

ARTICLE OPEN



Active learning to overcome exponential-wall problem for effective structure prediction of chemical-disordered materials

Xiaozhe Yuan^{1,2,3}, Yuwei Zhou^{1,2,4}✉, Qing Peng^{3,5}✉, Yong Yang^{1,2}, Yongwang Li^{1,2,4} and Xiaodong Wen^{1,2,4}✉

Chemical-disordered materials have a wide range of applications whereas the determination of their structures or configurations is one of the most important and challenging problems. Traditional methods are extremely inefficient or intractable for large systems due to the notorious exponential-wall issue that the number of possible structures increase exponentially for N-body systems. Herein, we introduce an efficient approach to predict the thermodynamically stable structures of chemical-disordered materials via active-learning accompanied by first-principles calculations. Our method, named LASou, can efficiently compress the sampling space and dramatically reduce the computational cost. Three distinct and typical finite-size systems are investigated, including the anion-disordered $\text{BaSc}(\text{O}_x\text{F}_{1-x})_3$ ($x = 0.667$), the cation-disordered $\text{Ca}_{1-x}\text{Mn}_x\text{CO}_3$ ($x = 0.25$) with larger size and the defect-disordered $\epsilon\text{-FeC}_x$ ($x = 0.5$) with larger space. The commonly used enumeration method requires to explicitly calculate 2664, 1033, and 10496 configurations, respectively, while the LASou method just needs to explicitly calculate about 15, 20, and 10 configurations, respectively. Besides the finite-size system, our LASou method is ready for quasi-infinite size systems empowering materials design.

npj Computational Materials (2023)9:12; <https://doi.org/10.1038/s41524-023-00967-z>

INTRODUCTION

Chemical-disordered materials are widely used in many areas including semiconductors, high-temperature superconductors, Li-ion batteries, metal alloys, ceramics, and heterogeneous catalysts due to their special properties and performances^{1–6}. Here, the term ‘chemical-disordered materials’ stems from the semi-ordered materials whose lattice is periodic (thus crystal) but the occupying atom species are non-periodic in space. From the point view of chemical compositions, the chemical-disordered materials can be classified into anionic, cationic, and defected counterparts, which can be simply considered as the anions, cations and defects occupy the non-periodic sites. For example, the $\text{RbBr}_x\text{Cl}_{1-x}$ ($x = 0.4$)⁷, $\text{SrNb}(\text{O}_x\text{N}_{1-x})_3$ ($x = 0.667$)⁸, and $\text{BaSc}(\text{O}_x\text{F}_{1-x})_3$ ($x = 0.667$)⁹ in anion-disordered materials, the $\text{La}(\text{Ca}_x\text{Zr}_{1-x}\text{O}_3)$ ($x = 0.5$)¹⁰, $\text{Ba}(\text{Mg}_x\text{Ti}_{1-x})_{12}\text{O}_{19}$ ($x = 0.5$)¹¹, and $\beta\text{-}(\text{Al}_x\text{Ga}_{1-x})_2\text{O}_3$ ($0 \leq x \leq 1$)¹² in cation-disordered materials, and the $\text{Ti}_{1-x}\square_x\text{O}_{1.11}\text{F}_{0.89}$ ($x = 0.22$)¹³, $[\text{La}_8\text{Sr}_2(\text{SiO}_4)_6]^{4+}(\text{O}_{2-x}\square_x)^{4-}$ ($0 < x \leq 2$)¹⁴, and $\text{Na}_{2-x}\square_x\text{Mn}_3\text{O}_7$ ($0 < x \leq 2$)¹⁵ in defect-disordered materials, where \square stands for the vacancy defect. In fact, the above systems are also alternatively called as fractional occupation, substitutional doping, and vacancy defect, which commonly used in both experimental characterizations and theoretical simulations.

Chemical-disordered materials have received substantial attention varied from experimental to theoretical research in many aspects. Among them, the determination of the structures, more exactly called as the atomic arrangements or site-occupied configurations, is one of the most important and challenging problems. The precise atomic structures with thermodynamic stability are crucial not only to the structural properties, characterization, and integrity, but also to the understanding of

underlying mechanism, the construction of relationship between structures and properties, and further the discovery and design of new materials. A variety of experimental methods have faced great challenge in determining the atomic structures/configurations of chemical-disordered materials due to uncertainties for the atomic occupying on lattice sites. The most commonly used X-ray diffraction (XRD) can only provide the averaged information of materials. Therefore, it can hardly obtain valuable data about the configurations. Some other methods that can provide local site information, such as nuclear magnetic resonance (NMR) and X-ray adsorption near-edge structure (XANES). However, they are still difficult to directly correlate with the configurations^{16,17}. Apart from the experimental approach, the computational approach can obtain the atomic structures more directly and efficiently. Many computational methods have been used to deal with chemical-disordered materials and determine the atomic structures and properties, such as virtual crystal approximation (VCA)¹⁸, coherent potential approximation (CPA)¹⁹, special quasirandom structures (SQS)^{20,21}, and supercell approximation, etc. The supercell approximation is one of the simplest and well-studied methods, while it has a huge obstacle in low-efficiency sampling and high computational cost even for a finite-size system. The grand challenge for structure determination is that the number of possible configurations increase exponentially with the number of atoms for N-body systems²². This is the notorious ‘exponential-wall’ problem of many-body systems, which becomes intractable with respect to the increase of system size and space size.

Many computational methods and programs have been proposed to deal with the aforementioned challenges. For the sampling, unlike the simply random methods (e.g. Monte Carlo algorithm), several enumeration methods and programs of

¹State Key Laboratory of Coal Conversion, Institute of Coal Chemistry, Chinese Academy of Sciences, Taiyuan 030001, China. ²National Energy Center for Coal to Clean Fuels, Synfuels China Technology Co. Ltd, Huairou District, Beijing 101400, China. ³The State Key Laboratory of Nonlinear Mechanics, Institute of Mechanics, Chinese Academy of Sciences, Beijing 100190, China. ⁴Beijing Advanced Innovation Center for Materials Genome Engineering, Industry–University Cooperation Base between Beijing Information S&T University and Synfuels China Technology Co. Ltd, Beijing, China. ⁵School of Engineering Sciences, University of Chinese Academy of Sciences, Beijing 100049, China.

✉email: zhouyuwei_kt@163.com; pengqing@imech.ac.cn; wxd@sxicc.ac.cn

SOD²³, enumlib^{24–26}, Supercell²⁷, and disorder²⁸ can enumerate all the inequivalent structures/configurations for the finite-size systems in a high efficient way. It should be noted that the enumerated structures usually generated with the combinatorics and crystal symmetries, hence it can be regarded as a ‘brute force’-based approach, which is still very difficult for the complex or quasi-infinite size systems. For the computational cost, the large amount of enumerated structures then usually calculated by first-principles methods or empirical potential functions. The first-principles methods can give accurate and reliable results after expensive calculations. The empirical potential functions are much faster, but the results are often inaccurate. Therefore, advanced algorithms are desirable for both high speed and accuracy in evaluating the substantial structures from the large phase spaces. The cluster expansion (CE) is an affordable and practical method, which is widely used to calculate the energies and properties, and construct phase diagrams of alloy systems^{29–34}. CE evaluates configuration energy using an expansion based on the effective cluster interactions (ECIs), which are fitted to the results of first-principles calculations upon a few configurations. However, CE has disadvantages in describing the complex systems with long-range interactions³⁵. It also degrades the reliability of energies and properties when there is a significant lattice deformation³⁶. Recently, an alternative approach, namely the machine learning (ML) or data-driven interatomic potential, has been attracted a great deal of attention^{37–46}. It successfully carried out many chemical-disordered materials, such as Au-Li³⁹, Ni-Mo⁴⁰, Ti-Al⁴¹, MgAl₂O₄⁴², and (Co_xMn_{1-x})₃O₄⁴³. To date, machine learning methods have received great success in material researches, especially in materials modeling, discovery and design^{44–46}. Generally, most of the machine learning interatomic potentials strongly depend on a large number of datasets (samples) to support the model training and validation. The problem of machine learning modeling with no data or with few data, also known as ‘small sample size problem’, is a major challenge^{47,48}. In brief, one should prepare as many highly reliable, diverse, and massive samples as possible before model construction, while it is apparently hard for the unexplored or unknown systems. Meanwhile, the machine learning-based potentials are usually constructed in an offline way. It means that a model is generally trained and validated only once, and the performance is strongly relied on the pre-prepared datasets.

In present work, we introduce an approach that combines first-principles calculations and active-learning algorithm to accelerate the prediction of thermodynamically atomic structures/configurations of the chemical-disordered materials. Here, machine learning interatomic potential has a good combination of speed and accuracy: it is much faster than first-principles methods with the accuracy comparable to the first-principles results, and much more accurate than empirical functions. It also overcomes the shortcomings of CE method. The active-learning algorithm is suitable for the situations with few labeled data (samples), which is frequently encountered in ‘small sample size problem’^{49–52}. With the assistance of machine learning interatomic potential and the active-learning algorithm, it can significantly enlarge the sampling space and dramatically reduce the first-principles calculations. Based on these features, our method is called as Large space sampling and Active labeling for searching (LASou, sou means searching in Chinese). For demonstration, three distinct finite-size systems are investigated, including the anion-disordered BaSc(O_xF_{1-x})₃ (x = 0.667), the cation-disordered Ca_{1-x}Mn_xCO₃ (x = 0.25) with larger size and the defect-disordered ε-FeC_x (x = 0.5) with larger space. It should be noted that a large size of dataset is not a prerequisite for the LASou method. Furthermore, the ML potential model will be re-trained/re-validated and improved on-the-fly. Compared with the enumeration approach, LASou

method can remarkably reduce the first-principles computational cost and significantly accelerate the structure/configuration prediction of chemical-disordered materials.

RESULTS AND DISCUSSION

We utilize the LASou method in combination with high accurate first-principles density functional theory (DFT) calculations to search and predict three categories of structures/configurations in chemical-disordered materials, including the anion-disordered BaSc(O_xF_{1-x})₃ (x = 0.667), the cation-disordered Ca_{1-x}Mn_xCO₃ (x = 0.25), and the defect-disordered ε-FeC_x (x = 0.5). We have calculated all the enumeration structures of BaSc(O_xF_{1-x})₃ (x = 0.667) and Ca_{1-x}Mn_xCO₃ (x = 0.25) with DFT as the benchmarks, and the DFT results of enumerated ε-FeC_x (x = 0.5) are provided by Liu et al.,⁵³ in which the thermodynamically most stable structure and the lowest energy are used as the ground-truth data. All the enumeration structures are employed here as the sampling space for testing.

Anion-disordered BaSc(O_xF_{1-x})₃ (x = 0.667)

S. Hariyani and J. Brgoch⁵⁴ reported an Eu²⁺ doped BaScO₂F perovskite material for highly efficient cyan emission. The refined crystal structure data showed that O atoms and F atoms are co-occupied with the fraction of 0.667 and 0.333, which can be represented as BaSc(O_xF_{1-x})₃ (x = 0.667). In order to obtain the local structure and electronic properties, the authors firstly studied the thermodynamically stable structure/configuration of O and F site-sharing. They constructed a 2 × 2 × 2 supercell containing 40 total number of atoms, and then used the Supercell program to generate all possible distributions of O and F atoms. After finishing the DFT calculations for all 2664 inequivalent enumeration structures, they obtained the most stable (lowest energy) structure with the formation of Ba-F tetrahedral chains, then the most stable structure was used to calculate the substituted model of Eu²⁺ and other properties. Here, we use the LASou method to search and predict the anion-disordered BaSc(O_xF_{1-x})₃ (x = 0.667) system. Following the article’s approach, we also construct a 2 × 2 × 2 supercell containing 16 O atoms and 8 F atoms on the 24 co-occupied sites, then 2664 inequivalent structures are enumerated by using Supercell program²⁷.

Figure 1a shows the complete ‘brute-force’ DFT results of total energy and the most stable structure within all the 2664 structures. Then, the LASou method was applied to search and predict iteratively until ten generations of runs reached. Figure 1b shows the searching process of total DFT energy against with the generation, in which the red line exhibits the best structure evolved in history. Clearly, the LASou method has successfully predicted the target (most stable) structure with lowest energy in the third generation and kept the goodness till to the end of the task. The number of structures for labeling and DFT relaxation is 5, which means we only spent about 15 DFT relaxed calculations to obtain the same results of 2664 enumeration DFT calculations in this task. The minimum in the first two generations is very close to the goal, which is mainly attributed to the sampling from large (enumeration) space. With more efficiently sampling, the LASou method can rapidly obtain the most stable structure, and thus greatly reduce the first-principles DFT computational cost and time demanding. The detailed performances of the machine learning (ML) potential within ten generations can be seen from Supplementary Figs. 1 and 2. With the increase of generations and datasets under the active learning algorithm, the ML model will be re-trained on-the-fly and get more reliable results for selection and labeling.

Cation-disordered Ca_{1-x}Mn_xCO₃ (x = 0.25)

Wang, Grau-Crespo and de Leeuw⁵⁵ have studied the thermodynamics of the disordered solid solution Ca_{1-x}Mn_xCO₃, which

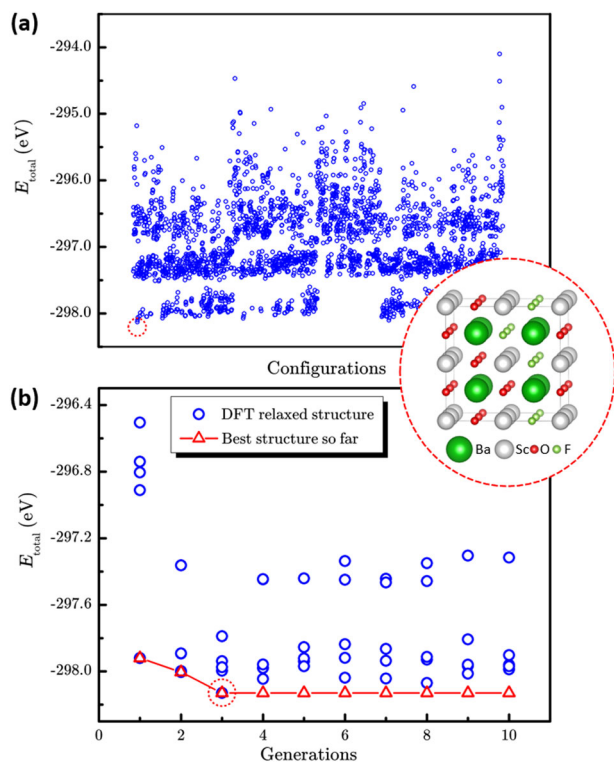


Fig. 1 The performance of enumeration method and LASou method for $\text{BaSc}(\text{O}_x\text{F}_{1-x})_3$ ($x = 0.667$). **a** The scatter plot of total energy of 2664 enumerated structures and the most stable structure. (The blue circles represent the energy for each structure, the red dashed circle represents the best structure with lowest energy.) **b** The searching process of LASou method for the total DFT energy against with generation. (The red triangles represent the lowest energy structures searched in history).

mixed from calcite (CaCO_3) and rhodochrosite (MnCO_3) in the full range of compositions ($0 \leq x \leq 1$). The arrangements of Ca^{2+} and Mn^{2+} can be simply modeled as the substitutional doped of Ca^{2+} by Mn^{2+} . The authors employed the SOD program²³ in several supercells of the hexagonal calcite structure to enumerate a large number of structures for each composition. Owing to the large number of structures, they used GULP program incorporated with empirical potential to calculate the thermodynamics properties. The most stable structure with lowest energy has been used to analyze and understand the homogeneity and heterogeneity of cation within and across layers. Here, we take $x = 0.25$ for $\text{Ca}_{1-x}\text{Mn}_x\text{CO}_3$ as a test case and employ first-principles DFT calculations to verify the efficiency and reliability of the LASou method for the cation-disordered systems. We adopt the disorder program²⁸ to enumerate 1033 inequivalent structures under the $2 \times 2 \times 1$ supercell of CaCO_3 , in which 6 Ca atoms of 24 Ca sites are replaced by 6 Mn atoms. Notably, it is much larger size for that there are 120 total number of atoms in the supercell, the system with larger size is very difficult for the first-principles DFT relaxation and searching.

Figure 2a shows the complete ‘brute-force’ DFT results of total energy and the most stable structure within all the 1033 structures. Then, the LASou method was carried out to iteratively search until ten generations of runs reached. Figure 2b shows the searching process of total DFT energy against with the generation. The LASou method also rapidly predicted the target structure with lowest energy in the fourth generation. We only spent 20 DFT relaxed calculations that can obtain the same results of 1033 enumeration DFT calculations. The rapid decline of the minimum in the first three generations may be raised from the increasing

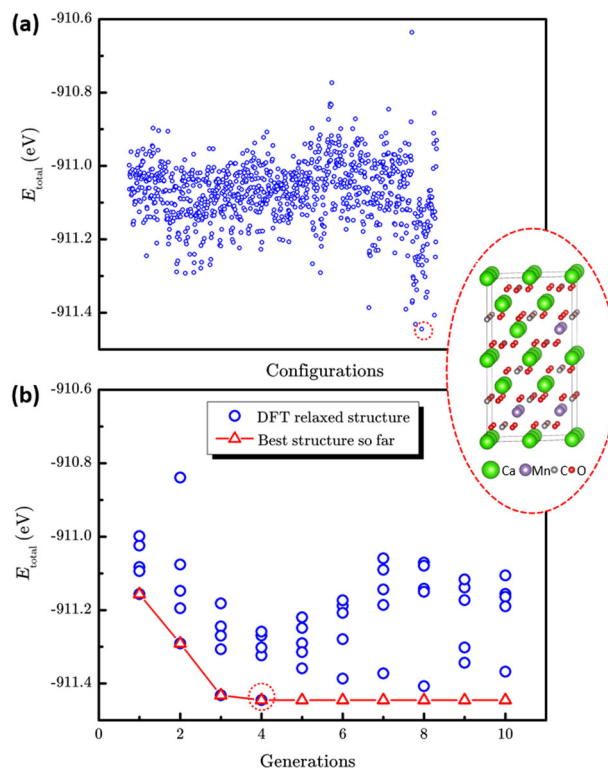


Fig. 2 The performance of enumeration method and LASou method for $\text{Ca}_{1-x}\text{Mn}_x\text{CO}_3$ ($x = 0.25$). **a** The scatter plot of total energy of 1033 enumerated structures and the most stable structure. **b** The searching process of LASou method for the total DFT energy against with generation.

accuracy of ML model. This is a clear demonstration of the efficiency, accuracy, and robust of the LASou method. The detailed performances of the ML potential within ten generations can be seen from Supplementary Figs. 3 and 4. With the increase of generations and datasets, the same behavior of ML model that occurs in the $\text{BaSc}(\text{O}_x\text{F}_{1-x})_3$ ($x = 0.667$) system can be obtained and get more reliable results for selection and labeling.

Defect-disordered $\epsilon\text{-FeC}_x$ ($x = 0.5$)

Iron carbides are the active phases of industrial catalysts in Fischer-Tropsch synthesis to produce liquid fuel. Among them, the phase identification of the $\epsilon\text{-Fe}_2\text{C}$ and the $\epsilon'\text{-Fe}_{2.2}\text{C}$ has been debated for half a century. Theoretically, Liu and coworkers⁵³ used SOD program²³ coupled with DFT calculations to investigate the equilibrium (most stable) structures and thermal stabilities of the $\epsilon\text{-FeC}_x$ phases. Starting from $\epsilon\text{-FeC}_x$ ($x = 1.0$) phase in a $2 \times 2 \times 3$ size of supercell, the authors built the $\epsilon\text{-Fe}_2\text{C}$ and $\epsilon'\text{-Fe}_{2.2}\text{C}$ structures with substitution the C atoms with vacancies, then enumerate all the independent configurations with the total number of 10496 and 9551 with SOD program. The results shown that $\epsilon\text{-Fe}_2\text{C}$ and $\epsilon'\text{-Fe}_{2.2}\text{C}$ are thermodynamically stable phases as observed in the experiment. Here, we take $x = 0.5$ for $\epsilon\text{-FeC}_x$ (namely $\epsilon\text{-Fe}_2\text{C}$ phase) as a test case coupled with DFT calculations to verify the efficiency and reliability of the LASou method for the defect-disordered systems. Following the article’s approach, we also construct a $2 \times 2 \times 3$ supercell for $\epsilon\text{-FeC}_x$ ($x = 1.0$) containing 24 octahedral sites (C atoms), and then employ the Supercell program²⁷ to enumerate all the inequivalent structures (configurations) with substitution the C atoms by 12 vacancies. Notably, it is much larger space for that there are 10496 total number of structures. The authors have successfully carried out all the 10496

DFT relaxed calculations, we quote the most stable structure with lowest energy as the ground-truth data.

Figure 3a shows the results of the complete ‘brute force’ calculated total energy and the most stable structure within all the 10496 structures via DFT calculations, which provided by Liu et al.⁵³. Then, the LASou method was used to iteratively search until ten generations of runs reached. Figure 3b shows the searching process of total DFT energy against with the generation. The LASou method has successfully predicted the target structure with lowest energy in the second generation. We spent about 10 DFT relaxed calculations that can obtain the same results of 10496 enumeration DFT calculations. It should be noted that there are some differences in the DFT running parameters used in present work and Liu et al., we get more lower energy for the same structure. As we can see from Fig. 3a, the energy of most stable structure is significantly lower than that of other structures, the target can be easily distinguished. Hence, the LASou method can find the target with great efficiency.

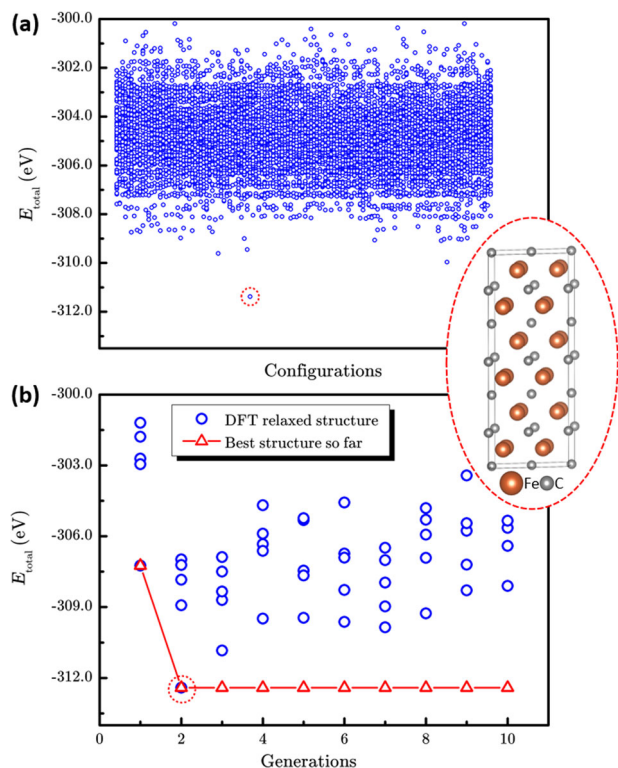


Fig. 3 The performance of enumeration method and LASou method for ϵ -FeC_x ($x = 0.5$). **a** The scatter plot of total energy of 10496 enumerated structures and the most stable structure, the DFT total energies are provided by Liu et al.⁵³. **b** The searching process of LASou method for the total DFT energy against with generation.

Robustness and performances of LASou method

For the practical structure/configuration prediction of the chemical-disordered materials, the choices of different parameters may lead to different results and efficiency for the LASou method. Here, we further examine the robustness of LASou method via testing several key factors and parameters, including (a) ensemble algorithm, (b) structure clustering for labeling, (c) structure relaxation before labeling, (d) other alternative ML models. The criterion, whether the target structure is successfully predicted within the maximal generations (e.g., 10 in this work), is employed to evaluate the performance. The target structure refers to the most stable structure with lowest energy obtained from the DFT calculations. On the top of ensemble-LBF/LRR under LASou method, we gradually turn off the above factors and other parameters remain unchanged. The ensemble-NN refers that the ensemble model is built with the neural network (NN) adopted by Yang et al.⁵⁶. The test results are listed in Table 1.

The robustness of LASou method is evidenced because the LASou method can efficiently predict the most stable structure/configuration within several generations, even if some factors are turned off. However, we recommend that all factors should be turned on to ensure the robustness as can be seen from Table 1. Firstly, using a single ML model instead of ensemble ML model will lead to slightly worse results. In practice, there are many unreasonable (unphysical) structures for structure relaxation when the single ML model is employed. The probable reason lies in the fact that ensemble ML model is capable to balance bias and variance well both for single point energy and structure relaxation. Secondly, structure clustering algorithm may have a significant impact on the efficiency for some cases. The structure clustering algorithm can improve the diversity, when the labeled structure might miss the target structure, and/or the energy difference of structure space is very small. Thirdly, the structure relaxation by ML model before labeling seems to have little effect on the overall efficiency. While after structure relaxation, it gives a much reasonable lower predicted energy, especially for the systems with apparently structural changes or distortions. The roles and performances of ensemble, clustering and relaxation in LASou method can be seen in Supplementary Notes of the Supplementary Information (SI). Furthermore, we have also examined the performances against with several combinations of the adjustable parameters on top of Ca_{1-x}Mn_xCO₃ ($x = 0.25$) system as presented in Supplementary Table 3. The results show that the adjustable parameters are insensitive to the overall efficiency. Finally, the LBF/LRR method used in present work to construct the ensemble model may not perform well for all systems. We can compatibly and flexibly use other ML methods to replace LBF/LRR, e.g., KRR, GPR, NN, and DL-based models, etc. Here we test the ensemble model built from neural network (NN) method of Yang et al.⁵⁶. As expected, the ensemble-NN can also achieve good results with high efficiency under LASou algorithm. However, compared with the essence of linear model for the LBF/LRR method, the nonlinear NN method requires more training cost and time demanding. Nevertheless, the time cost of model training is much lower than that of DFT calculations, which can be seen in Supplementary Figs. 11 and 12.

Table 1. Performances (found the target structure or not in 10 generations) of ensemble, clustering, relaxation, and alternative ML model under LASou method. The number in parentheses is the generation of finding the target structure.

Systems	Ensemble-LBF/LRR	Without ensemble	Without clustering	Without relaxation	Ensemble-Neural Network (NN)
BaSc(O _x F _{1-x}) ₃ ($x = 0.667$)	Found (3)	Found (8)	Found (3)	Found (2)	Found (2)
Ca _{1-x} Mn _x CO ₃ ($x = 0.25$)	Found (4)	Found (4)	Not Found (>10)	Found (2)	Found (2)
ϵ -FeC _x ($x = 0.5$)	Found (2)	Found (2)	Found (2)	Found (2)	Found (2)

There are several thumb-up tips for using the LAsou method. (a) It is priority to turn on the aforementioned factors and parameters. Due to the ML potential error and the existence of multiple structures with energy close to the target structure, our method has a certain probability of missing the target structure. We'd like to suggest to run at least twice to avoid missing the target structure as much as possible. (b) For the much larger or complex systems, the converged structure may not be the 'true' most stable, but it can be considered to be the 'putative' most stable. We can adjust some running parameters to obtain more reliable results, for example, the total number of iterations or generations (N_{iter}), the number of selected and labeled structures (N_s), and so on. (c) For the finite-size system as like in present work, the programs of SOD, Supercell, and disorder can successfully enumerate the sampling space for many kinds of chemical-disordered materials. But it is hardly to generate the complete structures for more complex or quasi-infinite size systems, we can still take the advantages of LAsou method incorporated with a variety of sampling operators⁵⁷ rather than the enumeration approach. It is worth noting that although the DFT calculations are applied for demonstration and validation, the energy assessment is not limited to DFT calculations. It can be performed through various methods including electronic structure calculations, empirical (forcefield) methods, and semi-empirical methods.

We have introduced the LAsou method, a simple yet highly efficient approach that combines the first-principles calculations and active-learning algorithm to search for the thermodynamically stable structures/configurations of chemical-disordered materials. The LAsou method shows great potential to solve the 'exponential-wall' problem for many-body systems. In LAsou method, the ML potential can largely reduce the DFT calculations via predicting and filtering from the large sampling space, the ensemble learning algorithm can significantly improve the stabilities for the prediction of energy and relaxation, and the active learning algorithm can gradually improve the accuracy of ML potential on-the-fly so that one doesn't have to pre-prepare a large amount of training data. With these advantages and features, the active learning-based LAsou method will be helpful for a wide range of applications for the larger, more complex, quasi-infinite size systems and the new materials that occurs in nanoparticles, catalysts, solid solutions, high-entropy alloys and high-entropy oxides, and so on.

METHODS

The computational technique of structure or configuration prediction has been widely used in material discovery and design for many various types of materials, such as zero-dimensional (0D) clusters and nanoparticles, two-dimensional (2D) layered films, three-dimensional (3D) bulks and high-pressure materials, etc. A lot of methods and programs (e.g., USPEX^{58–61}, CALYPSO⁶², and XtalOpt⁶³, etc) have been proposed and achieved great success in these systems. Our developed IMAGE program⁵⁷ also successfully predicted the bulk phases of iron carbides (Fe_xC_y , $1 \leq y \leq x \leq 7$, $0 < y/x \leq 1$) and found the unexpected magnetism properties of iron atoms. Among the algorithms, the most popular approach of 'global searching + local relaxation' is adopted. The global searching methods mainly include genetic algorithm (GA)⁶⁴, simulated annealing (SA)⁶⁵ and particle swarm optimization (PSO)⁶⁶, etc. The local relaxation methods mainly rely on first-principles calculations with the fact that it can give reliable results very well. Recently, there have been several reports that attempts to accelerate structure prediction using machine learning methods and active learning algorithm^{67–70}.

The large space sampling and active labeling for searching (LAsou) method

Here, we employ active learning algorithm coupled with first-principles calculations to build a simple yet highly efficient approach to predict the thermodynamically stable structures of chemical-disordered materials. As we know, machine learning (ML) potential in general strongly depends on a large size of datasets (e.g., structures/configurations, and/or properties for materials) before model construction. It is feasible for the well-studied systems or the data that can easily be labeled, while it is quite difficult for the unexplored system or without labeled data which often occurred in new materials discovery and design. Active learning (AL) algorithm provides an efficient online (on-the-fly) way to deal with the model constructed by using minimal size of labeled datasets. Taking the ML potential construction under the AL algorithm for example, the potential can be initially trained with a small number of datasets (structures ~ energies/forces) with low accuracy, then the AL algorithm will proactively pick out some favorable samples and add them into datasets. After that, the potential will be re-trained/re-validated and updated along with significant improvement of accuracy.

The flowchart of the active learning-based approach for the structure prediction of chemical-disordered materials is shown in Fig. 4. In present work, we mainly exhibit the simplest and minimal process of our approach for the structure prediction of finite-size systems, in which the sampling space is created by the 'brute

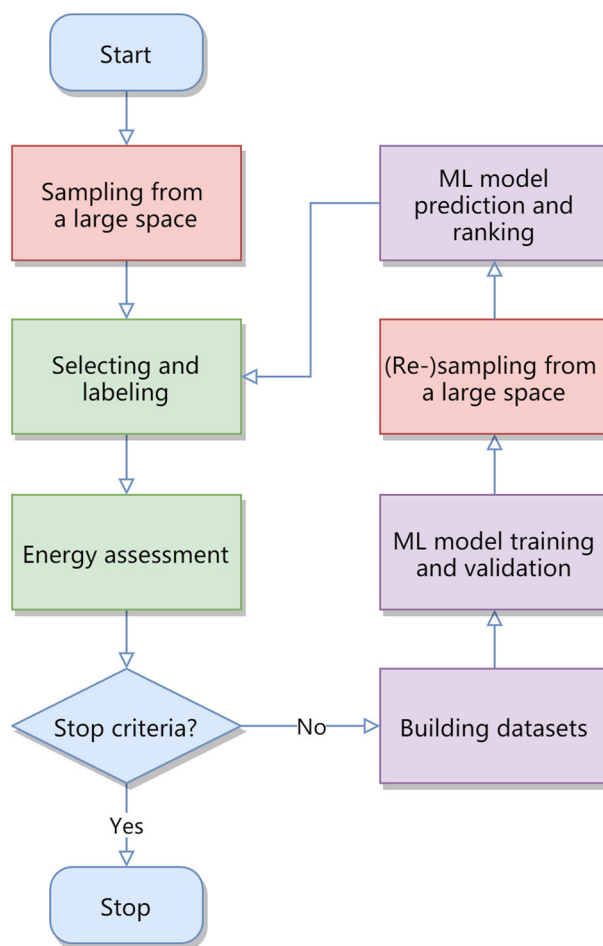


Fig. 4 Flowchart of the active learning-based LAsou method. An overview of the proposed approach for the structure prediction of finite-sized chemical-disordered materials coupled with first-principles calculations.

force-based enumeration method. Meanwhile, we employ a simply iterative search based on the greedy search (GS) algorithm and tabu search (TS) strategy.

Step 1: sampling from a large space. For the finite-size system, the enumerated structures compose the full sampling space for searching and prediction. Here, we can use the open-source programs (Supercell²⁷ and/or disorder²⁸) to generate a large space containing all the inequivalent structures based on the user-defined size of supercell and the percentage of occupied lattice sites. However, it is worth noting that the enumeration approaches are not adequate for the larger, complex, or quasi-infinite size systems.

Step 2: selecting and labeling. The candidates will be proactively selected and labeled from the large sampling space, and further passed to perform the energy assessment. This is a crucial step for that the selecting and labeling in active learning algorithm will recommend preferable structures and rapidly guide to the target. In practice, a considerable efficiency can be achieved with only a small amount of selected and labeled structures. We employ three criteria to obtain the candidates, including high thermodynamically stability with lower energy, high structural diversity with lower similarity, low sampling repeatability with tabu search (TS) strategy. After sampling from the large space, the three criteria are applied through the procedure as below. First, the ML potential will be used to calculate the energies for all structures, the lowest energy can be ensured high thermodynamically stability. Second, the K-means clustering method is used to divide all structures into several distinct classes, the structure with the lowest energy in each class will be selected and labeled. The selection from different class can reach high structural diversity. Third, each selected structure will be judged whether it has been emerged in the historically selected structures, the elimination of historical structure can realize low sampling repeatability. It should be noted that selection and labeling in the first generation only employs high structural diversity criterion, while the following generations include all the above criteria.

Step 3: energy assessment. In present work, the selected and labeled candidates will be performed the structural relaxation calculations in order to get the local minimum structures and energies. The structural relaxation calculations were accomplished by using first-principles density functional theory (DFT) methods, for that it can obtain reliable results for a wide range of systems. The details of calculations are described in next section. To compare with the enumeration approach, we have totally relaxed all the inequivalent structures, and the most stable structure with lowest energy is set as the target.

Step 4: building datasets. Again, the ML potential strongly depends on enormous datasets. Obviously, such a prerequisite is very difficult for the unexplored or unknown systems. With the advantages of active learning algorithm, here we adopt the online (on-the-fly) scheme to extract the dataset from the results of DFT calculations. Several frames of data are collected (i.e., structures ~ energies/forces) from each DFT relaxed trajectory with intermediate equal intervals. Such data explicitly takes into account the changes of structure, which is necessary to construct the ML potential and may be helpful for the improvement of structural relaxation. Furthermore, the datasets are continuously updated and enlarged after DFT calculations in each generation.

Step 5: machine learning model training and validation. Machine learning model will be used to construct the ML potential in this step. Generally, the total energy E_{tot} of a system containing N atoms can be expressed the summation of atomic energy E_i based on the additive model, $E_{\text{tot}} = \sum_{i=1}^N E_i$. Then the atomic energy can be fitted by ML model $E_i = f(X(\{\mathbf{R}\}), w)$, where $X(\{\mathbf{R}\})$ is the

structural features (also known as descriptors) associated with the structure or configuration, w is the model parameters, and f is the ML model. The total energy is trained and validated against with structures ~ energies/forces from the datasets. Currently, a variety of machine learning models^{71–77} have been successfully used in the establishing of interatomic potentials, such as, linear ridge regression (LRR), gaussian process regression (GPR), neural networks (NN), and deep learning (DL), etc. Meanwhile, many structural features^{78–80} based on local atomic environment (LAE) have been developed to reproduce the highly accuracy of potential energy surfaces, such as atom-centered symmetry function (ACSF), smooth overlap of atomic positions (SOAP), and many-body tensor representation (MBTR), etc. At present, a plenty of models and programs^{81–84} can be freely available for the construction of ML-based potentials. Here, we employ the linear basis function (LBF) model to construct the interatomic potential, in which the model is a simple LRR (namely linear regression + L_2 -regularization) method and the features are consist of Gaussian basis function as described in refs. ^{85,86}. This LBF/LRR model has been successfully used in our previous work⁵⁷ to obtain a high accuracy of relationship for local atomic structures and magnetic moments in iron carbide phases. It can be expected that the linear LBF/LRR model will be superior to the nonlinear models (e.g., NN, DL) in the efficiency of model training, validation, and prediction under the active learning algorithm. On the other hand, we employ the ensemble learning algorithm⁵⁶ to ensure the stability and reliability of the LBF/LRR potential. After training of M independent LBF/LRR models, we take the simple average as the ensemble model to get the results. The total interatomic potential, denoted as ensemble-LBF/LRR, can be expressed as,

$$E_{\text{tot}}^{\text{Ensemble}} = \frac{1}{M} \sum_{k=1}^M E_k^{\text{LBF/LRR}} \quad (1)$$

and the atomic forces can be given by

$$\mathbf{F}_i = -\frac{\partial E_{\text{tot}}^{\text{Ensemble}}}{\partial \mathbf{R}_i} = -\frac{1}{M} \sum_{k=1}^M \frac{\partial E_k^{\text{LBF/LRR}}}{\partial \mathbf{R}_i} \quad (2)$$

where $E_k^{\text{LBF/LRR}}$ is the total energy predicted by the k -th LBF/LRR model, $\mathbf{R}_i = (x_i, y_i, z_i)$ and $\mathbf{F}_i = (F_{ix}, F_{iy}, F_{iz})$ are the i -th atomic positions and forces, respectively. Next, we can use the least squares method to minimize the errors of energies or both energies and forces. In the following generations, each LBF/LRR model will be re-trained and the ensemble-LBF/LRR model will be updated along with the increase of datasets. Alternatively, any other ML models can replace the LBF/LRR model to obtain the more reliable ensemble model, such as ensemble-KRR, ensemble-GPR, ensemble-NN⁵⁶, etc. And the simple average approach can also be replaced with other Boosting or Bagging algorithms^{87,88}, such as AdaBoost, XGBoost, etc.

Step 6: machine learning model prediction and ranking. After completion of model training and validation, the final ensemble model will be used to predict and evaluate all the inequivalent structures in the large enumerated space. Alternatively, we can re-sample to get a large number of structures for other systems. Then, one can simply calculate the single-point energy, or structure relaxation for each structure. For the prediction of structure relaxation, a small uncertainty of ensemble model ($\sigma_{\text{uncertainty}}$) will be checked to ensure the consistency of each LBF/LRR model and then perform the calculation. The uncertainty is estimated by the standard deviation for each model, as given by

$$\sigma_{\text{uncertainty}} = \sqrt{\frac{\sum_{k=1}^N (E_k^{\text{LBF/LRR}} - \bar{E}_{\text{tot}}^{\text{Ensemble}})^2}{N}} \quad (3)$$

where $\bar{E}_{\text{tot}}^{\text{Ensemble}}$ is the averaged energy of the ensemble models. After that, the structure ranking should be applied to recommend

the candidates more reasonably rather than only the single-point energy. Further, the K-means clustering algorithm (implemented in scikit-learn⁸²) will be used to divide the enumerated structures into several parts, and then it goes to Step 2 for structure selecting and labeling.

Repeat Steps 2–6 until the approach terminates when the convergence criteria are reached. The converged structure with lowest energy among all DFT relaxed structures will be expected to the 'putative' most stable structure. The above active learning-based approach is called as the Large space sampling and Active labeling for searching (LASou, sou means searching in Chinese). In our approach, the sampling from enumeration space can significantly improve the efficiency that covers a huge space at once; the active labeling based on ML model and ranking can rapidly obtain a small amount of preferable candidate structures. In terms of probability, the candidates obtained from our approach will be significantly better than the traditional structure prediction algorithms.

The parameters of present work are listed as following: the total number of iterations or generations (N_{iter}) is 10, the number of selected and labeled structures (N_{sl}) is 5, the number of data points extracted from each trajectory is 5 with equidistance, the percentage of training set and validation set is 60 and 40% respectively, the cutoff radius (R_{cut}) is 6.0 Å, the Gaussian basis function in LBF/LRR model with taking 5 values uniformly in the range of $a \in [0.1, 2]$ and 10 values uniformly in the range of $b \in [0, 5]$ ^{85,86}, and 5 LBF/LRR models used to build the ensemble model. For the prediction of structural relaxation, the maximum atomic forces, maximum number of steps are set to 0.03 eV Å⁻¹, and 50, respectively.

First-principles calculations

We performed first-principles calculations using the planewave code Vienna ab initio simulation package (VASP)^{89,90}. The general gradient approximation of the Perdew-Burke-Ernzerhof parameterization (GGA-PBE)⁹¹ was adopted for the exchange and correlation functions. We use the MPRelaxSet module in Pymatgen software⁹² to generate input files for each structure, with slightly different parameter settings for different systems. For BaSc(O_xF_{1-x})₃ ($x = 0.667$), Ca_{1-x}Mn_xCO₃ ($x = 0.25$), the energy cutoff was set to 500 eV. The convergence criterion was set to 1×10^{-8} eV in energy and 1×10^{-6} eV Å⁻¹ in force. For ε-FeC_x ($x = 0.5$), the energy cutoff was set to 500 eV. We adopted the second-order Methfessel-Paxton⁹³ smearing scheme with a value of $\sigma = 0.2$ eV. Structure relaxations were performed with convergence criteria of 1×10^{-4} eV in energy and 0.03 eV Å⁻¹ in force.

DATA AVAILABILITY

The datasets include trajectory optimized by VASP for BaSc(O_xF_{1-x})₃ ($x = 0.667$) and Ca_{1-x}Mn_xCO₃ ($x = 0.25$) systems and test result files for different factors and several combinations of the adjustable parameters in the results and discussion section. They are available at <https://doi.org/10.6084/m9.figshare.21776579>.

CODE AVAILABILITY

Code developed in this study is available from the corresponding author upon reasonable request.

Received: 18 May 2022; Accepted: 11 January 2023;

Published online: 20 January 2023

REFERENCES

- Guevara, J., Vildosola, V., Milano, J. & Llois, A. M. Half-metallic character and electronic properties of inverse magnetoresistant Fe_{1-x}Co_xSi alloys. *Phys. Rev. B* **69**, 184422 (2004).

- Koga, E., Moriwake, H., Kakimoto, K.-I. & Ohsato, H. Raman spectroscopic evaluation and microwave dielectric property of order/disorder and stoichiometric/non stoichiometric Ba(Zn_{1/3}Ta_{2/3})O₃. *Ferroelectrics* **356**, 146–152 (2007).
- Davydov, S. A. et al. Effects of localisation in atomic-disordered high-Tc superconductors, in *Advances in Superconductivity* 463–468 (Springer, 1989).
- Shin, J. et al. Tetrahedral atom ordering in a zeolite framework: a key factor affecting its physicochemical properties. *J. Am. Chem. Soc.* **133**, 10587–10598 (2011).
- Allix, M. et al. Considerable improvement of long-persistent luminescence in germanium and tin substituted ZnGa₂O₄. *Chem. Mater.* **25**, 1600–1606 (2013).
- Robertson, A., Tukamoto, H. & Irvine, J. Li_{1+x}Fe_{1-3x}Ti_{1+2x}O₄ (0.0 ≤ x ≤ 0.33) Based Spinel: Possible Negative Electrode Materials for Future Li-Ion Batteries. *J. Electrochem. Soc.* **146**, 3958 (1999).
- Ahtee, M. Lattice constants of some binary alkali halide solid solutions (Suomalainen Tiedeakatemia, 1969).
- Marchand, R., Pors, F. & Laurent, Y. Préparation et caractérisation de nouveaux oxy-nitrides à structure perovskite. *Rev. Int. Hautes Temp. Refract.* **23**, 11–15 (1986).
- Needs, R. & Weller, M. A new 2+/3+ perovskite: The synthesis and structure of BaScO₂F. *J. Solid State Chem.* **139**, 422–423 (1998).
- Rabenau, A. Perowskit- und fluoritphasen in den systemen ZrO₂-LaO_{1.5}-MgO und ZrO₂-LaO_{1.5}-CaO. *Z. Anorg. Allg. Chem.* **288**, 221–234 (1956).
- Drobyshevskaya, N. D., Gindin, E. I., Kirillova, G. K. & Magamadova, T. BARIUM MAGNETITANATE BaMg₆Ti₆O₁₉ WITH THE MAGNETOPLUMBIC STRUCTURE. *Inorg. Mater.* **25**, 1641 (1989).
- Ota, Y. Band alignment of β-(Al_xGa_{1-x})₂O₃ alloys via atomic solid-state energy scale approach. *AIP Adv.* **10**, 125321 (2020).
- Li, H. et al. Vacancy-induced anion and cation redox chemistry in cation-deficient F-doped anatase TiO₂. *J. Mater. Chem. A* **8**, 20393–20401 (2020).
- Zhang, Y., Xiao, Z., Kamiya, T. & Hosono, H. Electron confinement in channel spaces for one-dimensional electrode. *J. Phys. Chem. Lett.* **6**, 4966–4971 (2015).
- Tsuchimoto, A. et al. Nonpolarizing oxygen-redox capacity without OO dimerization in Na₂Mn₃O₇. *Nat. Commun.* **12**, 1–7 (2021).
- Ashbrook, S. E. & Dawson, D. M. Exploiting periodic first-principles calculations in NMR spectroscopy of disordered solids. *Acc. Chem. Res.* **46**, 1964–1974 (2013).
- Charpentier, T. The PAW/GIPAW approach for computing NMR parameters: A new dimension added to NMR study of solids. *Solid State Nucl. Magn. Reson.* **40**, 1–20 (2011).
- Bellaïche, L. & Vanderbilt, D. Virtual crystal approximation revisited: Application to dielectric and piezoelectric properties of perovskites. *Phys. Rev. B* **61**, 7877 (2000).
- Velický, B. Theory of electronic transport in disordered binary alloys: coherent-potential approximation. *Phys. Rev.* **184**, 614 (1969).
- Wei, S.-H., Ferreira, L., Bernard, J. E. & Zunger, A. Electronic properties of random alloys: Special quasirandom structures. *Phys. Rev. B* **42**, 9622 (1990).
- Van de Walle, A. et al. Efficient stochastic generation of special quasirandom structures. *Calphad* **42**, 13–18 (2013).
- Waston, D. K. & Dunn, M. Rearranging the exponential wall for large N-body systems. *Phys. Rev. Lett.* **105**, 020402 (2010).
- Grau-Crespo, R., Hamad, S., Catlow, C. R. A. & De Leeuw, N. Symmetry-adapted configurational modelling of fractional site occupancy in solids. *J. Phys.: Condens. Matter* **19**, 256201 (2007).
- Hart, G. L. & Forcade, R. W. Algorithm for generating derivative structures. *Phys. Rev. B* **77**, 224115 (2008).
- Hart, G. L. & Forcade, R. W. Generating derivative structures from multilattices: Algorithm and application to hcp alloys. *Phys. Rev. B* **80**, 014120 (2009).
- Hart, G. L., Nelson, L. J. & Forcade, R. W. Generating derivative structures at a fixed concentration. *Comput. Mater. Sci.* **59**, 101–107 (2012).
- Okhotnikov, K., Charpentier, T. & Cadars, S. Supercell program: a combinatorial structure-generation approach for the local-level modeling of atomic substitutions and partial occupancies in crystals. *J. Cheminformatics* **8**, 1–15 (2016).
- Lian, J.-C., Wu, H.-Y., Huang, W.-Q., Hu, W. & Huang, G.-F. Algorithm for generating irreducible site-occupancy configurations. *Phys. Rev. B* **102**, 134209 (2020).
- Sanchez, J. M., Ducastelle, F. & Gratias, D. Generalized cluster description of multicomponent systems. *Phys. A* **128**, 334–350 (1984).
- Van De Walle, A., Asta, M. & Ceder, G. The alloy theoretic automated toolkit: A user guide. *Calphad* **26**, 539–553 (2002).
- Seko, A., Koyama, Y. & Tanaka, I. Cluster expansion method for multicomponent systems based on optimal selection of structures for density-functional theory calculations. *Phys. Rev. B* **80**, 165122 (2009).
- Sanchez, J. Foundations and practical implementations of the cluster expansion. *J. Phase Equilibria Diffus* **38**, 238–251 (2017).
- Wu, Q., He, B., Song, T., Gao, J. & Shi, S. Cluster expansion method and its application in computational materials science. *Comput. Mater. Sci.* **125**, 243–254 (2016).
- Chang, J. H. et al. CLEASE: a versatile and user-friendly implementation of cluster expansion method. *J. Phys.: Condens. Matter* **31**, 325901 (2019).

35. Seko, A. & Tanaka, I. Cluster expansion of multicomponent ionic systems with controlled accuracy: importance of long-range interactions in heterovalent ionic systems. *J. Phys.: Condens. Matter* **26**, 115403 (2014).
36. Nguyen, A. H., Rosenbrock, C. W., Reese, C. S. & Hart, G. L. Robustness of the cluster expansion: Assessing the roles of relaxation and numerical error. *Phys. Rev. B* **96**, 014107 (2017).
37. Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **145**, 170901 (2016).
38. Mishin, Y. Machine-learning interatomic potentials for materials science. *Acta Mater.* **214**, 116980 (2021).
39. Shimizu, K. et al. Phase stability of Au-Li binary systems studied using neural network potential. *Phys. Rev. B* **103**, 094112 (2021).
40. Li, X.-G. et al. Quantum-accurate spectral neighbor analysis potential models for Ni-Mo binary alloys and fcc metals. *Phys. Rev. B* **98**, 094104 (2018).
41. Seko, A. Machine learning potentials for multicomponent systems: The Ti-Al binary system. *Phys. Rev. B* **102**, 174104 (2020).
42. Kasamatsu, S. et al. Facilitating ab initio configurational sampling of multicomponent solids using an on-lattice neural network model and active learning. *J. Chem. Phys.* **157**, 104114 (2022).
43. Wallace, S. K. et al. Free energy of $(\text{Co}_x\text{Mn}_{1-x})_3\text{O}_4$ mixed phases from machine-learning-enhanced ab initio calculations. *Phys. Rev. Mater.* **5**, 035402 (2021).
44. Liu, Y., Zhao, T., Ju, W. & Shi, S. Materials discovery and design using machine learning. *J. Mater. Sci.* **3**, 159–177 (2017).
45. Schmidt, J., Marques, M. R., Botti, S. & Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *NPJ Comput. Mater.* **5**, 1–36 (2019).
46. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *NPJ Comput. Mater.* **3**, 1–13 (2017).
47. Vabalas, A., Gowen, E., Poliakoff, E. & Casson, A. J. Machine learning algorithm validation with a limited sample size. *PLoS One* **14**, e0224365 (2019).
48. Zhang, Y. & Ling, C. A strategy to apply machine learning to small datasets in materials science. *NPJ Comput. Mater.* **4**, 1–8 (2018).
49. Vandermause, J. et al. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. *NPJ Comput. Mater.* **6**, 1–11 (2020).
50. Zhang, L., Lin, D.-Y., Wang, H., Car, R. & Weinan, E. Active learning of uniformly accurate interatomic potentials for materials simulation. *Phys. Rev. Mater.* **3**, 023804 (2019).
51. Gubaev, K., Podryabinkin, E. V., Hart, G. L. & Shapeev, A. V. Accelerating high-throughput searches for new alloys with active learning of interatomic potentials. *Comput. Mater. Sci.* **156**, 148–156 (2019).
52. Kostichenko, T., Körmann, F., Neugebauer, J. & Shapeev, A. Impact of lattice relaxations on phase transitions in a high-entropy alloy studied by machine-learning potentials. *NPJ Comput. Mater.* **5**, 1–7 (2019).
53. Liu, X.-W. et al. Iron carbides in Fischer-Tropsch synthesis: Theoretical and experimental understanding in epsilon-iron carbide phase assignment. *J. Phys. Chem. C* **121**, 21390–21396 (2017).
54. Hariyani, S. & Brgoch, J. Local structure distortion induced broad band emission in the all-inorganic BaCo_2F_6 : Eu^{2+} perovskite. *Chem. Mater.* **32**, 6640–6649 (2020).
55. Wang, Q., Grau-Crespo, R. & de Leeuw, N. H. Mixing thermodynamics of the calcite-structured $(\text{Mn,Ca})\text{CO}_3$ solid solution: A computer simulation study. *J. Phys. Chem. B* **115**, 13854–13861 (2011).
56. Yang, Y., Jiménez-Negrón, O. A. & Kitchin, J. R. Machine-learning accelerated geometry optimization in molecular simulation. *J. Chem. Phys.* **154**, 234704 (2021).
57. Yuan, X. et al. Crystal structure prediction approach to explore the iron carbide phases: Novel crystal structures and unexpected magnetic properties. *J. Phys. Chem. C* **124**, 17244–17254 (2020).
58. Oganov, A. R. & Glass, C. W. Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. *J. Chem. Phys.* **124**, 244704 (2006).
59. Glass, C. W., Oganov, A. R. & Hansen, N. USPEX—Evolutionary crystal structure prediction. *Comput. Phys. Commun.* **175**, 713–720 (2006).
60. Oganov, A. R., Lyakhov, A. O. & Valle, M. How evolutionary crystal structure prediction works and why. *Acc. Chem. Res.* **44**, 227–237 (2011).
61. Lyakhov, A. O., Oganov, A. R., Stokes, H. T. & Zhu, Q. New developments in evolutionary structure prediction algorithm USPEX. *Comput. Phys. Commun.* **184**, 1172–1182 (2013).
62. Wang, Y., Lv, J., Zhu, L. & Ma, Y. CALYPSO: A method for crystal structure prediction. *Comput. Phys. Commun.* **183**, 2063–2070 (2012).
63. Lonie, D. C. & Zurek, E. Xtalopt: An open-source evolutionary algorithm for crystal structure prediction. *Comput. Phys. Commun.* **182**, 372–387 (2011).
64. Davis, L. Handbook of genetic algorithms (CuminCAD, 1991).
65. Kirkpatrick, S., Gelatt, C. D. Jr & Vecchi, M. P. Optimization by simulated annealing. *Science* **220**, 671–680 (1983).
66. Marini, F. & Walczak, B. Particle swarm optimization (PSO). a tutorial. *Chemom. Intell. Lab. Syst.* **149**, 153–165 (2015).
67. Podryabinkin, E. V. & Shapeev, A. V. Active learning of linearly parametrized interatomic potentials. *Comput. Mater. Sci.* **140**, 171–180 (2017).
68. Podryabinkin, E. V., Tikhonov, E. V., Shapeev, A. V. & Oganov, A. R. Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning. *Phys. Rev. B* **99**, 064114 (2019).
69. Tong, Q., Xue, L., Lv, J., Wang, Y. & Ma, Y. Accelerating CALYPSO structure prediction by data-driven learning of a potential energy surface. *Faraday Discuss* **211**, 31–43 (2018).
70. Huang, S.-D., Shang, C., Kang, P.-L., Zhang, X.-J. & Liu, Z.-P. LASP: Fast global potential energy surface exploration. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **9**, e1415 (2019).
71. Hoerl, A. E. & Kennard, R. W. Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics* **12**, 55–67 (1970).
72. Hoerl, A. E. & Kennard, R. W. Ridge regression: applications to nonorthogonal problems. *Technometrics* **12**, 69–82 (1970).
73. Deringer, V. L. et al. Gaussian process regression for materials and molecules. *Chem. Rev.* **121**, 10073–10141 (2021).
74. Castillo, I., Schmidt-Hieber, J. & Van der Vaart, A. Bayesian linear regression with sparse priors. *Ann. Stat.* **43**, 1986–2018 (2015).
75. Bishop, C. M. et al. Neural networks for pattern recognition (Oxford university press, 1995).
76. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
77. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
78. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).
79. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
80. Huo, H. & Rupp, M. Unified representation of molecules and crystals for machine learning. *Mach. Learn.: Sci. Technol.* **3**, 045017 (2022).
81. Himanen, L. et al. Dscribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **247**, 106949 (2020).
82. Pedregosa, F. et al. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
83. Khorshidi, A. & Peterson, A. A. Amp: A modular approach to machine learning in atomistic simulations. *Comput. Phys. Commun.* **207**, 310–324 (2016).
84. Zhang, Y. et al. DP-GEN: A concurrent learning platform for the generation of reliable deep learning based potential energy models. *Comput. Phys. Commun.* **253**, 107206 (2020).
85. Seko, A., Takahashi, A. & Tanaka, I. Sparse representation for a potential energy surface. *Phys. Rev. B* **90**, 024101 (2014).
86. Seko, A., Takahashi, A. & Tanaka, I. First-principles interatomic potentials for ten elemental metals via compressed sensing. *Phys. Rev. B* **92**, 054113 (2015).
87. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996).
88. Schapire, R. E. The Boosting Approach to Machine Learning: An Overview, in *Nonlinear Estimation and Classification 149–171* (Lecture Notes in Statistics vol. 171, Springer, 2003).
89. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558 (1993).
90. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50 (1996).
91. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
92. Ong, S. P. et al. Python Materials Genomics (pymatgen): A robust, opensource python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
93. Methfessel, M. & Paxton, A. High-precision sampling for Brillouin-zone integration in metals. *Phys. Rev. B* **40**, 3616 (1989).

ACKNOWLEDGEMENTS

The authors are grateful for the financial support from the National Key R&D Program of China (No. 2022YFA1604103), National Science Fund for Distinguished Young Scholars of China (Grant No. 22225206), the National Natural Science Foundation of China (Nos. 21972157, 21972160 and 21703272), CAS Project for Young Scientists in Basic Research (YSBR-005), Key Research Program of Frontier Sciences CAS (ZDBS-LY-7007), Major Research plan of the National Natural Science Foundation of China (92045303), CAS Project for Internet Security and Information Technology (CAS-WX2021SF0110), Science and Technology Plan Project of Inner Mongolia Autonomous Region of China (2021GG0309), and funding support from Beijing Advanced Innovation Center for Materials Genome Engineering, Synfuels China, Co. Ltd, and Institute of Coal Chemistry (CAS). Q. P. would like to acknowledge the support

provided by LiYing Program of the Institute of Mechanics, Chinese Academy of Sciences (Grant No. E1Z1011001).

AUTHOR CONTRIBUTIONS

X.-D.W., Y.-W.L., and Y.Y. designed research; X.-Z.Y., and Y.-W.Z. performed research and developed the LAsou Method; X.-Z.Y., Y.-W.Z., and Q.P. analyzed data; X.-Z.Y. and Y.-W.Z. wrote the paper; Q.P., and X.-D.W. polished the language and revised the paper. All the authors contributed to the idea and reviewed the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41524-023-00967-z>.

Correspondence and requests for materials should be addressed to Yuwei Zhou, Qing Peng or Xiaodong Wen.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023