R. G., Botti S., Marques M. A. L. Recent advances and applications of machine learning in solid-state materials science // npj Computational Materials. — 2019. — Vol. 5, № 1. — P. 83.
6. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. — 2011. — Vol. 12. — P. 2825–2830.
7. Ong S. P., Richards W. D., Jain A., Hautier G., Kocher M., Cholia S., et al. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis // Computational Materials Science. — 2013. — Vol. 68. — P. 314–319.
8. Butler K. T., Davies D. W., Cartwright H., Isayev O., Walsh A. Machine learning for molecular and materials science // Nature. — 2018. — Vol. 559, № 7715. — P. 547–555.
9. Xie T., Grossman J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties // Physical Review Letters. — 2018. — Vol. 120, № 14. — P. 145301.
10. Chen C., Ye W., Zuo Y., Zheng C., Ong S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals // Chemistry of Materials. — 2019. — Vol. 31, № 9. — P. 3564–3572.
11. Saal J. E., Kirklin S., Aykol M., Meredig B., Wolverton C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD) // JOM. — 2013. — Vol. 65, № 11. — P. 1501–1509.

12. Ward L., Agrawal A., Choudhary A., Wolverton C. A general-purpose machine learning framework for predicting properties of inorganic materials // npj Computational Materials. — 2016. — Vol. 2, № 1. — P. 16028.
13. Sendek A. D., Yang Q., Cubuk E. D., Duerloo K. A. N., Cui Y., Reed E. J. Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials // Energy & Environmental Science. — 2017. — Vol. 10, № 1. — P. 306–320.
14. Ramprasad R., Batra R., Pilania G., Mannodi-Kanakkithodi A., Kim C. Machine learning in materials informatics: recent applications and prospects // npj Computational Materials. — 2017. — Vol. 3, № 1. — P. 54.
15. Curtarolo S., Hart G. L. W., Nardelli M. B., Mingo N., Sanvito S., Levy O. The high-throughput highway to computational materials design // Nature Materials. — 2013. — Vol. 12, № 3. — P. 191–201.