Inorganic Chemistry

pubs.acs.org/IC

Article

TCSP: a Template-Based Crystal Structure Prediction Algorithm for Materials Discovery

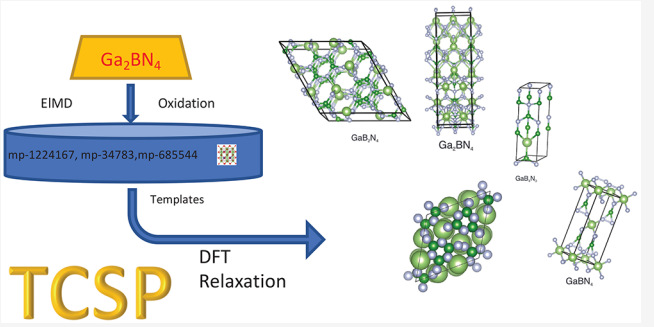
Lai Wei, Nihang Fu, Edirisuriya M. D. Siriwardane, Wenhui Yang, Sadman Sadeed Omee, Rongzhi Dong, Rui Xin, and Jianjun Hu*

Cite This: *Inorg. Chem.* 2022, 61, 8431−8439

Read Online

ACCESS | Metrics & More | Article Recommendations

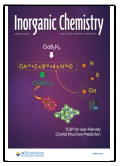
ABSTRACT: Fast and accurate crystal structure prediction (CSP) algorithms and web servers are highly desirable for the exploration and discovery of new materials out of the infinite chemical design space. However, currently, the computationally expensive first-principles calculation-based CSP algorithms are applicable to relatively small systems and are out of reach of most materials researchers. Several teams have used an element substitution approach for generating or predicting new structures, but usually in an ad hoc way. Here we develop a template-based crystal structure prediction (TCSP) algorithm and its companion web server, which makes this tool accessible to all materials researchers. Our algorithm uses elemental/chemical similarity and oxidation states to guide the selection of template structures and then rank them based on the substitution compatibility and can return multiple predictions with ranking scores in a few minutes. A benchmark study on the 98290 formulas of the Materials Project database using leave-one-out evaluation shows that our algorithm can achieve high accuracy (for 13145 target structures, TCSP predicted their structures with root-mean-square deviation < 0.1) for a large portion of the formulas. We have also used TCSP to discover new materials of the Ga−B−N system, showing its potential for high-throughput materials discovery. Our user-friendly web app TCSP can be accessed freely at www.materialsatlas.org/crystalstructure on our MaterialsAtlas.org web app platform.



1. INTRODUCTION

Crystal structure prediction (CSP) is increasingly becoming one of the most effective approaches for the discovery of new functional materials[1] because of the ease to obtain new compositions by either enumeration,[2] heuristic knowledge, or the latest deep-learning-based generative machine learning models.[3] While the peer protein structure prediction problem has recently been almost solved by the deep-learning-based AlphaFold and RossettaFold algorithms, the CSP problem remains elusive for a majority of categories of compositions. There are mainly three types of CSP approaches including the ab initio based global optimization[4−8] as reviewed in ref 9, machine-learning-based prediction,[7,10] and template-based elemental substitution.[11] The first approach instead depends on computationally expensive density functional theory (DFT) calculations and is applicable to only small chemical systems. The second approach is inspired by the AlphaFold family of deep-learning algorithms[12,13] but is only at the early stage of development. The last template-based CSP methods are the most widely used and easiest to implement. Even though this method cannot predict crystal structures of new prototypes, recent deep generative models can discover new prototype materials that can partially address this issue.[14] In a pioneering work,[11] Hautier et al. proposed a data-mining-based approach to identify the probabilities for different pairs of ionic substitutions, which can be applied to any prototype structures to generate new structures or used to select templates for template-based crystal structure prediction (TCSP). The difference of our algorithm from this substitution method is that their algorithm uses learned element substitution pairs to find formulas with existing structures that can be used as templates, without considering the overall composition similarity and oxidation state compatibility. Despite the wide usage of TCSP methods, there are many different ways to implement, and there is no working web app/server that is user-friendly enough to make it accessible to all materials scientists (the structure predictor of the Materials Project[15] web site was previously available, but it is not functional now).

Received: December 13, 2021
Published: April 14, 2022



https://doi.org/10.1021/acs.inorgchem.1c03879
Inorg. Chem. 2022, 61, 8431−8439

Here we propose a fast and user-friendly TCSP algorithm and related companion web server for broad adoption of CSP in the daily life of materials science. Our algorithm TCSP is based on the careful selection of template structures based on chemical formula similarity and the matching of oxidation states using an exhaustive enumeration strategy. Our predicted structures can be optimized by DFT or machine-learning-based structure relaxation. By using seven case studies, we have shown that our user-friendly and fast CSP web server has a high prediction performance when appropriate templates are available. We also apply our TCSP algorithm to predict the structures of all 98290 formulas using leave-one-out evaluation and have achieved good performances for a large portion of the targets: more than 13145 target structures have been found with maximum root-mean-square distances of less than 0.1. The good performance of this high-throughput CSP shows that the template/prototype-based element substitution CSP approach has big potential in exploratory materials discovery. With the development of large-scale prototype databases[16,17] and their applications in the generative design of new crystals,[18,19] the performance of our TCSP algorithm can be further improved.

2. METHOD

2.1. TCSP Algorithm. Our TCSP method is inspired by the fact that all known crystal structures, as deposited in ICSD and Materials Project databases, actually belong to a limited catalog of crystal structure prototypes,[16,20,21] each corresponding to one or more different compositions. This is because some ionic species substitutions for each other within these prototypes can retain the crystal structure, e.g., keeping the crystal symmetry/space group unchanged with minor adjustments of the unit cell constants or Wyckoff coordinates. The substitution patterns have been discovered using heuristics[22] and data-mining probabilistic models.[11] This general element substitution strategy has been used to find a variety of new compounds.[21,23,24]

Our TCSP algorithm is illustrated in Figure 1. Given an input formula (e.g., $SrTiO_3$), the user can choose to specify the expected space group number for predicted structures, which can be predicted by algorithms[25,26] or without constraining the space group. We then search structure templates with the same prototype (sometimes called an anonymous formula, e.g., ABC_3) and the same space group if specified. This step may retrieve too many matched templates, so we use Module A1, an Element's mover distance, to measure the composition similarity between the query formula and compositions of all of the template structures, which are then ranked by ascending order. We then pick the top K structures as template candidates with the smallest composition distances. For each of the candidate templates, we use the Pymatgen package to estimate its oxidation states and compare them to those of the query formula. If we find templates with identical oxidation states, we then add them to the final template list. If no such templates are found, we then neglect the oxidation match requirements and directly add them as final templates. The next step is to determine all of the possible element substitution pairs between the query and template formulas using the algorithm described in Algorithm 2. To further reduce the redundant template structures, we use the Pymatgen's StructureMatcher module to detect redundant (too similar) structure templates and keep only one for each such cluster, which can significantly remove duplicate similar

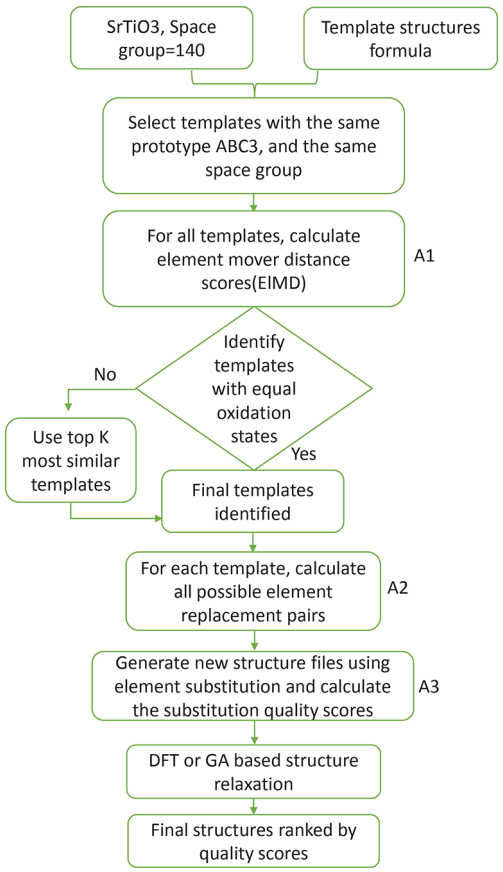


Figure 1. Flowchart of our TCSP. The space group specification is optional.

structure templates. Next, we will pick the template structure files and replace the elements according to the pair arrangements found by Algorithm 2. A replacement quality score is also calculated for each such element substitution arrangement using the procedure as described in Module 3. The resulting structures will then be subjected to DFT or machine-learning-based structure relaxation, which can be further used to calculate the formation energy, e-above-hull energy, and phonon dispersion for validation.

Module A1: Element Mover's Distance for Formula Similarity Calculation. We use the Element's Mover Distance measure ElMD[27] to select the most similar template structures. ElMD is a metric that allows measurement of the chemical similarity of two formulas in an explainable fashion. The ElMD is computed between two compositions from the ratio of each of the elements and the absolute distance between the elements on the modified Pettifor scale p (several other element similarities can also be used such as Mendeleev, Petti, Atomic, Mod_petti, Oliynyk, Oliynyk_sc, Jarvis, Jvarvis_sc, magpie,magpie_sc, CGCNN, Elemnet, mat2vec, Matscholar, megnet16, random). This metric shows clear strength in distinguishing compounds. It is shown that the ElMD distances have greater alignment with chemical understanding than the Euclidean distances. The ElMD is defined in formula (1).

$$\mathrm{ElMD}(X, Y) = \min \sum_{i=1}^{m} \sum_{j=1}^{n} q_{ij} |p_i - p_j|, \text{ subject to } q_{ij} \geq 0 \text{ for } \forall\ i,\ j \quad (1)$$

https://doi.org/10.1021/acs.inorgchem.1c03879
Inorg. Chem. 2022, 61, 8431−8439

subject to

$$\sum_{j=1}^{n} q_{ij} \le x_i, \quad \text{for } \forall\ 1 \le i \le m \tag{2}$$

$$\sum_{i=1}^{m} q_{ij} \le y_j, \quad \text{for } \forall\ 1 \le j \le n \tag{3}$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} q_{ij} = 1 \tag{4}$$

where the ElMD distance is first calculated by matching and pairing each of the m elements in an element fraction vector, X of formula A, to its most similar unmatched partner in the n elements of a second element fraction vector Y of formula B, until all have been paired. x_i indicates the fraction of the element at the ith position of formula A. y_j indicates the fraction of the element at the jth position of formula B. The quantity matched, q, from the ith element of X to the jth element of Y is given by q_{ij}. p_i and p_j are the element properties of the ith/jth element in formula A/B. Detailed calculation examples can be found in ref 27.

Algorithm A2: Element Replacement Pair Enumeration Algorithm. This algorithm is used to enumerate all possible element replacement strategies between two pairs of formulas.

Algorithm A2: Element replacement pair enumeration algorithm

Given two formulas X,Y, calculate theirs oxidation states then get the statelist and elementlist.
if $stateList_x == stateList_y$ **then**
    Replace element pairs
else
    Create elementGroupList to represent and distinguish elements of equal state
    for $i = 0, 1, \ldots$ **do**
        if $i == 0$ **then**
            $elementGroup = elementList_y[i]$;
        else if $(stateList_y[i] == stateList_y[i-1])$ **then**
            append $elementList_y[i]$ to $elementGroup$;
        else
            append $elementList_y[i]$ to $elementGroup$;
        end if
    end for
    Create $pre\_patterns$ to represent permutation and combination of elements in $elementGroupList$
    for $element = 0, 1, \ldots, j$ **do**
        if $j = 0$ **then**
            $pre\_patterns$ = permutations of $elementGroupList[j]$;
        else
            $patterns$ =permutations of $elementGroupList[j]$;
            for $p1 = 0, 1, \ldots, m$ **do**
                for $p2 = 0, 1, \ldots, n$ **do**
                    append $p1, p2$ to $new\_patterns$;
                end for
                $pre\_patterns = new\_patterns$;
            end for
        end if
    end for
    If the element in $pre\_patterns$ is different from the element in $elementList_x$, then replace the element
end if

Module A3: Element Substitution Scoring Function. To differentiate the resulting structures from the different element substitution arrangements between the query formula and template structure and rank the final output structures from different templates, we use the ElMD distance in Module A1 to calculate the similarity score between each pair of substitution elements for a given query formula and the final structure. Then we sum up these similarity scores for all of the element substitution pairs and use them to calculate the quality scores of the final structures.

$$S_{es} = \mathrm{ElMD}(e_1, e_2), \qquad S_r = \sum_{i=1}^{p} S_{es}^{i} \tag{5}$$

where S_{es} measures the element matching quality between two substitution elements e_1 and e_2. S_r is the replacement distance score for a given element substitution arrangement of a given query formula and template structure, which is equal to the sum of all of the element similarity scores of all substitution element pairs in the element substitution arrangement.

The final quality of the generated structures is then measured by the S_r scores, with lower scores corresponding to higher quality.

DFT or Genetic-Algorithm-Based Structure Relaxation. As with all predicted crystal structures, they usually need a fine-tuning or relaxation step to adjust the local atomic coordinates using either the DFT-based structural relaxation method or the recently developed machine-learning- and optimization-based relaxation approach,[28] which is much faster than the DFT approach. In this study, we used the DFT approach for evaluation purposes.

2.2. User Interface of Our Web Server. Our TCSP server has a user-friendly web interface, as shown in Figure 2.

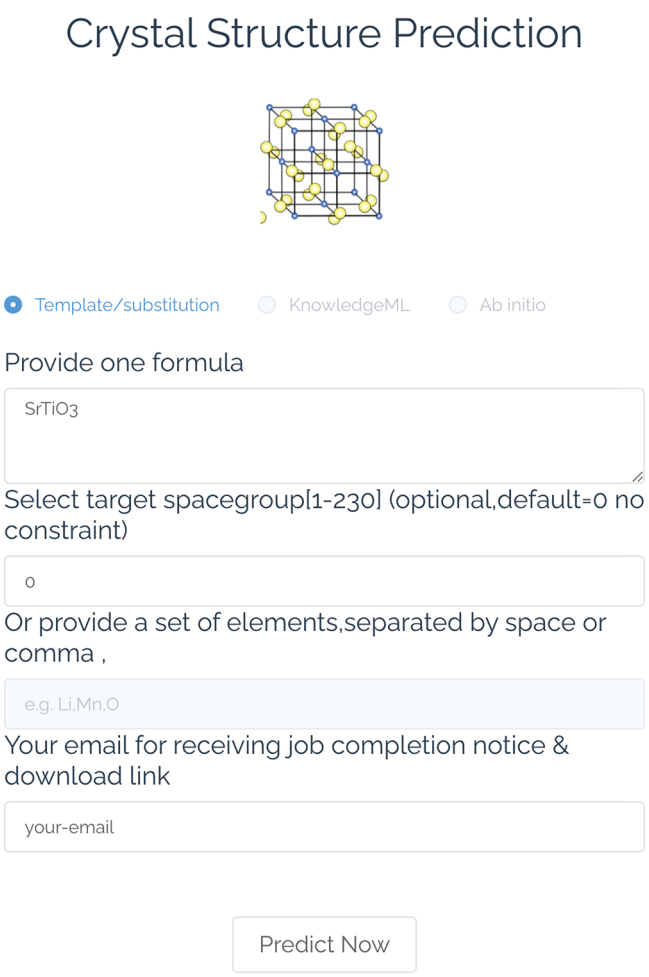


Figure 2. User interface of our TCSP web app for CSP.

Each time, a user can just put in a formula/composition, and then the target space group number from 1 to 230 can be set or just assigned to 0 to allow a template with any space group. Then the user types in their email in order to receive the job completion notification email with a downloadable URL link for the predicted structures. After a few minutes, an email will be sent to the user with the downloadable URL link for the

https://doi.org/10.1021/acs.inorgchem.1c03879
Inorg. Chem. 2022, 61, 8431−8439

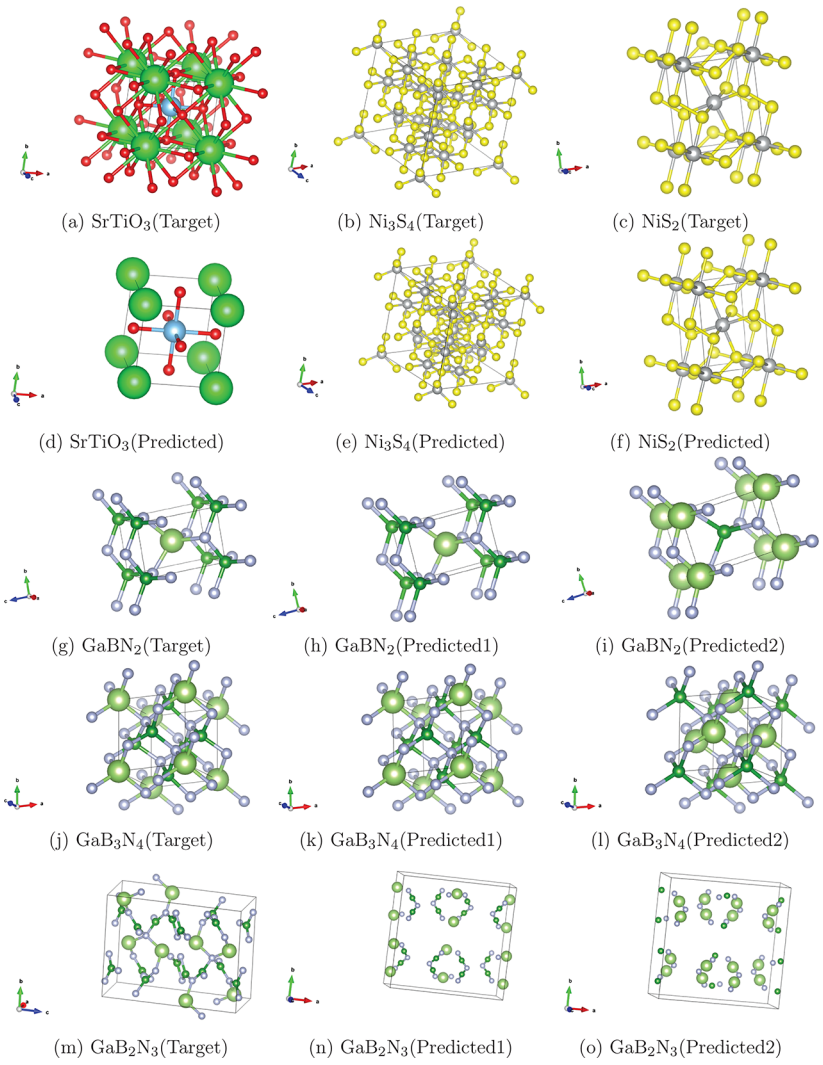


Figure 3. Structures predicted by TCSP compared to the targets: (a) $SrTiO_3$(Target), (b) Ni_3S_4(Target), (c) NiS_2(Target), (d) $SrTiO_3$(Predicted), (e) Ni_3S_4(Predicted), (f) NiS_2(Predicted), (g) $GaBN_2$(Target), (h) $GaBN_2$(Predicted), (i) $GaBN_2$(Predicted), (j) GaB_3N_4(Target), (k) GaB_3N_4(Predicted1), (l) GaB_3N_4(Predicted2), (m) GaB_2N_3(Target), (n) GaB_2N_3(Predicted1), and (o) GaB_2N_3(Predicted2).

predicted structures. After the zipped result file is downloaded and unzipped, the user can go into the folder and click the Name column to sort the files by filename. Then it shows several key files: (1) results.txt shows the template similarity scores, the templates with compatible oxidation states, and the element replace pairs for each template; (2) similar_formulas.csv file shows the distance scores of all templates to the query formula; (3) TemplateCandiates.csv shows the Materials Project IDs of the selected templates. (4) All of the remaining .cif files are predicted, which are sorted by their replacement quality score (the number before _mp of the filename), which is better when the number is smaller. However, it is strongly suggested to validate a couple of top-scored candidate structures because the candidate structure with the top quality score is not always the best one.

2.3. DFT Validation of Predicted Structures. The first-principles calculations based on DFT are carried out using the Vienna ab initio simulation package (*VASP*).[29−32] The projected-augmented-wave pseudopotentials, with 520 eV plane-wave cutoff energy, were used to treat the electron−ion interactions.[33,34] The exchange-correlation functional was considered with the generalized gradient approximation based on the Perdew−Burke−Ernzerhof method.[35,36] The energy convergence criterion was set as 10^{-5} eV, while the atomic positions were optimized with the force convergence criterion of 10^{-2} eV/Å. The Brillouin zone integration for the unit cells was computed using the Γ-centered Monkhorst−Pack k meshes. The formation energies (in eV/atom) of several materials were determined based on the expression in eq 6, where E[Material] is the total energy per unit formula of the considered structure, $E[A_i]$ is the energy of the ith element of the material, x_i indicates the number of A_i atoms in a unit formula, and n is the total number of atoms in a unit formula ($n = \sum_i x_i$).

$$E_{\text{form}} = \frac{1}{n}\left(E[\text{Material}] - \sum_i x_i E[A_i]\right) \tag{6}$$

2.4. Evaluation Criteria. To evaluate the reconstruction performance of the algorithm, we define the root-mean-square distance (RMSD) and mean absolute error (MAE) of two structures as

https://doi.org/10.1021/acs.inorgchem.1c03879
Inorg. Chem. 2022, 61, 8431−8439

Table 1. Prediction Performance (RMSD Error) of Top 10 Results for Each Sample in the Benchmark Set by TCSP

formula	metric	top 1	top 2	top 3	top 4	top 5	top 6	top 7	top 8	top 9	top 10
$SrTiO_3$	RMSD	0	0	0	0	0.1667	0.2832	0.4082	0.4410	0.4410	0.4410
	score	2	2	2	3	1	1	1	1	1	2
Ni_3S_4	RMSD	0.0007	0.2885	0.2888	0.2897						
	score	3	4	1	4						
NiS_2	RMSD	0.0049	0.0124	0.0777	0.2282	0.2424	0.2846				
	score	3	1	3	2	1	2				
$GaBN_2$	RMSD	0.0039	0.0039	0.0174	0.0209	0.3410	0.3412	0.3412	0.3886		
	score	3	1	2	1	9	1	5	31		
GaB_3N_4	RMSD	0.0023	0.0023	0.0152	0.0162	0.3335	0.3344	0.3344	0.3347	0.3347	
	score	1	3	22	1	1	1	3	4	0	
GaB_2N_3	RMSD	0.2475	0.2475								
	score	11	7								
Ga_3BN_4	RMSD	0.0040	0.0040	0.0206	0.3336	0.3337	0.3345	0.3345	0.3347	0.3347	
	score	1	3	1	1	22	1	3	0	4	

$$\mathrm{RMSD}(\mathbf{v}, \mathbf{w}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|v_i - w_i\right\|^2}$$

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[(v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2\right]} \quad (7)$$

$$\mathrm{MAE}(\mathbf{v}, \mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\left\|v_i - w_i\right\|$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\left\|v_{ix} - w_{ix}\right\| + \left\|v_{iy} - w_{iy}\right\| + \left\|v_{iz} - w_{iz}\right\|\right) \quad (8)$$

where n is the number of independent atoms in the target crystal structure. For symmetrized CIF structures, n is the number of independent atoms of the set of Wyckoff equivalent positions. For regular CIF structures, it is the total number of atoms in the compared structure. v_i and w_i are the corresponding atoms in the predicted crystal and target crystal structures, respectively.

In addition to the RMSD and MAE criteria based on the coordinates of the Wyckoff positions, other generic crystal structure similarity distances can also be used, including the geometry-based and symmetry-adapted similarity metrics, to compare the crystal structures.[37] Another possible evaluation method is to use the superpose3d-based RMSD,[38] which can translate and rotate the crystal structures (point clouds) to achieve maximum alignment, and then perform the RMSD calculation. This may provide a better measurement of the structural similarity.

3. RESULTS

3.1. Data Set. We used more than 130000 crystal structures deposited in the Materials Project database as our template sources. We picked seven test materials, including $SrTiO_3$ (mp-5229), Ni_3S_4 (mp-1050), NiS_2 (mp-849059), $GaBN_2$ (mp-1007823), GaB_3N_4 (mp-1019740), GaB_2N_3 (mp-1245554), and Ga_3BN_4 (mp-1019743). We then ran our algorithm and checked whether it could predict the correct structures that match the target structures.

3.2. Performance of TCSP. We selected seven formulas of target structures from the Materials Project database for evaluation of the capability of our TCSP algorithm for structure prediction. The first test target formula was $SrTiO_3$, which has three different phases corresponding to space groups of 140, 149, and 221. Its most famous structure is the cubic perovskite structure, as shown in Figure 3a. Our algorithm identified thousands of compatible templates and picked the top 10 as templates, including $BaZrO_3$ (mp-3834), $MgTiO_3$ (mp-1016830), $CaZrO_3$ (mp-542112), $MgZrO_3$ (mp-1017000), $BaTiO_3$ (mp-504715), $BaTiO_3$ (mp-5020), $SrZrO_3$ (mp-613402), $CaTiO_3$ (mp-5827), $BaTiO_3$ (mp-2998), and $SrHfO_3$ (mp-4551), among which all are cubic templates except for $BaTiO_3$ (mp-5020), which is a trigonal structure with space group 160. The top 4 predicted structures all have a zero RMSD error compared to the perovskite target structure: they all have fractional coordinates identical with those of the target structure except that the cubic lengths are different (the predicted cubic structure in Figure 3d has a lattice length of 4.256 Å, while the target structure has a lattice length of 3.945 Å), which may be fine-tuned using DFT-based relaxation.

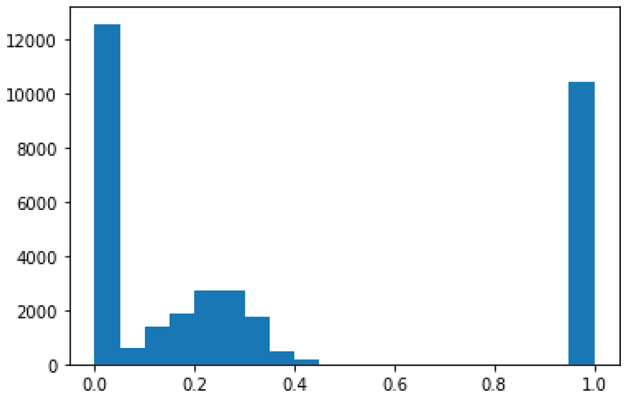


Figure 4. Distribution of RMSD errors of all Materials Project structures, as predicted by our TCSP using leave-one-out evaluation.

The second test sample is Ni_3S_4, which only has one cubic phase with space group 227. The structure is shown in Figure 3b. Our algorithm found the top 4 templates, including Co_3S_4 (mp-943), Co_3Se_4 (mp-20456), Ni_3Se_4 (mp-1120781), and Co_3O_4 (mp-18748), all of which are cubic structures with space group 227. The lowest RMSD is 0.000714, which is predicted by our algorithm using Co_3S_4 as the template. The RMSD errors of the structures from the other three templates are much larger, all around 0.288. We can see that the

https://doi.org/10.1021/acs.inorgchem.1c03879
Inorg. Chem. 2022, 61, 8431−8439

Table 2. Formation Energy and Corresponding Templates of the Top 10 Predictions for Each of the Four New Materials

GaB_2N_4		GaB_4N_3		Ga_2BN_4		$GaBN_4$	
mp-ID	E_{form} (eV)	mp-ID	E_{form} (eV)	mp-ID	E_{form} (eV)	mp-ID	E_{form} (eV)
mp-780282	−3.2957	mp-1224009	−2.4471	mp-532446	−2.9443	mp-30979	−3.29
mp-778103	−3.2414	mp-1225800	−2.2103	mp-698589	−2.8493	mp-20790	−2.9865
mp-13335	−3.2224	mp-1228436	−2.1092	mp-761314	−2.7696	mp-1224951	−2.7224
mp-780395	−3.1119	mp-1120750	−1.838	mp-5712	−2.5712	mp-1224810	−2.5462
mp-30161	−2.9124	mp-29672	−1.69	mp-1212041	−2.5703	mp-555538	−2.4165
mp-1194477	−2.4454	mp-1019378	−1.6865	mp-1255006	−2.5669	mp-1071955	−2.4012
mp-756317	−2.2645	mp-1228943	−0.3167	mp-765466	−2.5623	mp-1102285	−2.3984
mp-36866	−2.2639	mp-29672	−0.2741	mp-753397	−2.5622	mp-1071955	−2.398
mp-1208866	−2.1376	mp-1019508	−0.1409	mp-756649	−2.5621	mp-27462	−2.3971
N/A	N/A	mp-1223879	−0.0784	mp-1178203	−2.5621	mp-1102285	−2.3967

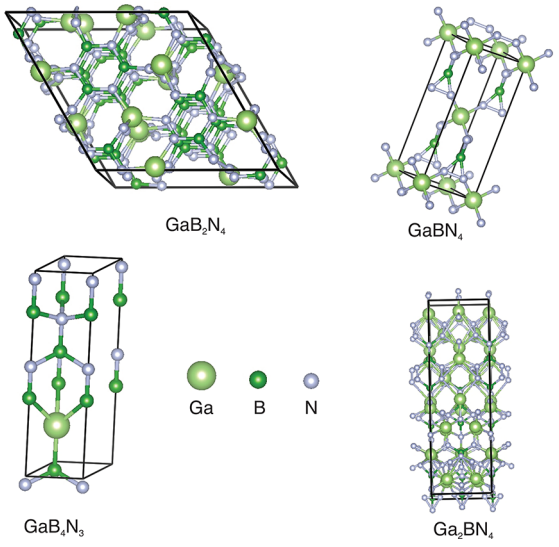


Figure 5. New Candidate structures with zero e-above-hull energy for GaB_2N_4 (using template mp-780282), GaB_4N_3 (using template mp-1224009), $GaBN_4$ (using template mp-30979), and Ga_2BN_4 (using template mp-698589).

predicted structure in Figure 3e is very close to the target structure in Figure 3b. We also found that the predicted structure of NiS_2 in Figure 3f also matches well with the target structure in Figure 3c, which has the smallest RMSD error of 0.004918.

We also tested four formulas of the chemical system Ga−B−N, including $GaBN_2$, GaB_3N_4, Ga_2BN_3, and Ga_3BN_4. For $GaBN_2$ with space group 115, the best eight templates are found by our algorithm. Five of them, including $AlBN_2$ (mp-1008557), $AlBN_2$ (mp-1008557), $AlGaP_2$ (mp-1228888), $AlGaN_2$ (mp-1008556), and B_2AsP (mp-1008528), have the same tetragonal crystal system as the target structure with space group 115. The remaining are trigonal with space group 166. The predicted structures $AlBN_2$ (mp-1008557) and $AlBN_2$ (mp-1008557) have the same lowest RMSD error of 0.003889. However, they have different structure patterns, as shown in Figure 3h,i. For the same template, our algorithm suggests two element replacement strategies. In the first strategy, the element Ga in the test formula is used to replace the element Al in the template $AlBN_2$; in the second strategy, the element Ga in the test formula $GaBN_2$ replaces the element B in the same template $AlBN_2$. For the formula GaB_3N_4 with space group 215, TCSP finds the top 9 most similar templates, AlB_3N_4 (mp-1019379), AlB_3N_4 (mp-1019379), $CrGa_3P_4$ (mp-985440), $AlGa_3N_4$ (mp-1019508), Al_3GaN_4 (mp-1019378), Al_3BN_4 (mp-1019380), Al_3BN_4 (mp-1019380), Ga_3BN_4 (mp-1019743), and Ga_3BN_4 (mp-1019743), of the same space group 215 as well. The lowest RMSD error with different structure templates AlB_3N_4 and AlB_3N_4 is 0.002336 by using two element replacement strategies. The element Ga in the first strategy is used to replace Al in the template, as shown in Figure 3k, and as shown in Figure 3l, Ga replaces B in AlB_3N_4 in the second strategy. For the formula GaB_2N_3 (mp-1245554) in Figure 3m, which has a monoclinic structure with space group 15, TCSP only finds two templates, AuC_2N_3 (mp-1245653) and AuC_2N_3 (mp-1245653), with the same space group. The lowest RMSD is 0.24746 for these two predicted structures. As shown in Figure 6a, our algorithm uses the first strategy, in which TCSP uses Ga in the test formula GaB_2N_3 to replace Au in the first template AuC_2N_3. In the second strategy, Ga replaces B in the second template, as shown in Figure 3o. For the formula Ga_3BN_4 (mp-1019743) with space group 215, TCSP finds the 9 most similar templates, Al_3BN_4 (mp-1019380), Al_3BN_4 (mp-1019380), Al_3GaN_4 (mp-1019378), $AlGa_3N_4$ (mp-1019508), $CrGa_3P_4$ (mp-985440), AlB_3N_4 (mp-1019379), AlB_3N_4 (mp-1019379), GaB_3N_4 (mp-1019740), and GaB_3N_4 (mp-1019740), with the same space group 215. The templates Al_3BN_4 (mp-1019380) and Al_3BN_4 (mp-1019380) with different structure patterns have the same lowest RMSD error of 0.004017781. Al_3BN_4, AlB_3N_4, and GaB_3N_4 all have different structure patterns using the two element replacement strategies.

Our TCSP algorithm can output multiple predictions using different templates. To understand this capability, Table 1 shows the RMSD and quality scores of the top 10 predictions for each input formula. For $SrTiO_3$, the top 4 structures all have zero RMSD errors for their fractional coordinates, with their replacement distance scores ranging from 2 to 3. The MAE errors of these four structures are also 0. We do find that their lattice lengths are different from those of the target structures, which, however, can be tuned by DFT-based structure relaxation. For Ni_3S_4, only the top 1 result is very close to the target structure with three much worse results. For NiS_2, the top 2 predicted structures have RMSD values of 0.0049 and 0.0124. Both are good compared to the target structures. For $GaBN_2$, the top 4 results all have small RMSD errors ranging from 0.0039 to 0.0209. The same is true for the predicted structures of GaB_3N_4. The worst prediction performance is on the formula GAB_2N_3, which can only find two compatible templates, both leading to structures very different from those of the target structure. Their MAE errors are 0.1853. For GA_3BN_4, the top 3 predictions are all of high

https://doi.org/10.1021/acs.inorgchem.1c03879
Inorg. Chem. 2022, 61, 8431−8439

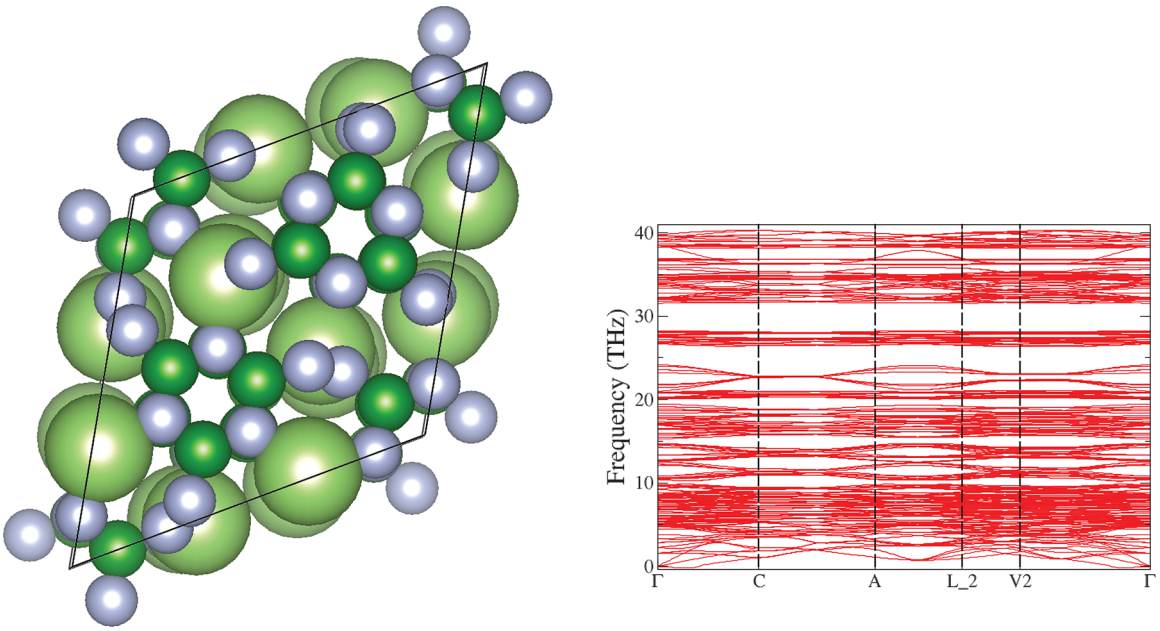


Figure 6. New material GaB_2N_4 discovered by our TCSP with zero e-above-hull energy, negative formation energy (−3.2957 eV), and dynamical stability: (a) predicted structure of GaB_2N_4; (b) phonon dispersion of GaB_2N_4 with the mp-780282 template structure.

quality, with RMSD values of only 0.004, 0.004, and 0.0206. In terms of the quality score distribution, we found that low replacement distance scores indicate predicted structures with good quality: for example, predictions with replacement distance scores greater than 5 are all low-quality results. However, low replacement distance scores do not always mean their structures are of high quality. For example, in the case of $SrTiO_3$, the structures with high RMSD values have lower replacement distance scores than those of the top 4 results.

To further evaluate the performance of our TCSP algorithm, we conducted comprehensive predictions of all 98290 formulas in the Materials Project database using the leave-one-out evaluation approach. For each formula, we predicted its structure using existing templates that did not have the same formula. Here we used the strict mode for finding the templates: only the top 10 templates with the same prototype and compatible oxidation states were used to predict new structures. For each formula, we first identified all of its corresponding mp-IDs and structures, and then for each of these target structures, we picked the structure with the lowest RMSD error out of all of the predicted structures and showed the distribution of these RMSD errors to see how our TCSP algorithm can recover the structures in the Materials Project database. The results are shown in Figure 4. We found that, for 34569 Materials Project structures, our algorithm identified templates for structure prediction. Out of these target structures, TCSP predicted hypothetical structures with a maximum RMSD of less than 0.01 for 11764 Materials Project structures or with a RMDS of less than 0.1 for 13145 Materials Project structures. We also found that, for 10433 structures, the algorithm could not find the correct templates that generate the same number of atoms in the unit cell for which we set the RMSD error at 1.0.

3.3. Discovery of New Materials and DFT Validation of the Predicted Structures. We were interested in how our TCSP algorithm could help to discover novel stable materials. We started with the Ga−B−N chemical system, for which we found four materials in the Materials Project database, as shown in our test set discussed in the previous section. According to the Materials Project database, those available Ga−B−N structures have nonzero energy-above-hull values, indicating that those materials are thermodynamically unstable. We wondered if thermodynamically and dynamically stable materials of this chemical system exist. We used the composition enumeration tool from our MaterialsAtlas.com toolbox to identify new Ga−B−N formula prototypes and their predicted formation energies. We picked the top 41 formulas that do not exist in the Materials Project database and used our TCSP to predict a set of candidate structures for each formula. We then conducted DFT relaxation and formation energy, e-above-hull energy, and phonon dispersion calculations to verify their thermodynamical and dynamical stability.

Our calculations (Table 2) showed that almost all candidate structures found by our TCSP have negative formation energies. This is immensely helpful for the discovery of new materials using first-principles calculations. If most of the candidate structures of a selected composition have positive formation energies, we have to waste computational hours to find the structures with negative formation energies. However, our ML model is able to filter out the unsuitable candidate structures to reduce the computational burden.

We further calculated the e-above-hull energy to investigate the stability against the Ga−B−N competing phases. As given in the Materials Project database, GaN, BN, Ga, B, and N2 are the stable competing phases that are available. Total energy calculations of the competing phases were done with the same *VASP* settings as those used for the Ga−B−N systems to determine the e-above-hull energy using the Pymatgen code. Our calculations suggest that 4 out of 41 materials exhibit zero e-above-hull energy, indicating that they are thermodynamically stable (Figure 5). Those materials and their candidate structures are shown in Table 2. We further carried out

https://doi.org/10.1021/acs.inorgchem.1c03879
Inorg. Chem. 2022, 61, 8431−8439

phonon calculations for the candidate structures with the lowest formation energies for those four materials. It is clear from Figure 6b that the GaB_2N_4 material with the mp-780282 template structure is dynamically stable at 0 K temperature. It has an interesting layered honeycomb structure.

4. CONCLUSION

CSP plays a key role in new materials discovery.[1] However, large-scale fast prediction of crystal structures is challenging, and user-friendly web apps are missing for such an important function despite the availability of some public software that needs expensive high-performance computing resources and an expertise of computational materials. We believe such fast CSP web apps are critical to the materials science community, as demonstrated by the bioinformatics field, which has more than 9000 web servers.[39] Here we propose a TCSP algorithm and its companion web server for fast and quick CSP. Because of the widely observed structure similarity across many materials families such as perovskites in the materials database, TCSP achieves a strong prediction performance, as benchmarked on the whole Materials Project structure using leave-one-out evaluation due to its flexible template selection algorithm using prototype and oxidation information. To our knowledge, this is the largest experiment for CSP. We believe our web-based TCSP algorithm will be of great interest to materials scientists for exploratory materials discovery. To further improve the speed of our algorithm, we will reduce the redundancy of the template structures by using only representative structural prototypes.[16,20]

AUTHOR INFORMATION

Corresponding Author

Jianjun Hu − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States;* orcid.org/0000-0002-8725-6660; Email: jianjunh@cse.sc.edu

Authors

Lai Wei − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States*

Nihang Fu − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States*

Edirisuriya M. D. Siriwardane − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States;* orcid.org/0000-0001-8960-5273

Wenhui Yang − *School of Mechanical Engineering, Guizhou University, Guiyang 550055, China*

Sadman Sadeed Omee − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States;* orcid.org/0000-0002-1016-8072

Rongzhi Dong − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States*

Rui Xin − *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29201, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.inorgchem.1c03879

Author Contributions

conceptualization, J.H.; methodology, J.H., L.W., W.Y., E.M.D.S., N.F., S.O., and R.D.; software, J.H., L.W., W.Y., and N.F.; resources, J.H.; writing—original draft preparation, J.H., L.W., E.M.D.S., N.F., and R.X.; writing—review and editing, J.H. and L.W.; visualization, J.H., L.W., and E.M.D.S.; supervision, J.H.; funding acquisition, J.H.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The research reported in this work was supported, in part, by National Science Foundation (NSF) under Grants 1940099, 1905775, and 2110033. The views, perspectives, and content do not necessarily represent the official views of the NSF. We thank Andrew Hughes for his help in proofreading the manuscript.

REFERENCES

(1) Oganov, A. R.; Pickard, C. J.; Zhu, Q.; Needs, R. J. Structure prediction drives materials discovery. *Nature Reviews Materials* **2019**, *4*, 331−348.

(2) Davies, D. W.; Butler, K. T.; Jackson, A. J.; Morris, A.; Frost, J. M.; Skelton, J. M.; Walsh, A. Computational screening of all stoichiometric inorganic materials. *Chem.* **2016**, *1*, 617−627.

(3) Dan, Y.; Zhao, Y.; Li, X.; Li, S.; Hu, M.; Hu, J. Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials. *npj Comput. Mater.* **2020**, *6*, 1−7.

(4) Glass, C. W.; Oganov, A. R.; Hansen, N. USPEX—Evolutionary crystal structure prediction. *Computer physics communications* **2006**, *175*, 713−720.

(5) Wang, Y.; Lv, J.; Zhu, L.; Lu, S.; Yin, K.; Li, Q.; Wang, H.; Zhang, L.; Ma, Y. Materials discovery via CALYPSO methodology. *J. Phys.: Condens. Matter* **2015**, *27*, 203203.

(6) Lee, I.-H.; Oh, Y. J.; Kim, S.; Lee, J.; Chang, K. J. Ab initio materials design using conformational space annealing and its application to searching for direct band gap silicon crystals. *Comput. Phys. Commun.* **2016**, *203*, 110−121.

(7) Lee, I.-H.; Chang, K. J. Crystal structure prediction in a continuous representative space. *Comput. Mater. Sci.* **2021**, *194*, 110436.

(8) Falls, Z.; Avery, P.; Wang, X.; Hilleke, K. P.; Zurek, E. The XtalOpt Evolutionary Algorithm for Crystal Structure Prediction. *J. Phys. Chem. C* **2021**, *125*, 1601−1620.

(9) Yin, X.; Gounaris, C. E. Search methods for inorganic materials crystal structure prediction. *Current Opinion in Chemical Engineering* **2022**, *35*, 100726.

(10) Hu, J.; Zhao, Y.; Yang, W.; Song, Y.; Siriwardane, E.; Li, Y.; Dong, R. AlphaCrystal: Contact map based crystal structure prediction using deep learning. *arXiv preprint arXiv:2102.01620* 2021.

(11) Hautier, G.; Fischer, C.; Ehrlacher, V.; Jain, A.; Ceder, G. Data mined ionic substitutions for the discovery of new compounds. *Inorganic chemistry* **2011**, *50*, 656−663.

(12) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; žídek, A.; Nelson, A. W.; Bridgland, A.; et al. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577*, 706−710.

(13) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.

(14) Zhao, Y.; Al-Fahdi, M.; Hu, M.; Siriwardane, E.; Song, Y.; Nasiri, A.; Hu, J. High-throughput discovery of novel cubic crystal materials using deep generative neural networks. *Adv. Sci.* **2021**, *8*, 2100566.

(15) Ceder, G.; Persson, K. *Materials Project: A materials genome approach* **2010**.
(16) Su, C.; Lv, J.; Li, Q.; Wang, H.; Zhang, L.; Wang, Y.; Ma, Y. Construction of crystal structure prototype database: methods and applications. *J. Phys.: Condens. Matter* **2017**, *29*, 165901.
(17) Hicks, D.; Toher, C.; Ford, D. C.; Rose, F.; De Santo, C.; Levy, O.; Mehl, M. J.; Curtarolo, S. AFLOW-XtalFinder: a reliable choice to identify crystalline prototypes. *npj Comput. Mater.* **2021**, *7*, 1−20.
(18) Bushlanov, P. V.; Blatov, V. A.; Oganov, A. R. Topology-based crystal structure generator. *Comput. Phys. Commun.* **2019**, *236*, 1−7.
(19) Sorkun, M. C.; Astruc, S.; Koelman, J. V. A.; Er, S. An artificial intelligence-aided virtual screening recipe for two-dimensional materials discovery. *npj Computat. Mater.* **2020**, *6*, 1−10.
(20) Mehl, M. J.; Hicks, D.; Toher, C.; Levy, O.; Hanson, R. M.; Hart, G.; Curtarolo, S. The AFLOW library of crystallographic prototypes: part 1. *Comput. Mater. Sci.* **2017**, *136*, S1−S828.
(21) Griesemer, S. D.; Ward, L.; Wolverton, C. High-throughput crystal structure solution using prototypes. *Physical Review Materials* **2021**, *5*, 105003.
(22) Olaf Muller, R. R. *The major ternary structural families*; Springer-Verlag: New York, 1974.
(23) Hautier, G.; Fischer, C. C.; Jain, A.; Mueller, T.; Ceder, G. Finding nature's missing ternary oxide compounds using machine learning and density functional theory. *Chem. Mater.* **2010**, *22*, 3762−3767.
(24) Fischer, C. C.; Tibbetts, K. J.; Morgan, D.; Ceder, G. Predicting crystal structure by merging data mining with quantum mechanics. *Nature materials* **2006**, *5*, 641−646.
(25) Zhao, Y.; Cui, Y.; Xiong, Z.; Jin, J.; Liu, Z.; Dong, R.; Hu, J. Machine Learning-Based Prediction of Crystal Systems and Space Groups from Inorganic Materials Compositions. *ACS omega* **2020**, *5*, 3596−3606.
(26) Li, Y.; Dong, R.; Yang, W.; Hu, J. Composition based crystal materials symmetry prediction using machine learning with enhanced descriptors. *Comput. Mater. Sci.* **2021**, *198*, 110686.
(27) Hargreaves, C. J.; Dyer, M. S.; Gaultois, M. W.; Kurlin, V. A.; Rosseinsky, M. J. The earth mover's distance as a metric for the space of inorganic compositions. *Chem. Mater.* **2020**, *32*, 10610−10620.
(28) Zuo, Y.; Qin, M.; Chen, C.; Ye, W.; Li, X.; Luo, J.; Ong, S. P. Accelerating Materials Discovery with Bayesian Optimization and Graph Deep Learning. *Mater. Today* **2021**, *51*, 126.
(29) Kresse, G.; Hafner, J. ab initio. *Phys. Rev. B* **1993**, *47*, 558−561.
(30) Kresse, G.; Hafner, J. ab initio. *Phys. Rev. B* **1994**, *49*, 14251−14269.
(31) Kresse, G.; Furthmüller, J. Efficiency of ab initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set. *Comput. Mater. Sci.* **1996**, *6*, 15−50.
(32) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for ab initio Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B* **1996**, *54*, 11169−11186.
(33) Blöchl, P. E. Projector Augmented-Wave Method. *Phys. Rev. B* **1994**, *50*, 17953−17979.
(34) Kresse, G.; Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys. Rev. B* **1999**, *59*, 1758−1775.
(35) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.
(36) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple [Phys. Rev. Lett. 77, 3865 (1996)]. *Phys. Rev. Lett.* **1997**, *78*, 1396−1396.
(37) Thomas, J. C.; Natarajan, A. R.; Van Der Ven, A. Comparing crystal structures with symmetry and geometry. *npj Computational Materials* **2021**, *7*, 1−11.
(38) Superpose3D for point cloud distance calculation. https://github.com/jewettaij/superpose3d (accessed: 2022−01−22).
(39) Hu, J.; Stefanov, S.; Song, Y.; Omee, S. S.; Louis, S.-Y.; Siriwardane, E.; Zhao, Y. MaterialsAtlas.org: A Materials Informatics Web App Platform for Materials Discovery and Survey of State-of-the-Art. *arXiv preprint arXiv:2109.04007* 2021.

https://doi.org/10.1021/acs.inorgchem.1c03879
Inorg. Chem. 2022, 61, 8431−8439