# DISTANCE MATRIX BASED CRYSTAL STRUCTURE PREDICTION USING EVOLUTIONARY ALGORITHMS

## A PREPRINT

**Jianjun Hu***
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC 29201
jianjunh@csc.sc.edu

**Wenhui Yang**
School of Mechanical Engineering
Guizhou University
Guiyang China 550050

**Edirisuriya M. Dilanga Siriwardane**
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC 29201

October 6, 2020

## ABSTRACT

Crystal structure prediction (CSP) for inorganic materials is one of the central and most challenging problems in materials science and computational chemistry. This problem can be formulated as a global optimization problem in which global search algorithms such as genetic algorithms (GA) and particle swarm optimization have been combined with first principle free energy calculations to predict crystal structures given only a material composition or only a chemical system. These DFT based ab initio CSP algorithms are computationally demanding and can only be used to predict crystal structures of relatively small systems. The vast coordinate space plus the expensive DFT free energy calculations make it inefficient and ineffective. On the other hand, a similar structure prediction problem has been intensively investigated in parallel in the protein structure prediction community of bioinformatics, in which the dominating predictors are knowledge based approaches including homology modeling and threading that exploit known protein structures. Surprisingly, the CSP field has mainly focused on ab initio approaches in the past decade. Inspired by the knowledge-rich protein structure prediction approaches, herein we explore whether known geometric constraints such as the pairwise atomic distances of a target crystal material can help predict/reconstruct its structure given its space group and lattice information. We propose DMCrystal, a genetic algorithm based crystal structure reconstruction algorithm based on predicted atomic pairwise distances. Based on extensive experiments, we show that the predicted distance matrix can dramatically help to reconstruct the crystal structure and usually achieves much better performance than CMCrystal, an atomic contact map based crystal structure prediction algorithm. This implies that knowledge of atomic interaction information learned from existing materials can be used to significantly improve the crystal structure prediction in terms of both speed and quality.

**Keywords** crystal structure prediction · distance matrix · machine learning · contact map · global optimization · genetic algorithms

# 1 Introduction

Discovery of novel materials has big potentials in transforming a variety of industries such as cell phones, electric vehicles,energy storage, quantum computing, and chemical catalysts [1]. Compared to traditional Edisonian experimental methods which usually strongly depends on the expertise of the scientists, computational materials discovery has the advantage of much more efficient search in the vast chemical design space. In the past few years, three main strategies have emerged as the most promising approaches for new materials discovery including inverse design [2, 3], generative machine learning models [4, 5, 3, 6, 7], and crystal structure predictions [8, 1, 9].

In a typical crystal structure prediction (CSP) problem [10], the goal is to find a stable and ideally synthesizable crystal structure with the lowest free energy for a given chemical composition (or a chemical system such as Mg-Mn-O with variable compositions) at given pressure–temperature conditions [1]. This is usually done by trying to adjust the coordinates of the given set of atoms while minimizing the free energy of the system. These kind of energy guided structure prediction approaches have been called as ab initio structure prediction. It is assumed that lower free energy corresponds to the more stable arrangement of atoms.

With the crystal structure of a chemical compound, many physichochemical properties can be predicted reliably and routinely using first-principle calculation or machine learning models [11]. The CSP approach for new materials discovery is especially appealing due to the efficient sampling algorithm that generates diverse chemically valid candidate materials compositions/formulas with low free energies [4]. In the past 10 years, crystal structure prediction algorithms based on evolutionary algorithms and particle swarm optimization have led to a series of new materials discoveries [12, 1, 13]. However, these ab initio free energy based global search algorithms have a major challenge that limits their successes to relative simple crystals [1, 14] (mostly binary materials with less than 20 atoms in the unit cell [1, 13]) due to their dependence on the costly DFT calculations of free energies for sampled structures. With limited DFT calculations budget, it is a key issue to efficiently sample the atom configurations. [12, 10]. To improve the sampling efficiency, a variety of strategies have been proposed such as exploiting symmetry [15, 16] and pseudosymmetry [10], smart variation operators, clustering[17], machine-learning interatomic potentials with active learning [18], designing chemically based swapping operators [19]. However, the scalability of these ab initio approaches remains an unsolved issue.

Recently, generative machine learning models have been emerging as a novel approach to generate new materials including generative adversarial networks (GAN) approach for both chemical composition discovery [20] and crystal structure generation for a given chemical system [21]. Compared to global free energy optimization based approaches in CSP, these methods can take advantage of the implicit compositional and atomic configuration rules and constraints embodied in the large number of known crystal structures which can be learned by the deep neural network models. Using neural networks to implicitly learn such rules can lead to more efficient sampling of the search space [20]. Actually, the idea of exploiting knowledge in existing structures is not new and has been widely used in the field protein structure prediction(PSP) [22, 23, 24, 25] and also in crystal structure prediction[25]. PSP algorithms can be largely divided into three categories: ab initio methods, homology modeling, and threading. Ab initio methods include energy or fragment based methods and also evolutionary covariation based approaches for contact map prediction and subsequent protein structure reconstruction. The latter two strategies are called comparative modeling [26], in which the homology modelling is based on the fact that homologous proteins will share similar structures. The protein threading approach uses a scoring function to compare a query protein sequence to the protein structure database to identify a protein structure template. While crystal structures have no coevolutionary history, the very shared chemical assembling rules and atomic interaction rules make that many crystals of the some class share a very similar structures such as $ABO_3$ perovskites and many oxides of the same space groups.

Herein we propose a new knowledge-rich approach for distance matrix based crystal structure prediction, which is inspired by the recent success of deep learning approaches for protein structure prediction (PSP) [27] in AlphaFold [28] which reconstructs the protein structure from a predicted distance matrix. In the protein structure prediction problem, one has to predict the 3D tertiary structure of a protein given only its amino acid sequence. The latest approach uses deep learning to predict the contact maps [29] or distance matrix [28], which can then be used to reconstruct the full three-dimensional (3D) protein structure with high accuracy [30]. In our prior work [31], we have shown that given a crystal's space group, its lattice constants, and the contact map, all of which can be predicted, we can reconstruct the crystal structure atom coordinates using global optimization algorithms such as genetic algorithms, particle swarm optimization or Bayesian optimization. However, it is known that pairwise distances between atoms within a crystal structure usually are conserved across different compounds so that machine learning models can be built to predict the distance matrix of a given composition. In this paper, we are exploring how we can use global optimization algorithms to reconstruct the atomic configuration for a given composition based on its space group and the distance matrix. The idea is that we can exploit the rich atom interaction distribution or other geometric patterns or motifs [23] existing in

the large number of known crystal structures to predict the atomic distance matrix. Compared to atomic contact map, the distance has much more information which should allow better crystal structure reconstruction. While in this paper, we use the true space group, the lattice constants, and the distance matrix derived from known materials structures, the space group of crystal structures can be predicted using a variety of prediction algorithms [32, 33] or be inferred from domain knowledge [34]. In [33], the top-3 accuracy for space group prediction ranges from 81% to 100% given its Bravais lattice, which can also be predicted using composition features with up to 84% accuracy. With the given distance matrix and the space group, we investigate whether global optimization algorithms such as GAs and CMA-ES methods can be used for reconstructing/predicting its crystal structure.

Our contributions can be summarized as follows:

- We propose a new knowledge rich approach for crystal structure prediction by using the pairwise atomic distance magtrix as a knowledge-rich strategy for solving the CSP problem. Our method is inspired by the success of corresponding algorithm in protein structure prediction, the AlphaFold.

- We define a series of benchmark test cases for testing global optimization algorithms to reconstruct the atomic configurations from atomic pairwise distances.

- we conduct extensive evaluations of how global optimization algorithms perform in distance matrix based crystal structure prediction, how the errors of the predicted distances may affect the reconstruction performance. We also confirm some of the predicted structures using DFT calculation of the formation energy of the predicted structures.

## 2 Materials and Methods

### 2.1 Problem formulation: distance matrix based crystal structure reconstruction

```
Title              Co4 As8

Lattice type       P
Space group name   P 21/c
Space group number 14
Setting number     1

Lattice parameters

    a        b        c       alpha    beta     gamma
 5.93826  5.89121  5.99374  90.0000  116.0027  90.0000

Unit-cell volume = 188.456919 Å^3

Structure parameters

                    x         y         z        Occ.
    1 Co  Co0    0.27182   0.50050   0.78556    1.000
    2 As  As1    0.15340   0.13397   0.86789    1.000
    3 As  As2    0.34307   0.63609   0.17457    1.000
========================================================
```
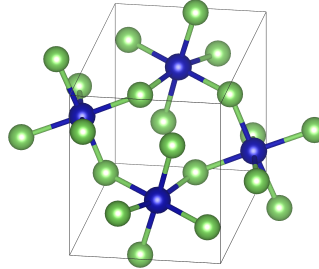
(a) Cif crysal structure file



(b) Graph representation

Figure 1: Cif and graph representation of crystal material $Co_4As_8$. (a) A cif file species the space group, the six lattice parameters (a,b,c,alpha,beta,gamma, and the fractional coordinates of each site of a crystal structure.(b) a crystal structure can be represented as a graph with bond connections.

A periodic crystal structure can be represented by its lattice constants a,b,c and angles $\alpha$, $\beta$, and $\gamma$, the space group, and the coordinates at unique Wyckoff positions. Using threshold for different atom pairs, the crystal structure can be converted into a graph, which can be represented as an adjacency matrix. The structure can also be looked as a point cloud, from which a pairwise distance can be derived or predicted based on known atom pair distance distribution. The pairwise distance matrix captures the interactions among atoms in the unit cell, which can be predicted by the know interaction patterns of these atom pairs in other known crystal materials structures. In this paper, we focus on the crystal structure reconstruction algorithm and assume that the pairwise atomic distance matrix has been obtained or predicted, and we'd like to check if the global optimization algorithms can help reconstruct the crystal structures in terms of the atom coordinates from the distance matrix, with or without adding other geometric or physical constraints. By formulating the distance matrix based CSP as an optimization problem, this allows us to evaluate how genetic algorithms (GA) can solve this optimization based structure reconstruction problem and how difficult this reconstruction problem for different crystal structures of varying complexity in terms of the number of unique Wyckoff positions(which determines the number of independent variables to be optimized), the level of symmetry as represented by the space

group, and also the number of atoms in the unit cell. For the example in Figure 1, the number of variables to optimize is 4x3=12, corresponding to 4 Wyckoff positions each with x,y,z three coordinate values. The crystal has 24 atoms in the unit cell, which can be mapped into a 24x24 distance matrix. The optimization problem is then how to search appropriate Wyckoff position atom coordinates so that after symmetry operations specified by space group 190, the generated crystal structure will have the same distance matrix. In this study, we assume the space group information, and the unit cell parameters of the target composition are all known, which is reasonable as they can be predicted using different approaches [35, 33, 36, 37]. While only distance matrix information is used as optimization target, other atomic interaction information such as limits of distances or preferential neighborhood relationships (e.g. atoms of some element pairs cannot stay too close to each other in known crystals) between some atom pairs can also be added as constraints in global search. The geometric constraint optimization objective can also be combined with the traditional free energy objective to achieve synergistic effect by e.g. reducing the number of DFT free energy calculations. Our work is in principle to the reconstruction algorithm used in contact map in protein structure prediction [30].
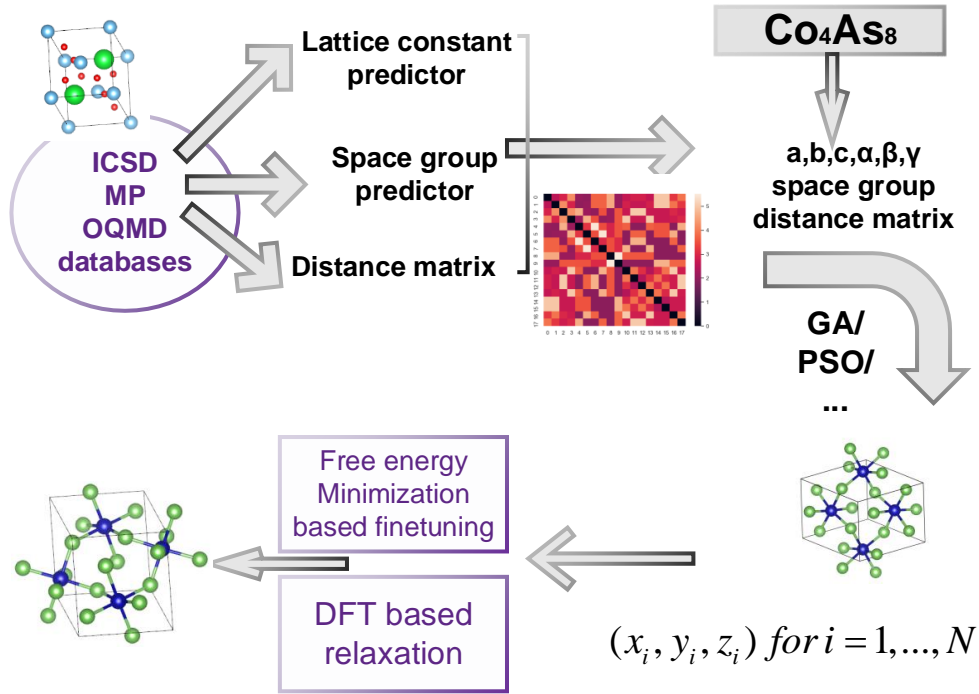


Figure 2: The DMCrystal algorithm for distance matrix based crystal structure prediction.

## 2.2 Distance matrix based crystal structure prediction using global optimization

In our problem formulation, the independent variables are a set of fractional coordinates $(x_i, y_i, z_i)$ for i=0,...,N, where N is the number of Wyckoff positions and $x_i, y_i, z_i$ are all real numbers in the range of [0,1]. To solve this crystal structure reconstruction problem, we propose to employ global optimization algorithms such as GAs to search the coordinates by maximizing the match between the distance matrix of the predicted structure and the distance matrix of the target crystal structure. This DMCrystal CSP framework is given in Figure 2. Basically, first, using the existing inorganic materials samples in the databases such as ICSD, Materials Project, or OQMD, three prediction models will be trained including a space group predictor [32, 33], a lattice constant predictor [38, 37, 36, 39, 40], and a distance matrix predictor. And then a global optimization algorithm such as the genetic algorithm will be used to search the atom coordinates such that the resulting structure's distance matrix matches the distance matrix of the predicted structure as much as possible. After that, the structures will then be fed to free energy minimization based DFT relaxation or refinement to generate the final structure prediction.

### 2.3 Genetic algorithms(GA)

In this work we focus on exploring how global optimization algorithm such as genetic algorithms can be used to search the atom coordinates guided by a given distance matrix. We evaluate the genetic algorithm on different problem instances. Genetic algorithms [41] are population based global optimization algorithms inspired by the biological evolution process. Candidate solutions (individuals) are encoded by binary or real-valued vectors. Starting with a random population of individuals, the population is then subject to generations of mutation, crossover, and selection to evolve the population toward individuals with high fitness, evaluated by the optimization objective functions. Compared to other heuristic search algorithms, GAs have proved to be suitable for large-scale global optimization problems [42] and has been used in several crystal structure prediction algorithms [8, 43, 44], and mainly for free energy minimization. The main hyper-parameters include the population size, crossover and mutation rates, and the number of generations. Here we apply the real-value encoded GA as the global optimization procedure for crystal structure reconstruction from distance matrix. Genetic algorithms also have the flexibility to achieve mixed variable-type optimization.

### 2.4 Objective function and Evaluation Criteria

The objective function for distance matrix based structure reconstruction is defined as the Euclidean distance of two distance matrices, which is shown in the following equation:

$$\text{fitness}_{opt} = \sqrt{\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \|A_{i,j} - B_{i,j}\|^2} \tag{1}$$

where $A$ is the distance matrix of the candidates of the search algorithm and $B$ is the predicted distance matrix of a given composition, and N is the number of atom in the unit cell.

To evaluate the reconstruction performance of different algorithms, we can use the RMSD error of distance matrices as one evaluation criterion, which however does not indicate the final structure similarity between the predicted structure and the true target structure. To address this, we define the root mean square distance (RMSD) and mean absolute error (MAE) of two structures as below:

$$\begin{aligned} \text{RMSD}(\mathbf{v}, \mathbf{w}) &= \sqrt{\frac{1}{n} \sum_{i=1}^{n} \|v_i - w_i\|^2} \\ &= \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left( (v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2 \right)} \end{aligned} \tag{2}$$

$$\begin{aligned} \text{MAE}(\mathbf{v}, \mathbf{w}) &= \frac{1}{n} \sum_{i=1}^{n} \|v_i - w_i\| \\ &= \frac{1}{n} \sum_{i=1}^{n} \left( \|v_{ix} - w_{ix}\| + \|v_{iy} - w_{iy}\| + \|v_{iz} - w_{iz}\| \right) \end{aligned} \tag{3}$$

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 the target crystal structure. It should be pointed out that in the experiments of this study, the only constraints for the optimization is the distance matrix, it is possible that the predicted atom coordinates are oriented differently from the target atoms in terms of of coordinate systems. To avoid this complexity, we compare the RMSD and MAE for all possible coordinate systems matching such as (x,y,z –>x,y,z), (x,y,z –>x,z,y), etc. and report the lowest RMSD and MAE.

We also use the contact map accuracy to evaluate how well a CSP algorithm can reconstruct the topology of the target material. It is defined as the dice coefficient, which is shown in the following equation:

$$\text{contactmapaccuracy} = \text{Dice} = \frac{2|A \cap B|}{|A| + |B|} \approx \frac{2 \times A \bullet B}{\text{Sum}(A) + \text{Sum}(B)} \tag{4}$$

where $A$ is the predicted contact map matrix and $B$ is the true contact map of a given composition, both only contain 1/0 entries. $A \cap B$ denotes the common elements of A and B, |g| represents the number of elements in a matrix, • denotes dot product, Sum(g) is the sum of all matrix elements. Dice coefficient essentially measures the overlap of two matrix samples, with values ranging from 0 to 1 with 1 indicating perfect overlap. We also call this performance measure as contact map accuracy.

## 3 Experiments

### 3.1 Test problems

We have selected a set of target crystal structures as test cases for evaluating the proposed distance matrix based crystal structure reconstruction algorithm using a genetic algorithm. The list of target materials are shown in Table 1. Here, the numbers of independent atom sites are 2 and 3 corresponding to 6 and 9 number of optimization variables. The space group numbers range from 4 to 61 corresponding to triclinic,monoclinic,orthorhombic structures (More symmetric structures are reported in Section 3.3.3.

Table 1: Statistics of target crystal structures

| Target | MP_id | No.of sites | #Atom in unit cell | Space Group | #variables |
|---|---|---|---|---|---|
| $Ag_4S_2$ | mp-560025 | 3 | 6 | 4 | 9 |
| $Bi_4Se_4$ | mp-1182022 | 2 | 8 | 14 | 6 |
| $B_4N_4$ | mp-569655 | 2 | 8 | 14 | 6 |
| $S_4N_4$ | mp-236 | 2 | 8 | 14 | 6 |
| $Pb_4O_4$ | mp-550714 | 2 | 8 | 29 | 6 |
| $Co_4As_8$ | mp-2715 | 3 | 12 | 14 | 9 |
| $Bi_8Se_4$ | mp-1102082 | 3 | 12 | 14 | 9 |
| $Te_4O_8$ | mp-561224 | 3 | 12 | 19 | 9 |
| $W_4N_8$ | mp-754628 | 3 | 12 | 33 | 9 |
| $Cd_4P_8$ | mp-402 | 3 | 12 | 33 | 9 |
| $Ni_8P_8$ | mp-27844 | 2 | 16 | 61 | 6 |

### 3.2 Experimental Setup

For all optimization algorithms, we set the lower boundary and upper boundary of all variables to be [0, 1] when optimizing fractional coordinates. The number of variables depends on the target materials, which is equal to the number of independent atom sites multiplied by 3. For GA, we set the population size to 100 and the number of generations to 1000 with cross-over probability of 0.8 mutation probability of 0.001.

### 3.3 Results

#### 3.3.1 Distribution analysis of pairwise atomic distances

One of the main assumptions of our distance matrix based crystal structure prediction is that the distance matrix of a new material with only composition information can be predicted. To show that the pairwise atom distance of a given pair of elements (A,B) can be predicted, we plot the histogram of the shortest distances of these atoms in all known crystals of ICSD structures including B-N and Co-As pairs which both have 341 and 136 pairs respectively. We also include the Pb-O and Te-O atom pairs which have 1175 and 1378 pairs in ICSD structures. From Figure 3, we can find that for atom pairs with both high frequencies and low frequencies, the pairwise atom distances are highly conserved, with two very high peaks in all cases. This means that by the two pairs of element information alone, one can achieve good distance prediction, which can be further improved by using more context information from the formula such as other elements for ternary materials.

(a) Atom pairs of B and N

(b) Atom pairs of Co and As

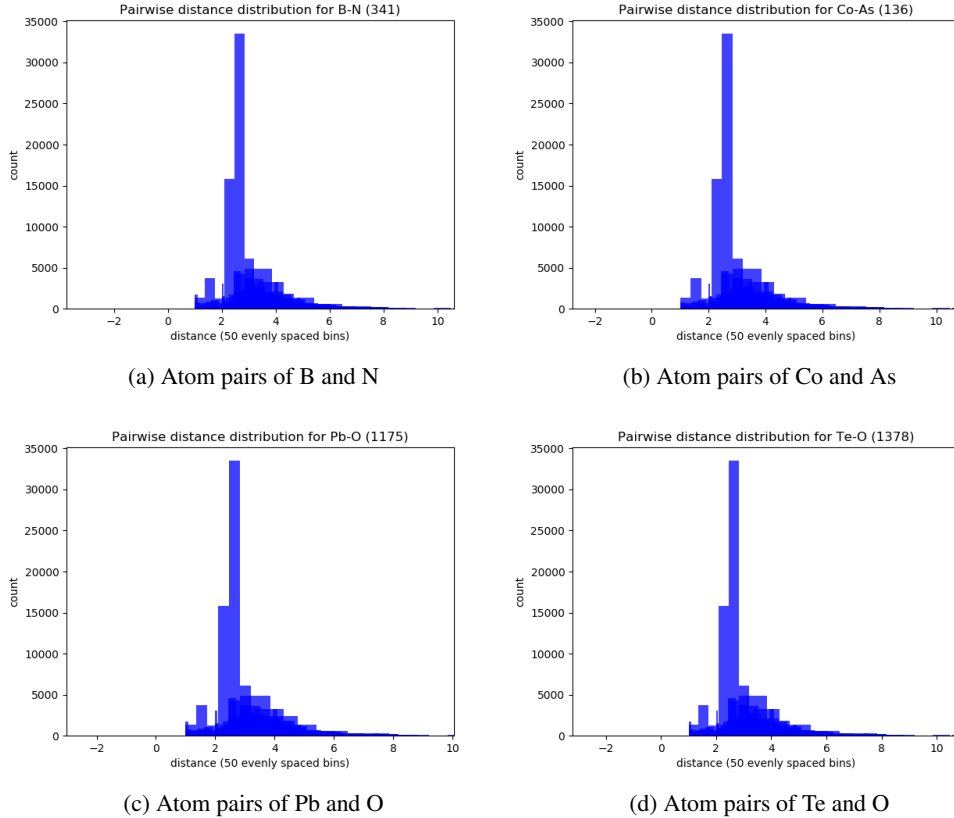(c) Atom pairs of Pb and O

(d) Atom pairs of Te and O

Figure 3: Pairwise atom distance distribution of atom pairs in ICSD. The highly conserved atom-pair distances mean that many of the atom pair distances may be predicted reliably.

### 3.3.2 Successful distance matrix based crystal structure prediction/reconstruction

To evaluate our DMCrystal method for crystal structure prediction, we apply it to a selected set of 11 target structures as shown in Table 1 with the number of atoms ranging from 6 to 16. The total number of objective evaluations is set as 100,000 or 1000 generations for population size of 100 for GA. The overall performance of of the DMCrystal algorithm for distance matrix based crystal structure reconstruction is shown in Table 2, which also includes the contact map based prediction results for comparison. We find that the contact prediction accuracy for 9 out of the 11 targets reach 100%, demonstrating the effectiveness of our method to find the target topology from random atom coordinates using the distance matrix as the optimization target. This performance scores are similar to those of CMCrystal which directly uses the contact map as the optimization objective [31]. Table 2 also shows the RMSD and MAE scores of the predicted structures compared to the target structures by CMCrystal and DMCrystal, both of which are calculated in terms of fractional coordinates of the independent atom sites. The RMSD values of DMCrystal range from 0.022 for $B_4N_4$ to 0.229 for $Co_4As_8$ while the MAE scores range from 0.017 for $B_4N_4$ to 0.135 for $Te_4O_8$. We found both scores of DMCrystal are much better than those of the CMCrystal algorithm. Indeed, when we compare the RMSD and MAE scores of these two methods for the 11 targets in columns 3/4 and 6/7 for CMCrystal and DMCrystal respectively, we find that for the 11 targets, the DMCrystal method has achieved on average 53% improvement in terms of RMSD/MAE reconstruction errors compared to CMCrystal algorithm. Out of the 11 target, DMCrystal works slightly worse for only one target structure ($Co_4As_8$) with RMSD of 0.229 versus 0.197 of CMCrystal. But its corresponding MAE of 0.125 is better than the RMSD (0.149) of CMCrystal. In terms of contact map accuracy, the DMCrystal achieves equal or better results for 10 out of 11 targets except $Ni_8P_8$, which is 5.26% worse. We also observe that from first row $Ag_4S_2$ to last row $Ni_8P$, the RMSD errors increase with the increasing number of atoms in the structures. These results demonstrate that distance matrix based crystal structure reconstruction/prediction can be much better, this is consistent with the conclusion in protein structure prediction in which the distance matrix based prediction algorithm achieved better performance for contact map based protein structure prediction. The better RMSD/MAE performance scores of DMCrystal is understandable as CMCrystal only aims to reconstruct the topology (the contact map matrix). In

7

Table 2: Performances of DMcrystal in terms of structure prediction accuracy

| material | Contact map(CMCrystal) | | | Distance matrix (DMCrystal) | | | Improvement % | | |
|---|---|---|---|---|---|---|---|---|---|
| | contact map accuracy | RMSD | MAE | contact map accuracy | RMSD | MAE | contact map accuracy | RMSD | MAE |
| $Ag_4S_2$ (mp-560025) | 1.000 | 0.320 | 0.233 | 1.0 | 0.076 | 0.057 | 0% | 76.19% | 75.57% |
| $Bi_4Se_4$ (mp-1182022) | 1.000 | 0.124 | 0.097 | 1.0 | 0.069 | 0.057 | 0% | 44.64% | 41.66% |
| $B_4N_4$ (mp-569655) | 1.000 | 0.070 | 0.054 | 1.0 | 0.022 | 0.017 | 0% | 68.03% | 68.94% |
| $Pb_4O_4$ (mp-550714) | 1.000 | 0.246 | 0.196 | 1.0 | 0.097 | 0.078 | 0% | 60.48% | 60.30% |
| $S_4N_4$ (mp-236) | 1.000 | 0.156 | 0.137 | 1.0 | 0.081 | 0.064 | 0% | 47.93% | 53.16% |
| $Te_4O_8$ (mp-561224) | 1.000 | 0.379 | 0.266 | 1.0 | 0.161 | 0.135 | 0% | 57.47% | 49.23% |
| $W_4N_8$ (mp-754628) | 1.000 | 0.368 | 0.214 | 1.0 | 0.149 | 0.130 | 0% | 59.41% | 39.11% |
| $Cd_4P_8$ (mp-402) | 1.000 | 0.320 | 0.204 | 1.0 | 0.129 | 0.105 | 0% | 59.67% | 48.23% |
| $Co_4As_8$ (mp-2715) | 0.923 | 0.197 | 0.149 | 1.0 | 0.229 | 0.125 | 8.3% | -16.34% | 16.15% |
| $Bi_8Se_4$ (mp-1102082) | 0.889 | 0.257 | 0.232 | 1.0 | 0.102 | 0.083 | 12.5% | 60.33% | 64.23% |
| $Ni_8P_8$ (mp-27844) | 1.000 | 0.381 | 0.335 | 0.947 | 0.128 | 0.104 | -5.26% | 66.53% | 69.01% |

Discussion Section, we explore whether adding a few atomic pairwise distance information can improve the performance of CMCrystal.

Figure 4 shows three sets of predicted and target crystal structures of $B_4N_4$, $Bi_4Se_4$, and $Co_4As_8$. For both $B_4N_4$ and $Bi_4Se_4$ (Figure 4 (a)-(d)), the contact map accuracy reaches 100% and the predicted structures are very close to the target structures. The RMSD of $B_4N_4$ is 0.022 which is smaller than the RMSD (0.124) of $Bi_4Se_4$, which is reflected by the higher similarity of the pairs of $B_4N_4$ than the pair of structures of $Bi_4Se_4$. The prediction accuracy for target structure of $Co_4As_8$ is lower with RMSD of 0.229. Compared to CMCrystal, we find that the DMCrystal algorithm usually can reach much more similar structures to the true structures due to the explicit distance constraints while both algorithms can achieve high accuracy in topology prediction.
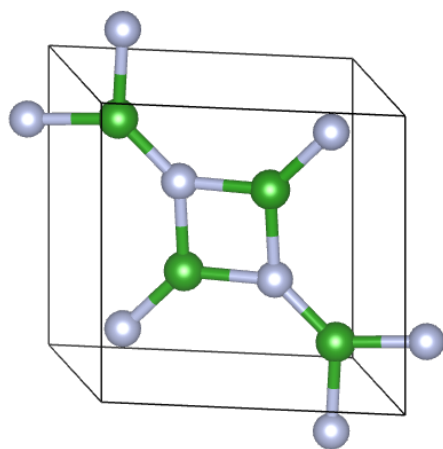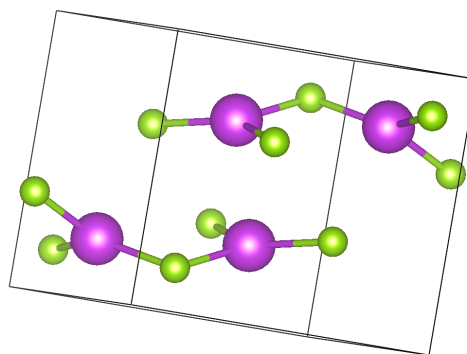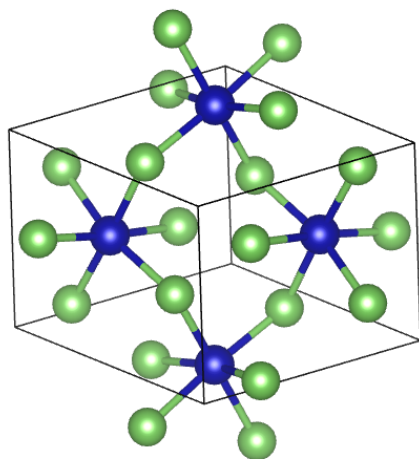
(a) Target structure $B_4N_4$
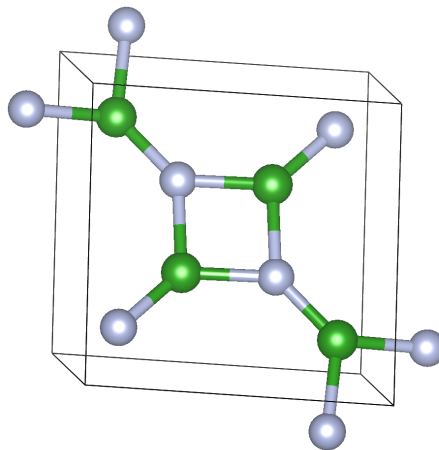
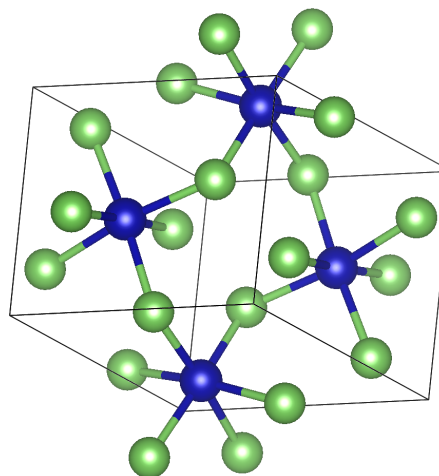(b) Predicted structure of $B_4N_4$ with contact map accuracy:100%, RMSD:0.022
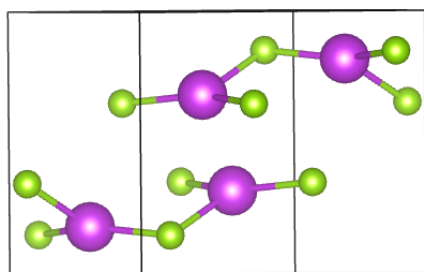


(c) Target structure of $Bi_4Se_4$

(d) Predicted structure of $Bi_4Se_4$ with contact map accuracy:100%, RMSD:0.069



(e) Target structure of $Co_4As_8$

(f) Predicted structure of $Co_4As_8$ with with contact map accuracy:92.3%, RMSD:0.197

Figure 4: Examples of predicted versus target crystal structures by DMCrystal.

### 3.3.3    Performance comparison of different reconstruction algorithms using distance information

The crystal structure prediction performance of our DMCrystal algorithms depends on the quality of the predicted distance matrix. In a realistic crystal structure prediction, the predicted distance matrices may have errors, especially for those non-local non-bond distances. Here we investigate how the different types of distance information may be used for crystal structure prediction using either DMCrystal or CMCrystal frameworks. For comparison purpose, we evaluate the following five algorithms sorted by decreasing amount of distance information used: 1) DMCrystal, which use all atom pair distance information; 2) DMCrystal with 10% noise on the distance matrix; 3) DMCrystal with only local bond distance information; 4) CMCrystal with local bond distance information; 5) CMCrystal: only contact map is used with no distance information. We ran all the algorithms over the 11 target structures in Table 1. The RMSD performance scores of all five algorithms are shown in Figure 5

First, we find that for all 11 targets, the DMCrystal achieves the lowest RMSD errors due to its exploitation of all pairwise atom distance information. When we add 10% of noise to the distance matrix, the performance degrades accordingly, but still better than the performance of the reconstruction algorithms using only local bond distances. The contact map based reconstruction algorithm achieves the worst performance in terms of RMSD for five out of 11 targets, reflecting its focus on topology(contact map) rather than the RMSD. However, its performance can be improved by adding the local bond distance information, which are usually easy to be predicted due to their high conservation (see Figure 3).
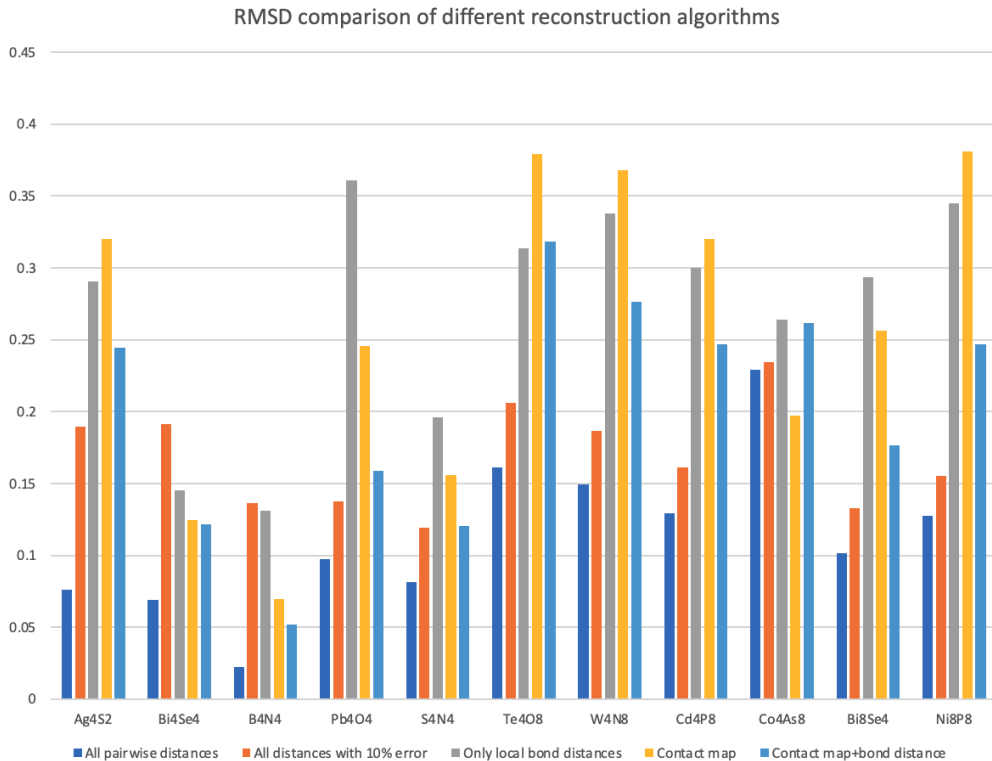


Figure 5: Performance comparison of different algorithms in terms of structure reconstruction RMSD.

### 3.3.4    Comparison with existing crystal structure prediction algorithms

To show the potential advantages of our knowledge-rich distance matrix based crystal structure prediction method compared to existing free energy minimization based ab initio method, we also apply CALYPSO [45] to the three targets shown in Figure 4. And then we calculate the formation energies of all their predicted structures as well of three predicted structures by the DMCrystal algorithm using density functional theory (DFT). DFT calculations are carried out based on the Vienna ab initio simulation package (VASP) [46, 47]. The interactions between the ions and electrons are treated by the projected augmented wave (PAW) method [48, 49]. The generalized gradient approximation (GGA) based exchange-correlation potential is considered with the Perdew-Burke-Ernzerhof (PBE) pseudopotentials [50]. A plane-wave energy cutoff is set to 500 eV. The convergence criteria for total energy and the Hellmann-Feyman

forces are $1.0 \times 10^{-6}$ eV and 0.01eV/ Å, respectively. The Γ-centered Monkhorst-Pack k-meshes are considered. The formation energies of the materials are computed based on the following equation.

$$E_{form} = \frac{1}{(x+y+z)}(E_{A_xB_yC_z} - x\mu_A - y\mu_B - z\mu_C), \tag{5}$$

where $E_{A_xB_yC_z}$ is the total energy of the material with the chemical formula $A_xB_yC_z$ and $\mu$ represents the energy of each atom in its respective ground-state bulk phase [51].

For CALYPSO, we let it run 24 hours on a 8-core Dell server with 3.7GHz multi-core CPU. DMCrystal is also run on the same machine but using only one core, which finishes each of the three running experiments for the three targets $B_4N_4$, $Bi_4Se_4$, $Co_4As_8$ within 45 minutes.

First, Figure 6 shows the predicted structures by CALYPSO for the three benchmark target structures. The corresponding RMSDs are 0.373, 0.358, and 0.423, which are significantly worse than those of the DMCrystal with values of 0.022, 0.069, and 0.229. When comparing the predicted structures by CALYPSO to those predicted by DMCrystal in Figure 4, we can find that the latter are much similar to the true target structures, which is reflected by the much lower RMSD scores. We have also compared the formation energy of the predicted structures by CALYPSO and DMCrystal as shown in Table 3. All three structures predicted by DMCrystal has negative formation energy values while those predicted by CALYPSO have positive formation energy indicating their failure to find stable structures close to the true targets.

Table 3: Prediction performance comparison with CALYPSO

| Method | Running time | Formation Energy (eV/atom) | | |
|---|---|---|---|---|
| | | $B_4N_4$ | $Bi_4Se_4$ | $Co_4As_8$ |
| DMCrystal | 45 min (using 1-core) | -0.909 | -0.110 | -0.319 |
| CALYPSO | 24 hours (using 8-core) | 2.037 | 0.286 | (N/A) Structure too bad to calculate its formation energy |
| | | RMSD | RMSD | RMSD |
| DMCrystal | | 0.022 | 0.069 | 0.229 |
| CALYPSO | | 0.373 | 0.358 | 0.423 |



(a) Predicted structure of $B_4N_4$ with RMSD:0.373

(b) Predicted structure of $Bi_4Se_4$ with RMSD:0.358

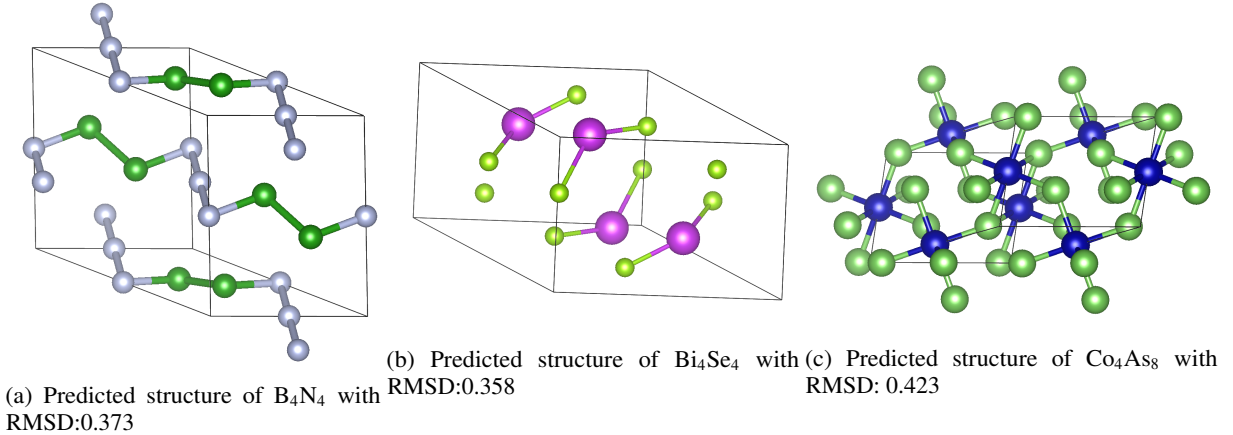(c) Predicted structure of $Co_4As_8$ with RMSD: 0.423

Figure 6: Structures of three targets predicted by CALYPSO after 24 hours running on a 8-core Dell server.

## 4 Discussion

In this work, our main focus is on crystal structure reconstruction from a predicted distance matrix along with the predicted space group and lattice parameters. In our experiments, we have used the real distance matrix, space group, and lattice parameters from the benchmark target structures. However, in reality, all these information are predicted for a given composition or materials formula.

Machine learning predictions of space groups and crystal systems from composition have been proposed in [33, 32]. It is reported that [33] that the top three space group prediction performance can reach to a range of 0.81 to 0.98 in

terms of $R^2$ scores based on different crystal systems. For cubic structures, the $R^2$ score is 0.96 on average. When evaluating their machine learning performance for 18 space group prediction, the F1 scores can reach 0.787 [32]. These two works show that we can predict the space group reasonably well, at least for top K performance. It is thus possible to predict e.g. top 10 candidate space groups for DMcrystal based structure reconstruction and then select the solutions with lowest formation energy. Composition based crystal lattice parameter prediction methods have also been proposed [52], which trained a support vector regressor using 1541 binary body centered cubic crystals and has achieved a $R^2$ score of 83.6% accuracy via cross-validation and maximum error of 4%. We have improved this result to $R^2$ of 0.97 for cubic structures.

We also examined additional factors that affect the structure reconstruction quality based on the target structure properties. From our extensive experiments, we find that there are several factors that affect the crystal structure prediction quality by our algorithm such as the number of independent atom sites, the number of atoms in the unit cell, and the number of bonds/topology constraints and etc. Here we report how two factors, the number of independent atomic sites in the unit cell of the material and its space group, affect the crystal structure reconstruction performance by DMCrystal. In Table 4 we compare the performance of DMCrystal for target structures with the same number of total atoms in the unit cells but different numbers of independent atom sites. It shows that in general, the contact map accuracy decreases when there are more atom sites in the structure. However the RMSD prediction error are relatively stable around 0.1 for all the structures.

Table 4: Prediction performance versus number of atom sites for DMCrystal

| Target | mp_id | atom site | contact map accuracy | RMSD |
|---|---|---|---|---|
| $La_{12}Se_{16}$ | mp-491 | 2 | 1.000 | 0.102 |
| $Bi_8Pd_4O_{16}$ | mp-29259 | 3 | 1.000 | 0.124 |
| $V_8O_{20}$ | mp-25280 | 4 | 1.000 | 0.083 |
| $Fe_{12}O_{16}$ | mp-1192788 | 5 | 0.970 | 0.120 |
| $Ge_8S_{12}I_8$ | mp-27928 | 6 | 0.923 | 0.091 |
| $Li_4V_4Si_4O_{16}$ | mp-1176508 | 7 | 0.889 | 0.149 |
| $Ba_2V_8O_{18}$ | mp-18910 | 8 | 0.881 | 0.091 |
| $Si_2H_{18}C_6Cl_2$ | mp-867818 | 9 | 0.857 | 0.116 |
| $Zr_8Cr_8F_{12}$ | mp-690241 | 10 | 0.842 | 0.164 |

Table 5 shows the prediction results of DMCrystal for a set of materials with five atom sites and similar numbers of atoms but different space groups. First, it can be observed that the higher the space group number, the higher the contact map accuracy, indicating that higher symmetry (with higher space group number) puts more constraints on the atom configurations and reduces the search space so that better performance can be achieved by the genetic algorithm in atom coordinate search. In the table, the contact map accuracy increases from 0.828 to 0.963 when the space group goes from 2 to 194. At the same time, the RMSD errors are relative stable around 0.12 with slight increase for targets of space groups 129 and 194. We also compare these performances with those of CMCrystal which only uses contact map as objectives for crystal structure reconstruction and we find the RMSD errors of DMCrystal is much lower.

Table 5: Prediction performance versus space group using DMCrystal

| Target | mp_id | atom site# | space group | contact map accuracy | RMSD |
|---|---|---|---|---|---|
| $Hg_8Cl_4O_4$ | mp-636805 | 5 | 2 | 0.828 | 0.115 |
| $Be_4B_2O_{10}$ | mp-1079124 | 5 | 5 | 0.833 | 0.149 |
| $Bi_6O_8F_2$ | mp-757162 | 5 | 13 | 0.875 | 0.086 |
| $Tl_6V_2O_8$ | mp-29047 | 5 | 44 | 0.900 | 0.120 |
| $La_4Sn_2S_{10}$ | mp-12170 | 5 | 55 | 0.900 | 0.132 |
| $Li_2V_2F_{12}$ | mp-753573 | 5 | 102 | 0.923 | 0.107 |
| $Yb_4H_4O_8$ | mp-625103 | 5 | 113 | 0.941 | 0.083 |
| $Mg_4Co_2H_{10}$ | mp-642660 | 5 | 129 | 0.966 | 0.158 |
| $K_{10}Cu_2As_4$ | mp-14623 | 5 | 194 | 0.963 | 0.132 |

# 5    Conclusion

We formulate the crystal structure prediction/reconstruction problem based on its space group symmetry and the atom pairwise distance matrix, and proposed a genetic algorithm based global optimization to solve this problem. Our experiments show that the genetic algorithm is able to reconstruct the crystal structure for a set of benchmark materials within 45 minutes by setting the pair-wise atomic distance matrix as the optimization objective given only their space group, lattice parameters and the stoichiometry. These predicted structures are close to the target crystal structures so that they can be used to seed the costly free energy minimization based ab initio crystal structure prediction algorithms for further structure refining. They may also be used for DFT based structure relaxation to obtain the correct crystal structures for some compositions. We also find that distance matrix can lead to much higher quality reconstruction compared to the contact map based crystal structure prediction. Our results show that our distance matrix based approach is a viable crystal structure prediction/reconstruction method, similar to their role in protein structure prediction.

# 6    Availability of data

The data that support the findings of this study are openly available in Materials Project database at http: www.materialsproject.org

# 7    Contribution

Conceptualization, J.H.; methodology, J.H. and W.Y.; software, W.Y. and J.H; validation, W.Y, J.H., D.S.; investigation, J.H., W.Y.; resources, J.H.; data curation, J.H. and W.Y.; writing–original draft preparation, J.H.; writing–review and editing, J.H, W.Y., D.S.; visualization, J.H. and W.Y; supervision, J.H.; funding acquisition, J.H.

# 8    Acknowledgement

# References

[1] Artem R Oganov, Chris J Pickard, Qiang Zhu, and Richard J Needs. Structure prediction drives materials discovery. *Nature Reviews Materials*, 4(5):331–348, 2019.

[2] Alex Zunger. Inverse design in search of materials with target functionalities. *Nature Reviews Chemistry*, 2(4):1–16, 2018.

[3] Baekjun Kim, Sangwon Lee, and Jihan Kim. Inverse design of porous materials using artificial neural networks. *Science advances*, 6(1):eaax9324, 2020.

[4] Yabo Dan, Yong Zhao, Xiang Li, Shaobo Li, Ming Hu, and Jianjun Hu. Generative adversarial networks (gan) based efficient sampling of chemical space for inverse design of inorganic materials. *arXiv preprint arXiv:1911.05020*, 2019.

[5] John Bradshaw, Brooks Paige, Matt J Kusner, Marwin Segler, and José Miguel Hernández-Lobato. A model to search for synthesizable molecules. In *Advances in Neural Information Processing Systems*, pages 7937–7949, 2019.

[6] Juhwan Noh, Jaehoon Kim, Helge S Stein, Benjamin Sanchez-Lengeling, John M Gregoire, Alan Aspuru-Guzik, and Yousung Jung. Inverse design of solid-state materials via a continuous representation. *Matter*, 1(5):1370–1384, 2019.

[7] Zekun Ren, Juhwan Noh, Siyu Tian, Felipe Oviedo, Guangzong Xing, Qiaohao Liang, Armin Aberle, Yi Liu, Qianxiao Li, Senthilnath Jayavelu, et al. Inverse design of crystals using generalized invertible crystallographic representation. *arXiv preprint arXiv:2005.07609*, 2020.

[8] Colin W Glass, Artem R Oganov, and Nikolaus Hansen. Uspex—evolutionary crystal structure prediction. *Computer physics communications*, 175(11-12):713–720, 2006.

[9] Alexander G Kvashnin, Zahed Allahyari, and Artem R Oganov. Computational discovery of hard and superhard materials. *Journal of Applied Physics*, 126(4):040901, 2019.

[10] Andriy O Lyakhov, Artem R Oganov, Harold T Stokes, and Qiang Zhu. New developments in evolutionary structure prediction algorithm uspex. *Computer Physics Communications*, 184(4):1172–1182, 2013.

[11] Steph-Yves Louis, Yong Zhao, Alireza Nasiri, Xiran Wong, Yuqi Song, Fei Liu, and Jianjun Hu. Global attention based graph convolutional neural networks for improved materials property prediction. *arXiv preprint arXiv:2003.13379*, 2020.

[12] Artem R Oganov, Andriy O Lyakhov, and Mario Valle. How evolutionary crystal structure prediction works and why. *Accounts of chemical research*, 44(3):227–237, 2011.

[13] Yanchao Wang, Jian Lv, Quan Li, Hui Wang, and Yanming Ma. Calypso method for structure prediction and its applications to materials discovery. *Handbook of Materials Modeling: Applications: Current and Emerging Materials*, pages 2729–2756, 2020.

[14] Lijun Zhang, Yanchao Wang, Jian Lv, and Yanming Ma. Materials discovery at high pressures. *Nature Reviews Materials*, 2(4):1–16, 2017.

[15] Xin Chen, Ya-Fan Zhao, Yang-Yang Zhang, and Jun Li. Tgmin: An efficient global minimum searching program for free and surface-supported clusters. *Journal of Computational Chemistry*, 40(10):1105–1112, 2019.

[16] Evan Pretti, Vincent K Shen, Jeetain Mittal, and Nathan A Mahynski. Symmetry-based crystal structure enumeration in two dimensions. *The Journal of Physical Chemistry A*, 124(16):3276–3285, 2020.

[17] Knud Haugaard Sørensen, Mathias S Jørgensen, A Bruix, and Bjørk Hammer. Accelerating atomic structure search with cluster regularization. *The Journal of Chemical Physics*, 148(24):241734, 2018.

[18] Evgeny V Podryabinkin, Evgeny V Tikhonov, Alexander V Shapeev, and Artem R Oganov. Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning. *Physical Review B*, 99(6):064114, 2019.

[19] Paul M Sharp, Matthew S Dyer, George R Darling, John B Claridge, and Matthew J Rosseinsky. Chemically directed structure evolution for crystal structure prediction. *Physical Chemistry Chemical Physics*, 22(32):18205–18218, 2020.

[20] Yabo Dan, Yong Zhao, Xiang Li, Shaobo Li, Ming Hu, and Jianjun Hu. Generative adversarial networks (gan) based efficient sampling of chemical composition space for inverse design of inorganic materials. *npj Computational Materials*, 6(1):1–7, 2020.

[21] Sungwon Kim, Juhwan Noh, Geun Ho Gu, Alán Aspuru-Guzik, and Yousung Jung. Generative adversarial networks for crystal structure prediction. *arXiv preprint arXiv:2004.01396*, 2020.

[22] Guo-Wei Wei. Protein structure prediction beyond alphafold. *Nature Machine Intelligence*, 1(8):336–337, 2019.

[23] Zizhong Zhu, Ping Wu, Shunqing Wu, Linhan Xu, Yixu Xu, Xin Zhao, Cai-Zhuang Wang, and Kai-Ming Ho. An efficient scheme for crystal structure prediction based on structural motifs. *The Journal of Physical Chemistry C*, 121(21):11891–11896, 2017.

[24] Brian Kuhlman and Philip Bradley. Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology*, 20(11):681–697, 2019.

[25] Kevin Ryan, Jeff Lengyel, and Michael Shatruk. Crystal structure prediction via deep learning. *Journal of the American Chemical Society*, 140(32):10158–10168, 2018.

[26] Dmitrii M Nikolaev, Andrey A Shtyrov, Maxim S Panov, Adeel Jamal, Oleg B Chakchir, Vladimir A Kochemirovsky, Massimo Olivucci, and Mikhail N Ryazantsev. A comparative study of modern homology modeling algorithms for rhodopsin structure prediction. *ACS omega*, 3(7):7555–7566, 2018.

[27] Wei Zheng, Yang Li, Chengxin Zhang, Robin Pearce, SM Mortuza, and Yang Zhang. Deep-learning contact-map guided protein structure prediction in casp13. *Proteins: Structure, Function, and Bioinformatics*, 87(12):1149–1164, 2019.

[28] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[29] Isaac Arnold Emerson and Arumugam Amala. Protein contact maps: A binary depiction of protein 3d structures. *Physica A: Statistical Mechanics and its Applications*, 465:782–791, 2017.

[30] Michele Vendruscolo, Edo Kussell, and Eytan Domany. Recovery of protein structure from contact maps. *Folding and Design*, 2(5):295–306, 1997.

[31] Jianjun Hu, Wenhui Yang, Rongzhi Dong, Yuxin Li, Xiang Li, and Shaobo Li. Contact map based crystal structure prediction using global optimization. *arXiv preprint arXiv:2008.07016*, 2020.

[32] Yong Zhao, Yuxin Cui, Zheng Xiong, Jing Jin, Zhonghao Liu, Rongzhi Dong, and Jianjun Hu. Machine learning-based prediction of crystal systems and space groups from inorganic materials compositions. *ACS omega*, 5(7):3596–3606, 2020.

[33] Haotong Liang, Valentin Stanev, A Gilad Kusne, and Ichiro Takeuchi. Cryspnet: Crystal structure predictions via neural network. *arXiv preprint arXiv:2003.14328*, 2020.

[34] Aurora J Cruz Cabeza, Elna Pidcock, Graeme M Day, WD Sam Motherwell, and William Jones. Space group selection for crystal structure prediction of solvates. *CrystEngComm*, 9(7):556–560, 2007.

[35] Yuqi Song, Joseph Lindsay, Yong Zhao, Alireza Nasiri, Steph-Yves Louis, Jie Ling, Ming Hu, and Jianjun Hu. Machine learning based prediction of noncentrosymmetric crystal materials. *Computational Materials Science*, 183:109792, 2020.

[36] LQ Jiang, JK Guo, HB Liu, M Zhu, X Zhou, P Wu, and CH Li. Prediction of lattice constant in cubic perovskites. *Journal of Physics and Chemistry of Solids*, 67(7):1531–1536, 2006.

[37] Menad Nait Amar, Mohammed Abdelfetah Ghriga, Mohamed El Amine Ben Seghier, and Hocine Ouaer. Prediction of lattice constant of a2xy6 cubic crystals using gene expression programming. *The Journal of Physical Chemistry B*, 2020.

[38] Yun Zhang and Xiaojie Xu. Machine learning lattice constants for cubic perovskite a2xy6 compounds. *Journal of Solid State Chemistry*, page 121558, 2020.

[39] Syed Gibran Javed, Asifullah Khan, Abdul Majid, Anwar M Mirza, and J Bashir. Lattice constant prediction of orthorhombic abo3 perovskites using support vector machines. *Computational materials science*, 39(3):627–634, 2007.

[40] Abdul Majid, Asifullah Khan, Gibran Javed, and Anwar M Mirza. Lattice constant prediction of cubic and monoclinic perovskites using neural networks and support vector regression. *Computational materials science*, 50(2):363–372, 2010.

[41] David E Goldberg and John Henry Holland. *Genetic algorithms and machine learning*. Kluwer Academic Publishers-Plenum Publishers; Kluwer Academic Publishers . . . , 1988.

[42] Darrell Whitley, Francisco Chicano, Gabriela Ochoa, A Sutton, and Renato Tinós. Next generation genetic algorithms. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 1113–1136, 2019.

[43] Farren Curtis, Xiayue Li, Timothy Rose, Alvaro Vazquez-Mayagoitia, Saswata Bhattacharya, Luca M Ghiringhelli, and Noa Marom. Gator: a first-principles genetic algorithm for molecular crystal structure prediction. *Journal of chemical theory and computation*, 14(4):2246–2264, 2018.

[44] Patrick Avery, Cormac Toher, Stefano Curtarolo, and Eva Zurek. Xtalopt version r12: An open-source evolutionary algorithm for crystal structure prediction. *Comput. Phys. Commun.*, 237:274–275, 2019.

[45] Yanchao Wang, Jian Lv, Li Zhu, and Yanming Ma. Crystal structure prediction via particle-swarm optimization. *Phys. Rev. B*, 82:094116, Sep 2010.

[46] G. Kresse and J. Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B*, 54:11169–11186, Oct 1996.

[47] Jürgen Hafner. Ab-initio simulations of materials using vasp: Density-functional theory and beyond. *Journal of computational chemistry*, 29(13):2044–2078, 2008.

[48] Peter E Blöchl. Projector augmented-wave method. *Physical review B*, 50(24):17953, 1994.

[49] Georg Kresse and Daniel Joubert. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical review b*, 59(3):1758, 1999.

[50] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical review letters*, 77(18):3865, 1996.

[51] Edirisuriya M. D. Siriwardane, Rajendra P. Joshi, Neeraj Kumar, and Deniz Çakır. Revealing the formation energy–exfoliation energy–structure correlation of mab phases using machine learning and dft. *ACS Applied Materials & Interfaces*, 12(26):29424–29431, 2020. PMID: 32495630.

[52] Keisuke Takahashi, Lauren Takahashi, Jakub D Baran, and Yuzuru Tanaka. Descriptors for predicting the lattice constant of body centered cubic crystal. *The Journal of chemical physics*, 146(20):204104, 2017.