npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-024-01471-8

Shotgun crystal structure prediction using machine-learned formation energies

Liu Chang [1], Hiromasa Tamaki[2], Tomoyasu Yokoyama [2], Kensuke Wakasugi [2], Satoshi Yotsuhashi[2], Minoru Kusaba[1], Artem R. Oganov [3] & Ryo Yoshida [1,4]

Stable or metastable crystal structures of assembled atoms can be predicted by finding the global or local minima of the energy surface within a broad space of atomic configurations. Generally, this requires repeated first-principles energy calculations, which is often impractical for large crystalline systems. Here, we present significant progress toward solving the crystal structure prediction problem: we performed noniterative, single-shot screening using a large library of virtually created crystal structures with a machine-learning energy predictor. This shotgun method (ShotgunCSP) has two key technical components: transfer learning for accurate energy prediction of pre-relaxed crystalline states, and two generative models based on element substitution and symmetry-restricted structure generation to produce promising and diverse crystal structures. First-principles calculations were performed only to generate the training samples and to refine a few selected pre-relaxed crystal structures. The ShotunCSP method is less computationally intensive than conventional methods and exhibits exceptional prediction accuracy, reaching 93.3% in benchmark tests with 90 different crystal structures.

The prediction of stable or metastable crystal structures from a given chemical composition has remained a fundamentally unsolved task in solid-state physics for several decades[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 an energy surface within a broad space of atomic configurations, in which the energy can be evaluated by first-principles density functional theory (DFT) calculations. The CSP problem can be solved by applying an exploratory algorithm to determine the crystal structure at the global or local minima by successively displacing the atomic configurations along the energy gradient.

A broad array of CSP methods have been developed to solve this problem, 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, the machine-learning interatomic potentials have attracted increasing attention because they can expedite 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 set of candidate crystal structures, whereupon their DFT energies are used as goodness-of-fit scores to prioritize 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 known as swarm shift[18]. However, these algorithms are time-consuming because of the need for ab initio structural relaxation of the candidate crystals at every step of the optimization process. In response, 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. This requires a large number of candidate structures to be relaxed at every step of the sequential search, which is impractical and computationally expensive for large systems that contain more than 30–40 atoms per unit cell.

A promising solution is to fully replace ab initio energy calculations with machine-learning surrogates. Energy predictors trained using DFT

[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]Skolkovo Institute of Science and Technology, Skolkovo Innovation Center, Moscow, 121205, Russia. [4] The Graduate Institute for Advanced Studies, The Graduate University for Advanced Studies (SOKENDAI), Tachikawa, 190-8562, Japan.
e-mail: yoshidar@ism.ac.jp

property databases, such as the Materials Project[21,22], AFLOW[23,24], OQMD[25,26], and GNoME[27], exhibit reasonably high prediction accuracy[28–30]. However, models trained on instances of stable or metastable structures in such databases are unsuitable for the prediction of pre-relaxed energies for a given system[31]. As shown later, although these models can predict energy differences between different crystalline systems, they cannot quantitatively discriminate between the energy differences of distinct conformations for the system of interest, which is a requirement 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 states from the Materials Project database. Subsequently, for a given chemical composition, the 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 are biased toward high-energy states and generally not applicable to the extrapolative domain of low-energy states in which optimal or suboptimal conformations exist. In CSP, a surrogate model must be able to predict the energies of various conformations, with high- to low-energy states corresponding to the pre- and post-relaxed crystal structures, respectively. We demonstrate that a surrogate model derived using transfer learning exhibits high prediction accuracy, even for low-energy states.

After creating candidate crystal structures, exhaustive virtual screening was performed using the transfer-learning-based energy predictor. The narrowed-down candidate crystals were relaxed by performing DFT calculations. A wide variety of structure generators can be applied to generate virtual crystal libraries, including (i) methods based on element substitution using existing crystal structures as templates[34–36], (ii) atomic coordinate generators that consider 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 previously synthesized crystals[40–42]. In this study, we validated our framework using two sets of virtual libraries created using methods related to (i) and (ii), that is, element substitution of template crystal structures and a Wyckoff position generator for de novo CSP. The search space was narrowed down in the latter by machine-learning-based prediction of the space groups and Wyckoff-letter assignments.

The ShotgunCSP workflow, which can be regarded as a high-throughput virtual screening of crystal structures, is perhaps the simplest among existing CSP methods to date. The entire workflow comprises first-principles single-point energy calculations for, at most, 3000 structures to create a training set for the transfer-learning-based energy predictor and 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 significantly less computationally demanding. Furthermore, the prediction performance is outstanding, as confirmed by experimental validation; specifically, the method accurately predicted more than 90% of the stable structures of 90 benchmark crystals with diverse space groups, structure types, constituent elements, system sizes, and application domains.

Results

Outline of methods

The stable crystal structures of atom assemblies with chemical composition X were predicted using the machine-learning workflow summarized in Fig. 1. This method involves two key technical components: a high-performance surrogate model for predicting DFT formation energies and two different generative models for producing candidate crystal structures.

For the energy calculation, a CGCNN with the same architecture as that in a previous paper[30] was pretrained from scratch on a set of 126,210 crystals for which DFT formation energies are available in the Materials Project database. This model, referred to as the global model, can accurately predict the baseline formation energies of diverse crystal structures; however, it is unable to discriminate between the local energy differences of different atomic conformations for a given target system. Therefore, the pretrained global model was localized to the target system X by transfer

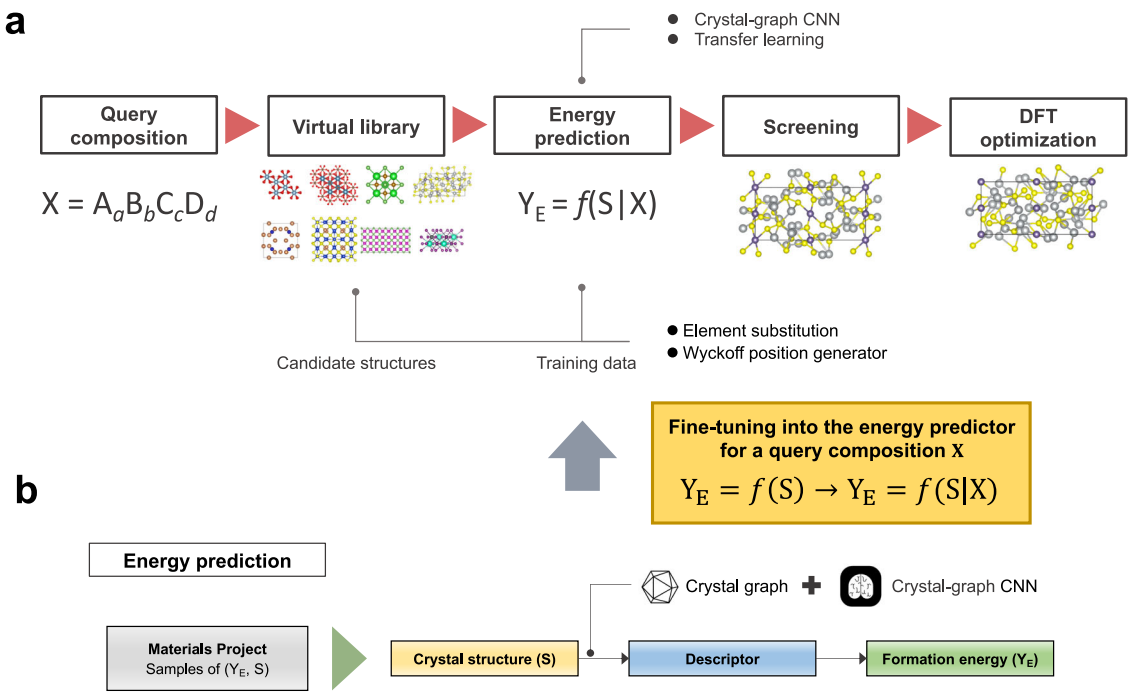


Fig. 1 | Workflow of the ShotgunCSP algorithm. a Virtual screening using a machine-learning-based energy predictor. A virtual library is created using a machine-learning-based element-substitution method (ShotgunCSP-GT) or a Wyckoff position generator (ShotgunCSP-GW). **b** Construction of formation-energy predictor based on a CGCNN. The CGCNN was trained with the Materials Project database and fine-tuned for the energy prediction of pre-relaxed candidate crystal structures for a query composition X.

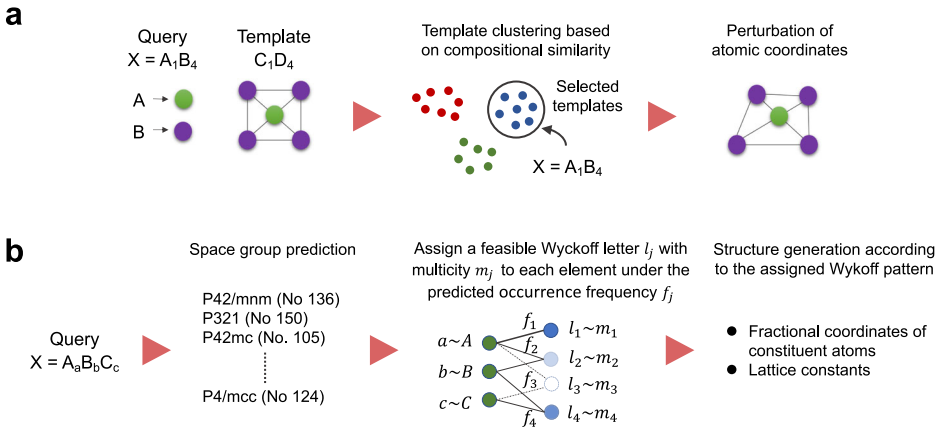


Fig. 2 | Two different crystal structure generators are used to generate training instances to refine the energy prediction model and to create virtual libraries to be screened. a ShotgunCSP-GT: element substitution of template crystal structures in the Materials Project database. **b** ShotgunCSP-GW: Wyckoff position generator to produce symmetry-restricted atomic coordinates.

learning. To this end, we randomly generated, at most, several thousand virtual crystal structures and calculated their formation energies by performing single-point energy calculations. The training structures were generated by either template-based substitution or the Wyckoff position generator, as described below. Using the dataset of generated structures, 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 Methods).

We developed and tested two algorithms to generate virtual crystals: element substitution (ShotgunCSP-GT) and a Wyckoff position generator (ShotgunCSP-GW).

Method 1 —Element substitution (ShotgunCSP-GT): Elements of already synthesized or theoretically possible crystals with the same composition as X are randomly substituted (Fig. 2(a)).

Method 2 —Wyckoff position generator (ShotgunCSP-GW): For a target composition, with the space group given a priori or predicted, the generator randomly creates symmetry-restricted atomic coordinates from all possible combinations of Wyckoff positions (Fig. 2(b)). A machine-learning predictor is employed to efficiently reduce the degrees of freedom in Wyckoff-letter assignment.

ShotgunCSP-GT cannot be applied unless a template is available for substitution, which limits its applicability. Therefore, we developed ShotgunCSP-GW, which can generate novel structures when no template is available. The configuration of space groups and assignment of Wyckoff letters were deduced based on machine learning, as described below. The two generators were used to produce training instances for fine-tuning the CGCNN as well as candidate structures during high-throughput virtual screening. The workflows for the two generators differed slightly, as described below.

For ShotgunCSP-GT, crystal structures were generated by replacing elements in existing crystals. This mimics the process whereby 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 composition ratio as that of X from the Materials Project database. Candidate structures were created by assigning the constituent elements of the query composition to the atomic coordinates of the selected templates. Elements with the same composition fraction in the template and query composition were substituted. When two or more elements had the same composition fraction, the assignment could not be uniquely determined. In this case, the most similar element pair was selected for substitution using the normalized Euclidean distance of the 58 element descriptors in the XenonPy library[33,43–46] as the similarity measure. The crystal structures with the substituted elements inherited the atomic coordinates of the template structures. The generated atomic coordinates were subjected to slight random perturbations as an additional refinement step. Considering that multiple crystals in the database have the same prototype structure (for example, 8005 compounds have the composition ratio $A_1B_1C_2$), a cluster-based template selection procedure was introduced to select highly relevant templates with query composition X while maintaining the diversity of the template structures. We applied DBSCAN[47,48] to classify the templates into clusters in which the chemical compositions were converted into 290-dimensional compositional descriptors using XenonPy. Then, only those templates belonging to the same cluster as the query composition X were selected to identify 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 the StructureMatcher module of pymatgen[49,50] to construct a unique set of templates without any identical prototype structures. The number of unique templates in the same cluster as the query composition is denoted by K_{temp}. A virtual library was created by generating 1000 structures from each of the K_{temp} selected templates by perturbing the atomic coordinates and lattice constants to form $1000 \times K_{\mathrm{temp}}$ candidate structures. This procedure was also used to randomly generate ten structures for each template, which served as a training dataset of $10 \times K_{\mathrm{temp}}$ structures for the fine-tuning step. (See Methods for additional details.)

By contrast, ShotgunCSP-GW produced random crystal structures with a prescribed space group for a given composition. The space group of the stable structure of a given composition X was predicted based on machine learning. The assignment of Wyckoff letters to the constituent atoms was narrowed down using a predictive model trained on a given set of crystal structures in the Materials Project database, as described later. This predictive model enabled us to randomly generate promising Wyckoff patterns while eliminating wasted search space. By restricting the Wyckoff-site multiplicity and symmetry, 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 with the objective of predicting and limiting the space group of the stable crystalline state for a given composition X. We compiled a list of the chemical compositions and 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 restricted to the top K_{SG} candidates (where K_{SG} was set as 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 (See Methods for additional details).

The transfer-learning-based energy predictor was then used for exhaustive virtual screening using each of the two generators separately. Finally, using DFT calculations, the promising structures that exhibited the lowest predicted energies were optimized with the Vienna Ab initio Simulation Package (VASP, version 6.1.2)[51] combined with projector

Table 1 | Results of ShotgunCSP-GT and ShotgunCSP-GW generators for the 40 crystals comprising Dataset I

Composition	Number of atoms	Space group	ShotgunCSP-GW	ShotgunCSP-GT
C	4	$R\bar{3}m$	✓ (✓ / ✓)	✓ (✓ / 37)
Si	2	$Fd\bar{3}m$	✓ (✓ / ✓)	✓ (✓ / 16)
GaAs	2	$F\bar{4}3m$	✓ (✓ / ✓)	✓ (✓ / 27)
ZnO	4	$P6_3mc$	✓ (✓ / ✓)	✓ (✓ / 141)
BN	4	$P6_3/mmc$	✓ (✓ / ✓)	✓ (✓ / 137)
$LiCoO_2$	16	$R\bar{3}m$	✓ (✓ / ✓)	✓ (✓ / 136)
Bi_2Te_3	5	$R\bar{3}m$	✓ (✓ / ✓)	✓ (✓ / 58)
$Ba(FeAs)_2$	5	$I4/mmm$	✓ (✓ / ✓)	✓ (✓ / 27)
SiO_2	6	$I\bar{4}2d$	✓ (✓ / ✓)	× (✓ / 12)
VO_2	6	$P4_2/mnm$	✓ (✓ / ✓)	✓ (✓ / 38)
La_2CuO_4	7	$I4/mmm$	× (✓ / ✓)	✓ (✓ / 33)
$LiPF_6$	8	$R\bar{3}$	✓ (✓ / ✓)	✓ (✓ / 12)
Al_2O_3	10	$R\bar{3}c$	✓ (✓ / ✓)	✓ (✓ / 91)
$SrTiO_3$	10	$I4/mcm$	✓ (✓ / ✓)	✓ (✓ / 108)
$CaCO_3$	10	$R\bar{3}c$	✓ (✓ / ✓)	✓ (✓ / 105)
TiO_2	12	$C2/m$	× (× / ✓)	× (✓ / 43)
ZrO_2	12	$P2_1/c$	✓ (✓ / ✓)	✓ (✓ / 44)
$ZrTe_5$	12	$Cmcm$	✓ (✓ / ✓)	✓ (✓ / 32)
V_2O_5	14	$Pmmn$	✓ (✓ / ✓)	× (× / 43)
Si_3N_4	14	$P6_3/m$	✓ (✓ / ✓)	✓ (✓ / 43)
Fe_3O_4	14	$Fd\bar{3}m$	✓ (✓ / ✓)	✓ (✓ / 47)
$Mn(FeO_2)_2$	14	$Fd\bar{3}m$	✓ (✓ / ✓)	✓ (✓ / 135)
ZnSb	16	$Pbca$	✓ (✓ / ✓)	✓ (✓ / 67)
$CoSb_3$	16	$Im\bar{3}$	✓ (✓ / ✓)	✓ (✓ / 4)
$LiBF_4$	18	$P3_121$	✓ (✓ / ✓)	✓ (✓ / 26)
Y_2Co_{17}	19	$R\bar{3}m$	✓ (✓ / ✓)	✓ (✓ / 2)
GeH_4	20	$P2_12_12_1$	✓ (✓ / ✓)	✓ (✓ / 48)
$CsPbI_3$	20	$Pnma$	× (× / ✓)	✓ (✓ / 170)
$NaCaAlPHO_5F_2$	24	$P2_1/m$	× (× / ✓)	–
$LiFePO_4$	28	$Pnma$	✓ (✓ / ✓)	✓ (✓ / 107)
$Cu_{12}Sb_4S_{13}$	29	$I\bar{4}3m$	✓ (✓ / ✓)	✓ (✓ / 2)
MgB_7	32	$Imma$	× (✓ / ✓)	× (× / 9)
Li_3PS_4	32	$Pnma$	✓ (✓ / ✓)	× (× / 64)
Cd_3As_2	80	$I4_1/acd$	✓ (✓ / ✓)	✓ (✓ / 32)
$Li_4Ti_5O_{12}$	42	$C2/c$	✓ (✓ / ✓)	✓ (✓ / 12)
$Ba_2CaSi_4(BO_7)_2$	46	$I\bar{4}2m$	× (× / ✓)	× (× / 12)
Ag_8GeS_6	60	$Pna2_1$	× (✓ / ✓)	✓ (✓ / 7)
$Nd_2Fe_{14}B$	68	$P4_2/mnm$	× (× / ✓)	✓ (✓ / 2)
$Y_3Al_5O_{12}$	80	$Ia\bar{3}d$	✓ (✓ / ✓)	✓ (✓ / 11)
$Ca_{14}MnSb_{11}$	104	$I4_1/acd$	× (✓ / ✓)	✓ (✓ / 2)
Overall			31/40 = 77.5%	33/40 = 82.5%

The top $5 \times K_{temp}$ or top $10 \times K_{SG}$ (where $K_{SG} = 30$) virtual structures with the lowest surrogate energies from ShotgunCSP-GT and ShotgunCSP-GW, respectively, were selected for DFT structural relaxation. The second column indicates the number of atoms in the primitive unit cells. In the fourth and fifth columns, symbols ✓ and × denote prediction success and failure, respectively, and dashes (–) indicate cases for which there was no template for element substitution or for which the calculation failed. For the fourth column, the left and right sides of the parentheses indicate the success or failure of Wyckoff-letter generation and space-group prediction, respectively. For the fifth column, the left and right sides of the parentheses indicate whether a sufficiently similar template structure ($r \le 0.2$) was found and the total number of templates selected (K_{temp}), respectively.

augmented wave (PAW) pseudopotentials[52] (see Methods for details). The top K lowest-energy structures were subjected to structural relaxation with DFT. In this study, we selected $K = 5 \times K_{temp}$ and $K = 10 \times K_{SG}$ structures for CSP using ShotgunCSP-GT and ShotgunCSP-GW, respectively. Generally, the top K candidates with minimized energies had similar structures, many of which converged to the same crystal structure during structural relaxation. To eliminate this redundancy, we considered structural similarity when selecting the top K candidate structures to maintain high structural diversity (see Methods).

Benchmark sets

The performance of the proposed CSP algorithm was evaluated on three benchmark sets. The first consisted of 40 stable crystals (Dataset I; Table 1) that were selected based on a literature survey using two criteria: (i) diversity of space groups, constituent elements, number of atoms, and element species; and (ii) diversity of applications such as battery and thermoelectric materials. Because the selection of crystal structures for Dataset I may have been biased due to the manual selection method, a second dataset of 50 stable crystals was randomly selected from the Materials Project database (Dataset II; Table 2). For Datasets I and II, the numbers of atoms in the unit cells of the selected crystals were 2–104 (mean ± standard deviation: 23.13 ± 24.09) and 2–288 (32.68 ± 45.41), respectively (Fig. S1). Of the benchmark crystals in Datasets I and II, 30% had more than 30 atoms; owing to the computational complexity and search performance, solving these structures was expected to be difficult with conventional heuristic searches based on iterative first-principles calculations.

As a more challenging benchmark, we randomly selected 30 stable structures from the Materials Project database for which no template exists (Dataset III; Table S1). Most of the crystal structures in Dataset III had considerably more atoms per unit cell (22–152; 66.50 ± 34.40) than those in Datasets I and II.

Space-group prediction

During virtual screening with ShotgunCSP-GW, 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 (Fig. 3(a)). To train and test the classifier, we used 33,040 instances of chemical compositions with stable crystalline states, including 213 distinct space groups, compiled from the Materials Project database. The 17 remaining space groups were not registered in the 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[33,43,44,46] (see Methods), 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 remaining instances were used for testing. To statistically evaluate the prediction accuracy, training and testing were repeated independently 100 times. Details of the model construction, including hyperparameter adjustment, are provided 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 ($K_{SG} \in \{1, \ldots, 40\}$). The average recall rates in the top 1, 10, 30, and 40 predictions were 60.22(±0.87)%, 85.35(±0.54)%, 92.61(±0.49)%, and 94.02(±0.43)%, respectively. This indicates that focusing on the top 30 predicted labels allows us to identify the space groups for 92.61(±0.49)% of the various crystalline systems. Using this model, the 213 space groups were narrowed down to the top 30 candidates, and for each of the selected candidates, a set of symmetry-restricted crystal structures was generated using ShotgunCSP-GW.

To examine the differences in prediction performance by space group, we visualized the relationship between the top 30 recall rates and the training

Table 2 | Results of ShotgunCSP-GT and ShotgunCSP-GW generators for the 50 crystals comprising Dataset II

Composition	Number of atoms	Space group	ShotgunCSP-GW	ShotgunCSP-GT
CsCl	2	$Fm\bar{3}m$	✓ (✓ / ✓)	✓ (✓ / 52)
MnAl	2	$P4/mmm$	✓ (✓ / ✓)	✓ (✓ / 37)
HoHSe	3	$P\bar{6}m2$	✓ (✓ / ✓)	✓ (✓ / 71)
$ErCdRh_2$	4	$Fm\bar{3}m$	✓ (✓ / ✓)	✓ (✓ / 113)
Eu_2MgTl	4	$Fm\bar{3}m$	✓ (✓ / ✓)	✓ (✓ / 114)
Pm_2NiIr	4	$Fm\bar{3}m$	✓ (✓ / ✓)	✓ (✓ / 113)
VPt_3	4	$I4/mmm$	✓ (✓ / ✓)	✓ (✓ / 60)
$Gd(SiOs)_2$	5	$I4/mmm$	✓ (✓ / ✓)	✓ (✓ / 28)
$LaAl_3Au$	5	$I4mm$	✓ (✓ / ✓)	✓ (✓ / 44)
U_2SbN_2	5	$I4/mmm$	✓ (✓ / ✓)	✓ (✓ / 120)
$MnGa(CuSe_2)_2$	8	$I\bar{4}$	✓ (✓ / ✓)	✓ (✓ / 11)
SmZnPd	9	$P\bar{6}2m$	✓ (✓ / ✓)	✓ (✓ / 2)
$Sn(TePd_3)_2$	9	$I4mm$	× (× / ×)	✓ (✓ / 96)
V_5S_4	9	$I4/m$	✓ (✓ / ✓)	✓ (✓ / 17)
Cs_3InF_6	10	$Fm\bar{3}m$	✓ (✓ / ✓)	✓ (✓ / 6)
$Eu(CuSb)_2$	10	$P4/nmm$	✓ (✓ / ✓)	✓ (✓ / 26)
$Rb_2TlAgCl_6$	10	$Fm\bar{3}m$	✓ (✓ / ✓)	✓ (✓ / 3)
$Ca_3Ni_7B_2$	12	$R\bar{3}m$	✓ (✓ / ✓)	✓ (✓ / 18)
$DyPO_4$	12	$I4_1/amd$	✓ (✓ / ✓)	✓ (✓ / 138)
LaSiIr	12	$P2_13$	✓ (✓ / ✓)	✓ (✓ / 33)
$SmVO_4$	12	$I4_1/amd$	✓ (✓ / ✓)	✓ (✓ / 136)
VCl_5	12	$P\bar{1}$	✓ (✓ / ✓)	✓ (✓ / 32)
YbP_5	12	$P2_1/m$	× (✓ / ✓)	✓ (✓ / 3)
$Eu(Al_2Cu)_4$	13	$I4/mmm$	✓ (✓ / ✓)	✓ (✓ / 1)
Zr_4O	15	$R\bar{3}$	× (× / ×)	× (× / 16)
$Ba_3Ta_2NiO_9$	15	$P\bar{3}m1$	✓ (✓ / ✓)	× (× / 39)
$K_2Ni_3S_4$	18	$Fddd$	✓ (✓ / ✓)	✓ (✓ / 78)
$Sr(ClO_3)_2$	18	$Fdd2$	✓ (✓ / ✓)	✓ (✓ / 76)
$LiSm_2IrO_6$	20	$P2_1/c$	× (✓ / ✓)	✓ (✓ / 83)
Pr_2ZnPtO_6	20	$P2_1/c$	× (✓ / ✓)	✓ (✓ / 85)
$Sc_2Mn_{12}P_7$	21	$P\bar{6}$	✓ (✓ / ✓)	✓ (✓ / 7)
$LaSi_2Ni_9$	24	$I4_1/amd$	✓ (✓ / ✓)	✓ (✓ / 2)
$CeCu_5Sn$	28	$Pnma$	✓ (✓ / ✓)	✓ (✓ / 2)
$LiP(HO_2)_2$	32	$Pna2_1$	× (✓ / ✓)	✓ (× / 99)
$Mg_3Si_2H_4O_9$	36	$P6_3cm$	✓ (× / ×)	× (× / 1)
$Y_4Si_5Ir_9$	36	$P6_3/mmc$	✓ (✓ / ✓)	–
$Na(WO_3)_9$	37	$R\bar{3}$	✓ (✓ / ✓)	–
$Sm_6Ni_{20}As_{13}$	39	$P\bar{6}$	✓ (✓ / ✓)	✓ (✓ / 2)
$BaCaGaF_7$	40	$P2/c$	✓ (✓ / ✓)	✓ (✓ / 6)
$Tm_{11}Sn_{10}$	42	$I4/mmm$	✓ (✓ / ✓)	✓ (✓ / 3)
$AlH_{12}(ClO_2)_3$	44	$R\bar{3}c$	✓ (× / ×)	✓ (✓ / 11)

Table 2 (continued) | Results of ShotgunCSP-GT and ShotgunCSP-GW generators for the 50 crystals comprising Dataset II

Composition	Number of atoms	Space group	ShotgunCSP-GW	ShotgunCSP-GT
$K_2ZrSi_2O_7$	48	$P2_1/c$	✓ (✓ / ✓)	✓ (× / 32)
$LiZr_2(PO_4)_3$	72	$P2_1/c$	× (× / ✓)	✓ (✓ / 21)
$K_5Ag_2(AsSe_3)_3$	76	$Pnma$	✓ (✓ / ✓)	–
$Be_{17}Ru_3$	80	$Im\bar{3}$	✓ (× / ✓)	✓ (✓ / 1)
$Cu_3P_8(S_2Cl)_3$	80	$Pnma$	× (× / ✓)	✓ (✓ / 2)
Al_2CoO_4	84	$P3m1$	× (× / ×)	✓ (✓ / 15)
$Li_6V_3P_8O_{29}$	92	$P1$	× (✓ / ✓)	✓ (× / 5)
$ReBi_3O_8$	96	$P2_13$	✓ (✓ / ✓)	✓ (✓ / 7)
$Na_5FeP_2(O_4F)_2$	288	$Pbca$	× (× / ✓)	–
Overall			39/50 = 78.0%	43/50 = 86.0%

See the caption of Table 1 for details.

data size (Fig. S2). The 25th, 50th, and 75th percentiles of the space-group-specific recall rates were 53.05%, 72.62%, and 84.25%, respectively. Thus, the variability in recall rates was partially correlated with the number of training instances for each space group.

The distribution of composition ratios in the dataset was highly skewed, raising the concern that the prediction performance may vary greatly from one composition ratio to another. To address this concern, we conducted a robustness test by evaluating changes in the prediction accuracy upon varying the upper limit on the number of samples with the same composition ratio in the training set. The results were similar to those in Fig. 3(b), demonstrating that the prediction performance did not vary significantly from one composition ratio to another (Fig. S3).

Wyckoff pattern prediction

After narrowing down the space groups with machine learning, Wyckoff letters were randomly assigned to each atom to generate atomic coordinates. As the number of combinations of atoms and Wyckoff letters increases, the complexity of de novo CSP grows. For example, for the space group *lmma* (No. 74) of 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 1755. If the assignments are incorrect, CSP usually fails. 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 possible assignments exist. Successfully narrowing down the space groups and accurately predicting the assignments of Wyckoff letters is, therefore, expected to improve the CSP task significantly.

This led us to construct a model to predict the occurrence frequencies of Wyckoff letters for stable structures based on the chemical composition (Fig. 4(a)). A model was created for each space group, with the input chemical composition represented by a 290-dimensional descriptor using XenonPy. The output is the probability distribution of the occurrence of Wyckoff letters. Using the 33,040 instances of stable structures in the Materials Project database (excluding the 120 benchmark crystals; 80% were used for training and 20% for testing), 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 the actual relative frequencies $q_1, \ldots, q_M$ of M Wyckoff letters were measured based on the Kullback–Leibler (KL) divergence:

$$\sum_{i=1}^{M} q_i \log \frac{q_i}{p_i} \quad (1)$$

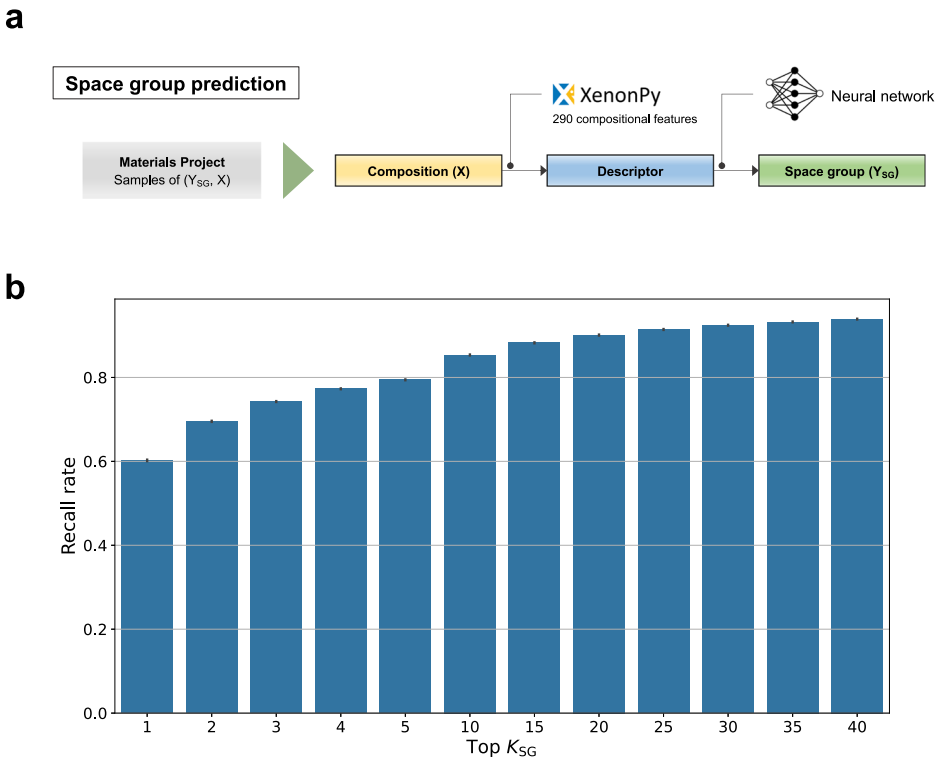


Fig. 3 | Space-group prediction. a Machine-learning workflow. **b** Change in recall rate across the top 1 to the top 40 predictions of the space group. Narrowing down to the top 30 predicted labels should, on average, enable the inclusion of the true space groups for 92.61% of the entire set of crystalline systems.

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

Using the Wyckoff-letter occurrence probability for a query composition, we randomly assigned Wyckoff letters using the procedure in the Methods section. The possible assignment of Wyckoff letters to elements is constrained by their multiplicity and query composition ratio. The sampling algorithm was designed to satisfy these constraints; the frequency of generated Wykoff labels generally matched the probabilities predicted by the random forest regressors.

Figure 4(c) shows the differences in the Wyckoff patterns generated with and without the Wyckoff-letter predictor by comparing the frequencies of actual and sampled Wyckoff letters for six randomly selected crystals. In all cases, the frequencies of Wyckoff letters generated from the predictor agreed well with the true frequencies. By contrast, those of randomly generated Wyckoff letters deviated significantly from the true frequencies.

Energy prediction

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

Note that the global model is unsuitable for predicting the energies of different randomly generated conformations for each composition X (Fig. 5(c)). In addition, it failed to discriminate between the energies of different randomly generated conformations for the 90 benchmark crystals. We tested the predictive capability of the global model on the DFT energies of 100 randomly generated pre-relaxed conformations for each of the 90 benchmark crystals. The average MAE decreased to 6.126 eV/atom with a standard deviation of 2.010. A similar result was obtained when the global model was trained with approximately 1,021,917 instances from the OQMD, including the formation energies of both relaxed and unrelaxed structures.

To overcome this limited predictive ability, the pretrained global model was localized to the target system X by transfer learning. For each X, the formation energies of a maximum of 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 Fig. 5(d), transfer learning improved the prediction performance for the formation energies of the 9,000 additional conformations. The average MAE reached 0.488 eV/atom with a standard deviation of 0.453 eV/atom. This is a 12.6-fold improvement compared to that of the pretrained global energy prediction model.

CSP using the library generator based on ShotgunCSP-GT

We used the fine-tuned surrogate energy predictor to sort the virtual crystals generated by ShotgunCSP-GT, narrowed them down to the top five structures for each template as described above, and then performed structural relaxation using DFT. The J relaxed structures with the lowest DFT energies were used as the final set of predicted structures. Figure 7 shows the top two (J = 2) predictions and the true structures for some selected cases. Figure S6 shows the top two predicted structures for all 120 benchmark crystals. Tables 1 and 2 summarize the success or failure of the top $5 \times K_{\mathrm{temp}}$ predictions for all crystal systems in Datasets I and II. The accuracies for Datasets I and II were 82.5% and 86.0%, respectively. No significant differences were observed in the accuracies between the two benchmark sets. Interestingly, increasing the number of atoms in the unit cell did not degrade the performance. Among the failure cases, five instances ($NaCaAlPHO_5F_2$, $K_{20}Ag_8As_{12}Se_{36}$, $Na_1W_9O_{27}$, $Na_{80}Fe_{16}P_{32}O_{128}F_{32}$, and $Y_8Si_{10}Ir_{18}$) did not have template structures in the Materials Project database with identical composition ratios. Excluding the five cases without a template, the accuracies for Datasets I and II reached 84.6% and 93.5%, respectively. For the other nine failure cases, for which a template was available but prediction failed, none of the Wyckoff-letter patterns of the true structures were included in the template structure set in the Materials Project database.

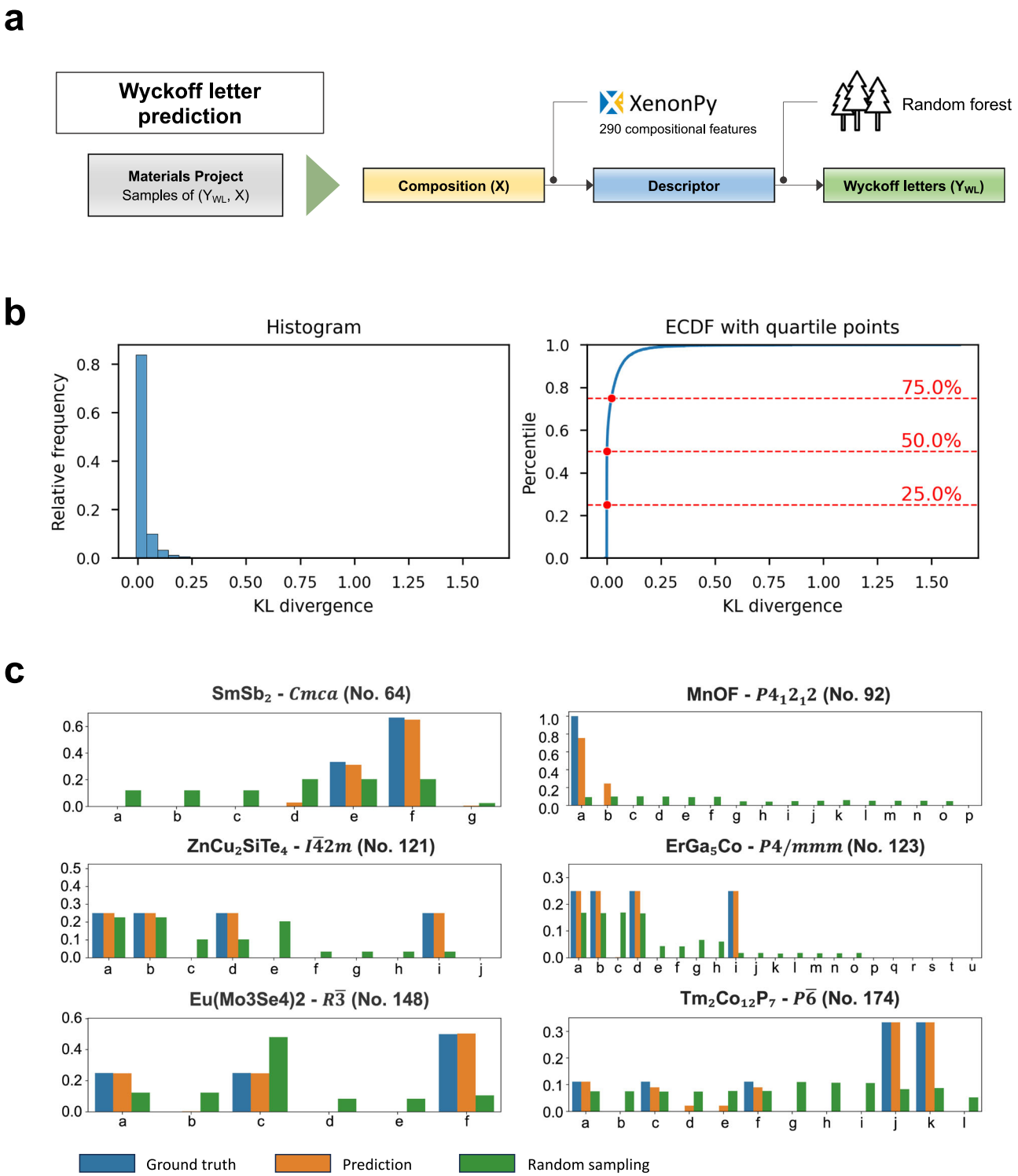


Fig. 4 | Prediction of Wyckoff-letter assignments. a Machine-learning workflow. **b** Histogram and empirical cumulative distribution function (ECDF) of KL divergence between the relative occurrence frequencies and predicted probability distributions of Wyckoff letters for the test set. **c** Histograms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

In summary, crystal systems with one or more replaceable template structures can be largely predicted by substituting the elements in existing crystals. For example, in the Materials Project database, the proportion of crystals with one or more interchangeable template structures was 98.0%. A similar conclusion was reached with another element-substitution-based CSP machine-learning algorithm called CSPML[36]. The prediction accuracy of ShotgunCSP-GT for Datasets I and II was higher than that of CSPML, which was 65.6% for the top ten predictions (see Table S4 for the testing results and the Supplementary Information for an outline of the methods).

CSP using ShotgunCSP-GW

The top ten candidate structures predicted using ShotgunCSP-GW with the lowest surrogate energies for each predicted space group were selected for structural relaxation using DFT ($J = 10 \times K_{SG}$ with $K_{SG} = 30$). Figure 7 displays the top two predicted and true structures for some selected

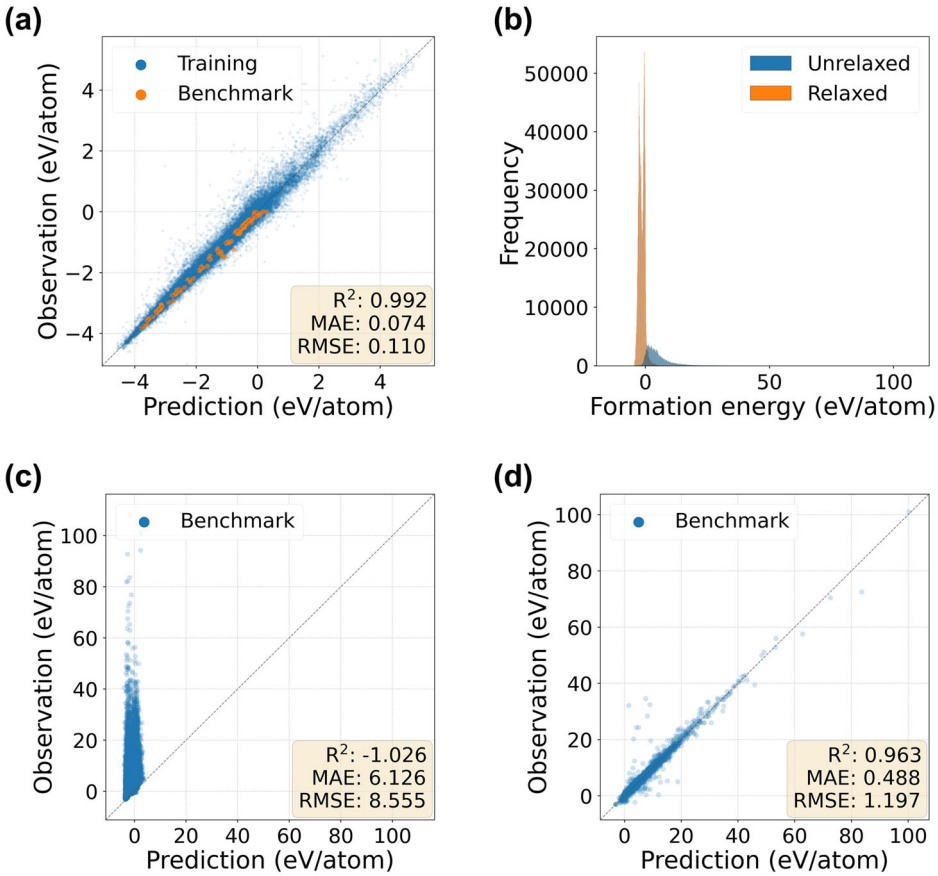


Fig. 5 | Performance of CGCNN for the prediction of DFT formation energies with and without transfer learning. The root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) for test instances are shown on each parity plot. **a** Results of global model for the prediction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

examples, and Fig. S6 shows the top two predicted structures for all 120 benchmark crystals. For the top ten predicted structures, 77.5% and 78.0% of the known stable structures were accurately predicted for Datasets I and II, respectively. Tables 1 and 2 summarize the performance (success or failure) of the top ten predictions for all crystal systems in Datasets I and II, respectively. The overall performance was lower than that of ShotgunCSP-GT. One reason for the decreased accuracy is the failure of space-group prediction. Specifically, the model failed to predict the space groups for approximately 5% of the 90 benchmark crystal structures in Datasets I and II. This is almost the same level of accuracy as reported above.

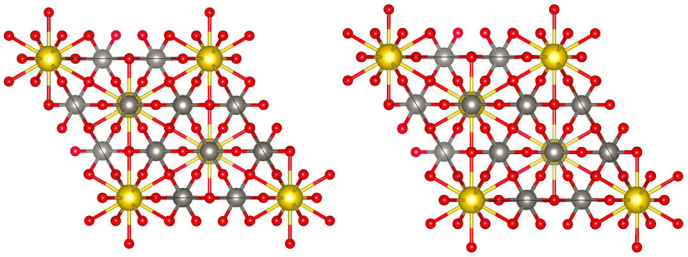
ShotgunCSP-GW successfully predicted 31 and 38 of the crystals in Datasets I and II, respectively. To highlight the prediction mechanism of the proposed method, we focus here on three crystals that were successfully predicted but for which no template exists in the Materials Project database: $Y_4Si_5Ir_9$, $K_5Ag_2(AsSe_3)_3$, and $Na(WO_3)_9$. These compounds all have relatively large numbers of atoms in the unit cells (36, 76, and 111, respectively). Nevertheless, for the space group $R\bar{3}$ (No. 148) of the stable structure of $Na(WO_3)_9$ (Fig. 6(a)), because the Wyckoff letters {*a*, *b*, *d*, *e*} are coordinate-fixed, the number of possible combinations of Wyckoff letters is reduced to approximately 48 owing to its multiplicity constraints. Consequently, the effective dimensions of the search space could be reduced by considering the crystal symmetry. This explains why ShotgunCSP-GW successfully predicted the complex stable structure of $Na(WO_3)_9$. For $K_5Ag_2(AsSe_3)_3$, which has 76 atoms in its unit cell (space group *Pnma*; No. 62) (Fig. 6(b)), the possibility of replacing the Wyckoff letter {*c*} with {*a*} or {*b*} increases the number of possible Wyckoff-letter combinations to over 300, even when considering multiplicity constraints. Nevertheless, ShotgunCSP-GW achieved success, primarily because the occurrence probability of Wyckoff letters {*a*} and {*b*} was successfully predicted to be extremely low by the Wyckoff-letter assignment predictor. This significantly narrowed the extensive search space during candidate structure generation, highlighting the significance of the Wyckoff-letter assignment refinement strategy in the success of the CSP task. The same mechanism facilitated the prediction of $Y_4Si_5Ir_9$.

On the other hand, of the 85 crystals for which the space groups were correctly identified, the ShotgunCSP-GW could not predict 9 and 11 of the true stable structures of crystals in Datasets I and II, respectively. Of these failure cases, true Wyckoff pattern generation failed in 3 and 3 cases for Datasets I and II, respectively, while ground-truth generation failed in 6 and 8 cases, respectively, despite correct true Wyckoff pattern generation. To elucidate the origin of these failures, we examined the generated structures in detail. 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, 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 *C*2/*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 an extensive search space. This was the case for TiO_2 (full formula: Ti_8O_{16}); despite the successful prediction of the Wyckoff-letter configuration Ti: 4*i*, Ti: 4*i*, O: 4*i*, O: 4*i*, O: 4*i*, O: 4*i*, the generation of precise atomic coordinates was unsuccessful (Fig. 6(c)).

Many of the predicted structures that were determined to be failures have been experimentally reported as metastable structures (Fig. 7). For example, for TiO_2[53], the anatase type is the most stable structure according to first-principles calculations, whereas the predicted structure was the rutile type, which is a known metastable state of TiO_2. Similarly, the true stable structure of Si_3N_4[54] is the hexagonal structure, β-Si_3N_4, whereas the

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



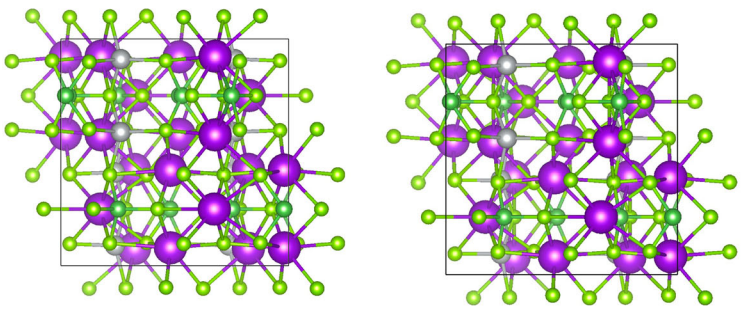
True Prediction

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



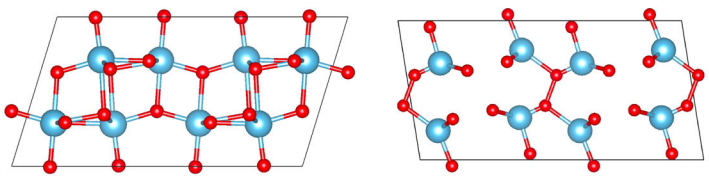
True Prediction

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), …

Fig. 6 | Three crystal structures predicted with ShotgunCSP-GW: $Na(WO_3)_9$ (full formula: $Na_3W_{27}O_{81}$, space group: $R\bar{3}$), $K_5Ag_2(AsSe_3)_3$ (full formula: $K_{20}Ag_8As_{12}Se_{36}$, space group: *Pnma*), and TiO_2 (full formula: Ti_8O_{16}, space group: *C2/m*). a For $Na(WO_3)_9$, the number of possible Wyckoff-letter combinations is limited to 48 when the space group is considered. **b** 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. **c** For Ti_8O_{16}, which ShotgunCSP-GW 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.

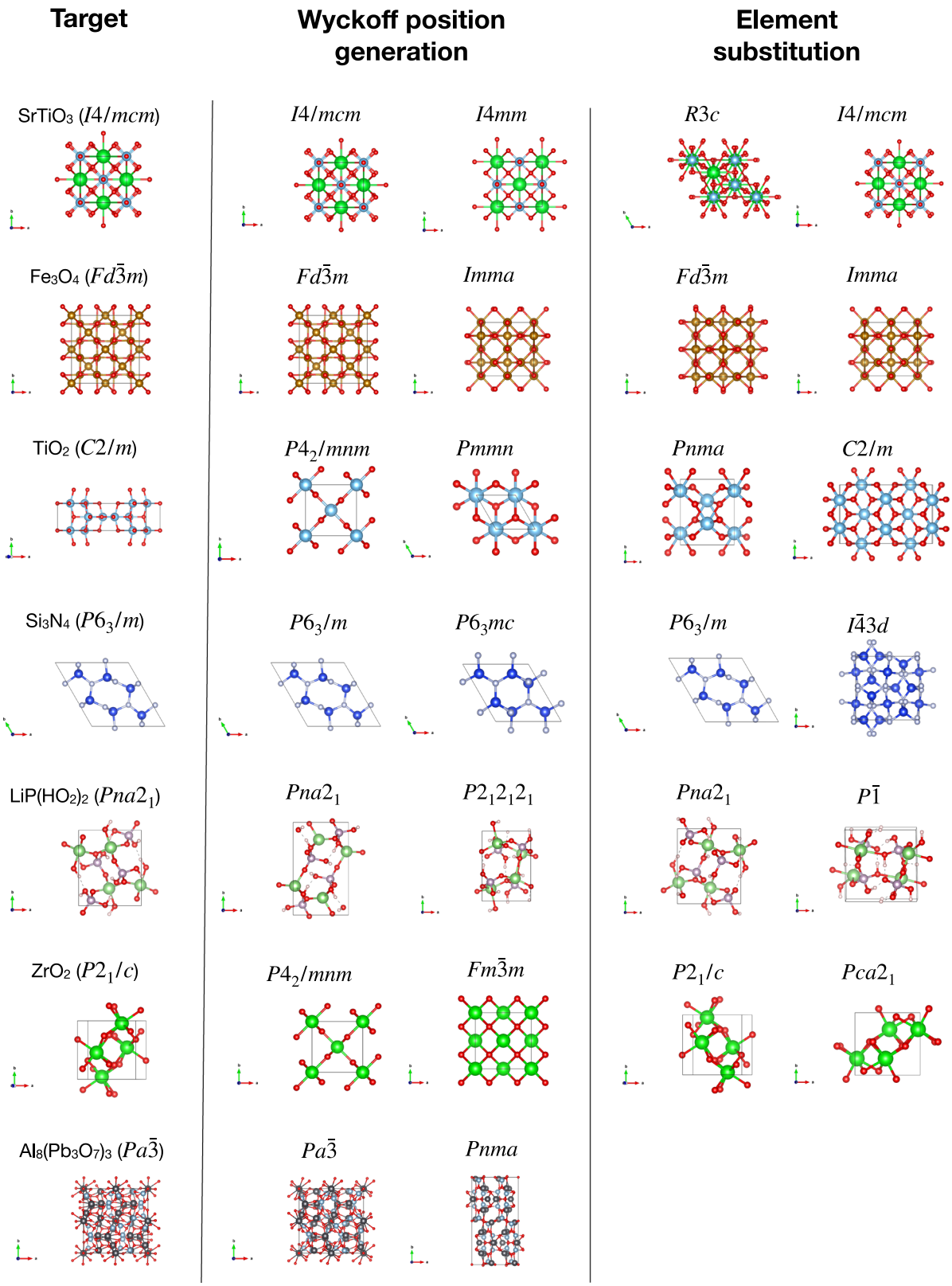


Fig. 7 | Examples of crystal structures predicted by the proposed CSP algorithms (depicted with VESTA[62] version 3.5.8). For each generator (ShotgunCSP-GT and ShotgunCSP-GW), the predicted structures with the two lowest DFT energies are shown. The true (target) stable structures are shown on the left.

predicted structure was the willemite-II type, which has been identified as a metastable structure of Si_3N_4 by DFT calculations. In many cases, even the predicted crystal structures that were judged to be failures partially captured structural features that were similar to those of the true structures. For example, the predicted structures of ZrO_2 and $LiP(HO_2)_2$ did not exactly match the true structures, but differed only slightly in atomic positions (Fig. 7). The energy difference between the true and predicted structures of these compounds was <5 meV/atom. Although the prediction of stable structures for low-symmetry compounds was often not entirely accurate, metastable structures or partial structural patterns can be predicted using this method.

For the cases where the prediction failed, the energy differences between the predicted and ground-truth structures are briefly discussed in the Supplementary Information. Although there were a few cases where the energy of the predicted structure was lower or almost the same as that of the ground-truth structure, we decided not to delve deeper into these cases.

We also evaluated the predictive performance of our method in quite challenging scenarios using Dataset III, which resulted in a notably low accuracy of 6.7% (see Table S1). Despite this outcome, the rather complex crystal structures of $Al_8(Pb_3O_7)_3$ and $Mg_{10}B_{16}Ir_{19}$ were accurately predicted. However, predicting crystal structures for more complex systems would require additional computational resources and enhanced methodologies.

Comparison with USPEX

The CSP tasks for Datasets I and II were conducted using USPEX by applying the calculation conditions outlined in the Methods section, in which the true space groups were given as initial parameters. 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[55]. The calculation of each crystal structure 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 to 76 otherwise.

Using the settings described in the Supplementary Information, only tasks involving small-unit-cell systems (comprising approximately 20 or fewer atoms in the primitive unit cell) were accomplished successfully. Specifically, within the allocated computational resources, 13 and 12 systems were completed for Datasets I and II, respectively. For the completed tasks, USPEX had prediction accuracies of 92.3% and 91.7% for the benchmark crystals in Datasets I and II, respectively (Table S5). The median number of structure relaxation calculations performed was 167, with computations taking 37.7 h on the designated supercomputer system. In comparison, our method utilizing the ShotgunCSP-GW generator for the same set of 25 benchmark structures yielded accuracies of 84.6% and 83.3% for Datasets I and II, respectively (Table S5). The median number of structure relaxation calculations performed was 172, with computations taking 21.4 h. While USPEX was conducted with the true space groups used to create initial random structures, all other space groups were sampled during the USPEX run. By contrast, our method searched for the top 30 space groups. Thus, our noniterative approach, utilizing an energy predictor to identify structures with sufficiently low energy followed by structure relaxation, reached almost the same level of accuracy as USPEX while significantly improving the computational speed.

Discussion

This paper presents a CSP workflow based on a machine-learning approach for the efficient prediction of stable crystal structures without iterative DFT calculations. The essence of the proposed method is a shotgun-type virtual screening of crystal structures. A surrogate model that predicts DFT energies is used to screen a large number of virtual crystal structures. The efficiently narrowed-down candidate structures are then relaxed using DFT calculations to predict stable crystal structures. The key technical components in this workflow are the surrogate model for energy prediction and the crystal structure generators. To train the surrogate model for DFT energy calculations, a pretrained CGCNN is fine-tuned for the prediction of the energies of virtual crystal structures in pre-relaxed states to decrease the number of training samples generated with DFT single-point energy calculations. Virtual libraries of candidate crystal structures are constructed by either element substitution of template crystals (ShotgunCSP-GT) or a Wyckoff generator involving space-group prediction and Wyckoff-label prediction (ShotgunCSP-GW). Of 90 known crystal structures (Datasets I and II) with a wide range of chemical compositions, symmetries, and structure types, the ShotgunCSP-GT- and ShotgunCSP-GW-based workflows successfully predicted 84.4% and 74.4% of the true structures, respectively.

The performance of the CSP algorithms was evaluated using a benchmark set of 120 cases (Datasets I–III). Although various CSP algorithms have been developed, they have often been assessed only in terms of success or failure in individual cases, with few comprehensive performance evaluations. In response, efforts to develop a benchmark set for CSP algorithms have recently emerged. For instance, CSPBench[56] consists of 180 cases of varying difficulty levels, carefully selected from the Materials Project database. Importantly, comprehensive testing with CSPBench has shown that most existing methods fall short of satisfactory predictive performance. Our benchmark cases and reported prediction performance are expected to facilitate systematic evaluations in subsequent research and developments of CSP algorithms.

For the 25 benchmark structures successfully predicted by USPEX, our ShotgunCSP-GW-based method achieved the same or better prediction accuracy and reduced the computation time by approximately 40%. To our knowledge, our method is the simplest CSP algorithm available today. A significant contribution of this study is in proving that such a straightforward approach can effectively predict numerous crystal structures that cannot be predicted using conventional methods, including those with low symmetry and large-unit-cell systems. However, our method remains incapable of predicting more complex crystal systems, such as those in Dataset III, which were selected to be more challenging. The bottleneck is expected to be partially eliminated by achieving the efficient prediction of the true Wyckoff labels of large and complicated crystalline systems.

Nonetheless, the simplicity of the proposed method makes it well-suited for parallel computing and managing a much larger database of candidate structures. Moreover, it can be integrated with deep generative models for crystal structure prediction, which are expected to advance further in the future.

Methods

ShotgunCSP-GT: template-based structure generation by element substitution

The calculation procedure for ShotgunCSP-GT is as follows:

1. Extract template structures 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 templates with elements that have the same number of atoms in X. If the substitution target is not uniquely determined, substitute the element that has 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[49,50], remove structurally redundant templates to obtain a unique template set (where K_{temp} is the number of unique templates in the same cluster as the query composition).
5. Estimate the lattice constants using a model that predicts the unit-cell volume from the composition X (see Supplementary Information).
6. Add perturbations to the atomic coordinates of each template with additive noise following the uniform distribution $U(-0.05, 0.05)$.
7. Add perturbations to the unit-cell volumes of each template with additive noise following the uniform distribution $U(-0.1, 0.1)$.

ShotgunCSP-GW: Wyckoff position generator

The calculation procedure for ShotgunCSP-GW is as follows:

1. Predict the space group and probabilities of Wyckoff letters of the query composition $X = X_{c_1}^1 X_{c_2}^2 \cdots X_{c_K}^K$ (where X^k denotes a chemical element $k = 1, \ldots, K$ and c^k is its composition ratio) for 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 X^k from composition X and a possible Wyckoff letter l_i with the predicted Wyckoff-letter-configuration probability p_i from set W, then assign l_i to X^k with its possible multiplicity m_i.

4. Remove the assigned atoms from composition X and define the remaining composition as new X.
5. Remove the used Wyckoff letter l_i from the set W if the Wyckoff position is exclusive, and re-normalize the probability of the remaining Wyckoff letters.

$$p_i \leftarrow p_i / \sum p_i$$

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

Compositional descriptor

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

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

The scalar quantity $\eta(X^k)$ on the right-hand side of Eq. (2) (where $k = 1, \ldots, K$) represents a feature value of the element X^k, such as the atomic weight, electronegativity, or polarizability. Using function g, the element features $\eta(X^1)$, …, $\eta(X^K)$ with compositions c^1, …, c^K are converted into a compositional feature. For g, we used five different summary statistics: weighted mean ϕ_{ave}, weighted variance ϕ_{var}, weighted sum ϕ_{sum}, max-pooling ϕ_{max}, and min-pooling ϕ_{min}:

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

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

Fine-tuning of CGCNN

A CGCNN localized to the energy prediction of a specific system with composition X was obtained by fine-tuning the pretrained global CGCNN model from Xie and Grossman[30] with randomly generated crystalline conformations and their DFT formation energies. We generated 100 training crystal structures for each candidate space group using the ShotgunCSP-GW generator or ten random structures from each selected template using the ShotgunCSP-GT generator. The pretrained model without the output layer was copied to the target model, to which a new output layer was subsequently added, and its parameters were randomly initialized. We then trained the target model on the target dataset. The hyperparameters, including the learning rate and gradient clipping value, were optimized by performing a grid search with the same range {0.01, 0.008, 0.006, 0.004, 0.002}, with early stopping based on the MAE of the validation set. The maximum number of epochs was fixed at 350.

DFT calculations

All DFT calculations were performed using VASP (version 6.1.2)[51] with PAW pseudopotentials[52]. The Perdew–Burke–Ernzerhof exchange–correlation functional[57] was considered for generalized gradient approximation. Brillouin zone integration of the unit cells was automatically determined using the Γ-centered Monkhorst–Pack mesh function implemented in the VASP code. Single-point energy calculations (also known as self-consistent field calculations) were performed on unrelaxed crystal structures that were created virtually 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[49] with significant modifications to generate the inputs for all VASP calculations (see Supplementary Information).

Structural similarity

To calculate the similarity between two structures, we encoded each query into a vector-type structural descriptor with its local coordination information (site fingerprint) from all sites[58]. Then, the structural similarity τ was calculated as the Euclidean distance between the two vectorized crystal structures. Note that the descriptor does not contain any information about the element species. The calculations were performed in matminer[59], which is an open-source toolkit for materials data mining, with the same configuration as that used officially in the Materials Project database. We also visually inspected the differences between structures with different τ. Structures with a dissimilarity of $\tau \leq 0.2$ were treated as similar structures.

USPEX calculations

USPEX calculations were performed using the official USPEX package (version 10.5)[60]. We specified the calculation parameters `calculationMethod`, `calculationType`, and `optType` as "USPEX," "300," and "enthalpy," respectively, to perform the CSP task for bulk crystals using an evolutionary algorithm. The related parameters were specified according to official recommendations[61]. 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 as soon as the best structure remained unchanged over M generations, where $M = N$ (rounded to the nearest ten; see Supplementary Information). Structure relaxation was executed automatically in VASP (version 6.1.2) combined with the PAW pseudopotentials. The VASP calculation settings were the same as those described in the DFT calculations section.

Data availability

The benchmark datasets are published on Figshare and can be accessed via the following https://doi.org/10.6084/m9.figshare.26536375. Additional data supporting the findings of this study are available from the corresponding author upon reasonable request.

Code availability

The ShotgunCSP-GT generator is published as a Python package on the GitHub website at https://github.com/TsumiNa/ShotgunCSP.

Received: 2 May 2023; Accepted: 24 November 2024;
Published online: 20 December 2024

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 (2006).

3. Pickard, C. J. & Needs, R. J. High-pressure phases of silane. *Phys. Rev. Lett.* **97**, 045504 (2006).
4. Pickard, C. J. & Needs, R. J. Structure of phase III of solid hydrogen. *Nat. Phys.* **3**, 473–476 (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 (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 (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 (2001).
9. Wang, Y., Lv, J., Zhu, L. & Ma, Y. Crystal structure prediction via particle-swarm optimization. *Phys. Rev. B* **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 (2017).
11. Oganov, A. R., Lyakhov, A. O. & Valle, M. How evolutionary crystal structure prediction works—and why. *Acc. Chem. Res.* **44**, 227–237 (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 (2013).
13. Yamashita, T. et al. Crystal structure prediction accelerated by Bayesian optimization. *Phys. Rev. Mater.* **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 (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 (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 (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 (2012).
19. Rasmussen, C. E. & Williams, C. K. I.*Gaussian processes for machine learning* (The MIT Press, 2005).
20. Mockus, J.Bayesian approach to global optimization: Theory and applications, **37** of *Mathematics and Its Applications* (Springer Netherlands, Dordrecht, 1989).
21. Jain, A. et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
22. The Materials Project. https://next-gen.materialsproject.org/. Accessed: 2024-05-03.
23. Curtarolo, S. et al. AFLOW: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226 (2012).
24. AFLOW: Automatic - FLOW for Materials Discovery. http://www.aflowlib.org. Accessed: 2024-05-03.
25. Kirklin, S. et al. The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **1**, 15010 (2015).
26. OQMD: The Open Quantum Materials Database. https://oqmd.org. Accessed: 2024-05-03.
27. Merchant, A. et al. Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (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**, 185 (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 (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 (2016).
33. Yamada, H. et al. Predicting materials properties with little data using shotgun transfer learning. *ACS Cent. Sci.* **5**, 1717–1730 (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 (2011).
35. Wang, H.-C., Botti, S. & Marques, M. A. L. Predicting stable crystalline compounds using chemical similarity. *npj Comput. Mater.* **7**, 12 (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 (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 (2021).
39. Hu, J. et al. Contact map based crystal structure prediction using global optimization. *CrystEngComm* **23**, 1765–1776 (2021).
40. Kim, B., Lee, S. & Kim, J. Inverse design of porous materials using artificial neural networks. *Sci. Adv.* **6**, eaax9324 (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. Sect. B: Struct. Sci.* **68**, 215–226 (2012).
42. Lee, I.-H. & Chang, K. J. Crystal structure prediction in a continuous representative space. *Comput. Mater. Sci.* **194**, 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 (2020).
44. Liu, C. et al. Machine learning to predict quasicrystals from chemical compositions. *Adv. Mater.* **33**, 2102507 (2021).
45. Liu, C. et al. Quasicrystals predicted and discovered by machine learning. *Phys. Rev. Materials* **7**, 093805 (2023).
46. XenonPy platform. https://github.com/yoshida-lab/XenonPy. Accessed: 2024-05-03.
47. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–231 (AAAI Press, Portland, Oregon, 1996).
48. 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 (2017).
49. 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).
50. pymatgen StructureMatcher. https://pymatgen.org/pymatgen.analysis.html#pymatgen.analysis.structure_matcher.StructureMatcher. Accessed: 2024-05-03.
51. 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 (1996).
52. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).

53. 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 (2016).
54. Kroll, P. Pathways to metastable nitride structures. *J. Solid State Chem.* **176**, 530–537 (2003).
55. SQUID ((Supercomputer for Quest to Unsolved Interdisciplinary Datascience). http://www.hpc.cmc.osaka-u.ac.jp/en/squid/. Accessed: 2024-05-03.
56. Wei, L. et al. CSPBench: A benchmark and critical evaluation of crystal structure prediction. *arXiv [cond-mat.mtrl-sci]*, https://doi.org/10.48550/arXiv.2407.00733 (2024).
57. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
58. 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. *Front. Mater.* **4**, 34 (2017).
59. Ward, L. et al. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60–69 (2018).
60. USPEX downloads. https://uspex-team.org/en/uspex/downloads. Accessed: 2024-05-03.
61. USPEX manual. https://uspex-team.org/online_utilities/tmp/uspex_manual_release/EnglishVersion/uspex_manual_english/index.html. Accessed: 2024-05-03.
62. Momma, K. & Izumi, F. *VESTA 3* for three-dimensional visualization of crystal, volumetric and morphology data. *J. Appl. Crystallogr.* **44**, 1272–1276 (2011).

Acknowledgements

This work was supported in part by a Ministry of Education, Culture, Sports, Science and Technology (MEXT) KAKENHI Grant-in-Aid for Scientific Research on Innovative Areas (grant number 19H05820); Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (A) (grant number 19H01132) and Early-Career Scientists (grant number 23K16955); and JST CREST (grant numbers JPMJCR19I3, JPMJCR22O3, and JPMJCR2332). Computational resources were partly provided by the supercomputer at the Research Center for Computational Science, Okazaki, Japan (projects 23-IMS-C113 and 24-IMS-C107).

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 L.C. designed and developed the machine-learning framework. L.C. developed the software and performed the experiments with the support of H.T., T.Y., K.W., S.Y., M.K. and A.R.O. H.T. and S.Y. designed and tested the benchmark crystal structures. R.Y., L.C., H.T. and T.Y. wrote and revised the manuscript. All authors discussed the results and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-024-01471-8.

Correspondence and requests for materials should be addressed to Ryo Yoshida.