

Lightweight Extendable Stacking Framework for Structure Classification in Atomistic Simulations

Yanhao Deng, Yangyang Wang, Ke Xu, and Yanming Wang*



Cite This: <https://doi.org/10.1021/acs.jctc.3c00838>



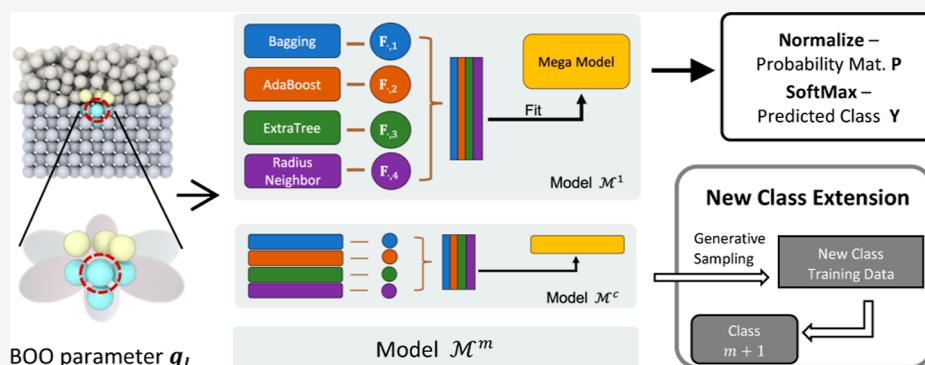
Read Online

ACCESS |

Metrics & More

Article Recommendations

Supporting Information



ABSTRACT: Identifying an atom's local crystal structure is one crucial step in many atomistic simulation analyses. However, many traditional methods are available to only a few limited types of structures, and their performance often relies on manually determined parameters, which may lead to poor classification results in complex material systems. Machine learning models can enhance accuracy and generalizability, but they typically require large amounts of data and computation. This issue could be more severe for deep-learning-based frameworks, especially when confronted with unfamiliar crystal structures. To address this challenge, we propose a lightweight and extendable stacked structure (LESS) classifier, which adopts bond orientational order parameters as features and assembles several efficient machine learning methods as based models. The LESS classifier can recognize a variety of crystal structures, e.g., amorphous, mono, and binary structures, with over 98.8% accuracy on our validation data set, outperforming many current methods even including some deep-learning methods. Our model can also conduct probabilistic classification that aids in the interpretation of atomic structures in complicated environments such as heterogeneous interfaces. Furthermore, when exposed to a completely unknown crystal structure, the LESS framework can efficiently incorporate this new knowledge with generative sampled data from the current model. Overall, our model exhibits great potential as an accurate and flexible atomic structure identification tool featuring high efficiency in both learning and retraining.

INTRODUCTION

Molecular dynamics (MD) simulation is capable of capturing motions of particles at the microscopic scale and predicting the dynamical properties of systems, making it a powerful scientific tool applicable to physics, biology, and chemistry research.^{1–3} In the analysis of MD simulations, crystal structure classification is often a fundamental first step. Accurate classification is vital for mechanistic studies on many important phenomena such as the formation of solid solution, nucleation and growth, ion transport across lattices, interface evolution, etc.^{4–6}

There are several traditional methods for classifying crystal structures.⁷ Some of these methods are based on the usage of coordination numbers and mapping patterns.⁸ For example, common neighbor analysis (CNA) considers the arranging characteristics of neighboring atoms around an atom.⁹ Ackland-Jones analysis assigns the radius information and bond angles to the closed-pack structures by measuring their deviations.¹⁰ The idea of describing topological symmetries of a particle's

neighborhood is also utilized in common neighborhood parameter (CNP),¹¹ cumulative CNP,¹² and polyhedral template matching (PTM).¹³ Another approach is to implement an order parameter derived from the location of neighboring particles.¹⁴ Ronchetti proposed a bond orientation order (BOO), which adopts the form of spherical harmonics with rotational invariance. Bond angle analysis¹⁵ uses the parameters calculated from bond pair analysis. The above methods typically require certain preknowledge of the target crystals, such as the number of nearest neighbors, for making meaningful predic-

Received: August 1, 2023

Revised: October 26, 2023

Accepted: October 26, 2023

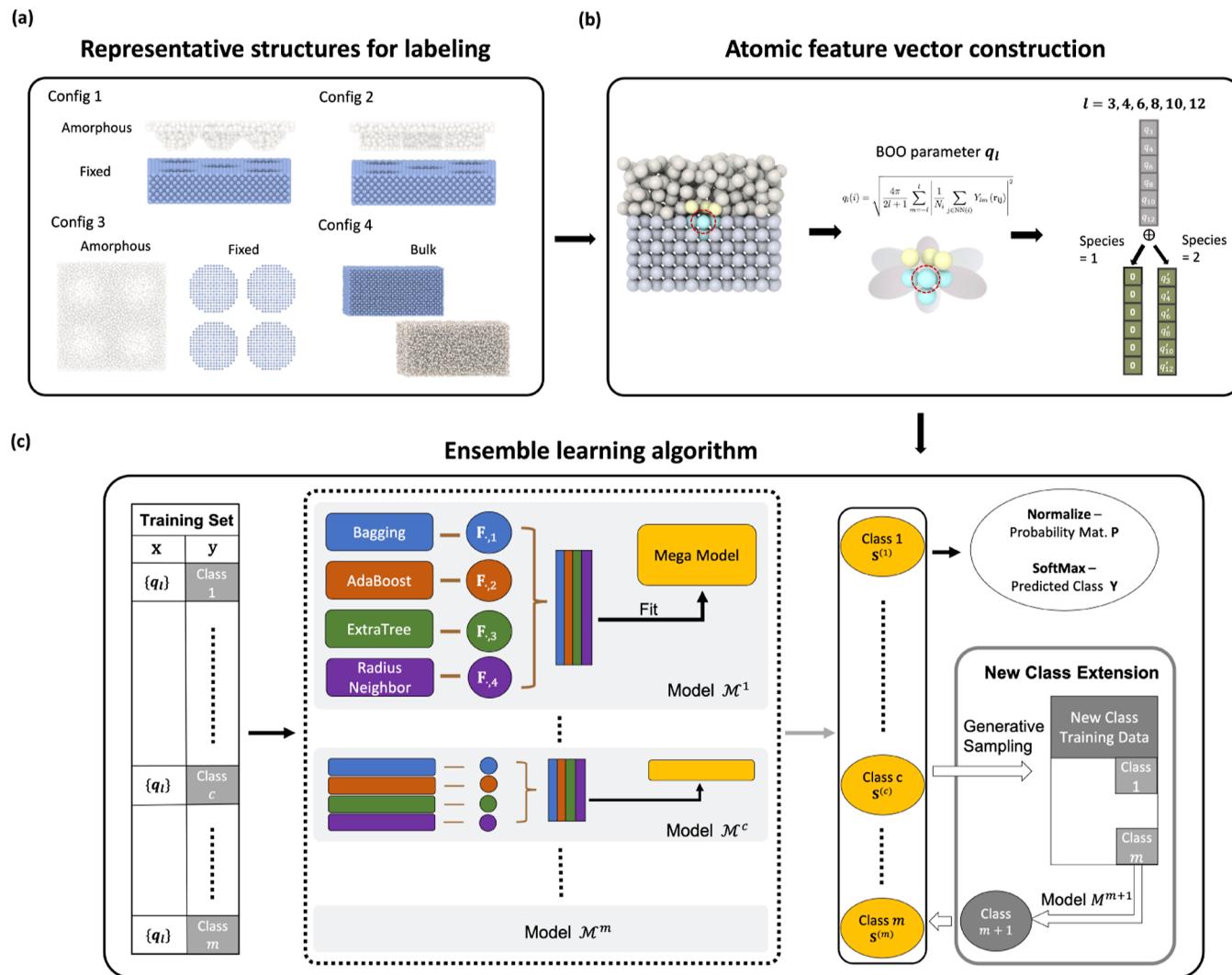


Figure 1. Workflow of our ensemble learning framework for the classification. (a) Four representative structures designed to generate the labeled data for each crystal structure. (b) Construction of the 12-dimension feature vector for the blue particle marked by the red circle. The feature vector consists of BOO parameters with a series of l . The prime sign denotes the BOO parameters calculated using neighbors with a different type from the central atom. (c) Workflow of the ensemble learning algorithm.

tions. However, for a complicated multiphase system or a completely new material system, usually it is challenging to provide an accurate estimation of these parameters beforehand. Thus, these traditional approaches may only be applicable to a limited number of crystal structures and are faced with great difficulties in their generalizability for classifying novel materials with unconventional and complex microstructures.¹⁷

This motivates the recent development of various machine learning (ML) methods.¹⁸ In general, ML algorithms can automatically mine the internal connections between atoms' local behaviors and their structural information. Therefore, combining ML with the conventional methods described above could be a promising approach for developing a high-precision crystal structure classifier. Along this direction, descriptors such as CNA signatures, radial structure order, and interface-induced order parameters^{5,19–21} have been converted to features fed to traditional ML methods, multilayer perception, and deep-learning architectures such as the crystal graph convolutional neural network (CGCNN).^{22,23} While deep-learning methods tend to become a mainstream ML solution, several common problems still exist. The typical long and complex features make

the model hard to interpret. Moreover, the learned deep features may barely be transferable if the ML model is designed for a specific material system.²⁴ Also, the training process could be highly data-demanding, with restrictive hardware requirements such as high-end graphics processing units. Clearly, there is a pressing need to develop a portable, flexible, and extendable ML model for identifying atomic structures from MD trajectories.

This work proposed an ensemble ML framework for efficient crystal structure identification, named the lightweight and extendable-stacked structure (LESS) classifier. Using BOO parameters as descriptors, the learned model can identify a wide range of crystal structures in various scenarios, including both single species and binary compounds, with satisfactory accuracy. Besides, when an extension of the built model is required to identify new structures, it is demonstrated that the retraining cost of our framework, in both time and training data size, can be significantly reduced by the design of a generative sampling scheme.

METHODS

As shown in Figure 1, our ensemble learning framework consists of three parts: training data set preparation, atomic feature construction, and implementation of the learning algorithm. A more detailed description of this framework is given as follows.

Data Preparation. We first generated the labeled data for each structure as the training data set. As it is challenging to differentiate between solid and amorphous atoms at the interface, we specially designed the data set to include such complicated interface shapes with four representatives shown in Figure 1a. Configurations 1–3 were constructed to represent common situations encountered in real-world applications such as etching and growth. For each of these configurations, in the region labeled as solid, the atoms were fixed at their perfect crystalline positions, while the amorphous regions were generated by a MD simulation at a temperature above the melting point. It should be noted that the near-interface liquid atoms were excluded from the training data set, as it is difficult to construct reliable microstructural features of these atoms. Configuration 4 represents the bulk cases for both crystal and amorphous structures. In addition to these four representative configurations, structures of similar shapes but different sizes were also included in the data set. Figures S1 and S2 are all the configurations used as training data and validation data for the face-centered cubic (fcc) structure and NaCl structure as representatives.

These atomic configurations were generated by ATOMSK²⁵ with manual adjustments. MD simulations were all performed using the large-scale atomic/molecular massively parallel simulator.²⁶ The interatomic potentials used for each structure are listed in Table 1. For all simulations, the time step was set to

Table 1. Materials and Potentials Chosen for Each Crystal Structure

structure	material	potential
fcc	Au	EAM ³¹
bcc	Fe	EAM/FS ³²
hcp	Mg	EAM/FS ³³
DC	Si	SW ³⁴
NaCl	NaCl	buck/coul/long ³⁵
CsCl	CsCl	EIM ³⁶
CaF ₂	CaF ₂	born/coul/long ³⁷
graphite	C	AIREBO ³⁸

0.001 ps. We simulated the system on Nosé-Hoover style non-Hamiltonian equations of motion as the isothermal–isobaric (NPT) ensemble for 10 ps in total.^{27–30}

Feature Construction. The 12-dimensional feature vectors contain local environment information for each atom. This vector is composed of BOO parameters, which integrate local structural information with rotational and translational invariance. BOO parameters are calculated according to eq 1, where N_i is the number of nearest neighbors for atom i within a cutoff radius r_{cut} , Y_{lm} is the spherical harmonics, and \mathbf{r}_{ij} is the spherical coordinate from atom i to its neighbor j , as illustrated in the middle of Figure 1b. A collection of BOO parameters with l values of 3, 4, 6, 8, 10, and 12 was found capable of representing the characteristics of most crystal structures, and thus, these values were used as the first six features. For binary structures with two species, another six BOO parameters were calculated, with the nearest neighbors being of a different species from the

center atom; while for single-species structures, these parameters were set to 0.

$$q_l(i) = \sqrt{\frac{4\pi}{2l+1} \sum_{m=-l}^l \left| \frac{1}{N_i} \sum_{j \in \text{NN}(i)} Y_{lm}(\mathbf{r}_{ij}) \right|^2}, r_{ij} < r_{\text{cut}}(i) \quad (1)$$

For a certain crystal structure, there may exist a specific number of neighbors that can best describe the character of this structure.³⁹ To determine this number for an arbitrary crystal structure, an adaptive cutoff radius scheme was adopted^{40,41} as eq 2, where P is a padding constant and N is the number of neighbors to calculate the radius. These two parameters were tuned such that the neighbors are the atoms within a distance of the first minimum of pair correlation function $g(r)$.

$$r_{\text{cut}}(i) = P \times \left(\frac{1}{N} \sum_{j=1}^N r_{ij} \right) \quad (2)$$

ML Workflow. For the task of atomic structure classification, we adopted an ensemble ML method called stacking to take advantage of different models, as each basic ML model may perform well but fail in certain corner cases. Our framework consists of Level-0 Models (based models) and Level-1 Models (a mega model). The former fits the training data and generates predictions to be compiled. The latter learns how to best combine the predictions of the based models and gives the final structure classification results.

We trained m different models for m classes of crystal structures, respectively. For testing, all results from each of the m separate models were computed and concatenated. By comparing the results for each structure class, we could decide on the final classification. The detailed process of model training is as follows.

Based Models. The based models for the m crystal structure classes would be denoted as \mathcal{M} , comprising four based models. Original data of atomic feature vectors, according to their labeled structure y , were sent to each of the models. The input data for x were split into k -fold for validation. One fold was used for prediction while the other $k - 1$ fold was adopted to train the model. In this way, each model \mathcal{M} would generate prediction results for all the data in a specific structure class. For instance, a model for target class c , named \mathcal{M}_c , has a loss function LF written as

$$\text{LF}_{\mathcal{M}_c}(\mathcal{M}_c(\mathbf{x}), I_{y,c}), \quad I_{y,c} = 1 \text{ when } y = c \text{ or } 0 \text{ otherwise} \quad (3)$$

We performed the ablation study to choose the based models and prioritized the ensemble models. The models emphasizing the neighborhood were also tried. Finally, we selected four different based models according to the performance in Figure 2b. They are bagging, ada-boosting, extra trees, and ·-radius neighbor. They have all been implemented in the Scikit-learn package⁴² as listed in Table S1. The tuned parameters are given in Table S2. The model selection details about the restriction of performance are listed in Table S3. Considering the generalizability of the framework, we also provided a flexible way for users to manipulate different alternative models.

For testing, all data were sent to those m models, and the results were called out-of-fold (OOF) features. Those features were compressed together into an $n \times m$ matrix \mathbf{F} , where n

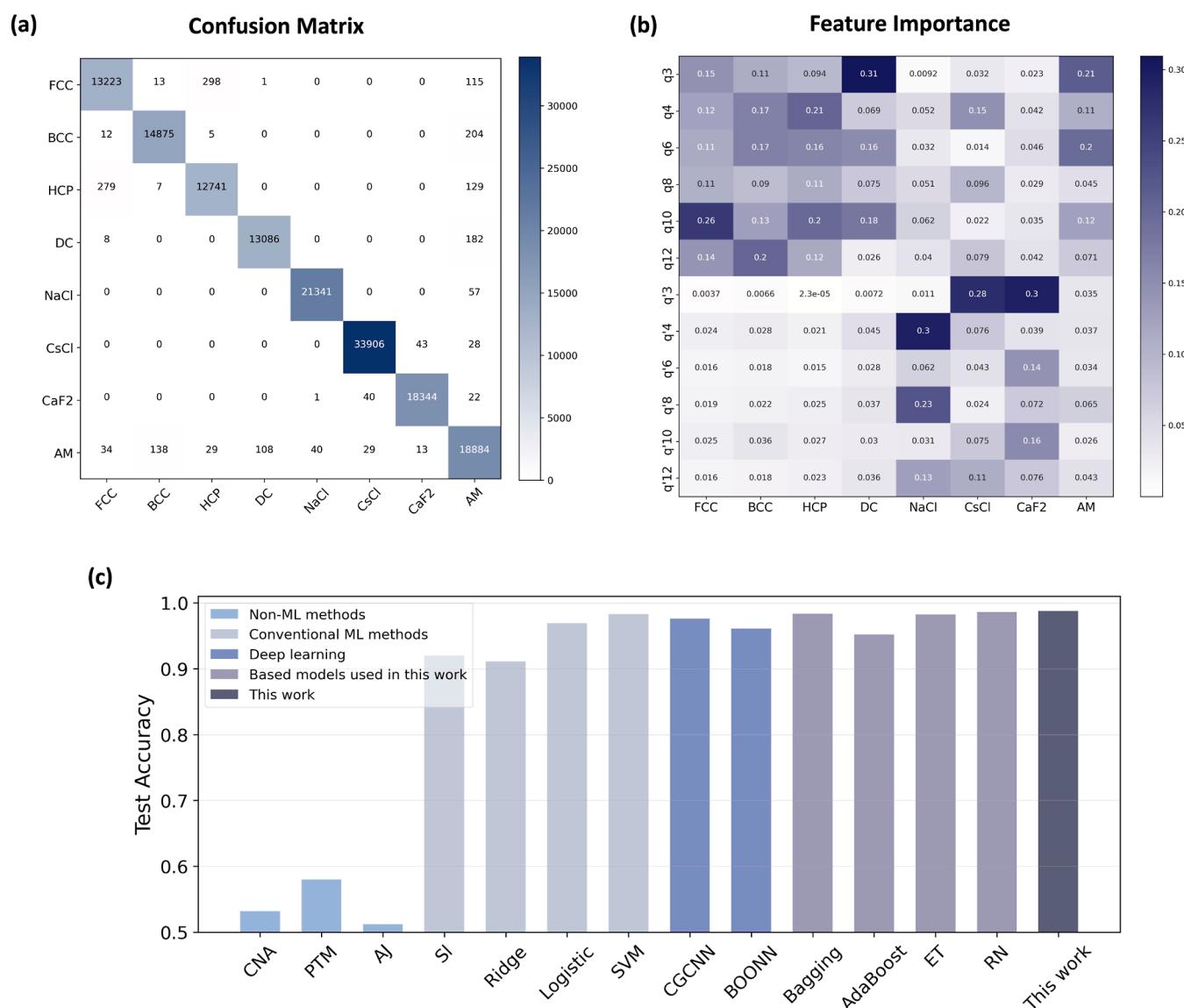


Figure 2. (a) Confusion matrix for the ensemble learning over the whole data set. Column: predictions; Row: true labels. (b) Importance of each feature for all of the crystal structure classes. (c) Accuracy of the test set for different models. For non-ML methods, nonavailable structures are all classified wrongly. The methods from left to right: CNA, PTM, Ackland-Jones, simple linear regression, ridge regression, logistic regression, support vector machine, CGCNN, bond order parameters neural network (BOONN), bagging, adaptive boosting, extra trees, radius neighbor, and this work.

denotes the test atom number and m denotes the number of models \mathcal{M} .

$$\mathbf{E}_{ij} = \mathcal{M}_j(\mathbf{x}_i), \quad i \in \{1, 2, \dots, n\}, j \in \{1, 2, \dots, m\} \quad (4)$$

Mega Model. The generated $n \times m$ OOF matrix \mathbf{F} would be sent to the mega model for training. The loss function would be written as

$$\text{LF}_{\mathcal{M}_{\text{mega}}}(\mathcal{M}_{\text{mega}}(\mathbf{F}), I_{y,c}), \quad I_{y,c}=1 \text{ when } y=c \text{ or } 0 \\ \text{otherwise} \quad (5)$$

We adopted light gradient-boosting machine (LGBM) regressor⁴³ as the mega model because the boosting process is a good fit for the selection of the OOF features F.

- For atom n labeled as structure class c , the corresponding value in $\mathbf{F}_{n,c}$ tends to be close to 1. Then the gradient-based one-side sampling drops the instances with value

$F_{n,c} \approx 1$ meaning that instances that have not been used well for training contribute more.

- For atom n' that is not labeled as structure class c , the corresponding value in $\mathbf{F}_{n',c}$ tends to be close to 0. Then the exclusive feature bundling handles sparse matrix well and drops redundant features when $\mathbf{F}_{n',c} \approx 0$ and improves the training speed without hurting accuracy.

For testing, we used different mega models to produce discrete results (classifier) or continuous results (regressors).

$$\mathbf{S}_i = \mathcal{M}_{\text{mega}}(\mathbf{E}_i), \quad i \in \{1, 2, \dots, n\} \quad (6)$$

Inference. Since there are m models, one could then extract the predicted class by comparing $S_i^{(1)}$, $S_i^{(2)}$, ..., $S_i^{(m)}$ and formulate a final prediction \hat{Y}_i for the i -th testing atom after the SoftMax layer.

$$\hat{\mathbf{Y}}_i = \text{SoftMax}(\mathbf{S}_i^{(1)}, \mathbf{S}_i^{(2)}, \dots, \mathbf{S}_i^{(m)}) \quad (7)$$

Combining all those results together as a $n \times m$ matrix \mathbf{S} , they would be normalized and could later be analyzed by users as the probability matrix \mathbf{P} .

$$\mathbf{P}_{i,j} = \frac{\mathbf{S}_{i,j} - E(\mathbf{S}_{i,:})}{\text{Var}(\mathbf{S}_{i,:})}, \quad i \in \{1, 2, \dots, n\}, j \in \{1, 2, \dots, m\} \quad (8)$$

Currently, we adopted the LightGBM regressor, and users can choose the mega model by themselves.

Extension. Our framework is highly flexible when exposed to data containing unknown crystal classes. This extendibility is granted by a generative sampling scheme. Specifically, we combined the Metropolis-Hastings and Gibbs sampling algorithms to overcome the risk of imbalanced data. We regulated that n atoms should be generated for each of the previous m classes. We then retrained those previous c models and one new model with $(m+1)n$ number of atoms, where n atoms are from each model.

The detailed Gibbs sampling algorithm was carried out as follows: to obtain n samples for the sampled features $\mathbf{Q} = (Q_4, Q_6, \dots, Q_{12})$ from a joint distribution $M(Q_4, Q_6, \dots, Q_{12}')$, denoting the i_{th} sample by $\mathbf{Q}^{(i)} = (Q_4^{(i)}, \dots, Q_{12}^{(i)})$, the procedures can be described as follows.

In this work, P_{ac} is set to be 0.8. $f_Q(q_i)$ is the normal distribution $\mathbf{N}(q_i, s_{Q_i}/\sqrt{n_{Q_i}})$, where s denotes sample variance and the value of $s_{Q_i}/\sqrt{n_{Q_i}}$ is calculated prior to computation. However, if too many permutations are rejected and cannot be accepted as new data, then the mean value of generated data would be appended.

Algorithm 1 Class Extension

Require: $N \geq 0 \wedge 0 < P_{ac} < 1$

Ensure: a list of sampled data D with number N .

```

 $\mathbf{Q}^{(0)} \leftarrow (\mathbb{E}[Q_4], \mathbb{E}[Q_6], \dots, \mathbb{E}[Q_{12}])$ 
 $D \leftarrow [\mathbf{Q}^{(0)}]$ 
while  $n \neq N$  do
   $\mathbf{Q}^{(n+1)} \leftarrow (q_4^{(n+1)}, q_6^{(n+1)}, \dots, q_{12}^{(n+1)}) \sim (f_{Q_4}(q_4^{(n)}), f_{Q_6}(q_6^{(n)}), \dots, f_{Q_{12}}(q_{12}^{(n)}))$ 
  if  $M(\mathbf{Q}^{(n+1)}) > P_{ac}$  then
     $D$  append  $\mathbf{Q}^{(n+1)}$ 
     $n \leftarrow n + 1$ 
  end if
end while

```

RESULTS AND DISCUSSION

Classification Performance. Following the workflow of the LESS classifier, validations have been achieved on the identification of several types of crystal structures. Besides the traditional structures with single species like the fcc structure, body-centered cubic (bcc) structure, hexagonal close-packed (hcp) structure, and diamond cubic (DC) structure, binary ionic structures including rock salt (NaCl) structure, cesium chloride (CsCl) structure, and fluorite (CaF_2) structure are also available, as well as the amorphous (AM) structure.

The crystal structure classification model generally performs very well on the validation data set, as demonstrated by the confusion matrix in Figure 2a. Major confusions only occasionally occur for atoms near the crystal interface, either between fcc and hcp, which may be attributed to their similar close-packed stacking mode, or related to the AM structures, as even fully randomized arrangement of atoms may yield symmetric local structure by chance. Moreover, the randomly generalized validation data set may also include the bulk data simulated at different temperatures from the training data, so the accuracy of

our model also proves its flexibility in dealing with the temperature issue. Figure S5 plots the classification results of our model compared with the CNA method regarding temperature fluctuation, showing the good robustness of our model. Unlike the traditional approach of utilizing BOO parameters with fixed thresholds based on human knowledge, the feature importance analysis in Figure 2b shows that our model comprehensively learns the distribution of each crystal structure over several selected BOO parameters, thereby improving the accuracy of the classification. More details are provided by the distribution of each feature in Figures S3 and S4.

Figure 2c shows that non-ML methods, such as CNA, PTM, and AJ, have very poor performance on our validation data set that contains a wide variety of crystal structures due to their limited applicability and flexibility. The detailed performance of each crystal structure is plotted in Figure S6. While most ML methods can achieve high accuracy, they have their own strengths and weaknesses with respect to different structures, as well as requiring large computational costs for both training and retraining. Our ensemble model combining several efficient-based models reaches an even slightly higher and more stable accuracy, making it suitable for many practical applications.

Thus, in Figure 3, beyond the validation data set, three additional test cases were designed to illustrate the framework's capability of analyzing complex structures in MD simulations. The first case involves an atomistic configuration of copper polycrystals created under nonperiodic boundary conditions, as shown in Figure 3a. The upper part of the figure depicts three different orientations colored in white, blue, and red, respectively, and the lower part shows that almost all of the atoms inside the grains are correctly recognized as fcc. The grain boundaries are clearly indicated by the non-fcc structures predicted by the model. Only some atoms on the surface may be classified as the hcp structure because of the similar packing modes. In the second case, the model's ability to capture solid/liquid (order/disorder) interfaces was demonstrated, as shown in Figure 3b. Here, a heat flux was introduced to melt the copper bar from its left surface. During this transient phase transition process, our model successfully traced the interface movement by distinguishing the crystalline atoms from the amorphous ones. To validate our classification results, we computed the temperature distribution along the horizontal direction, whose variation was described by color changing from red (higher temperature) to blue (lower temperature). It is observed that the position corresponding to the white spot presents the melting point of copper (1273 K), and this position is consistent with the solid–liquid interface identified by our model. The LESS classifier can also be operated in the regression mode, where the probability of an atom belonging to each crystal type is predicted to provide more information to interpret the classification results. This feature is illustrated by the third case that presents a Si (DC)/Al (fcc) interface obtained from a short MD simulation, as shown in Figure 3c. When we observe the average probability of each atomic layer belonging to the classes, those with the highest probability include DC, fcc, NaCl, and AM. Specifically, for layers near the interface, a high probability is seen for the NaCl and AM structures, suggesting binary and disordered features of this heterogeneous interface. For layers inside the bulk material, while in the classification mode, the atoms are undoubtedly sorted into the correct crystal types, in the regression mode, oscillation appears for the probability distribution functions, which may help reveal the subtle difference in the atoms' local environment between

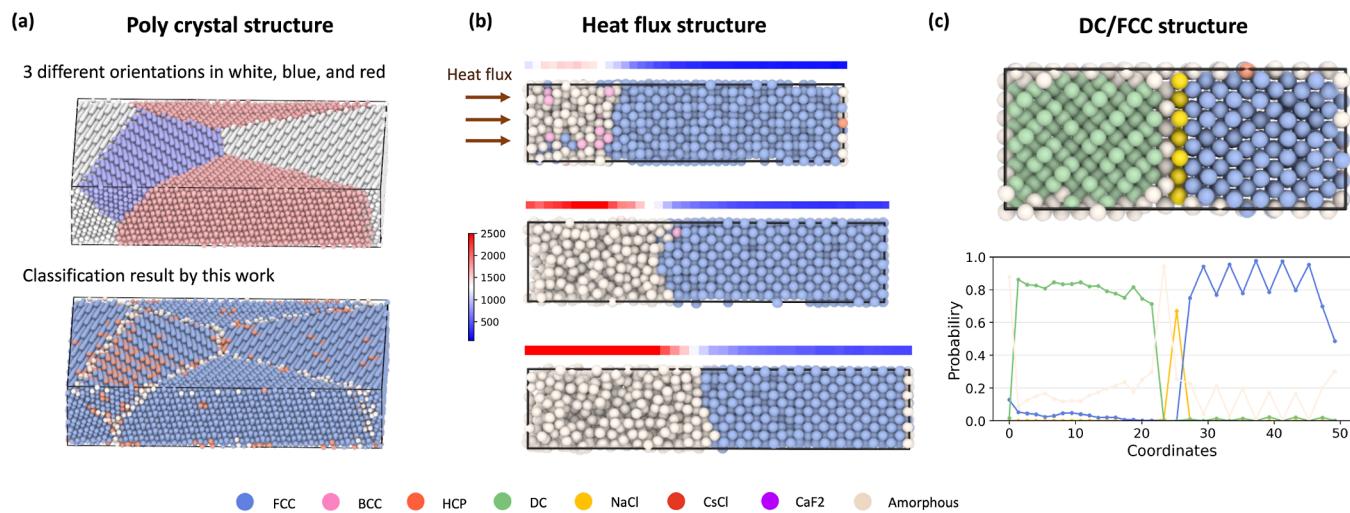


Figure 3. Structure predictions on test cases. (a) fcc structure with different orientations colored white, blue, and red. The figure below shows the classification results. (b) MD snapshots from the melting process of copper (Cu) crystal. The heat flux comes from the left side of the bar. The calculated temperature distribution is plotted above each snapshot, where the melting point of Cu (1273 K) is colored white. (c) Probability for each atom near the DC/fcc interface is predicted by the regression model. Along the vertical axis, the probability is averaged over each layer.

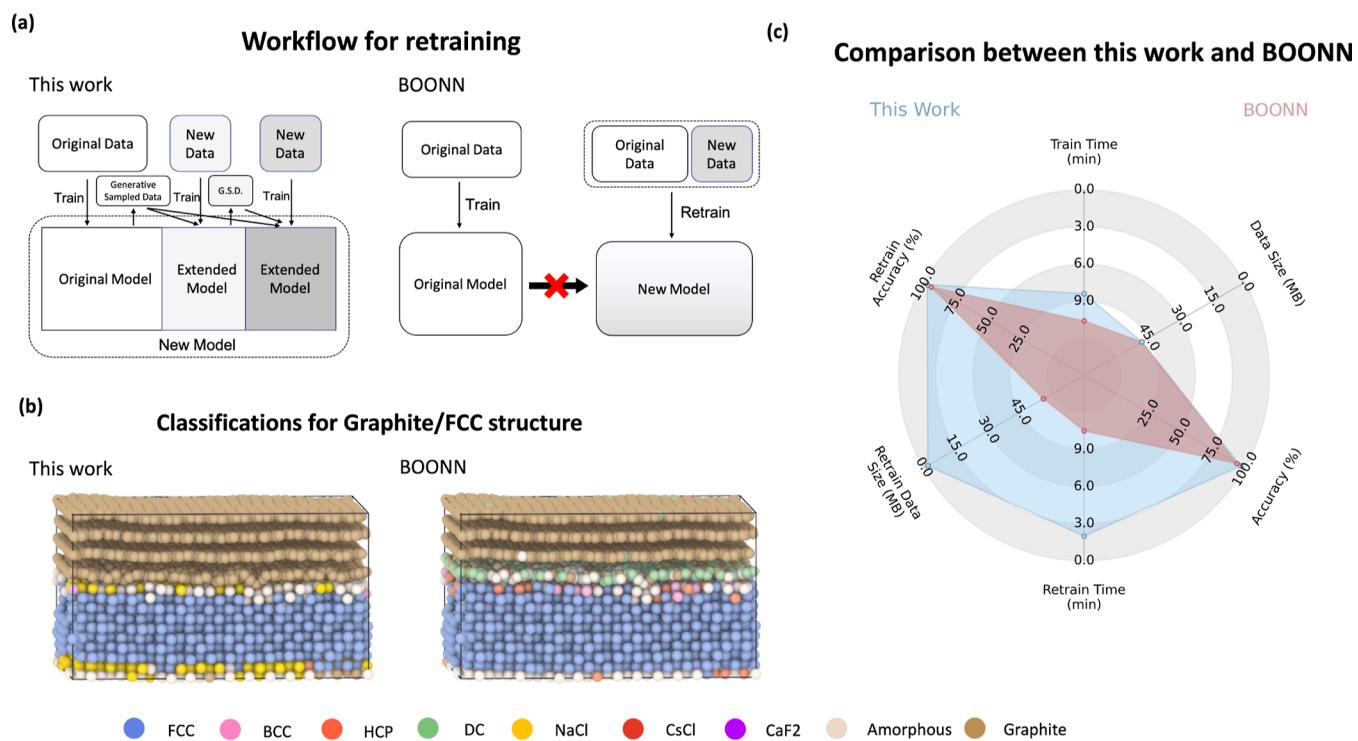


Figure 4. Comparison between the LESS classifier and the BOONN classifier on extendability. (a) Workflow to include one new structure type into the existing framework. (b) Predictions on the graphite/fcc structure by the extended LESS classifier. (c) Cost and efficiency comparison for both training and retraining processes. Both models are trained on the same CPU environment.

stacking layers. Overall, our model gave reasonable interpretations for these test cases, especially when compared with the performance of non-ML methods in Figures S7 and S8. These demonstrate the feasibility of our model for real applications.

Extendability. The flexibility and efficiency of the LESS classifier are evident in how easily it can be extended to recognize a new crystal structure. This task, for most ML methods, including BOONN as an example, requires retraining the entire model without any information from the original model (Figure 4a). In contrast, our model only needs a short

training process for the new class by storing the initial data point, allowing the original model to be maintained.

By retraining our model on only 2% of the original training data from 19,200 atoms in a graphite configuration, we achieved a prediction accuracy of 99.98% for a similar validation set. Figure 4b demonstrates the performance of our extended model on a graphite/fcc interface structure. Though the ratio of the new data is very small, our extended model still gives satisfactory predictions for the graphite part, delivering even better performance in comparison to the completely retrained BOONN model. Figure 4c shows that our statistical-learning-

based model (LGBM) in this work outperforms the deep-learning method BOONN (trained in 20 epochs) in terms of both time cost and required input data size, particularly for the retraining process.

CONCLUSIONS

In this work, we have proposed an efficient ensemble learning framework for crystal structure classification. Our LESS classifier leverages the BOO parameters and ensemble ML methods, resulting in the classification ability for a wide range of crystal structures and significantly improved prediction accuracy compared to many existing methods. Through comprehensive test cases, the feasibility of the LESS classifier has been demonstrated in interpreting atomic structures from a complicated local environment, with its flexibility in providing probabilistic information. Moreover, our framework integrates a modified Gibbs sampling algorithm for a low-cost extension of the classifier to understand completely new structures. This work may point out an alternative direction (in comparison with a deep-learning-based framework) in designing versatile atomic structure analyzers and inspire further developments in this area. Within a similar framework, some of the potential future work may include identification of defects, classification in multi-species systems, and consideration of long-range effects. There also exist applications to analyze the experimental results. Data from transmission electron microscopy, X-ray diffraction, and atom probe tomography have all been demonstrated to efficiently be used as the input to various ML models for crystal structure classification.^{5,44,45} The extendibility and flexibility of our work also enable the possibility to adopt these descriptors as inputs, contributing to more meaningful insights for various materials and systems. Overall, we believe that this lightweight extendable stacking framework has shown great potential as a powerful scientific tool for effective structure classification in atomistic simulations.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jctc.3c00838>.

Detailed packages and hyperparameters for this model, ablation study results, details when training with CGCNN and BOONN, sample configurations used for training and validation data, feature distribution on the training data, performance of this work and CNA at different temperatures, detailed classification results on each crystal structure, and performances of our work and non-ML methods on the test cases ([PDF](#))

AUTHOR INFORMATION

Corresponding Author

Yanming Wang — University of Michigan—Shanghai Jiao Tong University Joint Institute, Shanghai Jiao Tong University, Shanghai 200240, P. R. China;  orcid.org/0000-0002-0912-681X; Email: yanming.wang@sjtu.edu.cn

Authors

Yanhao Deng — University of Michigan—Shanghai Jiao Tong University Joint Institute, Shanghai Jiao Tong University, Shanghai 200240, P. R. China;  orcid.org/0000-0003-4370-983X

Yangyang Wang — University of Michigan—Shanghai Jiao Tong University Joint Institute, Shanghai Jiao Tong University, Shanghai 200240, P. R. China
Ke Xu — School of Materials Science and Engineering, Jilin University, Jilin 130022, P. R. China

Complete contact information is available at:
<https://pubs.acs.org/10.1021/acs.jctc.3c00838>

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The author would like to acknowledge the funding provided by the National Natural Science Foundation of China (no. 52102183) and the Natural Science Foundation of Shanghai (no. 22ZR1433400).

REFERENCES

- (1) Wang, X.; Xiao, S.; Zhang, Z.; He, J. Transportation of Janus nanoparticles in confined nanochannels: a molecular dynamics simulation. *Environ. Sci.: Nano* **2019**, *6*, 2810–2819.
- (2) Gatti, F. Molecular dynamics simulated by photons. *Nature* **2018**, *557*, 641–642.
- (3) Decherchi, S.; Cavalli, A. Thermodynamics and kinetics of drug-target binding by molecular simulation. *Chem. Rev.* **2020**, *120*, 12788–12833.
- (4) Yang, R.; Peng, S.; Lan, B.; Sun, M.; Zhou, Z.; Sun, C.; Gao, Z.; Xing, G.; Yu, L. Oxygen Defect Engineering of - MnO₂ Catalysts via Phase Transformation for Selective Catalytic Reduction of NO. *Small* **2021**, *17*, No. e2102408.
- (5) Ziletti, A.; Kumar, D.; Scheffler, M.; Ghiringhelli, L. M. Insightful classification of crystal structures using deep learning. *Nat. Commun.* **2018**, *9*, 2775–2810.
- (6) Smith, W. F.; Hashemi, J.; Presuel-Moreno, F. *Foundations of Materials Science and Engineering*; McGraw-Hill: New York, 2006; Vol. 509.
- (7) Ryan, K.; Lengyel, J.; Shatruk, M. Crystal Structure Prediction via Deep Learning. *J. Am. Chem. Soc.* **2018**, *140*, 10158–10168.
- (8) Zimmermann, N. E.; Jain, A. Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity. *RSC Adv.* **2020**, *10*, 6063–6081.
- (9) Honeycutt, J. D.; Andersen, H. C. Molecular dynamics study of melting and freezing of small Lennard-Jones clusters. *J. Phys. Chem.* **1987**, *91*, 4950–4963.
- (10) Ackland, G. J.; Jones, A. P. Applications of local crystal structure measures in experiment and simulation. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2006**, *73*, 054104.
- (11) Tsuzuki, H.; Branicio, P. S.; Rino, J. P. Structural characterization of deformed crystals by analysis of common atomic neighborhood. *Comput. Phys. Commun.* **2007**, *177*, 518–523.
- (12) Radhi, A.; Iacobellis, V.; Behdinan, K. A cumulative approach to crystalline structure characterization in atomistic simulations. *J. Phys. Chem. C* **2018**, *122*, 13156–13165.
- (13) Larsen, P. M. Robust Structural Identification via Polyhedral Template Matching. *Model. Simulat. Mater. Sci. Eng.* **2016**, *24*, 055007.
- (14) Dietz, C.; Kretz, T.; Thoma, M. Machine-learning approach for local classification of crystalline structures in multiphase systems. *Phys. Rev. E* **2017**, *96*, 011301.
- (15) Steinhardt, P. J.; Nelson, D. R.; Ronchetti, M. Bond-orientational order in liquids and glasses. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1983**, *28*, 784–805.
- (16) Ackland, G. J.; Jones, A. P. Applications of local crystal structure measures in experiment and simulation. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2006**, *73* (5), 054104.
- (17) Chung, H. W.; Freitas, R.; Cheon, G.; Reed, E. J. Data-centric framework for crystal structure identification in atomistic simulations using machine learning. *Phys. Rev. Mater.* **2022**, *6*, 043801.

- (18) Noé, F.; Tkatchenko, A.; Müller, K. R.; Clementi, C. Machine Learning for Molecular Simulation. *Annu. Rev. Phys. Chem.* **2020**, *71*, 361–390.
- (19) Takahashi, K.; Takahashi, L. Creating Machine Learning-Driven Material Recipes Based on Crystal Structure. *J. Phys. Chem. Lett.* **2019**, *10*, 283–288.
- (20) Reinhart, W. F.; Long, A. W.; Howard, M. P.; Ferguson, A. L.; Panagiotopoulos, A. Z. Machine learning for autonomous crystal structure identification. *Soft Matter* **2017**, *13*, 4733–4745.
- (21) Freitas, R.; Reed, E. J. Uncovering the effects of interface-induced ordering of liquid on crystal growth using machine learning. *Nat. Commun.* **2020**, *11*, 3260–3310.
- (22) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.
- (23) Fukuya, T.; Shibuta, Y. Machine learning approach to automated analysis of atomic configuration of molecular dynamics simulation. *Comput. Mater. Sci.* **2020**, *184*, 109880.
- (24) Lipson, J. Y. C. B. How transferable are features in deep neural networks? *Comput. Sci.* **2014**, *27*, 3320–3328.
- (25) Hirel, P. Atomsk: A tool for manipulating and converting atomic data files. *Comput. Phys. Commun.* **2015**, *197*, 212–219.
- (26) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in 't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.; Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **2022**, *271*, 108171.
- (27) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant pressure molecular dynamics algorithms. *J. Chem. Phys.* **1994**, *101*, 4177–4189.
- (28) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.
- (29) Tuckerman, M. E.; Alejandre, J.; López-Rendón, R.; Jochim, A. L.; Martyna, G. J. A Liouville-operator derived measure-preserving integrator for molecular dynamics simulations in the isothermal-isobaric ensemble. *J. Phys. A: Math. Gen.* **2006**, *39*, 5629–5651.
- (30) Shinoda, W.; Shiga, M.; Mikami, M. Rapid estimation of elastic constants by molecular dynamics simulation under constant stress. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2004**, *69*, 134103.
- (31) Foiles, S.; Baskes, M.; Daw, M. S. Embedded-atom-method functions for the fcc metals Cu, Ag, Au, Ni, Pd, Pt, and their alloys. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1986**, *33*, 7983–7991.
- (32) Mendelev, M.; Han, S.; Srolovitz, D.; Ackland, G.; Sun, D.; Asta, M. Development of new interatomic potentials appropriate for crystalline and liquid iron. *Phil. Mag.* **2003**, *83*, 3977–3994.
- (33) Sun, D.; Mendelev, M.; Becker, C.; Kudin, K.; Haxhimali, T.; Asta, M.; Hoyt, J.; Karma, A.; Srolovitz, D. J. Crystal-melt interfacial free energies in hcp metals: A molecular dynamics study of Mg. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2006**, *73*, 024116.
- (34) Stillinger, F. H.; Weber, T. A. Computer simulation of local order in condensed phases of silicon. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1985**, *31*, 5262–5271.
- (35) Khrapak, S. A.; Chaudhuri, M.; Morfill, G. E. Freezing of Lennard-Jones-type fluids. *J. Chem. Phys.* **2011**, *134*, 054120.
- (36) Zhou, X.; Doty, F.; Yang, P. Atomistic simulation study of atomic size effects on B1 (NaCl), B2 (CsCl), and B3 (zinc-blende) crystal stability of binary ionic compounds. *Comput. Mater. Sci.* **2011**, *50*, 2470–2481.
- (37) Dixon, M.; Gillan, M. Computer simulation of fast ion transport in fluorites. *Le Journal de Physique Colloques* **1980**, *41*, C6-24–C6-27.
- (38) Stuart, S. J.; Tutein, A. B.; Harrison, J. A. A reactive potential for hydrocarbons with intermolecular interactions. *J. Chem. Phys.* **2000**, *112*, 6472–6486.
- (39) Mickel, W.; Kapfer, S. C.; Schröder-Turk, G. E.; Mecke, K. Shortcomings of the bond orientational order parameters for the analysis of disordered particulate matter. *J. Chem. Phys.* **2013**, *138*, 044501.
- (40) Menon, S.; Leines, G. D.; Rogal, J. pscal: A python module for structural analysis of atomic environments. *J. Open Source Softw.* **2019**, *4*, 1824.
- (41) Stukowski, A. Structure identification methods for atomistic simulations of crystalline materials. *Model. Simulat. Mater. Sci. Eng.* **2012**, *20*, 045021.
- (42) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
- (43) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3146–3154.
- (44) Ra, M.; Boo, Y.; Jeong, J. M.; Batts-Etseg, J.; Jeong, J.; Lee, W. Classification of crystal structures using electron diffraction patterns with a deep convolutional neural network. *RSC Adv.* **2021**, *11*, 38307–38315.
- (45) Spannaus, A.; Law, K. J.; Luszczek, P.; Nasrin, F.; Micucci, C. P.; Liaw, P. K.; Santodonato, L. J.; Keffer, D. J.; Maroulas, V. Materials fingerprinting classification. *Comput. Phys. Commun.* **2021**, *266*, 108019.