Shotgun crystal structure prediction using machine-learned formation energies

Chang Liu[1], Hiromasa Tamaki[2], Tomoyasu Yokoyama[2], Kensuke Wakasugi[2], Satoshi Yotsuhashi[2], Minoru Kusaba[1], and Ryo Yoshida[1,3,4,*]

[1]The Institute of Statistical Mathematics, Research Organization of Information and Systems, Tachikawa, Tokyo 190-8562, Japan
[2]Technology Division, Panasonic Holdings Corporation, Kadoma, Osaka 571-8508, Japan
[3]National Institute for Materials Science, Research and Service Division of Materials Data and Integrated System, Tsukuba, 305-0047, Japan
[4]The Graduate University for Advanced Studies, Department of Statistical Science, Tachikawa, 190-8562, Japan
[*]yoshidar@ism.ac.jp

ABSTRACT

Stable or metastable crystal structures of assembled atoms can be predicted by finding the global or local minima of the energy surface defined on the space of the atomic configurations. Generally, this requires repeated first-principles energy calculations that are impractical for large systems, such as those containing more than 30 atoms in the unit cell. Here, we have made significant progress in solving the crystal structure prediction problem with a simple but powerful machine-learning workflow; using a machine-learning surrogate for first-principles energy calculations, we performed non-iterative, single-shot screening using a large library of virtually created crystal structures. The present method relies on two key technical components: transfer learning, which enables a highly accurate energy prediction of pre-relaxed crystalline states given only a small set of training samples from first-principles calculations, and generative models to create promising and diverse crystal structures for screening. Here, first-principles calculations were performed only to generate the training samples, and for the optimization of a dozen or fewer finally narrowed-down crystal structures. Our shotgun method proved to be computationally less demanding compared to conventional methods, which heavily rely on iterations of first-principles calculations, and achieved an exceptional prediction accuracy, reaching 92.2% in a benchmark task involving the prediction of 90 different crystal structures.

Introduction

Predicting the stable or metastable structures of a crystalline system with a given chemical composition is a fundamental unsolved problem that has been studied for several decades in solid-state physics[1,2]. In principle, the stable or metastable crystal structures of assembled atoms or molecules in the solid state can be determined using quantum mechanical calculations. Crystal structure prediction (CSP) is based on finding the global or local minima of the energy surface defined on a broad space of atomic configurations, in which the energy can be evaluated by first-principles density functional theory (DFT) calculations. To solve the CSP problem, we can apply an exploratory algorithm to determine the crystal structure at the global or local minimum by successively displacing the atomic configurations along the energy gradient.

To solve this hard problem, a broad array of CSP methods have been developed to date, including brute-force random search[3–5], simulated annealing[6,7], the Wang-Landau method[8], particle swarm optimization[9,10], genetic algorithms[2,11,12], Bayesian optimization[13], and look ahead based on quadratic approximation (LAQA)[14]. More recently, machine learning interatomic potentials have been attracting increasing attention because they can greatly speed up the optimization process by bypassing time-consuming ab initio calculations[15–17]. Conventionally, genetic manipulations such as mutation and crossover are performed to modify a current set of candidate crystal structures, and their DFT energies are used then as goodness-of-fit scores to prioritize promising candidates for survival in the new generation. This process is repeated until the energy minima are reached. For example, the pioneering software USPEX implements

a comprehensive set of genetic operations such as the mutation and crossover of crystal objects[2,11,12], while the CALYPSO code employs a genetic operation called the swarm shift[18]. However, such algorithms are time-consuming because of the need to perform ab initio structural relaxation of the candidate crystals at every step of the optimization process. CrySPY was developed to increase the computational efficiency by introducing a machine-learning energy calculator[14] based on the Gaussian process regressor[19]. The predictive performance is successively improved by accumulating a training set of candidate crystal structures and their relaxed energies via Bayesian optimization[20]. The surrogate energy predictor efficiently rules out unpromising candidates whose energies are unlikely to reach the minima. However, most existing methods utilize relaxed energy values to evaluate the goodness-of-fit in the selection process or to produce instances to train a surrogate model. Therefore, it is necessary to relax all candidate structures at every step of the sequential search. Such methods are impractical for large systems that contain more than 30–40 atoms in a unit cell, owing to their enormous computational cost.

To overcome this difficulty, a promising solution is to fully replace ab initio energy calculations with machine-learning surrogates. Energy prediction models trained using DFT property databases, such as the Materials Project[21,22], AFLOW[23,24], OQMD[25,26], and GNoME database[27] have been reported to exhibit a quite high prediction accuracy[28–30]. However, models trained on the instances from stable or metastable structures in such a database are inapplicable to the prediction of unrelaxed energies of varying atomic configurations for a given target system, as discussed in Gibson et al. (2022)[31]. As shown later numerically, such models can predict energy differences between different crystalline systems but cannot quantitatively discriminate energy differences of distinct conformations for the system of interest. This is the ability that is required for solving the CSP problem.

In this study, we employed a simple approach to building a predictive model for formation energies. First, a crystal-graph convolutional neural network (CGCNN)[30] was trained using diverse crystals with stable or metastable state energies from the Materials Project database. Subsequently, for a given chemical composition as the target, the DFT energies of a few dozen randomly generated unrelaxed structures were calculated by performing single-point energy calculations, and a transfer learning technique[32,33] was applied to fine-tune the pretrained CGCNN to the target system. Generally, limited data are available for model training, and randomly generated crystal conformations are distributed in high-energy regions. Models trained on such data that are biased towards high-energy states are generally not applicable to the extrapolative domain of low-energy states in which optimal or suboptimal conformers exist. In CSP, a surrogate model must be able to predict the energy of various conformations with high- to low-energy states corresponding to the pre- and post-relaxed crystal structures, respectively. We demonstrate that the surrogate model derived using the transfer learning method exhibited sufficiently high prediction accuracy, even in the domain of low-energy states.

After creating candidate crystal structures, exhaustive virtual screening was performed using the transferred energy predictor. The narrowed-down candidate crystals were relaxed by performing DFT calculations. Currently, a wide variety of structure generators can be applied to generate the virtual crystal libraries, e.g., (i) methods based on element substitution using existing crystal structures as templates[34–36], (ii) atomic coordinate generators that take into account crystallographic topology and symmetry[37,38], (iii) algorithms for reconstructing atomic configurations based on interatomic distance matrices (contact maps) predicted by machine learning[39], and (iv) deep generative models that mimic crystals synthesized to date[40–42]. In this study, we validated our framework using two sets of virtual libraries created using methods related to (i) and (ii), i.e., element substitution of template crystal structures and a Wyckoff position generator for de novo CSP. The search space is narrowed down in the latter by machine-learning prediction of space groups and Wyckoff letter assignments. Our workflow, which can be regarded as a high-throughput virtual screening of crystal structures, is perhaps the simplest among existing CSP methods to date. In the entire workflow, first-principles single-point calculations were performed for, at most, 3,000 structures to create a training set for the transferred energy predictor and for the structural relaxation of a dozen or fewer narrowed-down candidate crystals in the final stage. Compared to conventional methods such as USPEX, the present method is approximately two times or more less computationally demanding. Furthermore, outstanding prediction performance was also experimentally confirmed; in the prediction of the stable structures of 90 benchmark crystals that were chosen to obtain a set of materials with diverse space groups, structure types, constituent elements, system sizes, and application domains, we succeeded in accurately predicting approximately 90% of the benchmark structures.

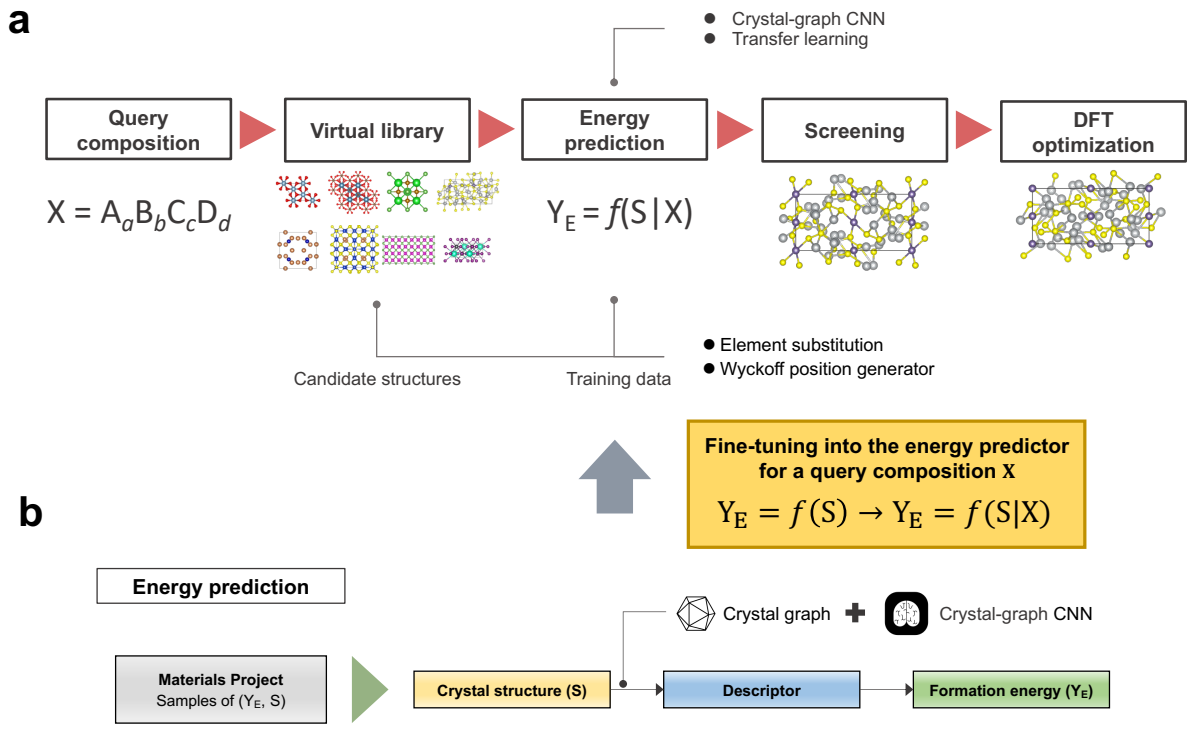


Figure 1. Workflow of the shotgun CSP. (a) Virtual screening using machine-learned formation energies. The virtual library is created using the template-based structure generator and Wyckoff position generator. (b) Construction of the formation energy predictor based on a CGCNN. The CGCNN trained with the Materials Project database is fine-tuned to the energy predictor of unrelated candidate crystal structures for a query composition X.

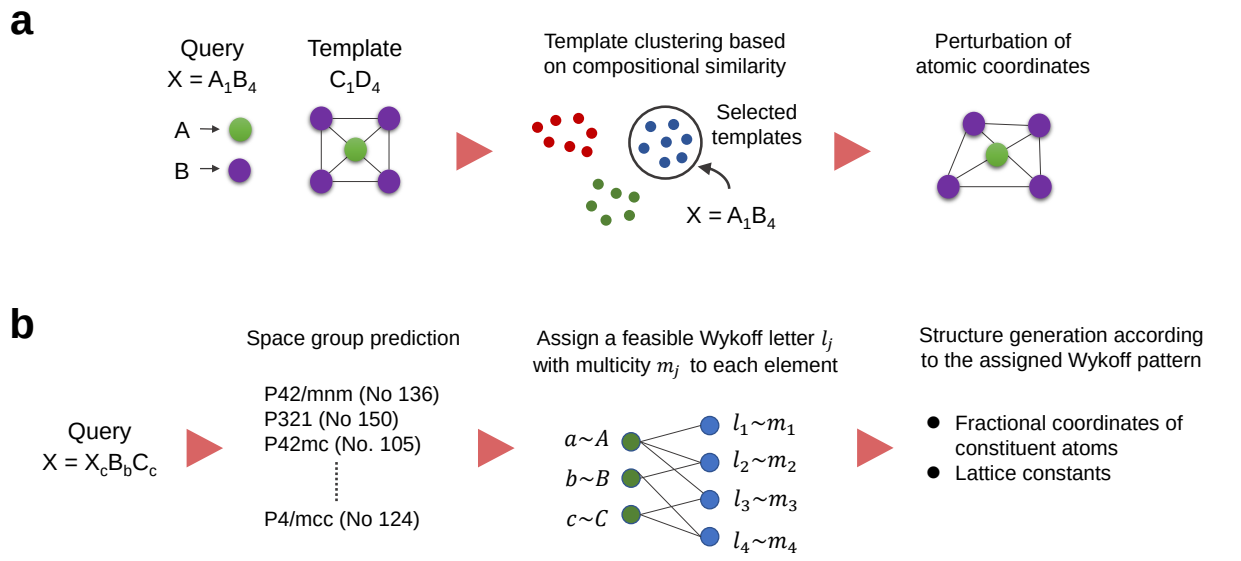


Figure 2. Two different crystal structure generators are used for the generation of training instances for the fine-tuning of the energy prediction model and the creation of virtual libraries to be screened. (a) Element substitution of template crystal structures in the Materials Project. (b) Wyckoff position generator to produce symmetry-restricted atomic coordinates.

Results

Methods outline

To predict the stable crystal structure formed by the assembled atoms with a given chemical composition X, we used a machine-learning workflow summarized in Figure 1. This method involves two key technical components: a high-performance surrogate model for DFT formation energies and two different generative models for candidate crystal structures.

For the energy calculation, a CGCNN with the same architecture as in the original paper[30] was first pretrained from scratch on a set of 126,210 crystals for which DFT formation energies are available in the Materials Project. We refer to this model as the global model; it can accurately predict the baseline formation energies of diverse crystal structures but lacks the ability to discriminate the local energy difference of different atomic conformations for a given target system. Therefore, the pretrained global model was transformed into a model localized to the target system X. To this end, we randomly generated, at most, several thousand virtual crystal structures and calculated their formation energies by performing single-point DFT calculations. The training structures were generated using either the random element substitution of existing crystals or the Wyckoff position generator, as described below. Using this dataset, we performed transfer learning to adapt the pretrained global model to a local model applicable to the energy evaluation of different configurations for X. Here, the output layer was trained from scratch, whereas the pretrained weight parameters in the other layers were retained and fine-tuned (see the Methods section).

In this study, we developed and tested two algorithms for the generation of virtual crystals.

Method 1—Element substitution: Element substitution of already synthesized or theoretically possible crystals with the same compositions as X is randomly performed (Figure 2(a)).

Method 2—Wyckoff position generator: For a target composition with the space group given a priori or predicted, the generator creates symmetry-restricted atomic coordinates randomly from all possible combinations of Wyckoff positions (Figure 2(b)). Here, we employ a machine learning predictor of Wyckoff label assignments.

The element-substitution method cannot be applied unless a template is available for substitution, limiting its applicability. Therefore, we developed the Wyckoff position generator to generate novel structural patterns. As described below, the space group and Wyckoff letter configuration to constituent atoms are inferred based on machine learning. These two generators were used to generate training instances in the fine-tuning of CGCNN and the candidate structure in the high-throughput virtual screening. We used slightly different workflows for these two generators, as described below.

The generation of crystal structures by replacing the elements in the existing crystals mimics the process by which humans synthesize new crystalline materials in a laboratory. For a given query composition X, we collected a set of template crystal structures with the same compositional ratio as that of the X from the Materials Project database. A candidate structure was created by assigning the constituent elements of the query composition to the atomic coordinates of a selected template. Elements with the same composition fraction in the template and in the query were substituted. When one or more elements have the same composition fraction, the assignment cannot be uniquely determined. In this case, substitution was performed on the most similar element pair with the normalized Euclidean distance of the 58 element descriptors in the XenonPy library[33,43–45] used as the similarity measure. The element-substituted crystal structure inherits the atomic coordinates of the template structure. A slight random perturbation was incorporated into the generated atomic coordinates as an additional refinement step. Considering that multiple crystals in the database belong to the same prototype structure (for example, 8,005 compounds have the same composition ratio $A_1B_1C_2$), a cluster-based template selection procedure was introduced. The objective is to select highly relevant templates with query composition X while maintaining the diversity of the template structures. We applied DBSCAN[46,47] to classify the templates into clusters in which the chemical compositions were converted into 290-dimensional compositional descriptors using XenonPy. Then, only the templates belonging to the same cluster as the query X were selected to narrow down to a set of templates with high compositional similarity (see the Supplementary Information for a brief explanation of the DBSCAN algorithm). In addition, to eliminate structurally redundant templates, we used pymatgen's StructureMatcher module to construct a unique set of templates that did not contain identical prototype structures. The number of the unique templates is denoted by K_{temp}. For virtual library creation, 1,000 structures were generated from each of the K_{temp} selected templates by perturbing the atomic coordinates and lattice constants, resulting in the $1000 \times K_{\mathrm{temp}}$ candidate structures. For the training dataset in the fine-tuning, we used 10 structures generated randomly for each template using the same procedure. For more details, see the Methods section.

The Wyckoff position generator produces random crystal structures with a prescribed space group for a given composition. For a given X, the space group of its stable structure is predicted based on machine learning. Furthermore, the assignment of Wyckoff letters to the constituent atoms is narrowed down efficiently using a predictive model trained on a given set of crystal structures in the Materials Project database as described later. With this predictive model, we can randomly generate promising Wyckoff patterns while significantly pruning wasted search space. With Wyckoff site multiplicity and symmetry restricted, the atomic coordinates and lattice parameters were generated uniformly from specific intervals. Structures generated with two or more atoms within a certain distance were excluded a posteriori. Here, a space group predictor was used to estimate the space group of X. The objective is to predict and limit the space group of the stable crystalline state for a given composition X. We compiled a list of chemical compositions and the space groups of 33,040 stable crystal structures from the Materials Project database for the training set. Using this model, the space group of the crystal system for X was narrowed to the top K_{SG} candidates. In this study, we set $K_{\mathrm{SG}} = 30$. Based on this setting, $100 \times K_{\mathrm{SG}}$ training instances and $15,000 \times K_{\mathrm{SG}}$ candidate crystals were generated for the fine-tuning and virtual screening steps, respectively. For more details, see the Methods section.

The transferred energy prediction model was then used to perform exhaustive virtual screening using each of the two generators separately. Finally, DFT calculations were performed to optimize the narrowed-down promising structures that exhibited the lowest predicted energies using the Vienna Ab initio Simulation Package (VASP)[48] version 6.1.2, combined with projector augmented wave (PAW) pseudopotentials[49] (see the Methods section for detailed procedures). The top K lowest-energy structures were subjected to structural relaxation with DFT. In this study, we set $K = 10 \times K_{\mathrm{SG}}$ and $K = 5 \times K_{\mathrm{temp}}$ for the Wyckoff position generation and for the element substitution generation, respectively. Generally, the top K candidates that reached the lowest energies consisted of significantly similar structures, many of which converged to the same crystal structure during the structural relaxation phase. To eliminate this redundancy, we considered structural similarity when selecting the top K candidate structures to maintain high structural diversity (the Methods section).

Benchmark sets

In this study, the performance of the proposed CSP algorithm was evaluated mainly on two benchmark sets. The first benchmark set (Dataset I) consists of 40 stable crystals selected based on a literature survey, as listed in Table 1, was selected based on two criteria: the diversity of space groups, constituent elements, number of atoms, and element species; and the diversity of applications such as battery and thermoelectric materials. Due to the presence of a certain bias in the selection of Dataset I, 50 additional stable crystals were randomly selected from the Materials Project (Table 2) (Dataset II). For Datasets I and II, the number of atoms in the unit cell of the selected crystals ranged from 2 to 104 and 2 to 288, with the mean $\pm$ standard deviation of 23.13 ± 24.09 and 32.68 ± 45.41, respectively (see also the histograms in the Supplementary Information). 30% of the benchmark crystals had more than 30 atoms; due to the computational complexity and search performance, these structures were expected to be difficult to solve with conventional heuristic searches based on first-principles calculations in the majority of cases.

As a more challenging benchmark, we randomly selected 30 stable structures for which no template exists from the Materials Project, designating as Dataset III (Table S1). As shown in Table S1, most of the crystal structures in Dataset III have a much larger number of atoms in the unit cell than in Dataset I and II (the average atomic number is 66.50 ± 34.40).

Space group prediction

In the virtual screening with the Wyckoff position generator, to narrow down the huge space of possible crystalline states, we introduced a multiclass discriminator to predict the space group Y_{SG} of a given chemical composition X (Figure 3(a)). To train and test the classifier, we used 33,040 instances of chemical compositions with 213 distinct space groups of the stable crystalline states compiled from the Materials Project database. The 120 benchmark crystal structures were removed from the training dataset. The compositional features of X were encoded into the 290-dimensional descriptor vector using XenonPy[33,43–45] (see the Methods section), and fully connected neural networks were trained to learn the mapping from the vectorized compositions to the 213 space groups. Of the total sample set, 80% of the instances were used for training, and the remainder was used for testing. To statistically evaluate the prediction accuracy, training, and testing were repeated 100 times independently. The details of the model construction, including hyperparameter adjustment, are described in the Supplementary Information.

Figure 3(b) shows the change in the recall rate from the top 1 to the top 40 predictions; that is, the change in the proportion of true labels included in the top K_{SG} most probable predicted class labels ($K_{\mathrm{SG}} = 1, \ldots, 40$). The recall rates in the top 1, 10, and 40 predictions were $60.22(\pm 0.87)\%$, $85.35(\pm 0.54)\%$, and $94.02(\pm 0.43)\%$ on average, respectively. This result indicates that by narrowing down to the top 40 predicted labels, we can identify the space groups for $94.02(\pm 0.43)\%$ of the various crystalline systems. Using this model, the 213 space groups were narrowed down to the top 40 candidates, and for each of the selected candidates, a set of symmetry-restricted crystal structures was generated using the Wyckoff generator. The top 40 recall rates for different space groups were varied from $0.00(\pm 0.00)\%$ to $99.22(\pm 0.31)\%$ as shown in Figure S1. The variability of the recall rate was partially correlated with the number of training instances in each space group.

Here, there was concern that the pattern of composition ratios in the dataset is highly biased. In such cases, the prediction performance would vary significantly from one composition ratio to another. We denote by S the upper bound on the number of samples with the same composition ratio in the training dataset. To address this concern, we evaluated the change in the prediction accuracy when the upper limit on the number of samples with the same composition was varied as $S \in \{10, 50, 100\}$, and as shown in Figure S3, it was verified that the prediction performance did not vary significantly.

Wyckoff pattern prediction

After narrowing down the space groups with machine learning, Wyckoff letters are randomly assigned to each atom to generate the atomic coordinates. As the number of combinations of atoms and Wyckoff letters increases, the complexity of de novo CSP grows. For example, in the case of the stable structure with space group *lmma* (No. 74) for Mg_8B_{56}, the multiplicity of Wyckoff letters $\{a, b, c, d, e, f, g, h, i, j\}$ is $\{4, 4, 4, 4, 4, 8, 8, 8, 8, 16\}$. In this case, the number of possible assignments for the Wyckoff letters exceeds 1,755. If the assignments are incorrect, the CSP fails to predict in most cases. On the other hand, for the space group $la\bar{3}d$ (No. 230) of $Y_{24}Al_{40}O_{96}$, the multiplicity of Wyckoff letters $\{a, b, c, d, e, f, g, h\}$ is $\{16, 16, 24, 24, 32, 48, 48, 96\}$. Despite the substantial number of atoms in the unit cell, only 27

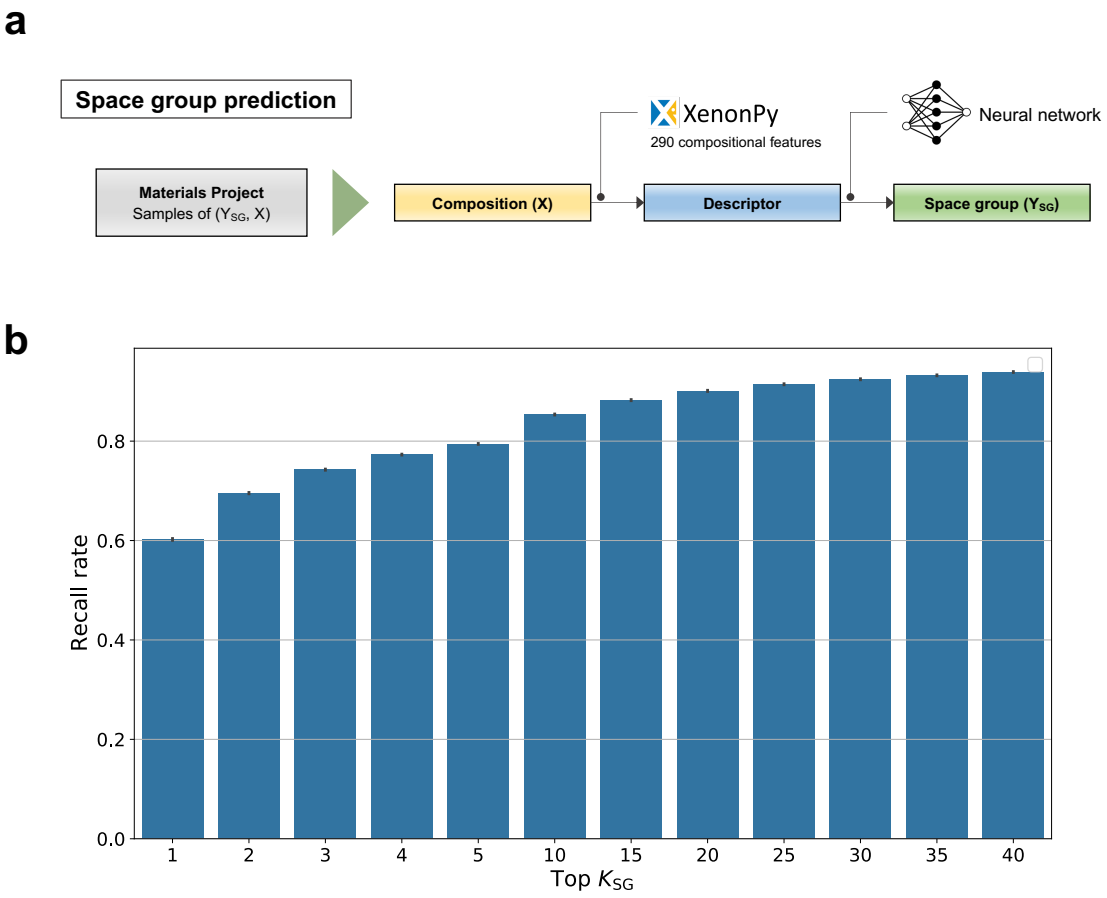


Figure 3. Space group prediction. (a) Machine learning workflow. (b) Change in the recall rate across the top 1 to the top 40 predictions of the space group. By narrowing down to the top 40 predicted labels, it is expected on average to enable the inclusion of the true space groups for 94.02% of the entire crystalline systems.

possible assignments exist. Successfully narrowing down the space groups and accurately predicting the assignments of Wyckoff letters is expected to improve the CSP task significantly.

Therefore, we constructed a model to predict the occurrence frequencies of Wyckoff letters for stable structures based on chemical composition (Figure 4(a)). The model was created for each space group, with the input chemical composition represented by the 290-dimensional descriptor using XenonPy. The output is the probability distribution of the occurrence of Wyckoff letters. Using 33,040 instances of the stable structures in the Materials Project, excluding the 120 benchmark crystals (80% for training and 20% for test), we trained a random forest regressor for each space group.

Figure 4(b) summarizes the prediction accuracy for the test set. The discrepancies between the output probability distribution $p_1, \ldots, p_M$ of the trained model and actual relative frequencies $q_1, \ldots, q_M$ of M Wyckoff letters were measured based on the Kullback-Leibler (KL) divergence:

$$\sum_{i=1}^{M} q_i \log \frac{q_i}{p_i} \tag{1}$$

The distribution of KL divergence for the test set was found to be highly concentrated around zero in most cases (Figure 4(b)). This observation indicates that the Wyckoff letters for stable structures are predictable from chemical composition.

Using the Wyckoff letter occurrence probability for a query composition, we randomly assigned Wyckoff letters according to the procedure shown in the Methods section. The possible assignments of the Wyckoff letters to elements are constrained by their multiplicity and query composition ratio. The sampling algorithm is designed to satisfy these constraints and generally aligns with the predicted probabilities from the random forest regressor.

Figure 4(c) shows the difference in Wyckoff pattern generation with and without employing the Wyckoff letter predictor, comparing the frequencies of actual and sampled Wyckoff letters for six randomly selected cases. In all cases, the frequencies of Wyckoff letters generated from the predictor agreed well with the true frequencies. In contrast, the randomly generated Wyckoff letters deviate significantly from the true frequencies.

Energy prediction

For the global energy prediction model, the CGCNN was trained on 126,210 stable and metastable crystal structures, with their formation energies retrieved from the Materials Project; the 120 benchmark crystals were excluded from the training set. To validate the prediction capability and uncertainty of the global model, we randomly extracted 80% of the overall dataset and created 100 bootstrap sets. The mean absolute error (MAE) with respect to the 25,249 test cases reached 0.074 eV/atom on average, with a standard deviation of 0.003, which is comparable to that in previous studies, for example, Xie & Grossman[30]. Figure 5(a) shows the prediction results for the 90 benchmark crystals in Datasets I and II.

Note that the global model is inapplicable to the energy prediction of different conformations randomly generated for each X, as shown in Figure 5(c), and clearly failed to discriminate the energies of different conformations randomly generated for the 90 benchmark crystals. Here, we tested the prediction capability of the global model on the DFT energies of 100 randomly generated conformations for each of the 90 benchmark queries. It was found that the MAE decreased to 6.126 eV/atom on average with a standard deviation of 2.010. A similar result was obtained from a global model trained with approximately 1,021,917 instances of the OQMD database, including the formation energies of both relaxed and unrelaxed structures.

To overcome this limited predictive ability, the pretrained global model was transferred to a model localized to the target system of X. For each X, the formation energies of, at most, 3,000 virtual crystals generated as described above were obtained by DFT single-point energy calculations, and the pretrained global model was fine-tuned to the target system. As shown in Figure 5(d), transfer learning successfully improved the prediction performance for the formation energies of the 9,000 additional conformations generated. The MAE reached 0.488 eV/atom on average with a standard deviation of 0.453, corresponding to a factor of 12.6 improvements compared to the MAE of the pretrained global energy prediction model.

CSP using the library generator based on element substitution

We used a fine-tuned surrogate energy predictor to sort the generated virtual crystals, narrowed them down to the top 5 structures for each template (as described above), and then performed structural relaxations using DFT. The J relaxed

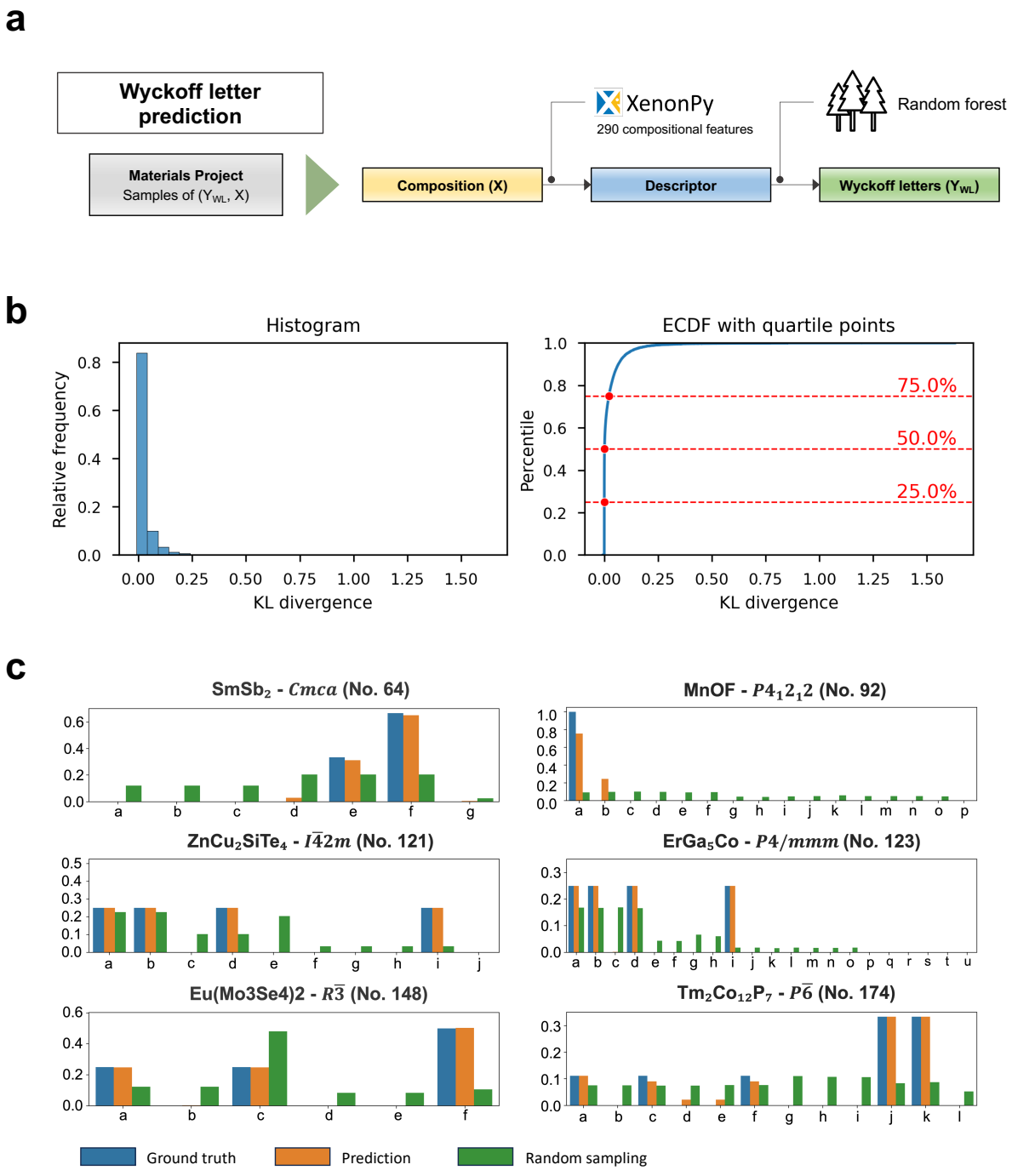


Figure 4. Prediction of Wyckoff letter assignments. (a) Machine learning workflow. (b) Histogram and empirical cumulative distribution function (ECDF) of KL divergence between the relative occurrence frequencies and predicted probability distribution of Wyckoff letters for the test set. (c) Histograms show the distributions of relative occurrence frequencies and predicted probabilities of Wyckoff letters for six randomly selected compounds, with their space group information shown in parentheses.

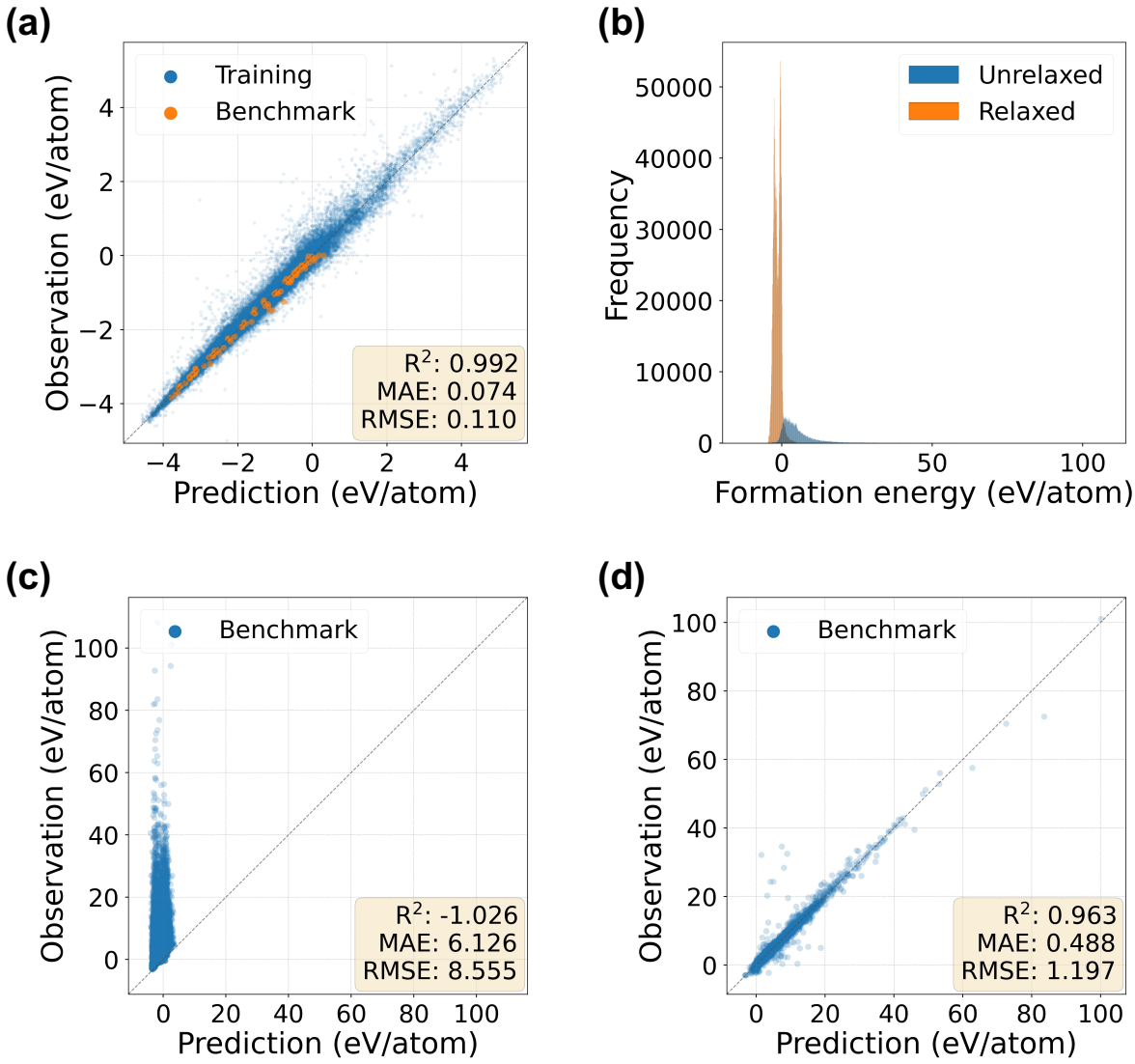


Figure 5. Performance of CGCNNs for the prediction of DFT formation energies with and without transfer learning. The root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) with respect to test instances are shown on each parity plot. (a) Results of the global model for the prediction of the relaxed formation energies of the 90 benchmark crystals (orange). (b) Histogram of the DFT formation energies of relaxed and randomly generated unrelaxed structures. (c)-(d) Prediction of unrelaxed formation energies of 100 randomly generated conformations for each of the 90 benchmark systems without and with fine-tuning of the pretrained global energy prediction model, respectively.

structures with the lowest DFT energies were used as the final set of predicted structures. Figure 7 shows the top 2 ($J = 2$) predictions and the true structures. Figure S3 shows the top 2 predicted structures for all of the 90 benchmark crystals. Tables 1 and 2 summarize the success or failure of the top 10 predictions for all crystal systems in Datasets I and II. The top 5 prediction accuracies ($J = 5$) for Datasets I and II were 75.0% and 76.0%, respectively, and the top 10 accuracies ($J = 10$) for Datasets I and II were 82.5% and 86.0%, respectively. No significant differences are observed in the resulting accuracies between the two benchmark sets. Interestingly, no performance degradation was observed due to the increasing number of atoms in the unit cell. Of the 90 crystals, the five ($NaCaAlPHO_5F_2$, $K_{20}Ag_8As_{12}Se_{36}$, $Na_1W_9O_{27}$, $Na_{80}Fe_{16}P_{32}O_{128}F_{32}$, and $Y_8Si_{10}Ir_{18}$) that could not be predicted had no template structures in the Materials Project with the same composition ratio. Excluding the five cases with no template, the top 10 accuracies for Datasets I and II reached 84.6% and 93.5%, respectively. In 9 of the 85 cases where there was a template but the prediction failed, none of the Wyckoff letter patterns of the true structures were included in the template structure set. Elemental substitution-based structure generation requires a one-to-one correspondence between structure and composition, making it difficult to predict the structure of compounds for which multiple structures are reported for a given composition. The number of data for SiO_2, V_2O_5, and TiO_2 in the Materials Project is 314, 58, and 42, respectively, and these compounds failed to be predicted. In summary, crystal systems with template structures that have the same composition ratio can be largely predicted by substituting the elements in the existing crystals. For example, in the Materials Project, the proportion of crystals with one or more interchangeable template structures was 98.0%. A similar conclusion was reached in a previous study[36] that proposed a CSP algorithm called CSPML based on elemental substitution using machine learning. As shown in Tables 1, 2, and S4, the shotgun CSP's prediction accuracy for Dataset I and II was significantly better than that of CSPML, which 65.6% for the top 10 candidates.

CSP using the Wyckoff position generator

The top 10 candidate structures with the lowest surrogate energies for each predicted space group were selected for structural relaxation using DFT. Similar to the results of the template-based method, Figure 7 displays the top 2 predicted and true structures for some selected examples, and Figure S2 shows the top 2 predicted structures for the 90 benchmark crystals. For the top 10 predicted structures, 77.5% and 68.0% of the known stable structures were predicted for Datasets I and II, respectively. Tables 1 and 2 summarize the success or failure of the top 10 predictions for all crystal systems in Datasets I and II. A significant decrease in performance was observed compared to the CSP algorithm using the template structure generator. One reason is the predictive performance of the space group. Considering the top 30 predictions, approximately 5% of the 90 benchmark crystal structures still have their space groups inaccurately predicted. This is almost the same level of accuracy as reported above.

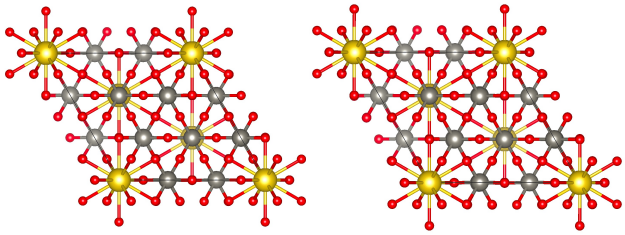
A total of 31 and 34 crystals were successfully predicted in Dataset I and II, respectively. Among these successful predictions, three crystals, $Y_4Si_5Ir_9$, $K_5Ag_2(AsSe_3)_3$, and $Na(WO_3)_9$ with no templates yet were successfully predicted. For these compounds, the number of atoms in the unit cells within their space group is quite large: 36, 76, and 111. Nevertheless, for the space group $R\bar{3}$ (No. 148) of the stable structure of $Na(WO_3)_9$ (Figure 6(a)), because of the Wyckoff letters $\{a,b,d,e\}$ are coordinate-fixed, the number of possible combinations of Wyckoff letters is reduced to approximately 48 due to its multiplicity restriction. Consequently, the effective dimension of the search space could be reduced by considering crystal symmetry in the structure generation. This explains why the shogun CSP successfully predicted the complex stable structure of $Na(WO_3)_9$.

Conversely, for $K_5Ag_2(AsSe_3)_3$, which has 76 atoms in its unit cell under space group *Pnma* (No. 62), as shown in Figure 6(b), the potential substitution of Wyckoff position c with a or b increases the number of possible Wyckoff letter combinations to over 300, even when considering multiplicity constraints. Nevertheless, the CSP method proved successful, largely because the frequencies of Wyckoff letters a and b, as predicted by the Wyckoff letter assignment predictor, were extremely low. This insight significantly narrowed the extensive search space during the candidate structure generation phase, exemplifying the impact of strategic considerations in CSP methodologies. The same improvement was encountered in $Y_4Si_5Ir_9$.

On the other hand, of the 85 crystals for which the space group was correctly identified, 9 and 11 true stable structures could not be predicted in Datasets I and II, respectively. Furthermore, of the 20 ($9+11$) failure cases, 3 and 3 structures failed to generate true Wyckoff patterns, and 6 and 8 failed to obtain the ground truth, even though the true Wyckoff patterns were correctly generated in Datasets I and II, respectively. To elucidate the origin of these failures, we examined the generated structures in detail. Consequently, it was found that the majority of structures that could not be predicted were characterized by low-symmetry structures with a space group number below 142, particularly below 15,

a

$Na(WO_3)_9$ - $R\bar{3}$ (No. 148)
Full formula: Na3 W27 O81
Wyckoff letter pattern:
- Na - a W - b, c, f O - e, f, f, f



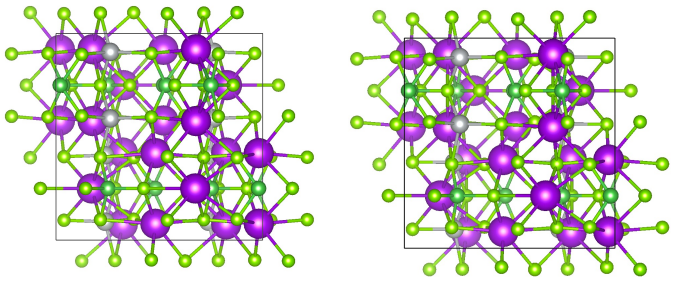
True Prediction

Wyckoff letters of No. 148

Wyckoff letter	Multiplicity	Coordinates
f	18	(x,y,z), …
e	9	(1/2,0,0), …
d	9	(1/2,0,1/2), …
c	6	(0,0,z), …
b	3	(0,0,1/2), …
a	3	(0,0,0), …

b

$K_5Ag_2(AsSe_3)_3$ - $Pnma$ (No. 62)
Full formula: K20 Ag8 As12 Se36
Wyckoff letter pattern:
- K - c, d, d Ag - d As - c, c, c Se - c, c, c, d, d, d



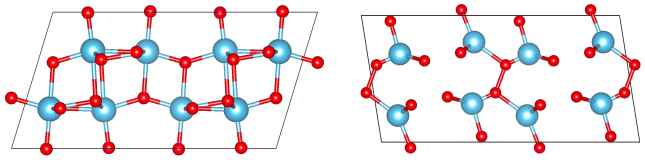
True Prediction

Wyckoff letters of No. 62

Wyckoff letter	Multiplicity	Coordinates
d	8	(x,y,z), …
c	4	(x,1/4,z), …
b	4	(0,0,1/2), …
a	4	(0,0,0), …

c

TiO_2 - $C2/m$ (No. 12)
Full formula: Ti8 O16
Wyckoff letter pattern:
- Ti - i, i Se - i, i, i, i



True Prediction

Wyckoff letters of No. 12

Wyckoff letter	Multiplicity	Coordinates
j	8	(x,y,z), …
i	4	(x,0,z), …
h	4	(0,y,1/2), …
g	4	(0,y,0), …
f	4	(1/4,1/4,1/2), …
e	4	(1/4,1/4,0), …
d	2	(0,1/2,1/2), …
c	2	(0,0,1/2), …
b	2	(0,1/2,0), …
a	2	(0,0,0), …

Figure 6. Three crystal structures predicted by the Wyckoff position random crystal structure generator: $Na(WO_3)_9$ (full formula: $Na_3W_{27}O_{81}$, number of atoms: 111, space group: $R\bar{3}$ (No. 148)), $K_5Ag_2(AsSe_3)_3$ (full formula: $K_{20}Ag_8As_{12}Se_{36}$, number of atoms: 76, space group: $Pnma$ (No. 62)) and TiO_2 (full formula: Ti_8O_{16}, number of atoms: 24, space group: $C2/m$ (No. 12)). In the case of $Na(WO_3)_9$, the number of possible Wyckoff letter combinations is limited to 48 when the space group is considered. For $K_5Ag_2(AsSe_3)_3$, despite the number of possible Wyckoff letter combinations exceeding 300, the Wyckoff letter assignment predictor reduces the search space considerably and effectively. For the case Ti_8O_{16}, where the CSP algorithm failed to predict, the high degree of freedom in the coordinate configurations prevented the generation of promising atomic coordinates despite the successful Wyckoff letter assignment.

such as orthorhombic, monoclinic, and triclinic structures. Because of their low symmetry, structures belonging to these space groups have high degrees of freedom in their coordinate configurations. Furthermore, the number of combinations of Wyckoff patterns with the same multiplicity is greater for lower-symmetry space groups. For instance, space group $C2/m$ (No. 12) has one coordinate-free Wyckoff letter $\{j\}$ with multiplicity 8, three coordinate-free Wyckoff letters $\{g,h,i\}$ and two coordinate-fixed Wyckoff letters $\{e,f\}$ with multiplicity 4, and four coordinate-fixed Wyckoff letters $\{a,b,c,d\}$ with multiplicity 2. Their possible combinations form the extensive search space of TiO_2 (full formula: Ti_8O_{16}). Despite the successful prediction of the Wyckoff letter configuration $Ti:4i, Ti:4i, O:4i, O:4i, O:4i, O:4i$ facilitated by the Wyckoff letter predictor, the generation of precise atomic coordinates remains unsuccessful (figure 6 (c)).

As illustrated in Figure 7, many of the predicted structures that were determined to be failures were metastable structures that have been reported experimentally. For example, for TiO_2[50], the anatase type is the most stable structure according to first-principles calculations, whereas the predicted structure is the rutile type, which is known to be a metastable state. The true stable structure of Si_3N_4[51] is known to be a hexagonal structure, beta-Si_3N_4, but the predicted structure is the willemite-II type, which was reported as a metastable structure by DFT calculations. In many cases, even the predicted crystal structures that were judged to be failures partially captured the structural features that are similar to the true structure. For example, the predicted structures of ZrO_2 and $LiP(HO_2)_2$ did not precisely match the true structures but differed only slightly in atomic positions, as shown in Figure 7. The energy difference between the true and predicted structures of these compounds is < 5 meV/atom. Although stable structures are often not predicted with full accuracy for low-symmetry compounds, metastable structures or partial structural patterns can be predicted using our method.

We also evaluated the predictive performance of our method in quite challenging scenarios using Dataset III, resulting in a notably low accuracy of 6.7% (see Table S1 in the Supplementary Information). Despite this low accuracy, it is worth mentioning that the rather complex crystal structures of $Al_8(Pb_3O_7)_3$ and $Mg_{10}B_{16}Ir_{19}$ could be accurately predicted. However, predicting crystal structures for more complex systems will require additional computational resources and enhanced methodologies.

Comparison with USPEX

The CSP tasks for Datasets I and II were conducted using USPEX, applying the calculation conditions outlined in the Methods section, with the assumption that the ground-truth space groups are known. The USPEX calculations were executed on the SQUID supercomputer system at Osaka University, which has two Intel Xeon Platinum 8368 CPUs with 76 cores running at 2.40 GHz at each node[52]. Each crystal calculation was allocated to one node, with the number of MPI cores set to 38 when the number of calculated atoms was less than 38, and set to 76 otherwise.

By using the settings described in the USPEX calculation inputs section in the Supplementary Information, only tasks involving small unit cell systems (containing approximately 20 atoms in the primitive unit cell) were successfully completed, with 13 and 12 finalized for Datasets I and II, respectively, within the allocated computing resources. For these completed tasks, the USPEX demonstrated a high prediction accuracy, achieving 76.9% and 75.0% for Datasets I and II, respectively. The median number of structural relaxation calculations performed was 167, with computations taking 37.7 hours on the designated supercomputer system. In comparison, our method utilizing the Wyckoff position generator for the same set of 25 benchmarks yielded an accuracy of 84.6% and 83.3% for Datasets I and II, respectively (Table S5). The median number of structural relaxation calculations performed was 172, with computations taking 21.4 hours. Note that while USPEX was conducted with the given true space group, our method searched for all 30 space groups. The enhancement in both accuracy and time efficiency stemmed from conducting structural relaxation exclusively on promising candidate structures with lower energy, guided by the machine learning energy predictor.

Table 1. Performance of the CSP algorithm with element substitution and Wyckoff position random crystal structure generators for the 40 crystals comprising Dataset I. The top 10 virtual structures with the lowest DFT energies from the two generators were proposed as the final candidate, respectively. The Number of atoms column shows the number of atoms in the primitive cell. The symbols of the ✓ and × indicate success and failure, respectively. The − denotes no template for element substitution or unfinished searching tasks. Additionally, in the Wyckoff position generation column, the symbols on either side of the slash within parentheses indicate whether the Wyckoff pattern was successfully generated and whether the space group prediction was successful, respectively, and in the element substitution column, the symbols in parentheses indicate whether a similar template structure ($\tau \leq 0.2$) was found.

Composition	Number of atoms	Space group	Wyckoff position generation	Element substitution
C	4	$R\bar{3}m$	✓ (✓ / ✓)	✓ (✓)
Si	2	$Fd\bar{3}m$	✓ (✓ / ✓)	✓ (✓)
GaAs	2	$F\bar{4}3m$	✓ (✓ / ✓)	✓ (✓)
ZnO	4	$P6_3mc$	✓ (✓ / ✓)	✓ (✓)
BN	4	$P6_3/mmc$	✓ (✓ / ✓)	✓ (✓)
$LiCoO_2$	16	$R\bar{3}m$	✓ (✓ / ✓)	✓ (✓)
Bi_2Te_3	5	$R\bar{3}m$	✓ (✓ / ✓)	✓ (✓)
$Ba(FeAs)_2$	5	$I4/mmm$	✓ (✓ / ✓)	✓ (✓)
SiO_2	6	$I\bar{4}2d$	✓ (✓ / ✓)	× (✓)
VO_2	6	$P4_2/mnm$	✓ (✓ / ✓)	✓ (✓)
La_2CuO_4	7	$I4/mmm$	× (✓ / ✓)	✓ (✓)
$LiPF_6$	8	$R\bar{3}$	✓ (✓ / ✓)	✓ (✓)
Al_2O_3	10	$R\bar{3}c$	✓ (✓ / ✓)	✓ (✓)
$SrTiO_3$	10	$I4/mcm$	✓ (✓ / ✓)	✓ (✓)
$CaCO_3$	10	$R\bar{3}c$	✓ (✓ / ✓)	✓ (✓)
TiO_2	12	$C2/m$	× (× / ✓)	× (✓)
ZrO_2	12	$P2_1/c$	✓ (✓ / ✓)	✓ (✓)
$ZrTe_5$	12	$Cmcm$	✓ (✓ / ✓)	✓ (✓)
V_2O_5	14	$Pmmn$	✓ (✓ / ✓)	× (×)
Si_3N_4	14	$P6_3/m$	✓ (✓ / ✓)	✓ (✓)
Fe_3O_4	14	$Fd\bar{3}m$	✓ (✓ / ✓)	✓ (✓)
$Mn(FeO_2)_2$	14	$Fd\bar{3}m$	✓ (✓ / ✓)	✓ (✓)
ZnSb	16	$Pbca$	✓ (✓ / ✓)	✓ (✓)
$CoSb_3$	16	$Im\bar{3}$	✓ (✓ / ✓)	✓ (✓)
$LiBF_4$	18	$P3_121$	✓ (✓ / ✓)	✓ (✓)
Y_2Co_{17}	19	$R\bar{3}m$	✓ (✓ / ✓)	✓ (✓)
GeH_4	20	$P2_12_12_1$	✓ (✓ / ✓)	✓ (✓)
$CsPbI_3$	20	$Pnma$	× (× / ✓)	✓ (✓)
$NaCaAlPHO_5F_2$	24	$P2_1/m$	× (× / ✓)	−
$LiFePO_4$	28	$Pnma$	✓ (✓ / ✓)	✓ (✓)
$Cu_{12}Sb_4S_{13}$	29	$I\bar{4}3m$	✓ (✓ / ✓)	✓ (✓)
MgB_7	32	$Imma$	× (✓ / ✓)	× (×)

Table 1 continued

Composition	Number of atoms	Space group	Wyckoff position generation	Element substitution
Li_3PS_4	32	$Pnma$	✓ (✓ / ✓)	× (×)
Cd_3As_2	80	$I4_1/acd$	✓ (✓ / ✓)	✓ (✓)
$Li_4Ti_5O_{12}$	42	$C2/c$	✓ (✓ / ✓)	✓ (✓)
$Ba_2CaSi_4(BO_7)_2$	46	$I\bar{4}2m$	× (× / ✓)	× (×)
Ag_8GeS_6	60	$Pna2_1$	× (✓ / ✓)	✓ (✓)
$Nd_2Fe_{14}B$	68	$P4_2/mnm$	× (× / ✓)	✓ (✓)
$Y_3Al_5O_{12}$	80	$Ia\bar{3}d$	✓ (✓ / ✓)	✓ (✓)
$Ca_{14}MnSb_{11}$	104	$I4_1/acd$	× (✓ / ✓)	✓ (✓)
Overall			$31/40 = 77.5\%$	$33/40 = 82.5\%$

Table 2. Performance of the CSP algorithm with element substitution and Wyckoff position random crystal structure generators for the 40 crystals comprising Dataset II. The top 10 virtual structures with the lowest DFT energies from the two generators were proposed as the final candidate, respectively. The Number of atoms column shows the number of atoms in the primitive cell. The symbols of the ✓ and × indicate success and failure, respectively. The − denotes no template for element substitution or unfinished searching tasks. Additionally, in the Wyckoff position generation column, the symbols on either side of the slash within parentheses indicate whether the Wyckoff pattern was successfully generated and whether the space group prediction was successful, respectively, and in the element substitution column, the symbols in parentheses indicate whether a similar template structure ($\tau \leq 0.2$) was found.

Composition	Number of atoms	Space group	Wyckoff position generation	Element substitution
CsCl	2	$Fm\bar{3}m$	✓ (✓ / ✓)	✓ (✓)
MnAl	2	$P4/mmm$	✓ (✓ / ✓)	✓ (✓)
HoHSe	3	$P\bar{6}m2$	✓ (✓ / ✓)	✓ (✓)
$ErCdRh_2$	4	$Fm\bar{3}m$	✓ (✓ / ✓)	✓ (✓)
Eu_2MgTl	4	$Fm\bar{3}m$	✓ (✓ / ✓)	✓ (✓)
Pm_2NiIr	4	$Fm\bar{3}m$	✓ (✓ / ✓)	✓ (✓)
VPt_3	4	$I4/mmm$	✓ (✓ / ✓)	✓ (✓)
$Gd(SiOs)_2$	5	$I4/mmm$	✓ (✓ / ✓)	✓ (✓)
$LaAl_3Au$	5	$I4mm$	✓ (✓ / ✓)	✓ (✓)
U_2SbN_2	5	$I4/mmm$	✓ (✓ / ✓)	✓ (✓)
$MnGa(CuSe_2)_2$	8	$I\bar{4}$	✓ (✓ / ✓)	✓ (✓)
SmZnPd	9	$P\bar{6}2m$	✓ (✓ / ✓)	✓ (✓)
$Sn(TePd_3)_2$	9	$I4mm$	× (× / ×)	✓ (✓)
V_5S_4	9	$I4/m$	✓ (✓ / ✓)	✓ (✓)
Cs_3InF_6	10	$Fm\bar{3}m$	✓ (✓ / ✓)	✓ (✓)
$Eu(CuSb)_2$	10	$P4/nmm$	✓ (✓ / ✓)	✓ (✓)
$Rb_2TlAgCl_6$	10	$Fm\bar{3}m$	✓ (✓ / ✓)	✓ (✓)
$Ca_3Ni_7B_2$	12	$R\bar{3}m$	✓ (✓ / ✓)	✓ (✓)

Table 2 continued

Composition	Number of atoms	Space group	Wyckoff position generation	Element substitution
$DyPO_4$	12	$I4_1/amd$	✓ (✓ / ✓)	✓ (✓)
LaSiIr	12	$P2_13$	✓ (✓ / ✓)	✓ (✓)
$SmVO_4$	12	$I4_1/amd$	✓ (✓ / ✓)	✓ (✓)
VCl_5	12	$P\bar{1}$	× (✓ / ✓)	✓ (✓)
YbP_5	12	$P2_1/m$	× (✓ / ✓)	✓ (✓)
$Eu(Al_2Cu)_4$	13	$I4/mmm$	✓ (✓ / ✓)	✓ (✓)
Zr_4O	15	$R\bar{3}$	× (× / ×)	× (×)
$K_2Ni_3S_4$	18	$Fddd$	✓ (✓ / ✓)	✓ (✓)
$Sr(ClO_3)_2$	18	$Fdd2$	✓ (✓ / ✓)	✓ (✓)
$LiSm_2IrO_6$	20	$P2_1/c$	× (✓ / ✓)	✓ (✓)
Pr_2ZnPtO_6	20	$P2_1/c$	× (✓ / ✓)	✓ (✓)
$Sc_2Mn_{12}P_7$	21	$P\bar{6}$	✓ (✓ / ✓)	✓ (✓)
$LaSi_2Ni_9$	24	$I4_1$/amd	✓ (✓ / ✓)	✓ (✓)
$CeCu_5Sn$	28	$Pnma$	✓ (✓ / ✓)	✓ (✓)
$LiP(HO_2)_2$	32	$Pna2_1$	× (✓ / ✓)	✓ (×)
$Mg_3Si_2H_4O_9$	36	$P6_3cm$	× (× / ×)	× (×)
$Y_4Si_5Ir_9$	36	$P6_3/mmc$	✓ (✓ / ✓)	–
$Na(WO_3)_9$	37	$R\bar{3}$	✓ (✓ / ✓)	–
$Sm_6Ni_{20}As_{13}$	39	$P\bar{6}$	✓ (✓ / ✓)	✓ (✓)
$BaCaGaF_7$	40	$P2/c$	× (✓ / ✓)	✓ (✓)
$Tm_{11}Sn_{10}$	42	$I4/mmm$	✓ (✓ / ✓)	✓ (✓)
$AlH_{12}(ClO_2)_3$	44	$R\bar{3}c$	× (× / ×)	✓ (✓)
$K_2ZrSi_2O_7$	48	$P2_1/c$	✓ (✓ / ✓)	✓ (×)
$Ba_3Ta_2NiO_9$	60	$P\bar{3}m1$	✓ (✓ / ✓)	× (×)
$LiZr_2(PO_4)_3$	72	$P2_1/c$	× (× / ✓)	✓ (✓)
$K_5Ag_2(AsSe_3)_3$	76	$Pnma$	✓ (✓ / ✓)	–
$Be_{17}Ru_3$	80	$Im\bar{3}$	× (× / ✓)	✓ (✓)
$Cu_3P_8(S_2Cl)_3$	80	$Pnma$	× (× / ✓)	✓ (✓)
Al_2CoO_4	84	$P3m1$	× (× / ×)	✓ (✓)
$Li_6V_3P_8O_{29}$	92	$P1$	× (✓ / ✓)	✓ (×)
$ReBi_3O_8$	96	$P2_13$	✓ (✓ / ✓)	✓ (✓)
$Na_5FeP_2(O_4F)_2$	288	$Pbca$	× (× / ✓)	–
Overall			$34/50 = 68.0\%$	$43/50 = 86.0\%$

Discussion

This paper presents a machine-learning workflow for the efficient prediction of stable crystal structures with no iterative calculations. The essence of the proposed method is the shotgun-type virtual screening of crystal structures, in which a surrogate model that predicts DFT energy is simply used to screen a large number of virtual crystal structures, and

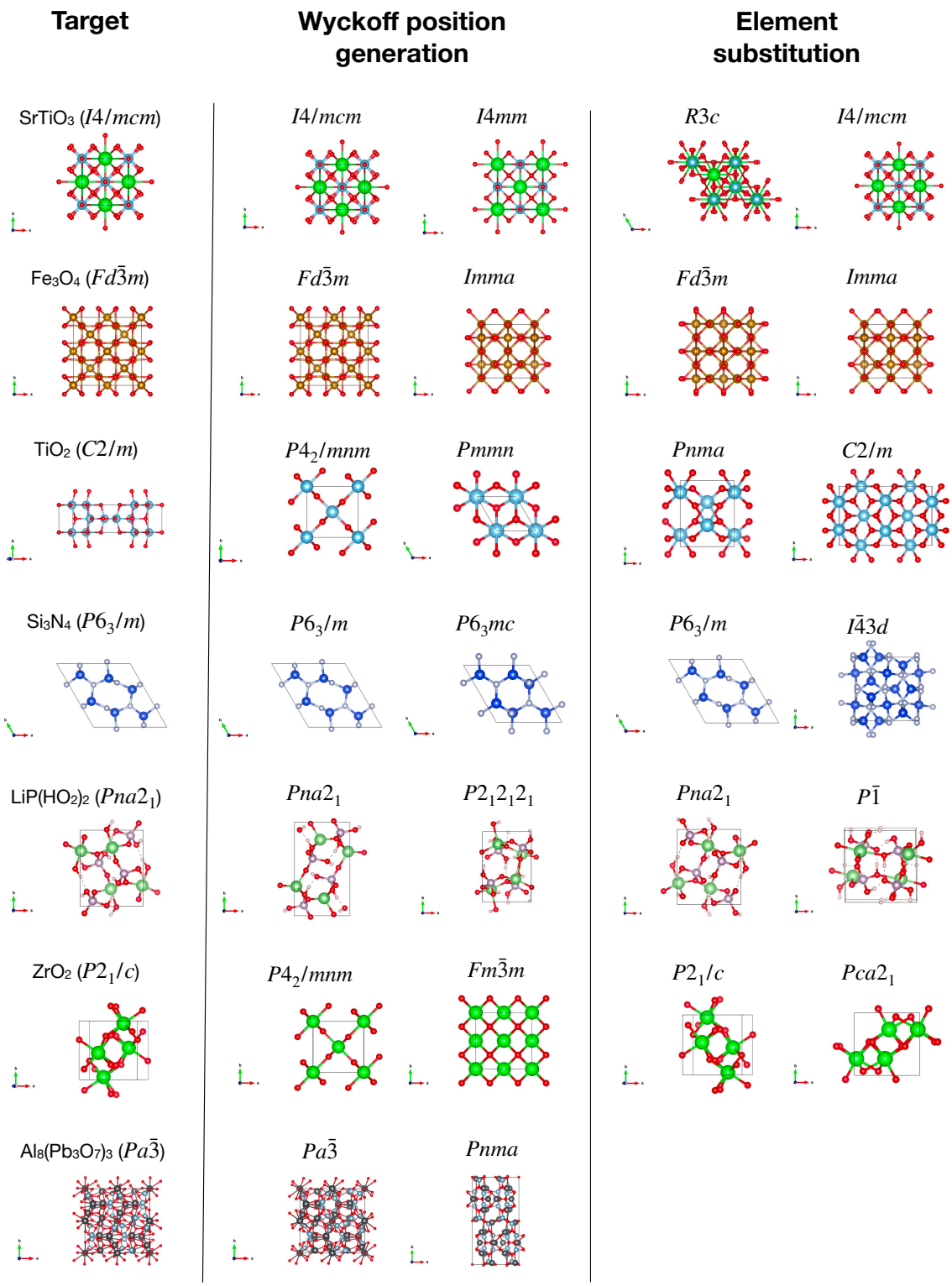


Figure 7. Examples of crystal structures predicted by the proposed CSP algorithm (depicted with VESTA[53] version 3.5.8). For each method (using the Wyckoff position generator or randomized element substitution), the predicted structures with the two lowest DFT energies are shown with the true stable structures.

the efficiently narrowed-down candidate structures are then relaxed by DFT calculations to predict the stable crystal structures. The technical components that play key roles in this workflow are the surrogate model for energy prediction and the crystal structure generators. To train the surrogate model for DFT energy calculations, transfer learning of the pretrained CGCNN was performed to decrease the number of training samples generated with DFT single-point calculations. To create virtual libraries of candidate crystal structures, a generator based on elemental substitutions in template crystals and a generator based on random combinations of Wyckoff positions were developed and tested. Of the 90 known crystal structures (Dataset I and II) with a wide range of chemical compositions, symmetries, and structure types, the workflows using the generator based on element substitution and the Wyckoff position generator predicted 84.4% and 74.4% of the true structures, respectively. For the 25 USPEX tested benchmarks, our Wyckoff position generator-based approach outperformed the prediction accuracy of USPEX by around 10%, which is the most widely used crystal structure prediction software that showed a prediction accuracy of 76.0%. Existing recursive algorithms that iterate over DFT calculations have not been thoroughly and comprehensively evaluated in terms of their prediction performance owing to their exceedingly high computational cost. To the best of our knowledge, our method is the simplest crystal structure prediction algorithm currently available. One of the contributions of this study is that this simple approach can efficiently solve many crystal structures that cannot be predicted using conventional methods, such as lower-symmetry structures and huge unit cell systems.

The comparison of the element-substitution-based generator and the Wyckoff position generator showed that the prediction performance of the former was significantly better than that of the latter. Given a known template structure, the element-substitution-based generator was shown to be able to predict the stable structure with almost perfect accuracy, and approximately 98.0% of stable or metastable crystals in the Materials Project are known to have a counterpart that can act as a template structure. Therefore, it is estimated that many crystal systems can be predicted using element-substitution-based crystal-structure predictions. By contrast, while, in principle, virtual screening based on the Wyckoff position generator can predict new crystal structures, its prediction accuracy in benchmarking was not very high. The bottlenecks of this approach are the prediction accuracy of the space group and the limited ability of the Wyckoff position generator. With the current workflow, the upper bound of the accuracy of the top 40 space group predictors was approximately 94%. Even if the space group can be appropriately identified, the number of possible combinations of Wyckoff positions becomes too large for large systems with a large number of atoms, making it significantly difficult to generate true structures.

Methods

Template-based structure generation by elemental substitution

The calculation procedure is as follows:

1. Extract a template structure with the same composition ratio as the query composition X from 33,040 stable structures in the Materials Project database
2. Replace elements in the template with elements that have the same number of atoms in X. If the substitution target is not uniquely determined, substitute the element with the smallest Euclidean distance in XenonPy's 58-dimensional element descriptors.
3. Convert the chemical compositions of the template structures to the 290-dimensional descriptors in XenonPy, and apply DBSCAN clustering to group the template structures
4. Extract template sets that belong to the same group as X. Furthermore, using the StructureMatcher module of pymatgen, remove structurally redundant templates to obtain a unique template set (the number of templates is K_{temp}).
5. Estimate the lattice constant using a model that predicts the volume from the composition X.
6. Add perturbations to the atomic coordinates of each template following the uniform distribution $U(-0.05, 0.05)$.
7. Add perturbations to the volume of the unit cell of each template following the uniform distribution $U(-0.1, 0.1)$.

Wyckoff position generator

The calculation procedure is as follows:

1. Predict the space group and the probabilities of Wyckoff letters of the query composition $X = X_{c_1}^1 X_{c_2}^2 \cdots X_{c_K}^K$ under each predicted space group.
2. Extract the set $W = \{(l_i, m_i) | i = 1, \ldots, j\}$ of Wyckoff letters l_i and multiplicity m_i for a predicted space group.
3. Randomly sample an element c from composition X and a possible Wyckoff letter l with the predicted probability p of composition X from set W, then assign l to c with its multiplicity m.
4. Remove the assigned atoms from composition X and define the remaining composition as new X.
5. Remove the used Wyckoff letter l from the set W if the Wyckoff position is exclusive and re-normalize the probability of remaining Wyckoff letters.

$$\tilde{p}_i = p_i / \sum p_i$$

6. Do 3–5 until all atoms are assigned.
7. Determine the fractional coordinate of atomic sites to which the same Wyckoff letter is assigned. If the Wyckoff position coordinate (x, y, z) of the atomic sites are allowed to vary, each coordinate position is sampled from a uniform distribution $U(0, 1)$.
8. Estimate the lattice constant using a model that predicts the volume from the composition X. Add perturbations to the volume of the unit cell of each template following the uniform distribution $U(-0.1, 0.1)$.

Fine-tuning of CGCNN

To obtain a CGCNN localized to the energy prediction of a specific system with composition X, the pretrained global CGCNN model from Xie & Grossman[30] was fine-tuned on a randomly generated conformations and their formation energies from each CSP algorithm. We generated 100 training crystal structures for each candidate space group given by the space group predictor in the Wyckoff position generation approach or 10 training crystal structures for each selected template in the element substitution approach. The pretrained model was copied to the target model, except for the output layer. Subsequently, a new output layer was added to the target model, and its parameters were randomly initialized. We then trained the target model on the target dataset. The hyperparameters learning rate and gradients clipping value were optimized by performing a grid search with range $\{0.01, 0.008, 0.006, 0.004, 0.002\}$, respectively, with the early stopping of the MAE of the validation set. The number of epochs was fixed to 350.

DFT calculation

All DFT calculations were performed using the Vienna Ab initio Simulation Package (VASP, version 6.1.2)[54] with projector-augmented wave pseudopotentials[49]. The exchange-correlation functional was considered using the Perdew–Burke–Ernzerhof formulation[55] of the generalized gradient approximation. The Brillouin zone integration for the unit cells was automatically determined using the Γ-centered Monkhorst–Pack mesh function implemented in the VASP code. Single-point calculations (also called self-consistent field calculations) were performed on unrelaxed crystal structures that were virtually created to produce a training set for fine-tuning the pretrained CGCNN. The geometry of the final selected candidate structure was locally optimized by performing DFT calculations. We used the *MPStaticSet* and *MPRelaxSet* presets implemented in pymatgen[56], with significant modifications to generate the inputs for all VASP calculations (See VASP calculation inputs section in the Supplementary Information).

Structural similarity

To calculate the similarity between two structures i and j, we encoded the given structures into a vector-type structural descriptor with their local coordination information (site fingerprint) from all sites[57]. Then, the structure similarity τ was calculated as the Euclidean distance between the crystal structure descriptors. Note that the criterion contains no elemental composition information and gets insensitivity for the low-symmetry structures. The calculations were performed by Matminer[58], which is an open-source toolkit for materials data mining, with the same configuration as the Materials Project officially used. We visually inspected the difference between structures for different τ. In this study, structures with dissimilarity $\tau \leq 0.2$ were treated as similar structures.

USPEX calculation

The USPEX calculations were performed using the official UPSEX package (version 10.5)[59]. We specified the calculation parameters `calculationMethod`, `calculationMethod`, and `calculationMethod` as "USPEX", "300", and "enthalpy", respectively, to perform the crystal structure prediction task for bulk crystal using the evolutionary algorithm (EA). The EA-related parameters were specified according to official recommendations[60]. For example, the number of structures in each generation was set to $2 \times N$ rounded to the nearest ten, where N is the number of atoms. The calculation was terminated if the best structure did not change over M generations, where $M =$ round N to the nearest ten (See the USPEX calculation inputs section in the Supplementary Information). Structural relaxation was executed automatically using the VASP (version 6.1.2) package combined with the projector-augmented wave pseudopotentials. The VASP calculation settings were the same as those described in the DFT calculation section.

Compositional descriptor

The chemical formula is $X = X_{c^1}^1 X_{c^2}^2 \cdots X_{c^K}^K$ where X_k denotes a chemical element and c_k is its composition ratio. Each element of the descriptor vector of length 290 takes the following form:

$$\phi_{g,\eta}(X) = g(c^1, \ldots, c^K, \eta(X^1), \ldots, \eta(X^K)). \tag{2}$$

The scalar quantity $\eta(X^k)$ on the right-hand side represents a feature value of the element X^k, such as the atomic weight, electronegativity, and polarizability. Using function g, the element features $\eta(X^1), \ldots, \eta(X^K)$ with compositions $c^1, \ldots, c^K$ are converted into compositional features. For g, we use five different summary statistics: weighted mean, weighted variance, weighted sum, max-pooling, and min-pooling as given by

$$\begin{aligned}
\phi_{\mathrm{ave},\eta}(S) &= \frac{1}{\sum_{k=1}^K c^k} \sum_{k=1}^K c^k \eta(S^k), \\
\phi_{\mathrm{var},\eta}(S) &= \frac{1}{\sum_{k=1}^K c^k} \sum_{k=1}^K c^k (\eta(S^k) - \phi_{\mathrm{ave},\eta}(S))^2, \\
\phi_{\mathrm{sum},\eta}(S) &= \sum_{k=1}^K c^k \eta(S^k), \\
\phi_{\mathrm{max},\eta}(S) &= \max\{\eta(S^1), \ldots, \eta(S^K)\}, \\
\phi_{\mathrm{min},\eta}(S) &= \min\{\eta(S^1), \ldots, \eta(S^K)\}.
\end{aligned}$$

We used 58 distinct elemental features implemented in XenonPy. The full list of the 58 features is given in Liu et al.[44], including the atomic number, covalent radius, van der Waals radius, electronegativity, thermal conductivity, band gap, polarizability, boiling point, melting point. In summary, composition X is characterized by a 290-dimensional descriptor vector ($= 58 \times 5$).

Data availability

The data supporting this study's findings are available from the corresponding authors upon reasonable request.

Code availability

The codes of element substitution-based CSP are available from the GitHub website (https://github.com/yoshida-lab/XenonPy/blob/master/samples/CSP_with_element_substitution.ipynb).

Acknowledgements

This work was supported in part by a MEXT KAKENHI Grant-in-Aid for Scientific Research on Innovative Areas (Grant Number 19H05820), a JSPS Grant-in-Aid for Scientific Research (A) 19H01132 and a JSPS Grant-in-Aid for Early-Career Scientists 20K19866 from the Japan Society for the Promotion of Science (JSPS), and JST CREST Grant Number JPMJCR19I3.

Author Contributions

R.Y. and H.T. designed and conceived the project, and R.Y. wrote the preliminary draft of the paper. R.Y. and C.L. designed and developed the machine learning framework. C.L. developed the software and performed the experiments with the support of H.T., H.T., T.Y., K.W., and S.Y., who designed and tested the benchmark crystal structures. C.L., H.T., T.Y., and R.T. wrote and revised the manuscript. All authors discussed the results and commented on the manuscript.

Additional Information

Conflicts of Interest: The authors declare no competing interests.

References

1. Martoňák, R., Laio, A. & Parrinello, M. Predicting crystal structures: the parrinello-rahman method revisited. *Phys. Rev. Lett.* **90**, 075503 (2003).
2. Oganov, A. R. & Glass, C. W. Crystal structure prediction using *ab initio* evolutionary techniques: Principles and applications. *J. Chem. Phys.* **124**, 244704, DOI: 10.1063/1.2210932 (2006).
3. Pickard, C. J. & Needs, R. J. High-Pressure Phases of Silane. *Phys. Rev. Lett.* **97**, 045504, DOI: 10.1103/PhysRevLett.97.045504 (2006).
4. Pickard, C. J. & Needs, R. J. Structure of phase III of solid hydrogen. *Nat. Phys.* **3**, 473–476, DOI: 10.1038/nphys625 (2007).
5. Pickard, C. J. & Needs, R. J. *Ab Initio* random structure searching. *J. Phys. Condens. Matter* **23**, 053201 (2011).
6. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. Optimization by Simulated Annealing. *Science* **220**, 671–680, DOI: 10.1126/science.220.4598.671 (1983).
7. Pannetier, J., Bassas-Alsina, J., Rodriguez-Carvajal, J. & Caignaert, V. Prediction of crystal structures from crystal chemistry rules by simulated annealing. *Nature* **346**, 343–345, DOI: 10.1038/346343a0 (1990).
8. Wang, F. & Landau, D. P. Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States. *Phys. Rev. Lett.* **86**, 2050–2053, DOI: 10.1103/PhysRevLett.86.2050 (2001).
9. Wang, Y., Lv, J., Zhu, L. & Ma, Y. Crystal structure prediction via particle-swarm optimization. *Phys. Rev. B* **82**, 094116, DOI: 10.1103/PhysRevB.82.094116 (2010).
10. Zhang, Y., Wang, H., Wang, Y., Zhang, L. & Ma, Y. Computer-Assisted Inverse Design of Inorganic Electrides. *Phys. Rev. X* **7**, 011017, DOI: 10.1103/PhysRevX.7.011017 (2017).
11. Oganov, A. R., Lyakhov, A. O. & Valle, M. How Evolutionary Crystal Structure Prediction Works—and Why. *Acc. Chem. Res.* **44**, 227–237, DOI: 10.1021/ar1001318 (2011).
12. 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, DOI: 10.1016/j.cpc.2012.12.009 (2013).
13. Yamashita, T. *et al.* Crystal structure prediction accelerated by Bayesian optimization. *Phys. Rev. Mater.* **2**, 013803, DOI: 10.1103/PhysRevMaterials.2.013803 (2018).
14. Terayama, K., Yamashita, T., Oguchi, T. & Tsuda, K. Fine-grained optimization method for crystal structure prediction. *npj Comput. Mater.* **4**, 32, DOI: 10.1038/s41524-018-0090-y (2018).
15. Jacobsen, T. L., Jørgensen, M. S. & Hammer, B. On-the-Fly Machine Learning of Atomic Potential in Density Functional Theory Structure Optimization. *Phys. Rev. Lett.* **120**, 026102, DOI: 10.1103/PhysRevLett.120.026102 (2018).
16. 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, DOI: 10.1103/PhysRevB.99.064114 (2019).
17. Takamoto, S. *et al.* Towards universal neural network potential for material discovery applicable to arbitrary combination of 45 elements. *Nat. Commun.* **13**, 2991 (2022).
18. Wang, Y., Lv, J., Zhu, L. & Ma, Y. CALYPSO: A method for crystal structure prediction. *Comput. Phys. Commun.* **183**, 2063–2070, DOI: 10.1016/j.cpc.2012.05.008 (2012).
19. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning (MIT Press, Cambridge, Mass, 2006).
20. Mockus, J. *Bayesian Approach to Global Optimization: Theory and Applications*. Mathematics and Its Applications. Soviet Series (Kluwer Academic, Dordrecht ; Boston, 1989).
21. Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002, DOI: 10.1063/1.4812323 (2013).

22. The Materials Project. https://materialsproject.org. Accessed: 2023-11-06.

23. Curtarolo, S. *et al.* AFLOW: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226, DOI: 10.1016/j.commatsci.2012.02.005 (2012).

24. AFLOW: Atomic - Flow for Materials Discovery. http://www.aflowlib.org. Accessed: 2023-11-06.

25. Kirklin, S. *et al.* The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **1**, 15010, DOI: 10.1038/npjcompumats.2015.10 (2015).

26. OQMD: The Open Quantum Materials Database. http://oqmd.org. Accessed: 2023-11-06.

27. Merchant, A. *et al.* Scaling deep learning for materials discovery. *Nature* 1–6 (2023).

28. Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **31**, 3564–3572 (2019).

29. Choudhary, K. & DeCost, B. Atomistic line graph neural network for improved materials property predictions. *npj Comput. Mater.* **7**, 1–8 (2021).

30. Xie, T. & Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **120**, 145301, DOI: 10.1103/PhysRevLett.120.145301 (2018).

31. Gibson, J., Hire, A. & Hennig, R. G. Data-augmentation for graph neural network learning of the relaxed energies of unrelaxed structures. *npj Comput. Mater.* **8**, 211 (2022).

32. Weiss, K., Khoshgoftaar, T. M. & Wang, D. A survey of transfer learning. *J. Big Data* **3**, 9, DOI: 10.1186/s40537-016-0043-6 (2016).

33. Yamada, H. *et al.* Predicting Materials Properties with Little Data Using Shotgun Transfer Learning. *ACS Cent. Sci.* **5**, 1717–1730, DOI: 10.1021/acscentsci.9b00804 (2019).

34. Hautier, G., Fischer, C., Ehrlacher, V., Jain, A. & Ceder, G. Data Mined Ionic Substitutions for the Discovery of New Compounds. *Inorg. Chem.* **50**, 656–663, DOI: 10.1021/ic102031h (2011).

35. Wang, H.-C., Botti, S. & Marques, M. A. L. Predicting stable crystalline compounds using chemical similarity. *npj Comput. Mater.* **7**, 12, DOI: 10.1038/s41524-020-00481-6 (2021).

36. Kusaba, M., Liu, C. & Yoshida, R. Crystal structure prediction with machine learning-based element substitution. *Comput. Mater. Sci.* **211**, 111496 (2022).

37. Bushlanov, P. V., Blatov, V. A. & Oganov, A. R. Topology-based crystal structure generator. *Comput. Phys. Commun.* **236**, 1–7, DOI: 10.1016/j.cpc.2018.09.016 (2019).

38. Fredericks, S., Parrish, K., Sayre, D. & Zhu, Q. PyXtal: A Python library for crystal structure generation and symmetry analysis. *Comput. Phys. Commun.* **261**, 107810, DOI: 10.1016/j.cpc.2020.107810 (2021).

39. Hu, J. *et al.* Contact map based crystal structure prediction using global optimization. *CrystEngComm* **23**, 1765–1776, DOI: 10.1039/d0ce01714k (2021).

40. Kim, B., Lee, S. & Kim, J. Inverse design of porous materials using artificial neural networks. *Sci. Adv.* **6**, eaax9324, DOI: 10.1126/sciadv.aax9324 (2020).

41. Zhu, Q., Oganov, A. R., Glass, C. W. & Stokes, H. T. Constrained evolutionary algorithm for structure prediction of molecular crystals: Methodology and applications. *Acta Crystallogr. B: Struct. Sci.* **68**, 215–226, DOI: 10.1107/S0108768112017466 (2012).

42. Lee, I.-H. & Chang, K. Crystal structure prediction in a continuous representative space. *Comput. Phys. Commun.* **194**, 110436, DOI: 10.1016/j.commatsci.2021.110436 (2021).

43. Wu, S., Lambard, G., Liu, C., Yamada, H. & Yoshida, R. iQSPR in XenonPy: A Bayesian Molecular Design Algorithm. *Mol. Inform.* **39**, 1900107, DOI: 10.1002/minf.201900107 (2020).

44. Liu, C. *et al.* Machine Learning to Predict Quasicrystals from Chemical Compositions. *Adv. Mater.* **33**, 2102507, DOI: 10.1002/adma.202102507 (2021).

45. Xenonpy platform. https://github.com/yoshida-lab/XenonPy. Accessed: 2023-11-06.

46. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, 226–231 (AAAI Press, Portland, Oregon, 1996).

47. Schubert, E., Sander, J., Ester, M., Kriegel, H. P. & Xu, X. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Trans. Database Syst.* **42**, 1–21, DOI: 10.1145/3068335 (2017).

48. Kresse, G. & Furthmüller, J. Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186, DOI: 10.1103/PhysRevB.54.11169 (1996).

49. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979, DOI: 10.1103/PhysRevB.50.17953 (1994).

50. Cui, Z.-H., Wu, F. & Jiang, H. First-principles study of relative stability of rutile and anatase TiO_2 using the random phase approximation. *Phys. Chem. Chem. Phys.* **18**, 29914–29922, DOI: 10.1039/C6CP04973G (2016).

51. Kroll, P. Pathways to metastable nitride structures. *J. Solid State Chem.* **176**, 530–537, DOI: 10.1016/S0022-4596(03)00300-1 (2003).

52. Squid (supercomputer for quest to unsolved interdisciplinary datascience). http://www.hpc.cmc.osaka-u.ac.jp/en/squid/. Accessed: 2023-11-06.

53. Momma, K. & Izumi, F. *VESTA3* for three-dimensional visualization of crystal, volumetric and morphology data. *J. Appl. Cryst.* **44**, 1272–1276, DOI: 10.1107/S0021889811038970 (2011).

54. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169 (1996).

55. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).

56. Ong, S. P. *et al.* Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).

57. Zimmermann, N. E. R., Horton, M. K., Jain, A. & Haranczyk, M. Assessing local structure motifs using order parameters for motif recognition, Interstitial Identification, and diffusion path characterization. **4**, 34, DOI: 10.3389/fmats.2017.00034.

58. Ward, L. *et al.* Matminer: An open source toolkit for materials data mining. **152**, 60–69, DOI: 10.1016/j.commatsci.2018.05.018.

59. Uspex downloads. https://uspex-team.org/en/uspex/downloads. Accessed: 2023-11-06.

60. Uspex manual. https://uspex-team.org/online_utilities/tmp/uspex_manual_release/EnglishVersion/uspex_manual_english/index.html. Accessed: 2023-11-06.

Supplementary Information

Shotgun crystal structure prediction using machine-learned formation energies

Chang Liu[1], Hiromasa Tamaki[2], Tomoyasu Yokoyama[2], Kensuke Wakasugi[2], Satoshi Yotsuhashi[2], Minoru Kusaba[1], and Ryo Yoshida[1,3,4,*]

[1]The Institute of Statistical Mathematics, Research Organization of Information and Systems, Tachikawa, Tokyo 190-8562, Japan
[2]Technology Division, Panasonic Holdings Corporation, Kadoma, Osaka 571-8508, Japan
[3]National Institute for Materials Science, Research and Service Division of Materials Data and Integrated System, Tsukuba, 305-0047, Japan
[4]The Graduate University for Advanced Studies, Department of Statistical Science, Tachikawa, 190-8562, Japan
[*]yoshidar@ism.ac.jp

Contents

List of Tables

List of Figures

Benchmark datasets

Dataset I consists of 40 stable crystal structures selected based on the diversity of the number of atoms in the unit cells, space groups, etc. Dataset II consists of 50 stable materials randomly selected from the Materials Project database. The chemical compositions and space groups are listed in Tables 1 and 2 in the main text. The crystal structures are visualized in Figure S5. Figure S1 shows histograms of the number of atoms in the crystal structures of Datasets I and II. The ranges of the number of atoms in Datasets I and II are [2, 104] and [2, 288], respectively. The average number of atoms $\pm$ standard deviation is 23.13 ± 24.09 and 32.68 ± 45.41, for Datasets I and II, respectively. In addition, as Dataset III, 30 stable structures for which no template exists were randomly selected from the Materials Project. As shown in Table S1, most of the crystal structures in Dataset III have a much larger number of atoms in the unit cell than in Dataset I and II, distributed in the rage [22, 152] with the average atomic number of 66.50 ± 34.40.

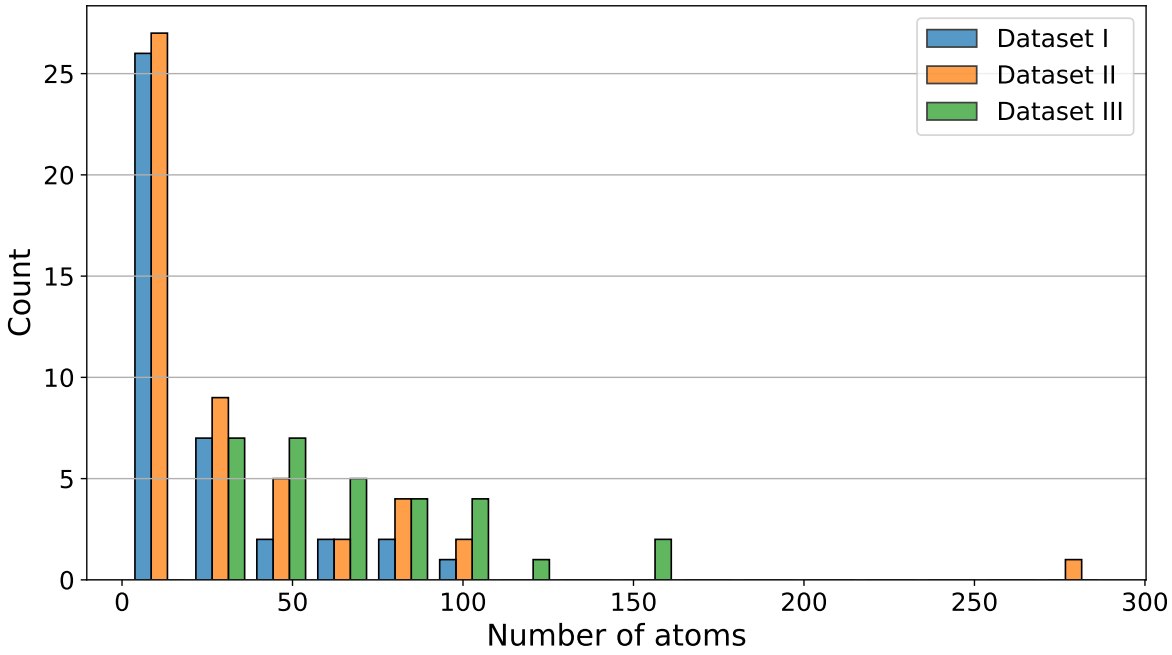


Figure S1. Histograms of the number of atoms in unit cells in Datasets I, II and III.

Energy predictions

We have employed transfer learning to fine-tune the crystal graph convolutional neural network (CGCNN) model, which was introduced by Xie et al.[2] for the prediction of formation energies of virtually generated crystal structures. The CGCNN model was first pre-trained on a substantial dataset comprising 126,210 crystal structures extracted from the Materials Project database. The 120 benchmark crystal structures were deliberately excluded from the training data. For a given composition, single-point DFT calculations were performed on 1,000 generated virtual structures to evaluate their energies. The energies were then used to fine-tune the composition-specific model by transfer learning.

All these calculations were executed on the supercomputer AI Bridging Cloud Infrastructure (ABCI)[3] utilizing 20 cores of an Intel Xeon Gold 6148 CPU per job. Table S2 provides a comprehensive summary of the computational cost associated with single-point DFT calculations for the 40 cases in Dataset I, showing the average, standard deviation, minimum, 25th, 50th, 75th percentiles, and maximum of computational time measured in seconds. In most cases, for example, composition with less than 30 atoms in the unit cell, calculations were finished in less than 2 minutes. Other compositions with large unit cells, such as $Y_3Al_5O_{12}$ and $Ca_{14}MnSb_{11}$, need more than 30 minutes to run.

Table S1. Performance of the CSP algorithm with Wyckoff position random crystal structure generators for 30 randomly selected stable crystals from the Materials Project for which no template exists that comprise Dataset III. The top 10 virtual structures with the lowest DFT energies were proposed as the final candidate. The ✓ and × symbols indicate success and failure, respectively. In the Wyckoff position generation column, the symbols on either side of the slash within parentheses indicate whether the Wyckoff pattern was successfully generated and whether the space group prediction was successful, respectively.

Composition	Number of atoms	Space group	Wyckoff position generation
$Ba_3Si_6N_4O_9$	22	$P3$	× (× / ×)
$Ho_{10}Te_7S_{10}$	27	$P1$	× (× / ✓)
$Ti_{13}Al_9Co_8$	30	$R3m$	× (× / ✓)
$K_{11}LiMn_4O_{16}$	32	$I\bar{4}2m$	× (× / ✓)
$BaSrMn_2Al_9PbO_{20}$	34	$P1$	× (× / ✓)
$RbNa_3Li_{12}Ti_4O_{16}$	36	$I4/m$	× (× / ×)
$B_{10}(Pb_2O_7)_3$	37	$P\bar{1}$	× (× / ✓)
$Nb_{12}Br_{17}F_{13}$	42	$P1$	× (× / ✓)
$Na_4PuH_7O_9$	42	$P\bar{1}$	× (× / ✓)
$RbLa_2C_6N_6ClO_6$	44	$P6_3/m$	× (× / ✓)
$Mg_{10}B_{16}Ir_{19}$	45	$I\bar{4}3m$	✓ (✓ / ✓)
$LiMn_3Al_2(HO_2)_6$	48	$P\bar{1}$	× (× / ✓)
$Na_8Al_6Si_6CO_{27}$	48	$R3$	× (× / ×)
$Sr_{16}V_8O_{31}$	55	$P\bar{1}$	× (× / ✓)
$Nd_{14}Zn_{43}Sn_3$	60	Pm	× (× / ✓)
$H_{18}SO_{12}$	62	Cc	× (× / ✓)
$Cs_{10}(Mo_2N_5)_3$	62	$R\bar{3}c$	× (× / ✓)
$La_{17}Al_4(Si_3N_{11})_3$	63	$F\bar{4}3m$	× (× / ×)
$Ba_2V_5(PO_6)_4$	70	Cm	× (× / ×)
$Ba_{10}Li_2Bi_4O_{21}$	74	$Pmc2_1$	× (× / ✓)
$Bi_8AsAuCl_9$	76	$P2_1/c$	× (× / ✓)
$Ba_8Nb_7S_{24}$	78	$P2_1/c$	× (× / ✓)
$KBa_6Zn_4(GaO_3)_7$	78	$P1$	× (× / ×)
$Al_2Si_3H_8(NO_5)_2$	100	$P2_1$	× (× / ✓)
$H_{13}S_2N_3O_8$	104	$P2/c$	× (× / ×)
$KNa_3Al_{12}H_{24}(SO_7)_8$	104	$P2/m$	× (× / ×)
$Er_6Al_{41}Cr_6$	106	$P\bar{3}1m$	× (× / ✓)
$K_3Fe_3P_4H_2O_{17}$	116	$Pnna$	× (× / ✓)
$Ta_4P_4S_{29}$	148	$P4_12_12$	× (× / ×)
$Al_8(Pb_3O_7)_3$	152	$Pa\bar{3}$	✓ (✓ / ✓)
Overall			$2/30 = 6.7\%$

Space group predictor

Using 33,040 stable crystal structures obtained from the Materials Project database and their space groups, we constructed a fully connected neural network model that classifies any composition into one of the 212 space groups.

All samples of the 120 benchmark crystal structures were removed from the dataset. The split ratios of the training and test sets were 80% and 20%, respectively. 5-fold cross-validation was looped with the training set to identify the hyperparameter pairs with the highest average prediction accuracy. Using the validation set, the hyperparameters of the models were selected by Bayesian optimization using the Optuna Python library (the number of trials was set to 200)[4] from the candidate set, the number of layers = $\{2,3,4\}$, the dropout ratio = $\{0,0.1,0.2\}$, and the number of neurons for each layer was set to $N \times \kappa$. Here, N denotes the number of neurons in the previous layer. $\kappa = 0.8 \sim 0.95$ is the hyperparameter. Finally, we calculated the predictive performance on the test set of the neural network trained using the hyperparameters selected by optuna. Random data partitioning was repeated 100 times independently to examine the mean and variance of prediction accuracy. The top 10 precision, recall, and F_1 scores are $0.8535(\pm 0.0053)$, $0.8535(\pm 0.0054)$, and $0.8535(\pm 0.0054)$, respectively. Figure S2 shows the recall rate versus the number of training data points for each space group. The top 10 recall rates varied widely from $0.0000(\pm 0.0000)$ to $0.9922(\pm 0.0031)$ among the different space groups. The variability of the recall rates was partially correlated with the number of training instances in each space group.

Here, there is a concern that if there is a significant bias in the pattern of composition ratios in the dataset, the prediction performance on the test instances having many identical composition ratios in the training set tends to be higher. We denote by S the upper bound on the number of samples with the same composition ratio in the training dataset. To address our concern, we evaluated the sensitivity in the prediction accuracy of the trained models when the upper bound on the number of samples, $S \in \{10,50,100\}$, was varied. As shown in Figure S3, it was confirmed that the prediction performance did not vary significantly.

Prediction of unit cell volumes

Using 33,040 stable crystal structures from the Materials Project database and their cell volumes, we constructed a fully connected neural network model for the prediction of the unit cell volume for any composition. All samples of the 120 benchmark crystal structures were removed from the dataset. The split ratios of the training and test sets were 80% and 20%, respectively. 5-fold cross-validation was performed with the training set to identify the hyperparameter pairs with the highest average prediction accuracy. Using the validation set, the hyperparameters of the models were selected by Bayesian optimization using the Optuna Python library (the number of trials was set to 200) from the candidate set, number of layers = $\{2,3,4\}$, dropout ratio = $\{0,0.1,0.2\}$, and the number of neurons in each layer was set to $N \times \kappa$. Here, N denotes the number of neurons in the previous layer. $\kappa = 0.8 \sim 0.95$ is the hyperparameter. Finally, we calculate the predictive performance of the test set of the MLP model trained on the training set using the hyperparameters selected by Optuna. The mean absolute error (MAE), root mean square error (RMSE), and R^2 of the predictions for the test set were 53.048 ($\pm$3.177) eV, 90.362 ($\pm$6.007) eV, and 0.973 ($\pm$0.004) eV, respectively. Performance metrics were averaged over 100 bootstrap sets.

VASP calculation inputs

In the DFT calculation workflow, the generation of VASP "INCAR" files was executed through a Python script, leveraging the *MPStaticSet* and *MPRelaxSet* presets from pymatgen[5], with modifications to suit our specific research needs. The script is used to iterate through a dataset of structures, implementing a series of relaxation steps, each uniquely parameterized to optimize the computational process and the accuracy of results.

Step 1: Initial Relaxation

- Utilizes *MPRelaxSet* with lower accuracy for coarse relaxation.
- Modified parameters include:
  - `ALGO=Fast` for quick electronic minimization.
  - `EDIFF=1e-2`, `EDIFFG=1e-1` for looser convergence criteria.
  - `ISIF=4` to relax ion positions, cell shape, and volume.
  - `PREC=LOW` and `POTIM=0.02`, `NSW=90` for ionic relaxation settings.

Step 2: Intermediate Relaxation

- Continues with *MPRelaxSet* for further relaxation.
- Parameters include `EDIFF=1e-3`, `EDIFFG=1e-2` for tighter convergence, and `IBRION=1` for conjugate gradient algorithm.
- `PREC=Normal` and `POTIM=0.3` set for moderate accuracy.

Step 3: Further Relaxation

- Utilizes *MPRelaxSet* to balance accuracy and computational efficiency.
- Includes `IBRION=2`, `ISIF=3`, and `SIGMA=0.1` for consistent relaxation and electronic convergence.

Step 4: Pre-static Calculation

- Employing *MPRelaxSet* with a focus on higher accuracy.
- Key parameters include `EDIFF=1e-4`, `EDIFFG=1e-3` for tight convergence and `PREC=Accurate` for enhanced precision.

Step 5: Static Calculation

- Uses *MPStaticSet* for final static computations.
- Parameters such as `ALGO=Fast`, `EDIFF=1e-4` for electronic minimization, and `IBRION=-1`, `ISMEAR=-5` tailored for static calculations.
- `PREC=Accurate`, `SIGMA=0.05`, and `NSW=0` ensure high precision and no ionic relaxation.

It is important to note that for all calculations, we use a consistent plane-wave cutoff energy `ENCUT=520` eV. This parameter choice ensures sufficient accuracy across all types of calculations, from initial relaxation to final static analysis, and is a crucial factor in achieving reliable and consistent results in our simulations.

The configuration of these VASP settings is strategically designed to ensure the stability and applicability of the structure optimization process, while also considering computational efficiency. For simpler compounds like carbon (C) and silicon (Si), comprising fewer atoms, we optimized computational speed by strategically adjusting parameters such as "NELM", "NSW", and "EDIFF". These adjustments facilitate a faster calculation process without compromising the integrity and reliability of the simulation results.

USPEX calculation inputs

In this study, we developed a Python script for the systematic generation of input files for USPEX calculations, adhering closely to the official guidelines. A template file, meticulously prepared, serves as the foundation for this process. It encompasses placeholders representing essential parameters for USPEX calculations, including aspects of the evolutionary algorithm, population settings, and variation operators.

For each specific composition and its corresponding ground-truth structure, the script performs a detailed extraction of elemental types and quantities, along with space group information. These extracted values are then meticulously substituted for the respective placeholders in the template, namely `atomType`, `numSpecies`, and `symmetries`.

The parameter `populationSize` is meticulously calculated by aggregating the quantities of each constituent element within a structure and rounding this sum to the nearest ten. This calculation is subject to an upper limit of 60, a constraint implemented to maintain computational efficiency. Parallelly, the `stopCrit` parameter is derived in a congruent manner, employing the total atom count, rounded up to the nearest ten, to define the termination criterion for the evolutionary exploration.

Furthermore, the `fracAtomsMut` parameter is judiciously set to 0.20 for compositions comprising a single element, reflecting a higher mutation rate apt for systems of lesser complexity. Conversely, for compositions featuring multiple elements, this parameter is reduced to 0.10, acknowledging the intricate complexity and potential variations in the energy landscape of such systems.

Lastly, the `keepBestHM` parameter is calculated as 15% of the `populationSize`, and this value is then rounded to the nearest whole number. This approach is strategically designed to strike a balance between preserving the most

promising structures from the current generation and fostering the exploration of novel configurations in ensuing generations.

This methodical and rigorous parameterization process ensures that each generated "INPUT.txt" file is exquisitely tailored to the unique attributes of each individual structure, thereby optimizing the efficacy and precision of the evolutionary algorithm within USPEX. To exemplify this process, we present the "INPUT.txt" file for the composition $Ag_{32}Ge_4S_{24}$ as a demonstrative case.

```
************************
TYPE OF RUN AND SYSTEM
************************
USPEX : calculationMethod
300 : calculationType
1 : AutoFrac


% optType
1
% EndOptType


% atomType
Ag Ge S
% EndAtomType


% numSpecies
32 4 24
% EndNumSpecies

% symmetries
33
% EndSymmetries


*************
POPULATION
**************
60 : populationSize
60 : initialPopSize
100 : numGenerations
0.00 : reoptOld
0.60 : bestFrac
9 : keepBestHM
60 : stopCrit


************************
VARIATION OPERATORS
************************
0.40 : fracGene
0.20 : fracRand
0.00 : fracRotMut
0.20 : fracTopRand
0.10 : fracAtomsMut
0.00 : fracLatMut
```

0.10 : fracPerm

****
DETAILS OF AB INITIO CALCULATIONS
****
% abinitioCode
1 1 1 1 1
% ENDabinit

% KresolStart
0.14 0.12 0.11 0.09 0.06
% Kresolend

DBSCAN clustering

DBSCAN is a clustering method that forms clusters based on the density of data points. The procedures can be summarized as follows:

1. **Definition of ε-neighborhood**: Define the ε-neighborhood for each data point using a given distance threshold ε as a parameter. This neighborhood is the set of other data points that are within a distance of ε from a given data point.

2. **Identification of core points**: If the ε-neighborhood of a data point contains at least $N_{\min}$ (another parameter) data points, that data point is considered a core point.

3. **Direct density reachability**: Data point i is considered directly density-reachable from data point j if there is a chain of continuous core points from i to j. This indicates that the core points are densely connected, implying they belong to the same cluster.

4. **Cluster formation**: If core points are directly density-reachable, they belong to the same cluster. Consequently, the entire dataset is divided into several clusters.

5. **Identification of noise points**: Data points that are not directly density-reachable from core points are considered noise points. These points do not belong to any cluster and are isolated.

A key advantage of DBSCAN is that it does not require the number of clusters to be specified beforehand. It can detect clusters of arbitrary shapes and identify outliers (noise points). However, selecting appropriate values for ε and $N_{\min}$ can be challenging, particularly in datasets with varying densities. In this study, we set $\varepsilon = 9$ and $N_{\min} = 10$.

Details of the CSPML model

We describe the training details and model architecture of the CSPML model[6]. We utilized the same compositional descriptor as explained in the Methods section. As a training set, we used a dataset consisting of 126,210 crystals from the Materials Project, identical to the one used for training the global energy predictor in the shotgun CSP.

The dataset exhibited a significant bias towards specific composition ratios. To handle this bias, we trained and evaluated the CSPML model in a manner slightly different from the previous paper[6]. First, we excluded all crystals with the chemical compositions listed in Datasets I and II. Then, we extracted all stable structures (defined as energy above hull = 0) from the remaining data, resulting in 33,064 stable structures, each with a unique chemical composition. These stable structures were randomly divided into training, validation, and test datasets in approximately a 6:2:2 ratio. Due to the significant bias in the dataset towards certain composition ratios—for instance, there are 3,892 compositions with a 2:1:1 ratio out of 33,064 compositions—an upper limit was set at 1% of the data size (e.g., 100 if the data size is 10,000) to be included in the dataset. Specifically, data with composition ratios exceeding this limit underwent down-sampling. This down-sampling process was applied to each training, validation, and test dataset.

Finally, the following procedures were conducted; For each composition ratio with two or more compositions, stable crystal structures with that composition ratio were selected from the dataset, and the structural dissimilarity of all pairs of those crystal structures was calculated using the methods described in the Methods section. Pairs with a structural dissimilarity smaller than 0.3 were considered to have similar stable structures and were divided into two groups: pairs with a dissimilarity smaller than 0.3 (similar pairs) and pairs with a dissimilarity of 0.3 or greater (dissimilar pairs). Since similar pairs are generally fewer relative to the total number of pairs, down-sampling was performed until the number of dissimilar pairs equaled the number of similar pairs. If the number of similar pairs exceeded the number of dissimilar pairs, all dissimilar pairs were selected.

To construct an model ensemble, we repeated the aforementioned procedure five times independently using different random seeds, resulting in five distinct sets of training, validation, and test data. The number of data points for each set, respectively, is $\{83,460,8,862,9,242\}$, $\{76,860,8,390,8,414\}$, $\{79,842,8,684,8,396\}$, $\{84,472,8,357,8,461\}$, and $\{79,016,8,739,9,124\}$. While the number of similar pairs and dissimilar pairs should ideally be approximately 1:1 due to down-sampling, in practice, the number of similar pairs is only slightly higher than the number of dissimilar pairs due to certain cases where the former exceeds the latter. For instance, in the training data of the first dataset, there were 42,875 similar pairs and 40,585 dissimilar pairs. Similar and dissimilar pairs were labeled as 1 and 0, respectively. Since chemical composition is represented by the 290-dimensional descriptor vector as described above, for each composition pair, 290-dimensional composition pair descriptors were created by taking the absolute difference between the paired composition descriptors.

The architecture of the model utilized for mapping the 290-dimensional chemical composition pair descriptors to a binary classification of structural similarity is depicted in Figure S4. This model adopts a forward and fully connected neural network architecture, comprising densely connected layers (Dense), rectified linear units (ReLU), dropout layers (Dropout), batch normalization layers (Batch norm), and softmax function as the final activation function for binary classification (Softmax). The network consists of one or more repeating structures with a dropout layer and an output layer without dropout. All intermediate layers have an identical number of units.

The hyperparameters of the models were selected using Bayesian optimization with the Optuna Python library, employing the validation set (with 50 trials). The hyperparameter search space includes the number of layers $\in \{2,3,4\}$, dropout rate $\in \{0,0.1,0.2\}$, the number of neurons in each layer $\in \{200,400,600,800,1000\}$, the batch size $\in \{1024,2048\}$, and the patience for early stopping $\in \{50,75,100\}$. Each model was trained until the validation error converged (with a patience hyperparameter varying from 50 to 100 epochs) or until reaching 1,000 epochs. The Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) was employed for back-propagating gradients. The training was performed using the TensorFlow-macOS v2.9.0 library, with TensorFlow-metal v0.5.1 plug-in utilized for GPU calculations (Apple M1 Max, GPU 32 cores).

To reduce computational costs, Bayesian optimization of hyperparameters was performed using only the first training and validation dataset. The obtained hyperparameters were then applied to train the models on all five training and validation datasets. The hyperparameters obtained by Bayesian optimization were: number of layers = 3, dropout rate = 0.2, number of neurons in each layer = 400, batch size = 1,024, and patience = 50.

Performance metrics such as mean average precision (MAP), mean accuracy (MACC), average precision (AP), and accuracy (ACC) with respect to the test sets are presented in Table S3. MAP and MACC were calculated by averaging the AP and ACC calculated for each composition ratio. The performance metrics were averaged over the five trials, with the numbers in parentheses representing the standard deviations. These results demonstrate that the models can accurately predict structural similarity from chemical composition pairs with identical composition ratios.

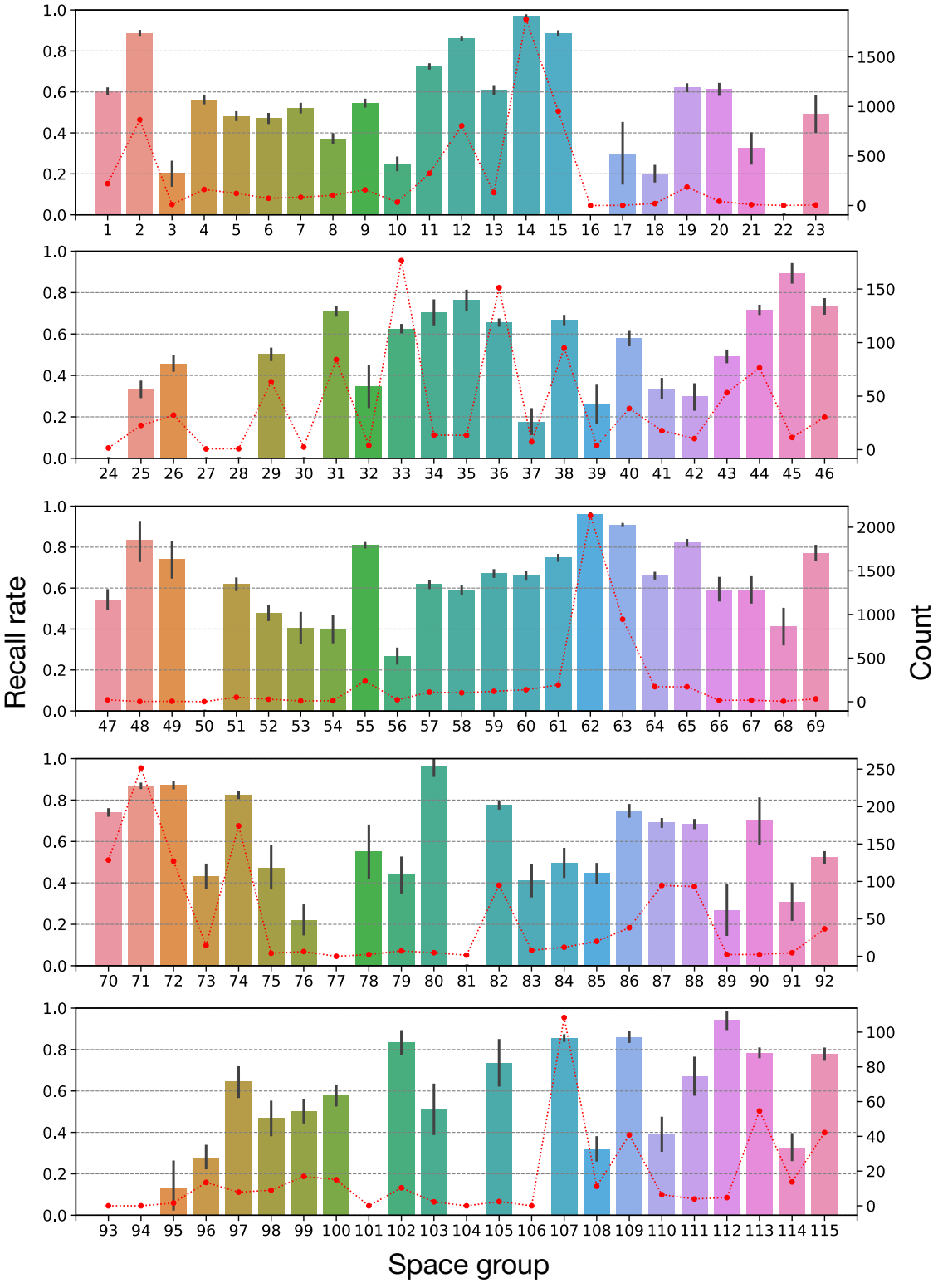
The ensemble of the five models, denoted as $f_1,\ldots,f_5$, was leveraged to generate the predicted class label. The probability of being classified into similar pairs is determined by $\hat{f}(|\phi(C_i)-\phi(C_j)|) = \frac{1}{5}\sum_{b=1}^{5} f_b(|\phi(C_i)-\phi(C_j)|)$. For the set of candidate templates, all 33,064 stable structures were utilized. Utilizing these models and the candidate templates, crystal structure prediction was conducted on the 90 query compositions listed in Datasets I II. The crystal structure prediction procedure aligns with the description in the Methods section of the previous paper[6]. The computational environment for DFT calculations, which optimize the proposed structure through elemental replacement, remained consistent with the details outlined in the main paper.

Up to the top 10 structures were predicted for each query composition. However, this method could not propose any templates for eight out of 90 query compositions: $NaCaAlPHO_5F_2$, $K_5Ag_2(AsSe_3)_3$, $Na(WO_3)_9$, $Li_6V_3P_8O_{29}$, and $Mg_3Si_2H_4O_9$. None of the candidates shared the same composition ratio in the pool of 33,064 candidates. Additionally,

for MgB_7, $Ba_2CaSi_4(BO_7)_2$, and $Y_4Si_5Ir_9$, none of the candidates exhibited class probabilities greater than 0.5. The results of the crystal structure prediction are summarized in Table XX of the main paper. The data and code utilized to train this CSPML model are available on GitHub[?], enabling the reproduction of all aforementioned results.

Table S2. Computational time measured in seconds for single-point DFT calculations of Dataset I.

Composition	Number of atoms	Average	Standard deviation	Minimum	25%	50%	75%	Maximum
C	4	14	5	5	11	13	15	39
Si	2	5	1	4	4	5	6	8
GaAs	2	10	1	7	9	10	10	15
ZnO	4	12	6	7	9	11	13	39
BN	4	10	3	6	8	9	11	26
$LiCoO_2$	16	17	7	9	12	14	19	46
Bi_2Te_3	5	16	4	11	14	15	17	37
$Ba(FeAs)_2$	5	18	5	12	16	17	20	40
SiO_2	6	207	133	90	118	162	243	749
VO_2	6	33	16	12	22	30	42	103
La_2CuO_4	7	51	27	26	36	41	48	154
$LiPF_6$	8	17	6	9	12	15	20	39
Al_2O_3	10	12	4	6	10	12	14	27
$SrTiO_3$	10	33	14	11	24	31	41	76
$CaCO_3$	10	26	12	12	20	23	28	80
TiO_2	12	29	15	11	20	27	34	128
ZrO_2	12	32	11	9	25	31	36	81
$ZrTe_5$	12	48	24	15	37	41	51	140
V_2O_5	14	45	18	24	33	40	52	133
Si_3N_4	14	20	6	10	17	19	22	41
Fe_3O_4	14	75	35	14	50	69	90	173
$Mn(FeO_2)_2$	14	59	29	28	40	50	64	180
ZnSb	16	45	14	7	39	45	51	86
$CoSb_3$	16	42	27	11	31	38	49	185
$LiBF_4$	18	66	31	25	39	57	81	102
Y_2CO_{17}	19	187	22	177	181	183	206	212
GeH_4	20	33	10	20	27	31	37	87
$CsPbI_3$	20	130	73	37	99	124	157	339
$NaCaAlPHO_5F_2$	24	124	33	92	101	123	145	198
$LiFePO_4$	28	164	89	66	105	130	197	615
$Cu_{12}Sb_4S_{13}$	29	134	26	107	126	128	131	201
MgB_7	32	38	12	9	35	39	45	59
Li_3PS_4	32	151	74	75	100	126	169	405
Cd_3As_2	80	216	31	191	199	208	217	288
$Li_4Ti_5O_{12}$	42	375	206	153	228	318	434	1077
$Ba_2CaSi_4(BO_7)_2$	46	249	69	131	210	245	272	439
Ag_8GeS_6	60	489	36	453	461	466	517	544
$Nd_2Fe_{14}B$	68	1026	178	678	823	911	1223	1785
$Y_3Al_5O_{12}$	80	1489	233	998	1221	1366	1467	2648
$Ca_{14}MnSb_{11}$	104	1912	591	1005	1451	1859	2103	3424


Recall rate
Count
Space group

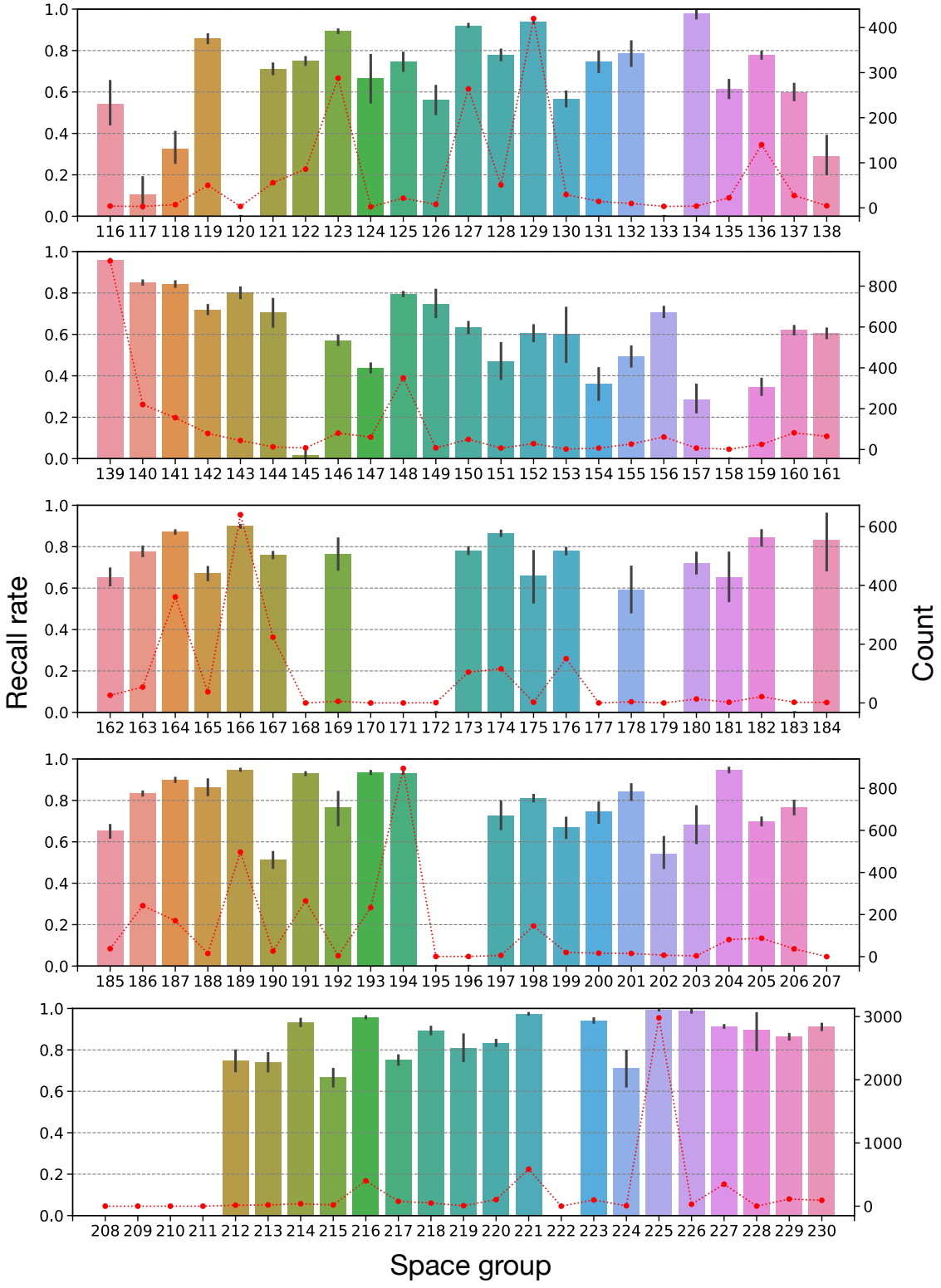


Figure S2. The recall rate of the top 10 predictions and the number of training instances for each space group. The training and testing were repeated 100 times independently. Color bars show the prediction recall rate, with error bars representing the standard deviation for each space group. The Red dashed line shows the average number of training instances.

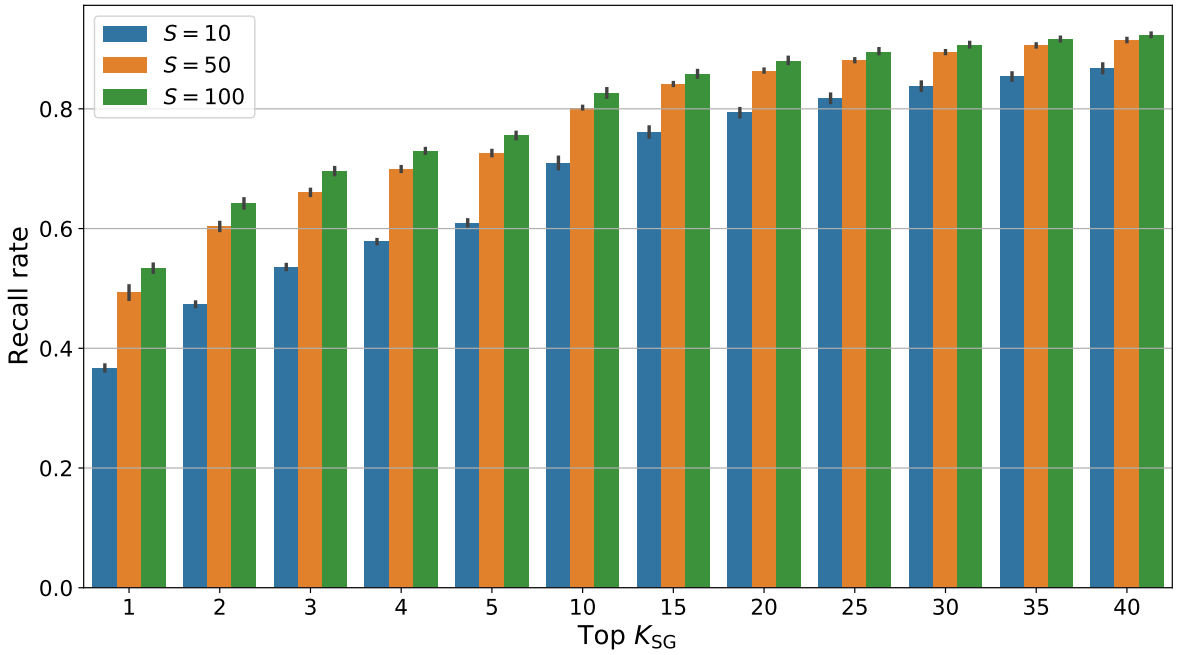


Figure S3. Change in the recall rates of the space group prediction in which the upper bound on the number of samples with the same composition ratio in the training dataset, S, was varied as $S \in \{10, 50, 100\}$.

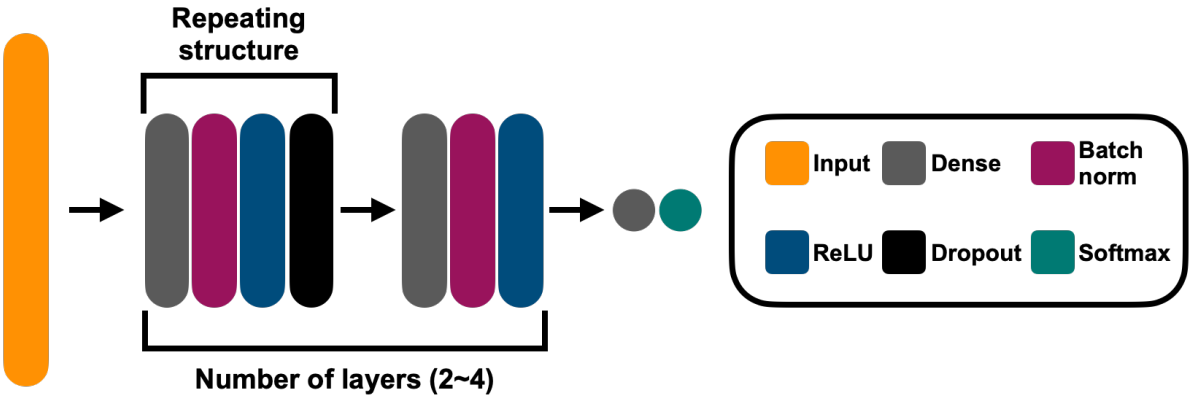


Figure S4. Schematic view of the architecture of a conventional MLP model for CSPML.

Table S3. Prediction performance of CSPML test dataset.

MAP	MACC	AP	ACC
0.941 (±0.007)	0.890 (±0.007)	0.913 (±0.009)	0.864 (±0.011)

Table S4. Prediction performance of CSPML for the 90 crystals comprising Dataset I and II. The top 10 virtual structures were proposed as the final candidate. The Number of atoms column shows the number of atoms in the primitive cell. The symbols of the ✓ and × indicate success and failure, respectively. The − denotes no template for element substitution. Additionally, the symbols in parentheses indicate whether a similar template structure ($\tau \leq 0.2$) was found.

Composition	Number of atoms	Space group	Element substitution
C	4	$R\bar{3}m$	× (✓)
Si	2	$Fd\bar{3}m$	✓ (✓)
GaAs	2	$F\bar{4}3m$	✓ (✓)
ZnO	4	$P6_3mc$	× (✓)
BN	4	$P6_3/mmc$	✓ (✓)
$LiCoO_2$	16	$R\bar{3}m$	✓ (✓)
Bi_2Te_3	5	$R\bar{3}m$	× (✓)
$Ba(FeAs)_2$	5	$I4/mmm$	✓ (✓)
SiO_2	6	$I\bar{4}2d$	✓ (✓)
VO_2	6	$P4_2/mnm$	✓ (✓)
La_2CuO_4	7	$I4/mmm$	✓ (✓)
$LiPF_6$	8	$R\bar{3}$	✓ (✓)
Al_2O_3	10	$R\bar{3}c$	✓ (✓)
$SrTiO_3$	10	$I4/mcm$	× (✓)
$CaCO_3$	10	$R\bar{3}c$	✓ (✓)
TiO_2	12	$C2/m$	× (✓)
ZrO_2	12	$P2_1/c$	✓ (✓)
$ZrTe_5$	12	$Cmcm$	✓ (✓)
V_2O_5	14	$Pmmn$	−
Si_3N_4	14	$P6_3/m$	× (✓)
Fe_3O_4	14	$Fd\bar{3}m$	✓ (✓)
$Mn(FeO_2)_2$	14	$Fd\bar{3}m$	✓ (✓)
ZnSb	16	$Pbca$	× (✓)
$CoSb_3$	16	$Im\bar{3}$	✓ (✓)
$LiBF_4$	18	$P3_121$	× (✓)
Y_2Co_{17}	19	$R\bar{3}m$	✓ (✓)
GeH_4	20	$P2_12_12_1$	× (✓)
$CsPbI_3$	20	$Pnma$	✓ (✓)
$NaCaAlPHO_5F_2$	24	$P2_1/m$	−
$LiFePO_4$	28	$Pnma$	✓ (✓)
$Cu_{12}Sb_4S_{13}$	29	$I\bar{4}3m$	✓ (✓)
MgB_7	32	$Imma$	−
Li_3PS_4	32	$Pnma$	× (✓)
Cd_3As_2	80	$I4_1/acd$	× (✓)
$Li_4Ti_5O_{12}$	42	$C2/c$	× (✓)

Table S4 continued

Composition	Number of atoms	Space group	Element substitution
$Ba_2CaSi_4(BO_7)_2$	46	$I\bar{4}2m$	–
Ag_8GeS_6	60	$Pna2_1$	✓ (✓)
$Nd_2Fe_{14}B$	68	$P4_2/mnm$	× (✓)
$Y_3Al_5O_{12}$	80	$Ia\bar{3}d$	–
$Ca_{14}MnSb_{11}$	104	$I4_1/acd$	✓ (✓)
CsCl	2	$Fm\bar{3}m$	✓ (✓)
MnAl	2	$P4/mmm$	✓ (✓)
HoHSe	3	$P\bar{6}m2$	✓ (✓)
$ErCdRh_2$	4	$Fm\bar{3}m$	✓ (✓)
Eu_2MgTl	4	$Fm\bar{3}m$	✓ (✓)
Pm_2NiIr	4	$Fm\bar{3}m$	✓ (✓)
VPt_3	4	$I4/mmm$	✓ (✓)
$Gd(SiOs)_2$	5	$I4/mmm$	✓ (✓)
$LaAl_3Au$	5	$I4mm$	✓ (✓)
U_2SbN_2	5	$I4/mmm$	✓ (✓)
$MnGa(CuSe_2)_2$	8	$I\bar{4}$	✓ (✓)
SmZnPd	9	$P\bar{6}2m$	✓ (✓)
$Sn(TePd_3)_2$	9	$I4mm$	✓ (✓)
V_5S_4	9	$I4/m$	✓ (✓)
Cs_3InF_6	10	$Fm\bar{3}m$	✓ (✓)
$Eu(CuSb)_2$	10	$P4/nmm$	✓ (✓)
$Rb_2TlAgCl_6$	10	$Fm\bar{3}m$	✓ (✓)
$Ca_3Ni_7B_2$	12	$R\bar{3}m$	✓ (✓)
$DyPO_4$	12	$I4_1/amd$	✓ (✓)
LaSiIr	12	$P2_13$	✓ (✓)
$SmVO_4$	12	$I4_1/amd$	✓ (✓)
VCl_5	12	$P\bar{1}$	✓ (✓)
YbP_5	12	$P2_1/m$	✓ (✓)
$Eu(Al_2Cu)_4$	13	$I4/mmm$	✓ (✓)
Zr_4O	15	$R\bar{3}$	× (✓)
$K_2Ni_3S_4$	18	$Fddd$	✓ (✓)
$Sr(ClO_3)_2$	18	$Fdd2$	✓ (✓)
$LiSm_2IrO_6$	20	$P2_1/c$	✓ (✓)
Pr_2ZnPtO_6	20	$P2_1/c$	✓ (✓)
$Sc_2Mn_{12}P_7$	21	$P\bar{6}$	✓ (✓)
$LaSi_2Ni_9$	24	$I4_1$/amd	✓ (✓)
$CeCu_5Sn$	28	$Pnma$	✓ (✓)
$LiP(HO_2)_2$	32	$Pna2_1$	–

Table S4 continued

Composition	Number of atoms	Space group	Element substitution
$Mg_3Si_2H_4O_9$	36	$P6_3cm$	–
$Y_4Si_5Ir_9$	36	$P6_3/mmc$	–
$Na(WO_3)_9$	37	$R\bar{3}$	–
$Sm_6Ni_{20}As_{13}$	39	$P\bar{6}$	✓ (✓)
$BaCaGaF_7$	40	$P2/c$	✓ (✓)
$Tm_{11}Sn_{10}$	42	$I4/mmm$	–
$AlH_{12}(ClO_2)_3$	44	$R\bar{3}c$	✓ (✓)
$K_2ZrSi_2O_7$	48	$P2_1/c$	× (✓)
$Ba_3Ta_2NiO_9$	60	$P\bar{3}m1$	✓ (✓)
$LiZr_2(PO_4)_3$	72	$P2_1/c$	✓ (✓)
$K_5Ag_2(AsSe_3)_3$	76	$Pnma$	–
$Be_{17}Ru_3$	80	$Im\bar{3}$	–
$Cu_3P_8(S_2Cl)_3$	80	$Pnma$	✓ (✓)
Al_2CoO_4	84	$P3m1$	× (✓)
$Li_6V_3P_8O_{29}$	92	$P1$	–
$ReBi_3O_8$	96	$P2_13$	× (✓)
$Na_5FeP_2(O_4F)_2$	288	$Pbca$	× (✓)
Overall			$59/50 = 86.0\%$

Table S5. Performance of the CSP algorithm with Wyckoff position random crystal structure generators and USPEX for the 25 crystals from Dataset I and II. The top 10 virtual structures with the lowest DFT energies were proposed as the final candidates for the Wyckoff position random crystal structure generation. The Number of atoms column shows the number of atoms in the primitive cell. The symbols of the ✓ and × indicate success and failure, respectively. Additionally, in the Wyckoff position generation column, the symbols on either side of the slash within parentheses indicate whether the Wyckoff pattern was successfully generated and whether the space group prediction was successful, respectively

Composition	**Number of atoms**	**Space group**	**Dataset**	**Wyckoff position generation**	**USPEX**
C	4	$R\bar{3}m$	I	✓ (✓ / ✓)	✓
GaAs	2	$F\bar{4}3m$	I	✓ (✓ / ✓)	✓
ZnO	4	$P6_3mc$	I	✓ (✓ / ✓)	✓
BN	4	$P6_3/mmc$	I	✓ (✓ / ✓)	✓
$LiCoO_2$	16	$R\bar{3}m$	I	✓ (✓ / ✓)	✓
Bi_2Te_3	5	$R\bar{3}m$	I	✓ (✓ / ✓)	×
$Ba(FeAs)_2$	5	$I4/mmm$	I	✓ (✓ / ✓)	✓
La_2CuO_4	7	$I4/mmm$	I	× (✓ / ✓)	✓
Al_2O_3	10	$R\bar{3}c$	I	✓ (✓ / ✓)	✓
$SrTiO_3$	10	$I4/mcm$	I	✓ (✓ / ✓)	×
$CaCO_3$	10	$R\bar{3}c$	I	✓ (✓ / ✓)	✓
Fe_3O_4	14	$Fd\bar{3}m$	I	✓ (✓ / ✓)	✓
$CoSb_3$	16	$Im\bar{3}$	I	✓ (✓ / ✓)	✓
$CsPbI_3$	20	$Pnma$	I	× (× / ✓)	×
MnAl	2	$P4/mmm$	II	✓ (✓ / ✓)	✓
HoHSe	3	$P\bar{6}m2$	II	✓ (✓ / ✓)	✓
$ErCdRh_2$	4	$Fm\bar{3}m$	II	✓ (✓ / ✓)	✓
Eu_2MgTl	4	$Fm\bar{3}m$	II	✓ (✓ / ✓)	✓
Pm_2NiIr	4	$Fm\bar{3}m$	II	✓ (✓ / ✓)	×
$LaAl_3Au$	5	$I4mm$	II	✓ (✓ / ✓)	✓
$Ca_3Ni_7B_2$	12	$R\bar{3}m$	II	✓ (✓ / ✓)	✓
LaSiIr	12	$P2_13$	II	✓ (✓ / ✓)	✓
$SmVO_4$	12	$I4_1/amd$	II	✓ (✓ / ✓)	✓
Zr_4O	15	$R\bar{3}$	II	× (× / ×)	✓
$LiSm_2IrO_6$	20	$P2_1/c$	II	× (✓ / ✓)	×
$Ba_3Ta_2NiO_9$	60	$P\bar{3}m1$	II	✓ (✓ / ✓)	×
Overall				$21/25 = 84.0\%$	$19/25 = 76.0\%$

Visualization of solved structures

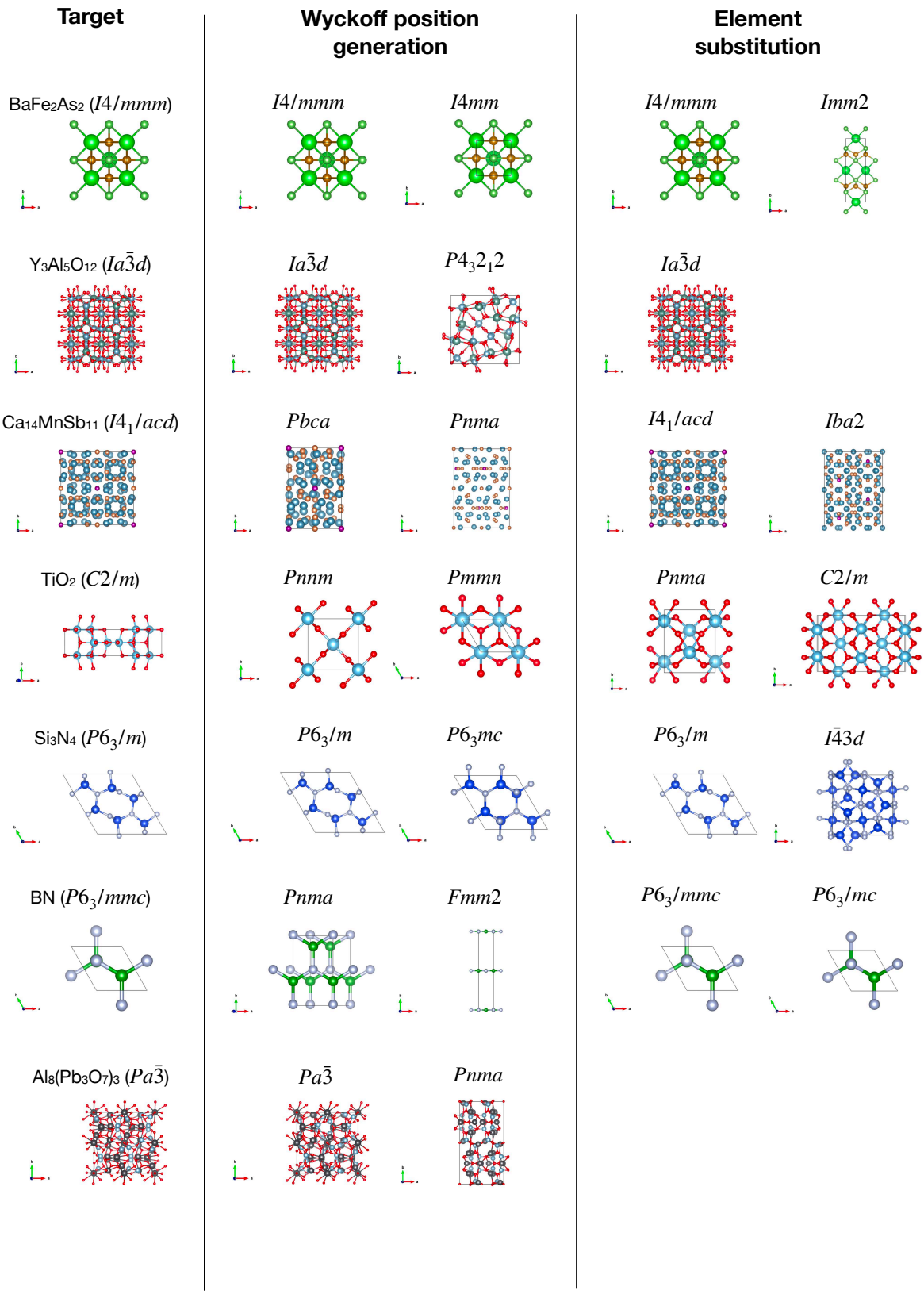
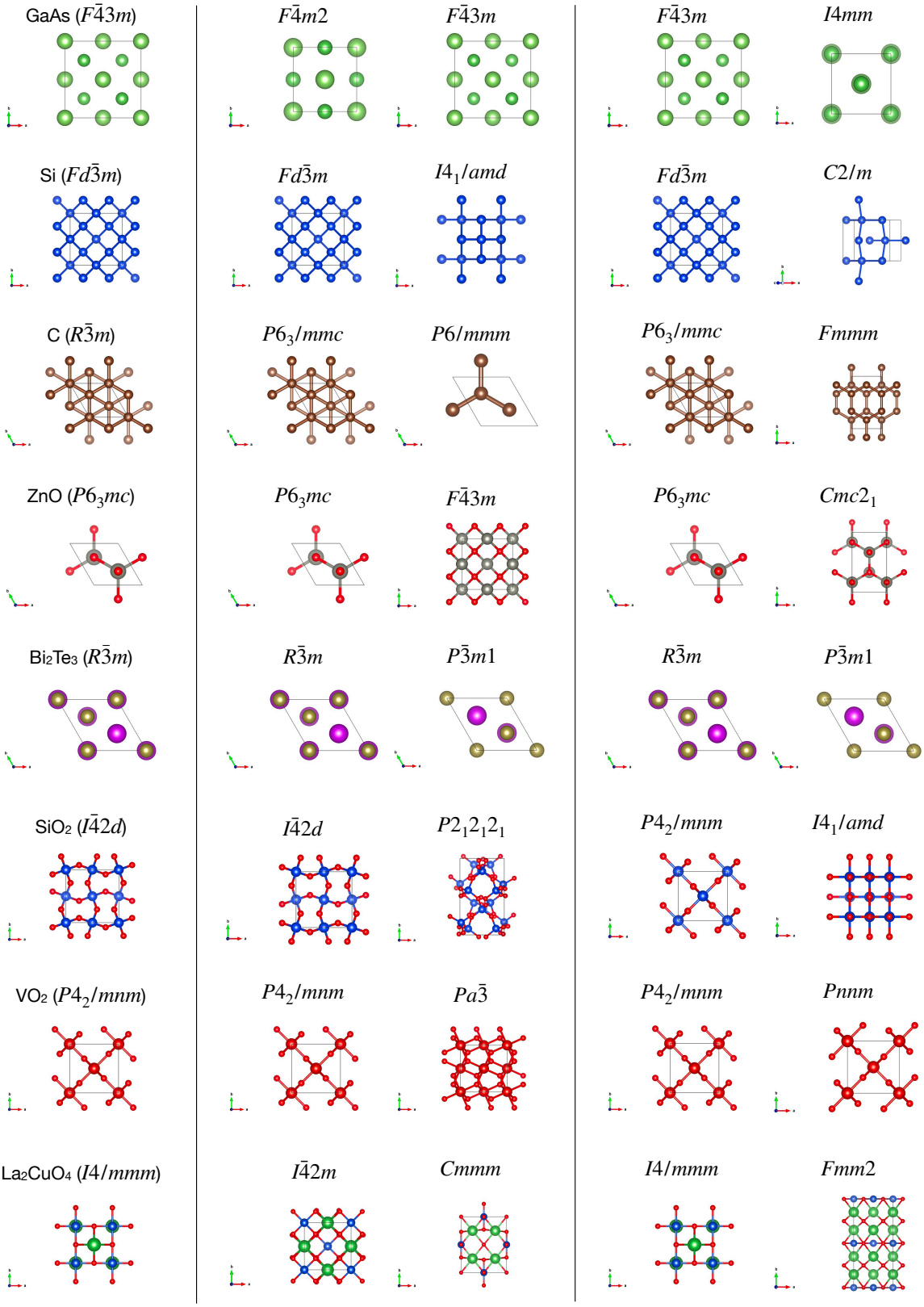
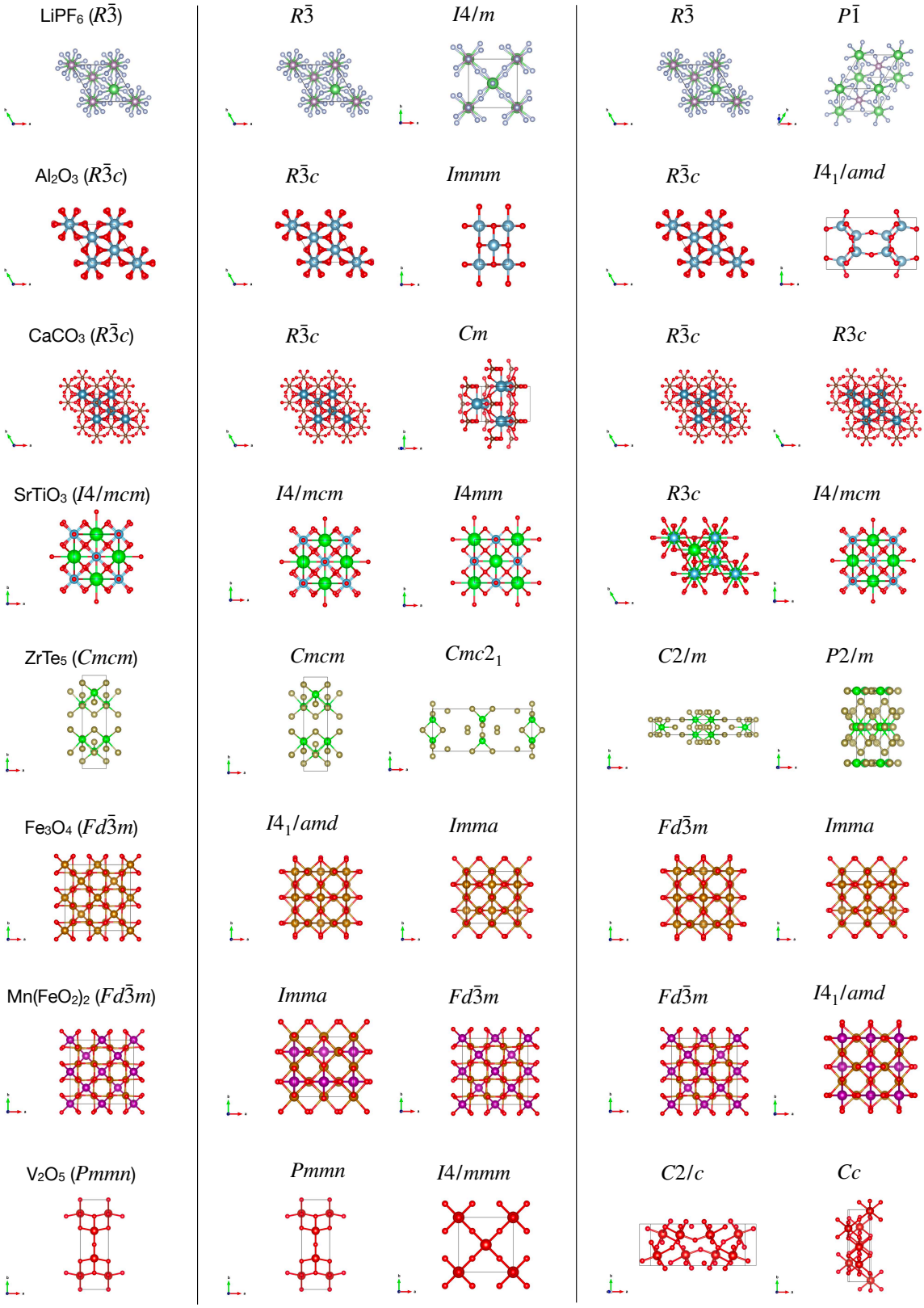
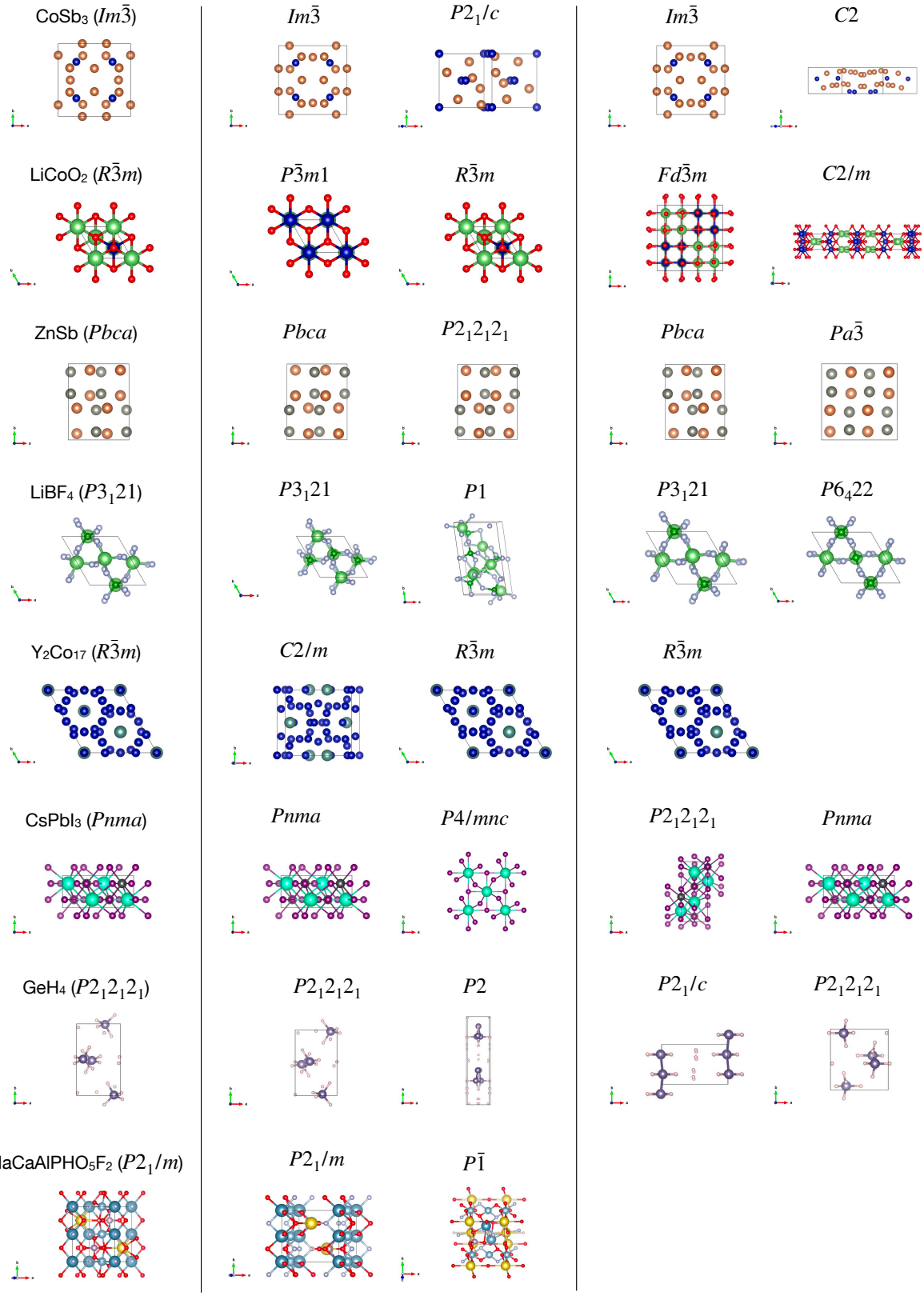
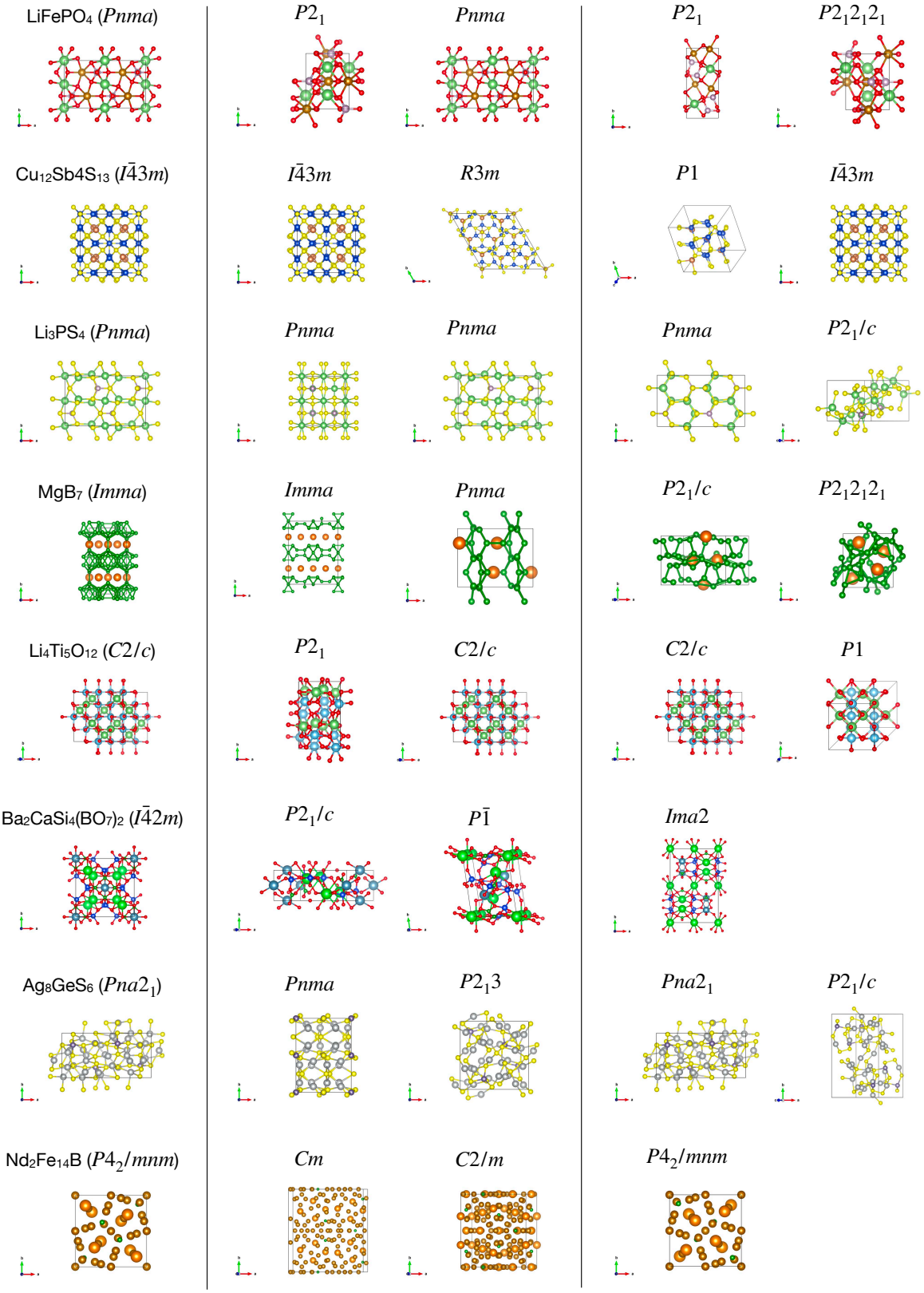

Target
Wyckoff position generation
Element substitution
BaFe2As2 (I4/mmm)
I4/mmm
I4mm
I4/mmm
Imm2
Y3Al5O12 (Ia3̄d)
Ia3̄d
P4₃2₁2
Ia3̄d
Ca14MnSb11 (I4₁/acd)
Pbca
Pnma
I4₁/acd
Iba2
TiO2 (C2/m)
Pnnm
Pmmn
Pnma
C2/m
Si3N4 (P6₃/m)
P6₃/m
P6₃mc
P6₃/m
Ī43d
BN (P6₃/mmc)
Pnma
Fmm2
P6₃/mmc
P6₃/mc
Al8(Pb3O7)3 (Pa3̄)
Pa3̄
Pnma

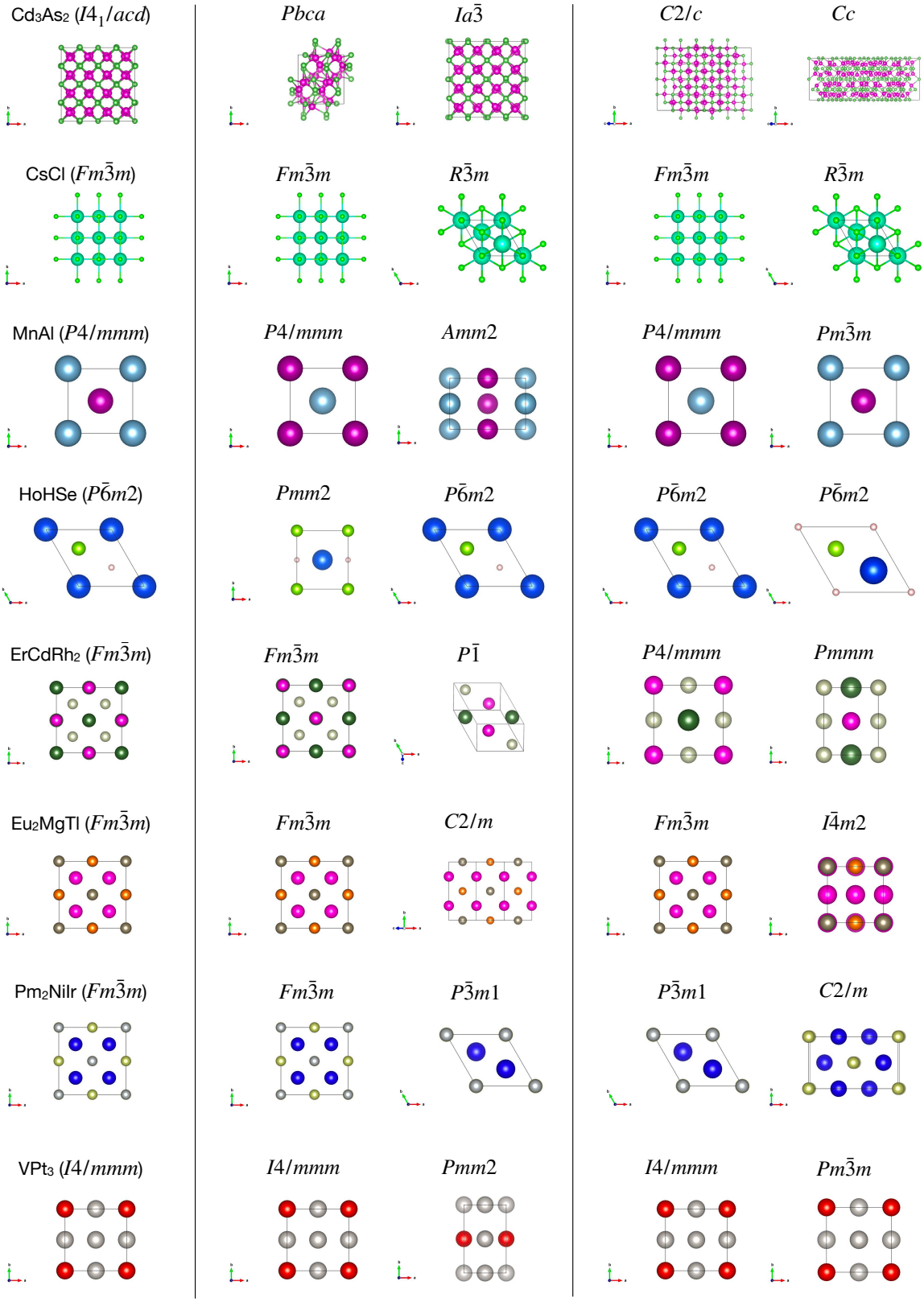
Figure S5

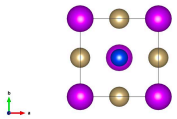
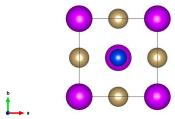
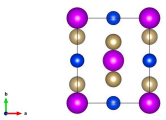
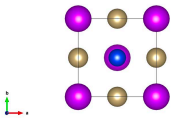
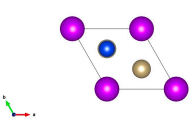
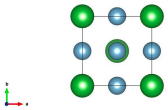
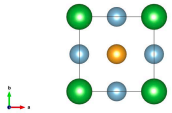
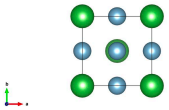
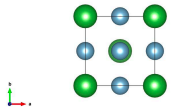
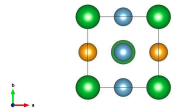
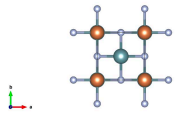
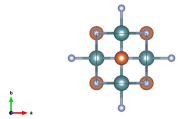
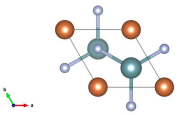
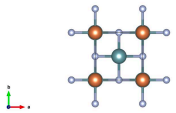
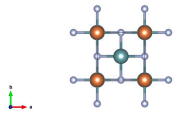
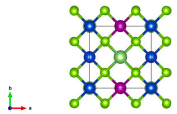
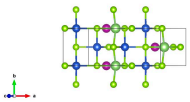
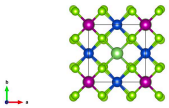
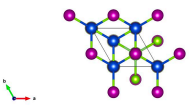
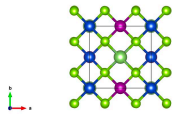
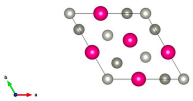
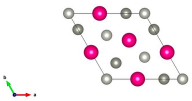
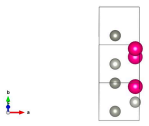
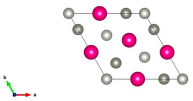
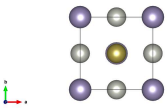
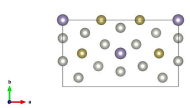
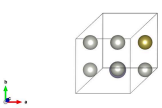
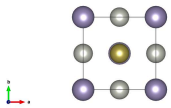
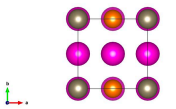
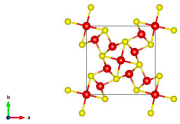
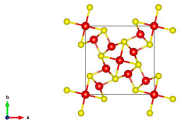
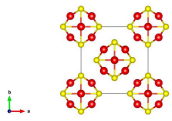
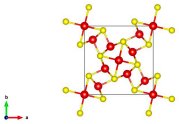
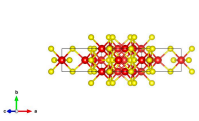
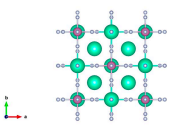
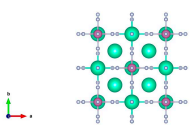
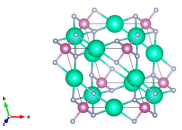
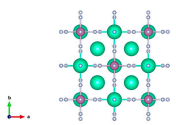
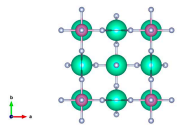

GaAs ($F\bar{4}3m$)
$F\bar{4}m2$
$F\bar{4}3m$
$F\bar{4}3m$
$I4mm$
Si ($Fd\bar{3}m$)
$Fd\bar{3}m$
$I4_1/amd$
$Fd\bar{3}m$
$C2/m$
C ($R\bar{3}m$)
$P6_3/mmc$
$P6/mmm$
$P6_3/mmc$
$Fmmm$
ZnO ($P6_3mc$)
$P6_3mc$
$F\bar{4}3m$
$P6_3mc$
$Cmc2_1$
Bi_2Te_3 ($R\bar{3}m$)
$R\bar{3}m$
$P\bar{3}m1$
$R\bar{3}m$
$P\bar{3}m1$
SiO_2 ($I\bar{4}2d$)
$I\bar{4}2d$
$P2_12_12_1$
$P4_2/mnm$
$I4_1/amd$
VO_2 ($P4_2/mnm$)
$P4_2/mnm$
$Pa\bar{3}$
$P4_2/mnm$
$Pnnm$
La_2CuO_4 ($I4/mmm$)
$I\bar{4}2m$
$Cmmm$
$I4/mmm$
$Fmm2$


LiPF6 (R3̄)
R3̄
I4/m
R3̄
P1̄
Al2O3 (R3̄c)
R3̄c
Immm
R3̄c
I41/amd
CaCO3 (R3̄c)
R3̄c
Cm
R3̄c
R3c
SrTiO3 (I4/mcm)
I4/mcm
I4mm
R3c
I4/mcm
ZrTe5 (Cmcm)
Cmcm
Cmc21
C2/m
P2/m
Fe3O4 (Fd3̄m)
I41/amd
Imma
Fd3̄m
Imma
Mn(FeO2)2 (Fd3̄m)
Imma
Fd3̄m
Fd3̄m
I41/amd
V2O5 (Pmmn)
Pmmn
I4/mmm
C2/c
Cc

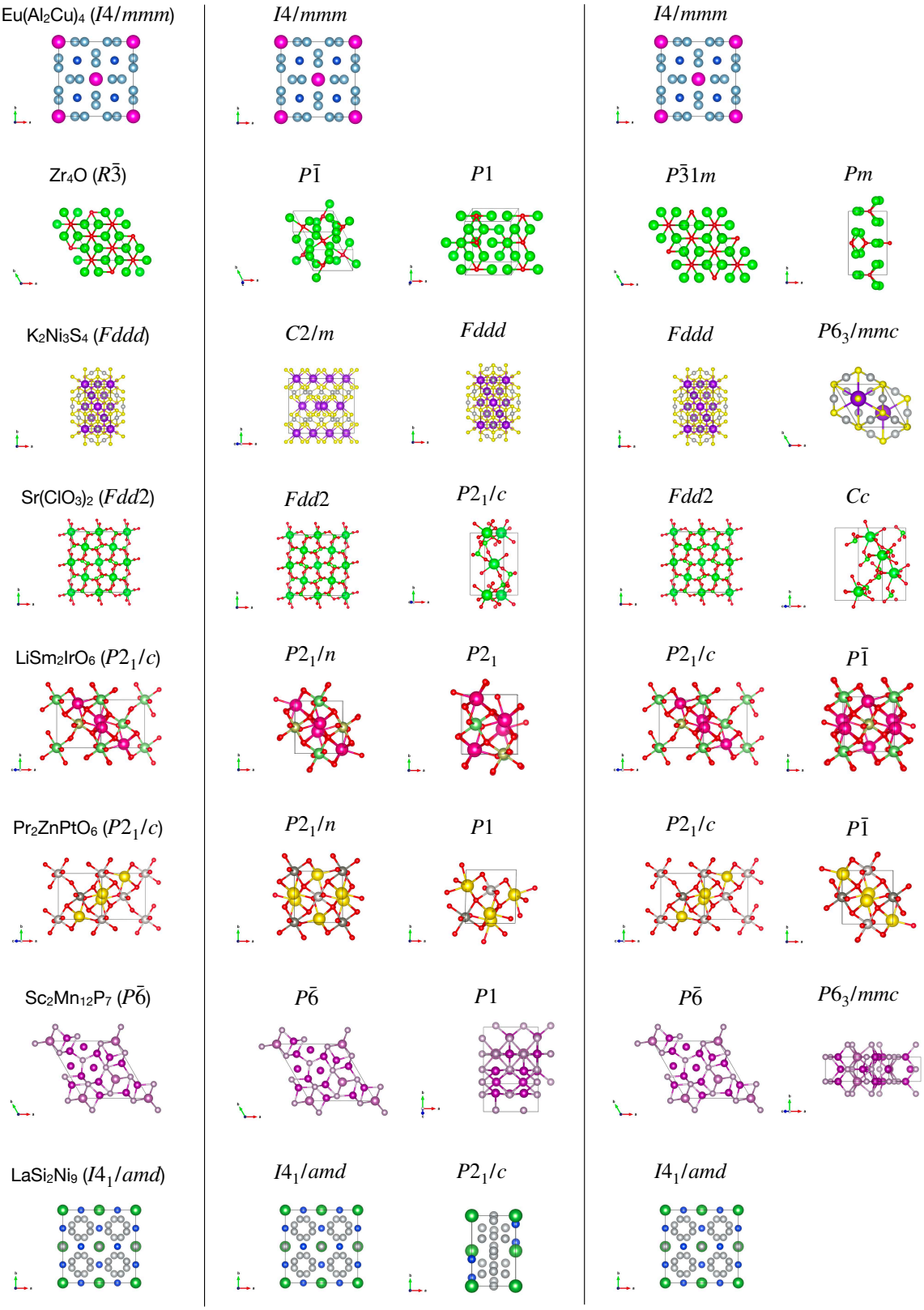

CoSb3 (Im3̄)
Im3̄
P21/c
Im3̄
C2
LiCoO2 (R3̄m)
P3̄m1
R3̄m
Fd3̄m
C2/m
ZnSb (Pbca)
Pbca
P212121
Pbca
Pa3̄
LiBF4 (P3121)
P3121
P1
P3121
P64 22
Y2Co17 (R3̄m)
C2/m
R3̄m
R3̄m
CsPbI3 (Pnma)
Pnma
P4/mnc
P212121
Pnma
GeH4 (P212121)
P212121
P2
P21/c
P212121
NaCaAlPHO5F2 (P21/m)
P21/m
P1̄

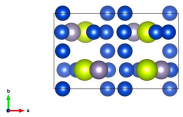
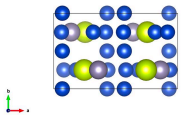
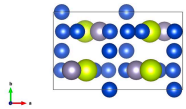
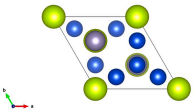
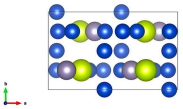
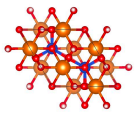
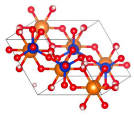
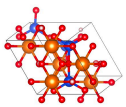
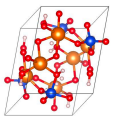
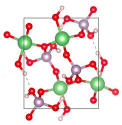
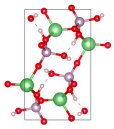
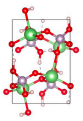
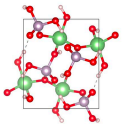
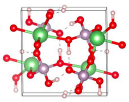
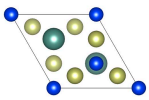
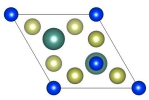
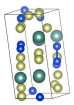
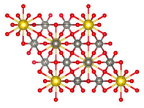
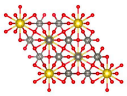
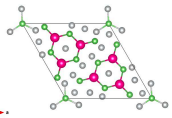
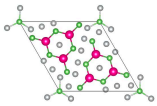
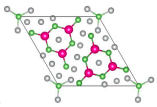
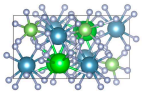
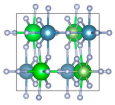
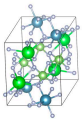
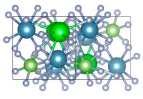
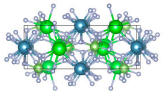
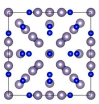
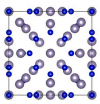
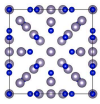
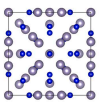

LiFePO4 (Pnma)
P21
Pnma
P21
P21212121
Cu12Sb4S13 (I43m)
I43m
R3m
P1
I43m
Li3PS4 (Pnma)
Pnma
Pnma
Pnma
P21/c
MgB7 (Imma)
Imma
Pnma
P21/c
P21212121
Li4Ti5O12 (C2/c)
P21
C2/c
C2/c
P1
Ba2CaSi4(BO7)2 (I42m)
P21/c
P1
Ima2
Ag8GeS6 (Pna21)
Pnma
P213
Pna21
P21/c
Nd2Fe14B (P42/mnm)
Cm
C2/m
P42/mnm

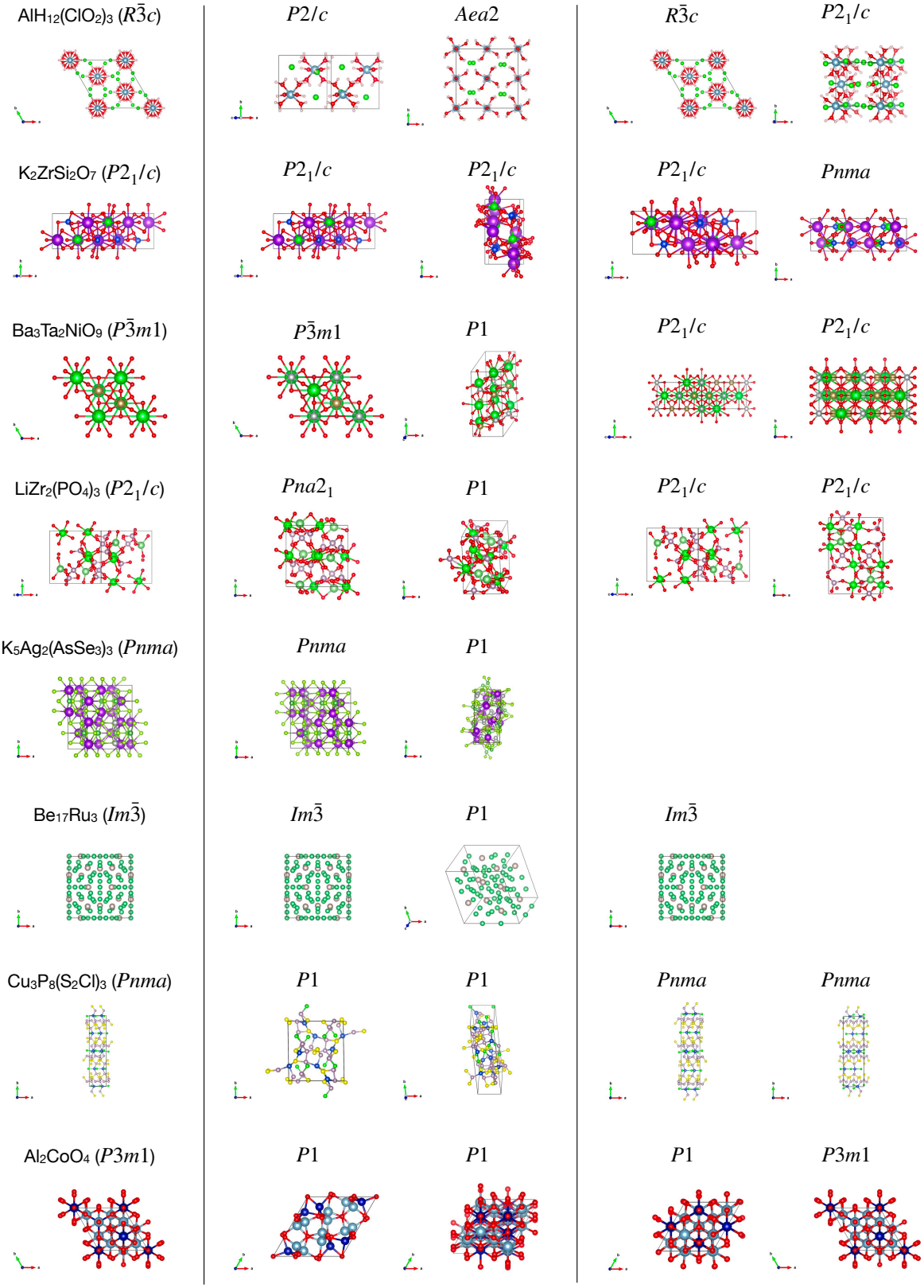

Cd3As2 (I41/acd)
Pbca
Iā3
C2/c
Cc
CsCl (Fm3̄m)
Fm3̄m
R3̄m
Fm3̄m
R3̄m
MnAl (P4/mmm)
P4/mmm
Amm2
P4/mmm
Pm3̄m
HoHSe (P6̄m2)
Pmm2
P6̄m2
P6̄m2
P6̄m2
ErCdRh2 (Fm3̄m)
Fm3̄m
P1̄
P4/mmm
Pmmm
Eu2MgTl (Fm3̄m)
Fm3̄m
C2/m
Fm3̄m
Ī4m2
Pm2Nilr (Fm3̄m)
Fm3̄m
P3̄m1
P3̄m1
C2/m
VPt3 (I4/mmm)
I4/mmm
Pmm2
I4/mmm
Pm3̄m

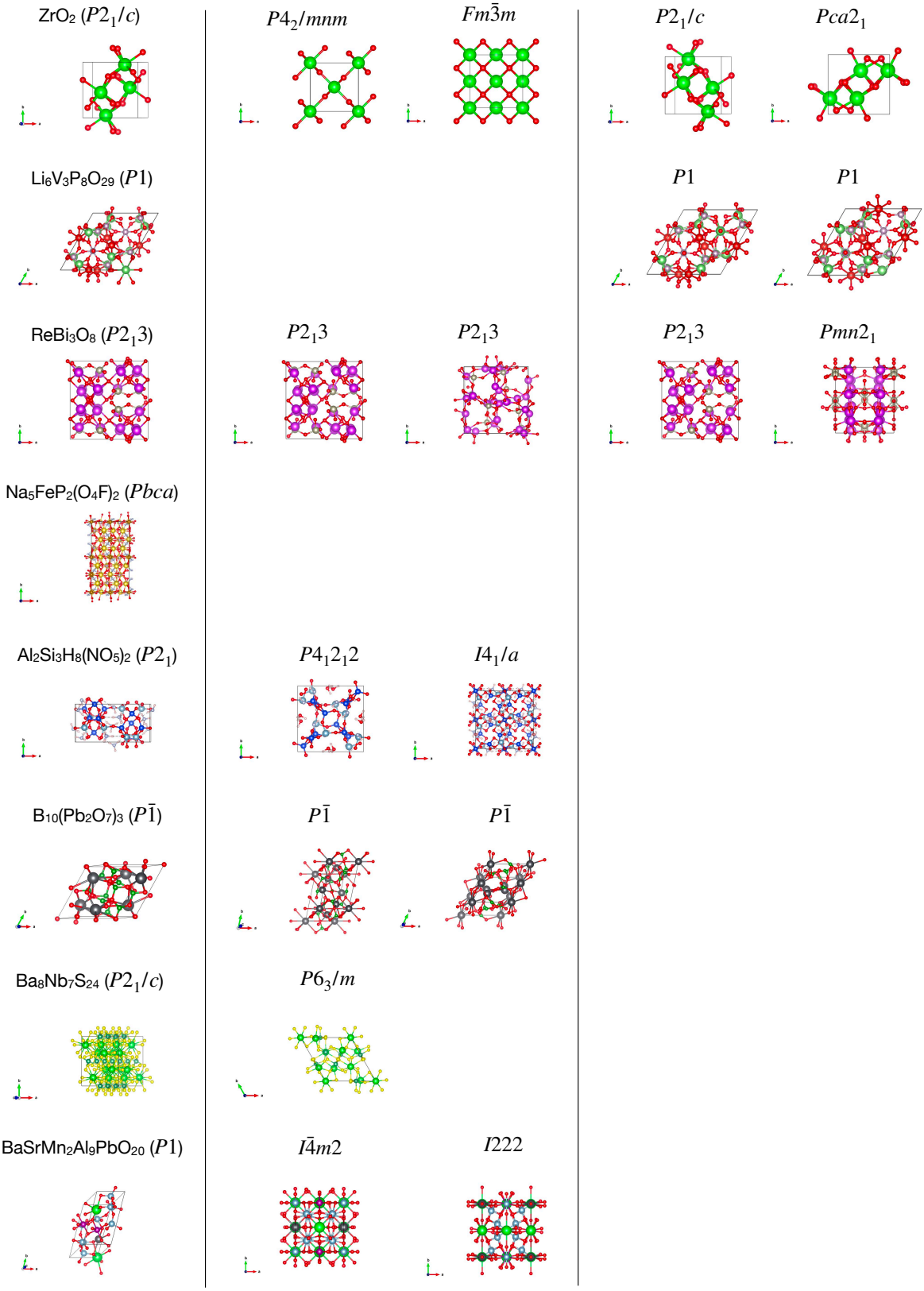
$Gd(SiOs)_2$ ($I4/mmm$)	$I4/mmm$	$Immm$	$I4/mmm$	$P\bar{3}m1$	
					
$LaAl_3Au$ ($I4mm$)	$P4mm$	$I4mm$	$I4mm$	$I\bar{4}m2$	
					
U_2Sb1N_2 ($I4/mmm$)	$Immm$	$P\bar{3}m1$	$I4/mmm$	$Immm$	
					
$MnGa(CuSe_2)_2$ ($I\bar{4}$)	Cm	$I\bar{4}2m$	$R3m$	$I\bar{4}$	
					
$SmZnPd$ ($P\bar{6}2m$)	$P\bar{6}2m$	$P1$	$P\bar{6}2m$		
					
$Sn(TePd_3)_2$ ($I4mm$)	$C2$	$P1$	$I4mm$	$I\bar{4}m2$	
					
V_5S_4 ($I4/m$)	$I4/m$	$I4/mmm$	$I4/m$	$C2/m$	
					
Cs_3InF_6 ($Fm\bar{3}m$)	$Fm\bar{3}m$	$P\bar{1}$	$Fm\bar{3}m$	$I4/mmm$	
					

$Eu(CuSb)_2$ ($P4/nmm$)	$I4/mmm$	$P4mm$	$I4/mmm$	$P4/nmm$	
$Rb_2TlAgCl_6$ ($Fm\bar{3}m$)	$Fm\bar{3}m$	$C2/m$	$Fm\bar{3}m$	$I4/mmm$	
$Ca_3Ni_7B_2$ ($R\bar{3}m$)	$Immm$	$R\bar{3}m$	$R\bar{3}m$	$C2/m$	
$DyPO_4$ ($I4_1/amd$)	$I4_1/amd$	$Cmcm$	$I4_1/amd$	$Fddd$	
$LaSiIr$ ($P2_13$)	$P2_13$	$P1$	$P2_13$	$Pnma$	
$SmVO_4$ ($I4_1/amd$)	$P1$	$I4_1/amd$	$Fddd$	$P1$	
VCl_5 ($P\bar{1}$)	$P\bar{1}$	$P1$	$Pmmn$	$P6_3/mmc$	
YbP_5 ($P2_1/m$)	$P2_1/m$	$P2_1/m$	$P2_1/m$	$P2_1/m$	


Eu(Al2Cu)4 (I4/mmm)
I4/mmm
I4/mmm
Zr4O (R3̄)
P1̄
P1
P3̄1m
Pm
K2Ni3S4 (Fddd)
C2/m
Fddd
Fddd
P63/mmc
Sr(ClO3)2 (Fdd2)
Fdd2
P21/c
Fdd2
Cc
LiSm2IrO6 (P21/c)
P21/n
P21
P21/c
P1̄
Pr2ZnPtO6 (P21/c)
P21/n
P1
P21/c
P1̄
Sc2Mn12P7 (P6̄)
P6̄
P1
P6̄
P63/mmc
LaSi2Ni9 (I41/amd)
I41/amd
P21/c
I41/amd

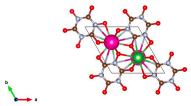
$CeCu_5Sn$ ($Pnma$)	$Pnma$	$Pnma$	$P6_3/mmc$	$Pnma$
				
$Mg_3Si_2H_4O_9$ ($P6_3cm$)	$P1$	$P1$	$P\bar{1}$	
				
$LiP(HO_2)_2$ ($Pna2_1$)	$Pna2_1$	$P2_12_12_1$	$Pna2_1$	$P\bar{1}$
				
$Y_4Si_5Ir_9$ ($P6_3/mmc$)	$P6_3/mmc$	$P1$		
				
$Na(WO_3)_9$ ($R\bar{3}$)	$R\bar{3}$			
				
$Sm_6Ni_{20}As_{13}$ ($P\bar{6}$)	$P\bar{6}$		$P\bar{6}$	$P\bar{1}$
				
$BaCaGaF_7$ ($P2/c$)	$Pnma$	$P1$	$P2/c$	$C2/c$
				
$Tm_{11}Sn_{10}$ ($I4/mmm$)	$I4/mmm$		$Immm$	$I4/mmm$
				


AlH12(ClO2)3 (R3̄c)
P2/c
Aea2
R3̄c
P21/c
K2ZrSi2O7 (P21/c)
P21/c
P21/c
P21/c
Pnma
Ba3Ta2NiO9 (P3̄m1)
P3̄m1
P1
P21/c
P21/c
LiZr2(PO4)3 (P21/c)
Pna21
P1
P21/c
P21/c
K5Ag2(AsSe3)3 (Pnma)
Pnma
P1
Be17Ru3 (Im3̄)
Im3̄
P1
Im3̄
Cu3P8(S2Cl)3 (Pnma)
P1
P1
Pnma
Pnma
Al2CoO4 (P3m1)
P1
P1
P1
P3m1

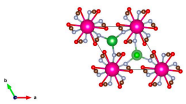

ZrO2 (P21/c)
P42/mnm
Fm3̄m
P21/c
Pca21
Li6V3P8O29 (P1)
P1
P1
ReBi3O8 (P213)
P213
P213
P213
Pmn21
Na5FeP2(O4F)2 (Pbca)
Al2Si3H8(NO5)2 (P21)
P41212
I41/a
B10(Pb2O7)3 (P1̄)
P1̄
P1̄
Ba8Nb7S24 (P21/c)
P63/m
BaSrMn2Al9PbO20 (P1)
I4̄m2
I222

$Bi_8AsAuCl_9$ ($P2_1/c$)	$P2_1/c$	$Pnma$	
$H_{18}SO_{12}$ (Cc)	$P\bar{4}2_1c$	$P2_1/c$	
$K_3Fe_3P_4H_2O_{17}$ ($Pnna$)	$P2_13$	$Pnma$	
$K_{11}LiMn_4O_{16}$ ($I\bar{4}2m$)	$I\bar{4}2m$	$I4/mmm$	
$LiMn_3Al_2(HO_2)_6$ ($P\bar{1}$)	$P2_1/c$	$Cmc2_1$	
$Mg_{10}B_{16}Ir_{19}$ ($I\bar{4}3m$)	$I\bar{4}3m$		
$Na_4PuH_7O_9$ ($P\bar{1}$)	$Fdd2$	$C2/c$	
$Nb_{12}Br_{17}F_{13}$ ($P1$)	$Fmm2$	$P\bar{1}$	

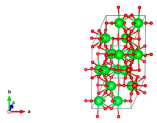
$RbLa_2C_6N_6ClO_6$ ($P6_3/m$)



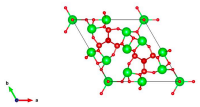
$P6_3/m$



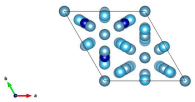
$Sr_{16}V_8O_{31}$ ($P\bar{1}$)



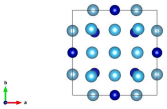
$P\bar{6}_2m$



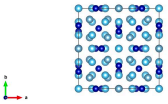
$Ti_{13}Al_9Co_8$ ($R3m$)



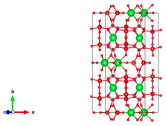
$P4/mmm$



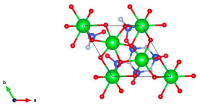
$Fm\bar{3}m$



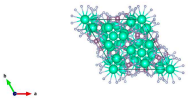
$Ba_2V_5(PO_6)_4$ (Cm)



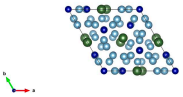
$Ba_3Si_6N_4O_9$ ($P3$)



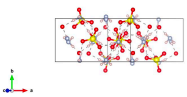
$Cs_{10}(Mo_2N_5)_3$ ($R\bar{3}c$)



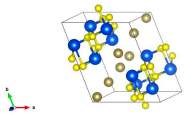
$Er_6Al_{41}Cr_6$ ($P\bar{3}1m$)



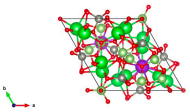
$H_{13}S_2N_3O_8$ ($P2/c$)



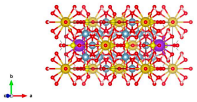
$Ho_{10}Te_7S_{10}$ ($P1$)



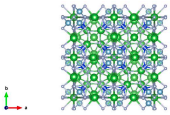
$KBa_6Zn_4(GaO_3)_7$ ($P\bar{1}$)



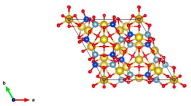
$KNa_3Al_{12}H_{24}(SO_7)_8$ ($P2/m$)



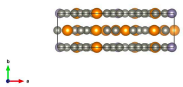
$La_{17}Al_4(Si_3N_{11})_3$ ($F\bar{4}3m$)



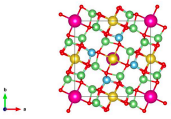
$Na_8Al_6Si_6CO_{27}$ ($R3$)



$Nd_{14}Zn_{43}Sn_3$ (Pm)



$RbNa_3Li1_2Ti_4O_{16}$ ($I4/m$)



$Ta_4P_4S_{29}$ ($P4_32_12$)

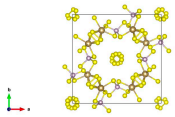


Figure S5. 120 stable structures solved by the crystal structure prediction algorithm (depicted with VESTA[1] version 3.5.8). For each prediction algorithm, the structures with the two lowest DFT energies are shown.

References

1. Momma, K. & Izumi, F. *VESTA3* for three-dimensional visualization of crystal, volumetric and morphology data. *J. Appl. Cryst.* **44**, 1272–1276, DOI: 10.1107/S0021889811038970 (2011).
2. Xie, T. & Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **120**, 145301, DOI: 10.1103/PhysRevLett.120.145301 (2018).
3. AI Bridging Cloud Infrastructure (ABCI). https://abci.ai. Accessed: 2023-11-06.
4. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631 (2019).
5. Ong, S. P. *et al.* Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
6. Kusaba, M., Liu, C. & Yoshida, R. Crystal structure prediction with machine learning-based element substitution. *Comput. Mater. Sci.* **211**, 111496 (2022).