Workflow of Materials Machine Learning for Perovskite Materials

Subjects: Others

Contributor: Junya Wang, Pengcheng Xu, Xiaobo Ji, Minjie Li, Wencong Lu

Perovskite materials have been one of the most important research objects in materials science due to their excellent photoelectric properties as well as correspondingly complex structures. Machine learning (ML) methods have been playing an important role in the design and discovery of perovskite materials, while feature selection as a dimensionality reduction method has occupied a crucial position in the ML workflow.

Keywords: perovskites ; materials design ; machine learning

1. Introduction

Machine learning (ML), as an interdisciplinary technique covering multiple fields of statistics, computer science, and mathematics, has been widely used in the medical, bioinformatics, financial, and agriculture fields [1][2][3][4][5]. Especially in the materials field, ML technology has accelerated the design and discovery of new materials by constructing models for the prediction of their properties [6][7]. In recent years, perovskite materials have drawn the attention of many scholars due to their excellent properties, such as excellent electrical conductivity, ferroelectricity, superconductivity, longer carrier diffusion lengths, a tunable bandgap (E_g), and high light absorption that can be applied in solar cells, light-emitting diodes, lasers, and photocatalysis materials fields [8][9][10][11].

2. Workflow of Materials Machine Learning

As shown in **Figure 1**, the workflow of ML in materials could be divided into four steps: data preparation, feature engineering, model evaluation and selection, and model application [12].

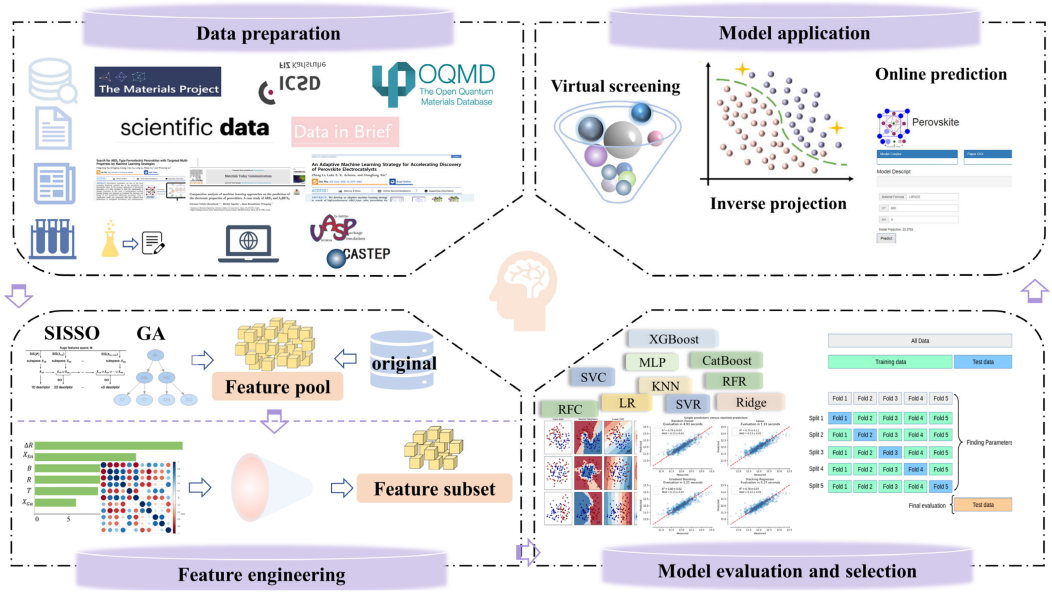

Figure 1. The general workflow of materials ML.

Data preparation includes data collection and data preprocessing. Materials data could be generally obtained through publicly available materials databases, published papers, experimental data of the same standard, data journals, and density functional theory (DFT) calculations [13][14][15][16][17]. The latest data can be obtained by searching the publications, but it is time-consuming and laborious. Data from data journals and databases can be obtained in a short time, but the latest data are generally not available in a timely manner. *Scientific Data* by Springer Nature and *Data in Brief* by Elsevier

are the more representative data journals. **Table 1** lists the commonly used material databases, including perovskites. Experimental data may be a good source of data, but it could be costly. DFT calculations are susceptible to material systems, which may lead to the doubling of time and computing resources. Data preprocessing is essential due to the characteristics of multi-source data and the high noise of the material data. To ensure the availability of data, common preprocessing operations include filling in missing values, removing duplicates and outliers, dimensionless processing, treating data imbalances, and rationally dividing data [18][19].

Table 1. Commonly used materials databases, including perovskites.

Name	URL	Data Type
The Perovskite Database Project (PDP)	https://www.perovskitedatabase.com (accessed on 19 March 2023)	Exp.
Open Quantum Materials Database (OQMD)	http://www.oqmd.org/ (accessed on 19 March 2023)	Comp.
Materials Project (MP)	https://materialsproject.org/ (accessed on 19 March 2023)	Comp.
Computational Materials Repository (CMR)	https://cmr.fysik.dtu.dk/ (accessed on 19 March 2023)	Comp.
The Inorganic Crystal Structure Database (ICSD)	https://icsd.fiz-karlsruhe.de/index.xhtml (accessed on 19 March 2023)	Exp.
Materials Platform for Data Science (MPDS)	https://mpds.io/#modal/menu (accessed on 19 March 2023)	Comp. and Exp.
Automatic-FLOW for Materials Discovery (AFLOW)	http://www.aflowlib.org/ (accessed on 19 March 2023)	Comp.

Feature engineering, including feature construction and feature selection, is an extremely important part of the ML workflow. In most ML processes, the quality of the data related to the sample size and feature dimensionality, as well as the validity of the features, determines the upper limit of the model's performance. In general, a high ratio of sample size to feature dimension would lead to better model performance. When the existing features do not contain enough valid information to cause low model performance, new features can be either constructed based on domain knowledge or generated by simple mathematical transformation of existing features through algorithms such as the Sure Independence Screening Sparsifying Operator (SISSO) and genetic algorithm (GA) to improve model performance [20][21]. The properties of materials are influenced by their composition, structure, experimental conditions, and environmental factors, but there may be weakly correlated, uncorrelated, or redundant features in the data. For the original set of features in the data, feature selection can remove the redundant features and keep the key features that are easily accessible and have a significant impact on the target variable to further improve the model's performance while increasing the computational efficiency.

Before the model construction, it is necessary to confirm the type of models corresponding to classification or regression according to the target variables are discrete or continuous, respectively. There are many ML algorithms, but no perfect algorithm exists. Although for a specific classification or regression task, the researchers could choose linear, nonlinear, or ensemble algorithms preliminary based on their understanding or guessing of the potential "structure-property relationship" of the materials. It is still difficult to determine the most suitable algorithm based on the limited data volume. Even with the same data and algorithm, the trained model will not be the same with the different hyperparameters. Therefore, it is necessary to evaluate a series of models to select the relatively optimal one. Model performance and model complexity are the key factors that determine model selection. Model performance can be measured by evaluation metrics calculated based on the true and predicted values of the target variable. Common evaluation metrics for regression tasks include coefficient of determination (R^2), correlation coefficient (R), mean square error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and average relative error (MRE), while common evaluation metrics for classification tasks include accuracy (ACC), area under the curve (AUC), recall, precision, and F_1 score. To ensure the reliability of the results, the hold-out method and cross-validation method are generally used to evaluate the models after the evaluation metrics are determined. Common methods of cross validation include 5-fold cross validation (5-fold CV), 10-fold cross validation (10-fold CV), and leave-one-out cross validation (LOOCV). Furthermore, researchers tend to choose the model with better performance and lower model complexity. After selecting a specific ML algorithm, hyperparameter optimization is usually performed to further improve the performance of the model, and the final model is determined after the determination of hyperparameters. Contemporary hyperparametric optimization algorithms can be mainly classified into various types, including grid-search, Bayesian-based optimization algorithms, gradient-based optimization, and population-based optimization.

The final aim of ML is to predict the target variables of unknown samples based on the trained model. The three major scenarios of model application are high-throughput screening (HTS), inverse design, and the development of online prediction programs. HTS uses the constructed model to predict the target variables of a huge number of virtual samples in order to filter out samples with high performance potential to guide experimental synthesis [22][23]. The inverse design can be used to obtain the features of designed samples via the inverse projection method, which is an effective way to realize the material from properties to composition [24][25]. The prediction of designed samples helps screen out candidates with breakthrough performance and improves computational efficiency. The development of an online prediction program makes it possible to quickly achieve the prediction of target properties by simply inputting the necessary information, such as a chemical formula, on the input page, which facilitates the extension of model application and effectively realizes model sharing [26].

References

1. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. Science 2015, 349, 255–260.
2. Shehab, M.; Abualigah, L.; Shambour, Q.; Abu-Hashem, M.A.; Shambour, M.K.Y.; Alsalibi, A.I.; Gandomi, A.H. Machine learning in medical applications: A review of state-of-the-art methods. Comput. Biol. Med. 2022, 145, 105458.
3. Henrique, B.M.; Sobreiro, V.A.; Kimura, H. Literature review: Machine learning techniques applied to financial market prediction. Expert Syst. Appl. 2019, 124, 226–251.
4. Liakos, K.G.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. Machine Learning in Agriculture: A Review. Sensors 2018, 18, 2674.
5. Larranaga, P.; Calvo, B.; Santana, R.; Bielza, C.; Galdiano, J.; Inza, I.; Lozano, J.A.; Armananzas, R.; Santafe, G.; Perez, A.; et al. Machine learning in bioinformatics. Brief. Bioinform. 2006, 7, 86–112.
6. Butler, K.T.; Davies, D.W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. Nature 2018, 559, 547–555.
7. 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 Comput. Mater. 2019, 5, 83.
8. Tao, Q.; Xu, P.; Li, M.; Lu, W. Machine learning for perovskite materials design and discovery. NPJ Comput. Mater. 2021, 7, 23.
9. Min, K.; Cho, E. Accelerated discovery of potential ferroelectric perovskite via active learning. J. Mater. Chem. C 2020, 8, 7866–7872.
10. Gok, E.C.; Yildirim, M.O.; Haris, M.P.U.; Eren, E.; Pegu, M.; Hemasiri, N.H.; Huang, P.; Kazim, S.; Uygun Oksuz, A.; Ahmad, S. Predicting Perovskite Bandgap and Solar Cell Performance with Machine Learning. Sol. RRL 2021, 6, 2100927.
11. Yin, W.-J.; Weng, B.; Ge, J.; Sun, Q.; Li, Z.; Yan, Y. Oxide perovskites, double perovskites and derivatives for electrocatalysis, photocatalysis, and photovoltaics. Energy Environ. Sci. 2019, 12, 442–462.
12. Xu, P.; Chen, H.; Li, M.; Lu, W. New Opportunity: Machine Learning for Polymer Materials Design and Discovery. Adv. Theory Simul. 2022, 5, 2100565.
13. Xu, P.; Chang, D.; Lu, T.; Li, L.; Li, M.; Lu, W. Search for ABO3 Type Ferroelectric Perovskites with Targeted Multi-Properties by Machine Learning Strategies. J. Chem. Inf. Model. 2022, 62, 5038–5049.
14. Zhou, Q.; Lu, S.; Wu, Y.; Wang, J. Property-Oriented Material Design Based on a Data-Driven Machine Learning Technique. J. Phys. Chem. Lett. 2020, 11, 3920–3927.
15. Belsky, A.; Hellenbrandt, M.; Karen, V.L.; Luksch, P. New developments in the Inorganic Crystal Structure Database (ICSD): Accessibility in support of materials research and design. Acta Crystallogr. Sect. B-Struct. Sci.Cryst. Eng. Mat. 2002, 58, 364–369.
16. 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, 65, 1501–1509.
17. Jain, A.; Ong, S.P.; Hautier, G.; Chen, W.; Richards, W.D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. APL Mater. 2013, 1, 011002.
18. Dong, Y.; Zhang, Y.; Ran, M.; Zhang, X.; Liu, S.; Yang, Y.; Hu, W.; Zheng, C.; Gao, X. Accelerated identification of high-performance catalysts for low-temperature NH3-SCR by machine learning. J. Mater. Chem. A 2021, 9, 23850–23859.

19. Lu, T.; Li, H.; Li, M.; Wang, S.; Lu, W. Predicting Experimental Formability of Hybrid Organic-Inorganic Perovskites via Imbalanced Learning. J. Phys. Chem. Lett. 2022, 13, 3032–3038.
20. Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L.M. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. Phys. Rev. Mater. 2018, 2, 083802.
21. Liu, S.; Wang, J.; Duan, Z.; Wang, K.; Zhang, W.; Guo, R.; Xie, F. Simple Structural Descriptor Obtained from Symbolic Classification for Predicting the Oxygen Vacancy Defect Formation of Perovskites. ACS Appl. Mater. Interfaces 2022, 14, 11758–11767.
22. Mai, J.; Lu, T.; Xu, P.; Lian, Z.; Li, M.; Lu, W. Predicting the maximum absorption wavelength of azo dyes using an interpretable machine learning strategy. Dyes Pigment. 2022, 206, 110647.
23. Tao, Q.; Lu, T.; Sheng, Y.; Li, L.; Lu, W.; Li, M. Machine learning aided design of perovskite oxide materials for photocatalytic water splitting. J. Energy Chem. 2021, 60, 351–359.
24. Lu, T.; Li, H.; Li, M.; Wang, S.; Lu, W. Inverse Design of Hybrid Organic–Inorganic Perovskites with Suitable Bandgaps via Proactive Searching Progress. ACS Omega 2022, 7, 21583–21594.
25. Yang, C.; Ren, C.; Jia, Y.; Wang, G.; Li, M.; Lu, W. A machine learning-based alloy design system to facilitate the rational design of high entropy alloys with enhanced hardness. Acta Mater. 2022, 222, 117431.
26. Shi, L.; Chang, D.; Ji, X.; Lu, W. Using Data Mining To Search for Perovskite Materials with Higher Specific Surface Area. J. Chem. Inf. Model. 2018, 58, 2420–2427.
