Contact map based crystal structure prediction using global optimization

Journal:	*CrystEngComm*
Manuscript ID	CE-ART-11-2020-001714.R1
Article Type:	Paper
Date Submitted by the Author:	23-Dec-2020
Complete List of Authors:	Hu, Jianjun; University of South Carolina Yang, Wenhui; Guizhou University, School of mechanical engineering Dong, Rongzhi; Guizhou University, School of Mechanical Engineering Li, Yuxin; Guizhou University, School of Mechanical Engineering Li, Xiang; Guizhou University, School of Mechanical Engineering Li, Shaobo; Guizhou University, School of Mechanical Engineering Siriwardane, Edirisuriya; University of South Carolina, Department of Computer Science and Engineering

SCHOLARONE™
Manuscripts

Contact map based crystal structure prediction using global optimization

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

Wenhui Yang, Rongzhi Dong, Yuxin Li, Xiang Li, Shaobo Li
School of Mechanical Engineering
Guizhou University
Guiyang China 550050

Edirisuriya MD Siriwardane
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC 29201

December 23, 2020

Abstract

Crystal structure prediction is now playing an increasingly important role in the discovery of new materials or crystal engineering. Global optimization methods such as genetic algorithms (GA) and particle swarm optimization have been combined with first principle free energy calculations to predict crystal structures given composition or only a chemical system. While these approaches can exploit certain crystal patterns such as symmetry and periodicity in their search process, they usually do not exploit the large amount of implicit rules and constraints of atom configurations embodied in the large number of known crystal structures. They currently can only handle crystal structure prediction of relatively small systems. Inspired by the knowledge-rich protein structure prediction approach, herein we explore whether known geometric constraints such as the atomic contact map of a target crystal material can help predict its structure given its space group information. We propose a global optimization based algorithm, CMCrystal, for crystal structure (atomic coordinates) reconstruction based on atomic contact maps. Based on extensive experiments using six global optimization algorithms, we show that it is viable to reconstruct the crystal structure given the atomic contact map for some crystal materials but more geometric or physicochemical constraints are needed to achieve the successful reconstruction of other materials

Keywords crystal structure prediction · machine learning · contact map · global optimization · implicit rules

1 Introduction

Accurate computational prediction of crystal materials structures have a variety of important applications. It can be used for computational discovery of novel functional materials has big potential in transforming a variety of industries such

as cell phone batteries, electric vehicles, quantum computing hardware, catalysts[1]. Compared to traditional Edisonian or trial-and-error approaches which usually strongly depend on the expertise of the scientists, computational materials discovery has the advantage of efficient search in the vast chemical design space. Togethre with inverse design[2, 3] and generative machine learning models[4, 5, 3, 6, 7], crystal structure prediction (CSP) [8, 9, 1, 10] is among the most promising approaches for new materials discovery. Crystal structure prediction also has important applications in crystal engineering, which is concerned with design and synthesis of molecular solid state structures with desired properties. For example, CSP allows us to conduct mutagenesis experiments to examine how composition changes may affect the structural mutations in terms of lattice constant changes or symmetry breaking. Crystal structure prediction can also be used to augment the X-ray diffraction (XRD) based crystal structure determination via space group identification [11] or providing initial parameters for the XRD based Rietveld refinement method for structure determination [12]

In a standard crystal structure prediction (CSP) problem[13], one has to find a crystal structure with the lowest free energy for a given chemical composition (or a chemical system such as Mg-Mn-O with variable composition) at given pressure–temperature conditions [1]. With the crystal structure of a chemical substance, many physicochemical properties can be predicted reliably and routinely using first-principle calculation or machine learning models [14]. It is assumed that lower free energy corresponds to the more stable arrangement of atoms. The CSP approach for new materials discovery is especially appealing due to the efficient sampling algorithm that generates diverse chemically valid candidate compositions with low free energies[4]. CSP algorithms based on evolutionary algorithms [15] and particle swarm optimization [16] have led to a series of new materials discoveries [17, 1, 18]. However, these global free energy search based algorithms have a major obstacle that limits their successes to relative simple crystals [1, 19] (mostly binary materials with less than 20 atoms in the unit cell[1, 18]) due to their dependence on the costly DFT calculations of free energies for sampled structures. With limited DFT calculations budget, how to efficiently sample the atom configurations becomes a key issue [17, 13]. To improve the sampling efficiency, a variety of strategies have been proposed such as exploiting symmetry[20] and pseudosymmetry[13], smart variation operators, clustering, machine-learning interatomic potentials with active learning [21]. Several heuristic ideas that exploit known structures in the databases have also been proposed [22]. However, the scalability of these 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 [23] and crystal structure generation for a given chemical system [24] and autoencoder based models for crystal structure generation[6, 7]. Compared to global free energy approaches in CSP, these methods can take advantage of the implicit composition, atomic configuration rules, and constraints embodied in the large number of known crystal structures that can be learned by the deep neural network models. Using neural networks to implicitly learn such rules may lead to more efficient sampling of the search space [23].

Herein we explore a new knowledge-rich approach for crystal structure prediction, which is inspired by the recent success of deep learning approaches for protein structure prediction (PSP)[25] led by the famous AlphaFold [26] algorithm from Google DeepMind. In the PSP 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[27] or distance matrix[26], which can then be used to reconstruct the full three-dimensional (3D) protein structure with high accuracy[28]. 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 atomic contact map. The idea is that we can exploit the rich atom interaction distribution or other geometric patterns or motifs [29] existing in the large number of known crystal structures to predict the atomic contact map. The space group of crystal structures can also be predicted using a variety of prediction algorithms [30, 31] or be inferred from domain knowledge [32]. In [31], 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 predicted contact map and the space group, we investigate whether global optimization algorithms such as GAs and CMA-ES methods can be used for predicting its crystal structure. The comparison of the main differences between conventional CSP and knowledge-rich CSP is shown in Figure1.

Our contributions can be summarized as follows:

- We propose a new approach for crystal structure prediction using the atomic contact map as a knowledge-rich methodology for solving CSP problems.
- We define a series of benchmark test cases for testing global optimization algorithms to reconstruct the atomic configurations from atomic contact maps
- we conduct extensive evaluations of how different global optimization algorithms perform in contact map based crystal structure prediction.

(a) Conventional free energy minimization based CSP
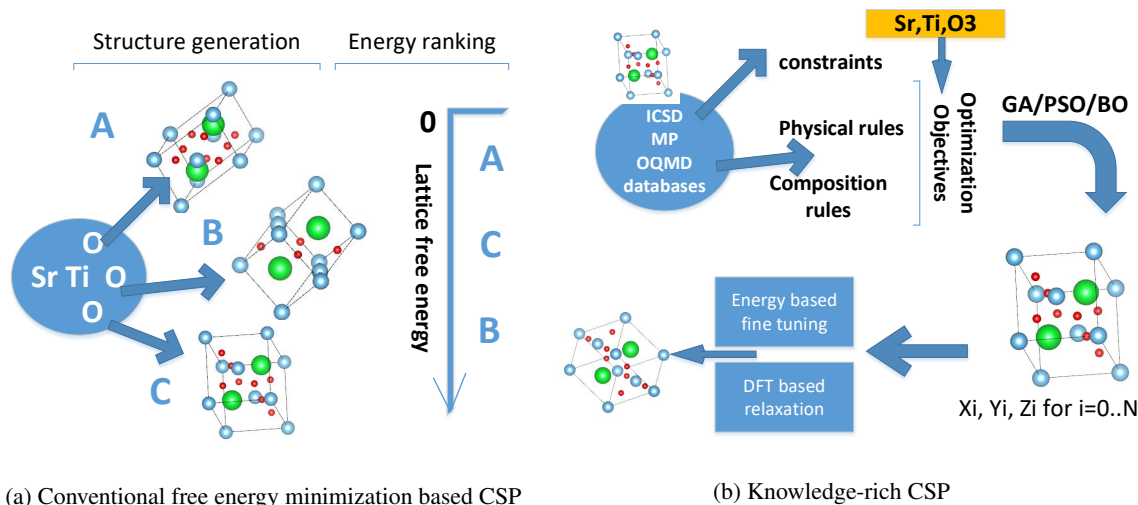
(b) Knowledge-rich CSP

Figure 1: Comparison of traditional crystal structure prediction (CSP) and knowledge-guided CSP. Conventional CSP is limited by its dependence on expensive DFT calculations of free energies while knowledge-rich CSP exploits chemical rules and geometric or physical constraints from known crystals to guide the structure search.

2 Materials and Methods

2.1 Problem formulation: knowledge-rich contact map based CSP

(a) Cif file of crystal material $Fe_{12}O_{12}$
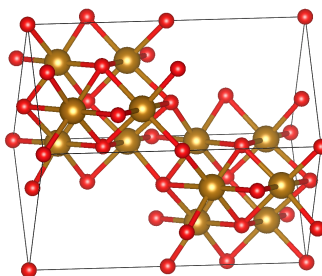
(b) Graph representation of the crystal $Fe_{12}O_{12}$

Figure 2: Cif and graph representation of crystal materials

A periodic crystal structure can be represented by its lattice constants a,b,c and angles α, β, and γ, the space group, and the coordinates at unique Wyckoff positions. Using a threshold 3.0 Å, the crystal structure can be converted into a graph, which can be represented as an adjacency matrix, or contact map. To get more controlled experiments, in our experiments, we used the thresholds used by the VESTA software to define the contact map. The contact map captures the interactions among atoms in the unit cell, which can be predicted by the known interaction patterns of these atom pairs in other known crystal materials structures. Here we assume that the perfect atom contact maps have been obtained, 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 contact map, with or without adding other geometric or physical constraints. By formulating the contact map based CSP as an optimization problem, it allows us to evaluate how different global optimization algorithms such as genetic algorithms (GA), particle swarm optimization (PSO), differential evolution (DE) can solve this problem and how difficult this reconstruction problem is for different crystal structures of varying complexity in terms of the number of unique Wyckoff positions(which determine the number of independent variables to optimize), the level of symmetry as represented by the space group, and also the number of atoms in the unit cell, which determines the number of contact constraints. For the example in Figure2, 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 contact map 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 contact map 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 [33, 31, 34, 35]. While only contact map 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 a synergistic effect by e.g. reducing the number of DFT free energy calculations.
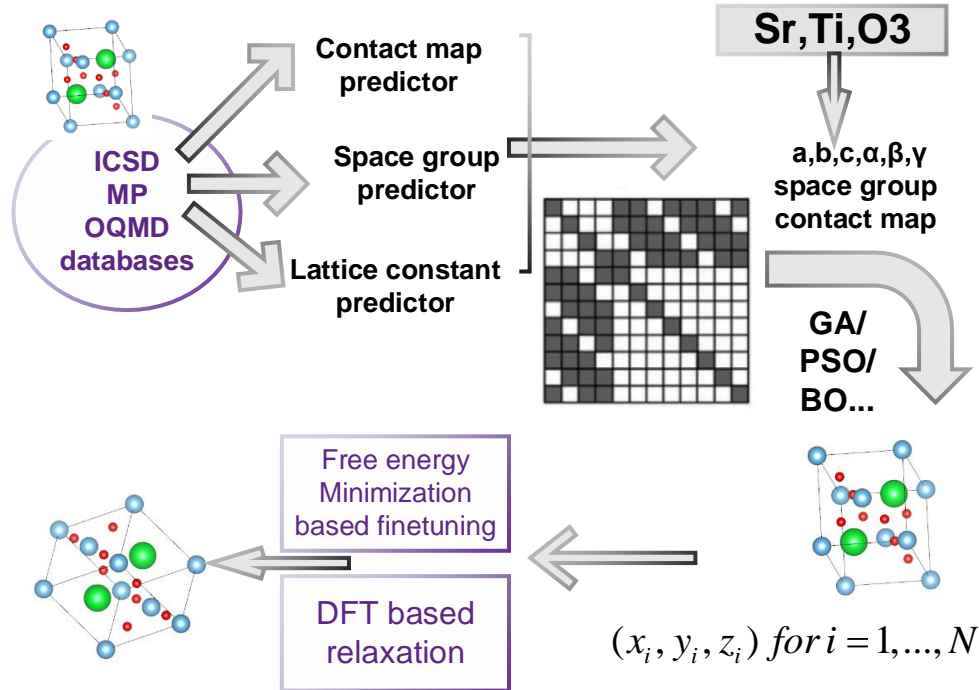

Figure 3: The CMCrystal algorithm for contact map based crystal structure prediction.

2.2 Contact map based CSP 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 and PSO to search the coordinates by maximizing the match between the contact map of the predicted structure and the contact map of the target crystal structure. This CMCrystal CSP framework is given in Figure3. Basically, first, using the existing inorganic materials samples in the databases such as ICSD, Materials Project, and OQMD, three prediction models will be trained including a space group predictor [30, 31], a lattice constant predictor[36, 35, 34, 37, 38], and a contact map predictor. And then given these information a global optimization algorithm, such as the genetic algorithm, particle swarm optimization, or Bayesian optimization, to search the atom coordinates such that the resulting structure's topology(contact map) matches the predicted contact map 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.

In this work, we focus on exploring how global optimization can be used to search the atom coordinates guided by a given contact map. We apply a set of six state-of-the-art global optimization algorithms to different problem instances to evaluate and compare their performance. Here, we summarize the main ideas of the selected optimization algorithms, their advantages, and key hyper-parameters.

2.2.1 Genetic algorithms(GA)

Genetic algorithms[39] are population-based search 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[40] and has been used in several crystal structure prediction algorithms [8, 41, 42], and mainly for free energy minimization. The main hyper-parameters include the population size, crossover and mutation rates. Here we apply the real-value encoded GA as the global optimization procedure for crystal structure reconstruction.

2.2.2 Differential evolution (DE)

Differential evolution [43] is a stochastic, population-based evolutionary optimization algorithm designed for optimize real parameter, real-valued functions, many of which are nondifferentiable, non-continuous, non-linear, noisy, flat, multi-dimensional or have many local minima, constraints or stochasticity. While genetic algorithms more focus on the crossover operator, DE mainly uses its special mutation operator, which generates new candidates by adding a weighted difference between two population members to a third member. This mutation operator has an inherent adaptive characteristic to make smaller mutations when the population approaches global or local optima. It is thus usually robust and has fast convergence. It has three main parameters: the population size (usually 5-10 times of the number of variables), the scaling factor F, and the crossover rate.

2.2.3 Particle Swarm Optimization (PSO)

Particle swarm optimization (PSO) [44] is a population-based stochastic optimization algorithm inspired by the social behavior of some animals such as flocks of birds or schools of fish to solve nonlinear global optimization problems. The algorithm seeks the optimal solution through collaboration and information sharing between searching individuals in the group. Each individual updates its movement through the search space by combining some aspect of its own history of current and best (best-fitness) locations with those of one or more members of the swarm. Although PSO can fall into the local optimum for complexity problems, its search speed is fast, efficient, and the algorithm is simple, so it is used in the article to optimize the target.

2.2.4 Bayesian Optimization (BO)

Bayesian optimization (BO) [45, 46] is an algorithm for optimizing expensive objective functions that take a long time to evaluate. It is good for optimization over continuous domains of less than 20 dimensions[47]. Bayesian optimization is one of the most efficient approaches to optimization in terms of the number of function evaluations by incorporating problem belief about the problem to help direct the sampling and by employing an automated mechanism to trade-off exploration and exploitation of the search space based on its acquisition function based sampling. Common acquisition functions include expected improvement, entropy search, and knowledge gradient. It usually uses Gaussian process regressor or deep neural networks[48] to build a surrogate model [49] for the expensive objective function and uses Bayesian estimation to calculate the prediction uncertainty at each sampling points. BO has been widely used in tuning hyper-parameters for machine learning algorithms [50]and active learning for materials design [51].

2.2.5 Covariance Matrix Adaptation-Evolution Strategy (CMA-ES)

Covariance Matrix Adaptation-Evolution Strategy (CMA-ES)[52] is a random, fast and robust local search algorithm that does not need to calculate gradients. It samples new candidate solutions from the multivariate normal distribution of its mean, and adapts after each iteration. CMA-ES is mainly used to solve nonlinear and non-convex optimization problems. It belongs to a category of evolutionary algorithms and has randomness. Compared with most other evolutionary algorithms, it is a quasi-parameterless algorithm. CMA-ES is one of the most effective methods to deal with difficult numerical optimization problems [53]. CMA-ES has been widely used in practical problems [54, 55, 56]. This algorithm is superior to all other similar learning algorithms in the benchmark multimodal functions. Good results with CMA-ES can be achieved when given a very large evaluation budget [57].

2.2.6 RBF Model-based optimization (RBFOpt)

RBFOpt [58] is a continuous optimization algorithm based on the Radial Basis Function method. It constructs and iteratively refines a surrogate model of the unknown objective function and exploits a noisy but less expensive surrogate model to accelerate convergence to the optimum of the exact oracle. In this aspect, it shares some principles with the Bayesian optimization approach. It also introduces an automatic model selection phase during the optimization process.

One of its key ideas is to use RBF interpolation to build a surrogate model, and define a measure of "bumpiness". Given a target objective function value at a sampling point, its bumpiness measures the likelihood that the target function value occurs there, based on the interpolation points. The assumption is that the unknown function f does not oscillate too much so that a model that can explain the data and minimizes the bumpiness can be found. Previous benchmark studies show that this algorithm has high efficiency in terms of the number of evaluations and robustness.

2.3 Objective function and Evaluation Criteria

The objective function for contact map based structure reconstruction is defined as the dice coefficient, which is shown in the following equation:

$$\text{fitness}_{opt} = \text{Dice} = \frac{2|A \cap B|}{|A| + |B|} \approx \frac{2 \times A \bullet B}{\text{Sum}(A) + \text{Sum}(B)} \tag{1}$$

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, $\bullet$ 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 the contact map accuracy.

To evaluate the reconstruction performance of different algorithms, we can use the dice coefficient 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:

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

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

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 contact map, it is possible that the predicted atom coordinates are oriented differently from the target atoms in terms 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.

3 Experiments

3.1 Test problems

We have selected a set of target crystal structures as test cases for evaluating the proposed contact map based crystal structure reconstruction algorithm using different global optimization algorithms. 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, and orthorhombic structurs (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 three. For GA and DE, we set the population size to 100 and the number of generations to 1000 with a mutation probability of 0.001. For PSO, the number of particles is set as 100. For CMA-ES, we set the population size to be 300 and the generation number to be 1000. For RBFOpt, we set the max_iterations to be 1000 and the maximum number of function evaluations in accurate mode to be 300. The population size and generation number are set based on our empirical experiments which allows us to find reasonably good structures. large population size or longer running time may further improve the results if premature convergence issue is controlled well.

3.3 Results

3.3.1 Successful contact map based crystal structure predictions

To evaluate our CMCrystal method for crystal structure prediction, we apply it to a selected set of 11 target structures as shown in Table1 with the number of atoms ranging from 6 to 16. The total number of objective evaluations is set as 100,000. The overall performance of different global optimization algorithms for contact map based crystal structure reconstruction is shown in Table 2. 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 contact map as the target. Table2 also shows the RMSD and MAE of the predicted structures compared to the target structures, both of which are calculated in terms of fractional coordinates of the independent atom sites. The RMSD values range from 0.07 to 0.381 with MAE ranging from 0.054 (for B_4N_4) to 0.335 (for Ni_8P_8).

Figure4 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 (Figure4(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 0.07 which is smaller than the RMSD (0.124) of Bi_4N_4, which is reflected by the higher similarity of the pairs of B_4N_4 than the pair of structures of Bi_4N_4. The contact map accuracy for the target structure of Co_4As_8 is lower with a value of 92.3% and higher RMSD of 0.197. We note that the topology of the predicted structure in general can reach the target topology while the precise coordinates can be different.

3.3.2 Performance comparison of different algorithms

To compare the performance of different optimization algorithms for crystal structure reconstruction from contact maps, we run all six optimization algorithms for a set of target structures of different complexity. Figure5 shows the performance of six algorithms. For easy cases of Bi_4Se_4 and B_4N_4, all algorithms reach the 100% accuracy for contact map prediction. For the more complex one Ni_8P_8, only DE achieves 100% accuracy in the given computing budget (100,000 evaluations) while PSO and BO fall behind the most. For the most complex target Co_4As_8, no algorithms have achieved an accuracy of 100% while GA, DE, CMA-ES, and RBFOpt all achieve 92% accuracy.

We also compared the RMSD performance for the six algorithms as shown in Figure6. Here we find that the CMA-ES achieves the best RMSD performance (0.07 and 0.12 respectively) for B_4N_4 and Co_4As_8. For Bi_4Se_4, the best result is obtained by GA with a RMSD of 0.12. For Ni_8P_8, the best result is obtained by RBFOpt with a value of 0.19. However, we must note that the objective function in our study here contains only the topology information, the contact

map. So algorithms with better contact map accuracy do not necessarily have better RMSD performances. In terms of computational complexity, our experiments are run on a Dell Precision workstation using a single CPU core with 1.7GHz. For most of the global optimization experiments here, each experiment takes about 40 minutes for results in
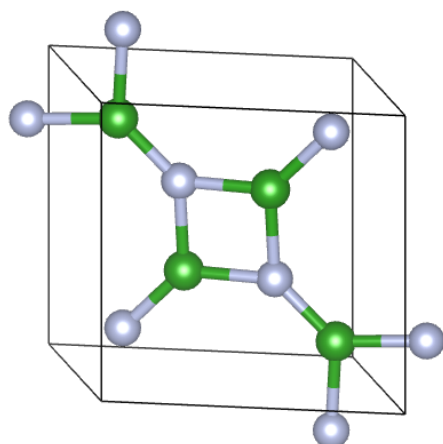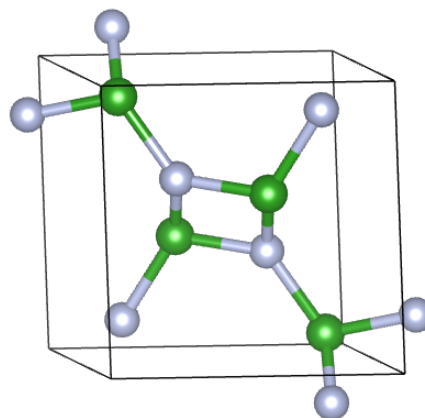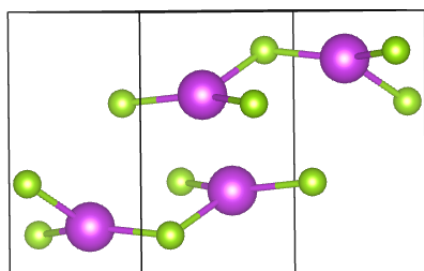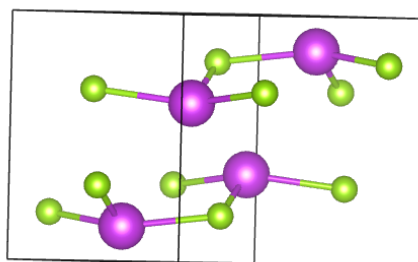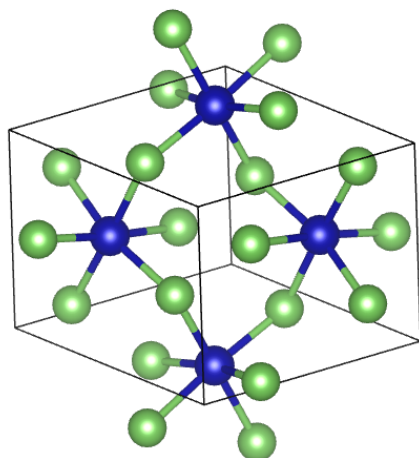

(a) Target structure B_4N_4

(b) Predicted structure of B_4N_4 with contact map accuracy:100%, RMSD:0.07

(c) Target structure of Bi_4Se_4

(d) Predicted structure of Bi_4Se_4 with contact map accuracy:100%, RMSD:0.124

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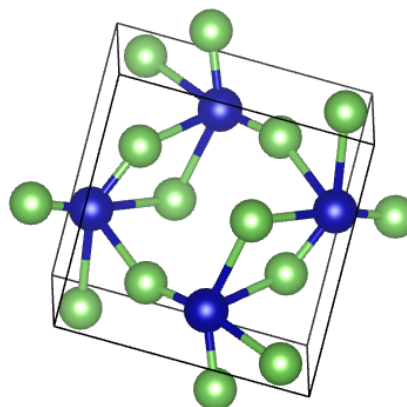
(f) Predicted structure of Co_4As_8 with with contact map accuracy:92.3%, RMSD:0.197

Figure 4: Examples of crystal structures predicted by CMCrystal(with GA) versus true target structures.

Table 2: Performances of genetic algorithms in terms of contact map prediction accuracy

Target material	contact map accuracy	RMSD	MAE
Ag_4S_2	1.000	0.320	0.233
Bi_4Se_4	1.000	0.124	0.097
B_4N_4	1.000	0.070	0.054
Pb_4O_4	1.000	0.246	0.196
S_4N_4	1.000	0.156	0.137
Te_4O_8	1.000	0.379	0.266
W_4N_8	1.000	0.368	0.214
Cd_4P_8	1.000	0.320	0.204
Co_4As_8	0.923	0.197	0.149
Bi_8Se_4	0.889	0.257	0.232
Ni_8P_8	1.000	0.381	0.335

Table 2, 70 minutes for results in Table 4 and 120 minutes for results in Table 3, which are marginal compared to the computationally demanding DFT based search algorithms.

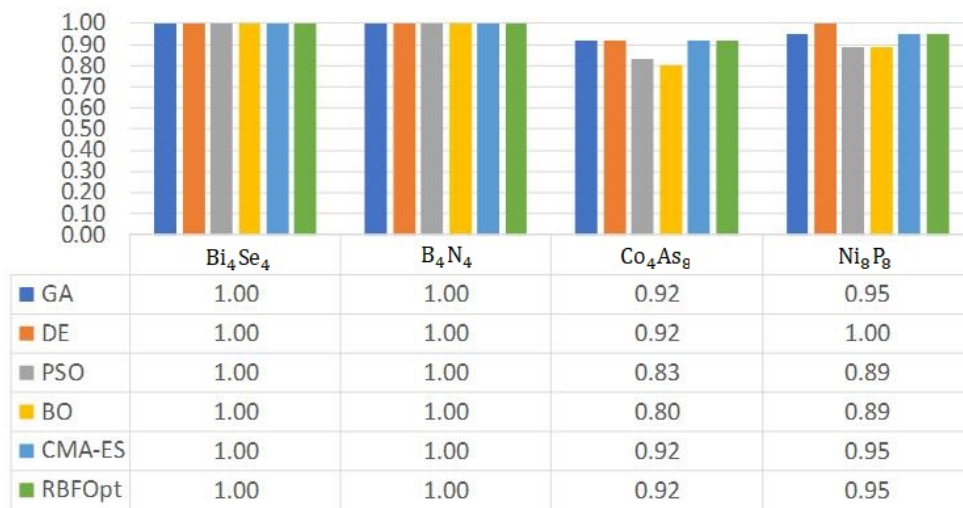|	Bi_4Se_4	B_4N_4	Co_4As_8	Ni_8P_8
■ GA	1.00	1.00	0.92	0.95
■ DE	1.00	1.00	0.92	1.00
▥ PSO	1.00	1.00	0.83	0.89
■ BO	1.00	1.00	0.80	0.89
■ CMA-ES	1.00	1.00	0.92	0.95
■ RBFOpt	1.00	1.00	0.92	0.95

Figure 5: Performance comparison of different algorithms in terms of contact map prediction accuracy over four target structures

3.3.3   Factors that affect the optimization difficulty

From our extensive experiments, we find that there are several factors that affect the crystal structure prediction performance of our algorithms such as the number of independent atom sites, the number of atoms in the unit cell, the space group, the number of bonds/topology constraints, etc. Here we report how two factors, the number of independent atomic sites and the space group, affect the crystal structure reconstruction performance by the CMA-ES algorithm. To gain a more intuitive comparison, we plot both results in Figure7.

In Figure7(a), we compare the performance of CMA-ES for problem instances 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 prediction accuracy gradually drops with an increasing number of atom sites, which corresponds to more optimization variables for the optimization problem. This trend is also reflected by the corresponding RMSD errors as shown in Table3.

Figure7(b) and Table4 show the performance results of CMA-ES structure reconstruction for a set of materials with the same number of five atom sites and similar numbers of atoms but different space groups. It is found that in general the higher the space group, the contact map accuracy is higher, indicating that higher symmetry puts more constraints on the atom configurations and reduces its search space so that better performance can be found. To be more specifically, as Table4 shows, the contact map accuracy increases from 0.811 to 0.923 when the space group goes from 2 to 194 for Mg4Co2H10, a hexagonal structure. At the same time, the RMSD error has no consistent trend as it goes up to 0.37 and
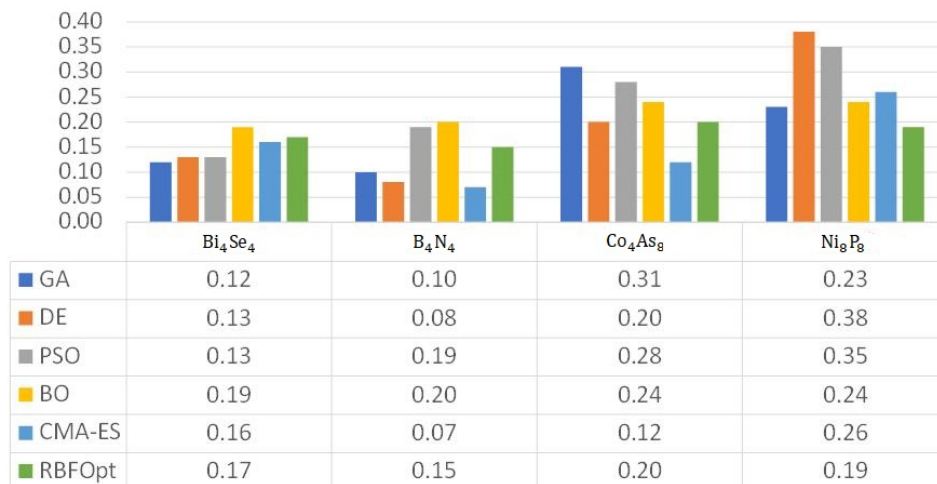
Figure 6: Performance comparison of different algorithms in terms of contact map prediction root mean square distance (RMSD) over four target structures

drops to 0.182 and then goes up to 0.380 and goes down to 0.276. As we discuss above, since we have only included the contact map without any distance information into our objective function, it is understandable that the RMSD have alternating up and downs.

(a) Contact map accuracy vs # of atom sites

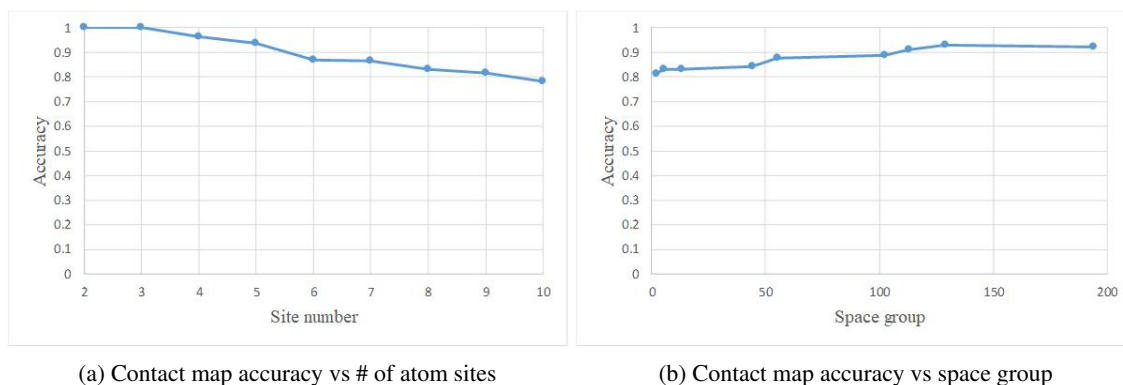(b) Contact map accuracy vs space group

Figure 7: Problem difficulty based on space group and number of atom sites

Table 3: Prediction performance versus number of atom sites for CMA-ES[52].

Target	mp_id	atom site#	atom #	contact map accuracy	RMSD
$La_{12}Se_{16}$	mp-491	2	28	1.000	0.193
$Bi_8Pd_4O_{16}$	mp-29259	3	28	1.000	0.257
V_8O_{20}	mp-25280	4	28	0.963	0.216
$Fe_{12}O_{16}$	mp-1192788	5	28	0.938	0.206
$Ge_8S_{12}I_8$	mp-27928	6	28	0.870	0.248
$Li_4V_4Si_4O_{16}$	mp-1176508	7	28	0.865	0.358
$Ba_2V_8O_{18}$	mp-18910	8	28	0.831	0.296
$Si_2H_{18}C_6Cl_2$	mp-867818	9	28	0.818	0.253
$Zr_8Cr_8F_{12}$	mp-690241	10	28	0.783	0.349

Table 4: Prediction performance versus space group using CMA-ES

Target	mp_id	atom site#	space group	contact map accuracy	RMSD
$Hg_8Cl_4O_4$	mp-636805	5	2	0.811	0.293
$Be_4B_2O_{10}$	mp-1079124	5	5	0.833	0.371
$Bi_6O_8F_2$	mp-757162	5	13	0.833	0.182
$Tl_6V_2O_8$	mp-29047	5	44	0.842	0.344
$La_4Sn_2S_{10}$	mp-12170	5	55	0.878	0.335
$Li_2V_2F_{12}$	mp-753573	5	102	0.889	0.380
$Yb_4H_4O_8$	mp-625103	5	113	0.909	0.310
$Mg_4Co_2H_{10}$	mp-642660	5	129	0.929	0.376
$K_{10}Cu_2As_4$	mp-14623	5	194	0.923	0.276

4   Discussion

Our systematic experiments have demonstrated the potential of our proposed knowledge-rich approach for crystal structure prediction. In this approach, machine learning models can be trained to predict the physical, chemical or geometric constraints such as the atomic pair contact maps and space group, and lattice parameter, which can then be used to reconstruct the coordinates of the atoms. Compared to existing ab initio free energy calculation based evolutionary search approaches, our method may achieve significant speed-up as it does not use the expensive first principle energy calculations. In many cases, the existing CSP approaches just get stuck and fail to find the stable structures. This shows the advantages of our algorithm to exploit the large number of known structures in crystal materials databases. One of the potential challenges of our algorithm is that its performance depends on the prediction accuracy of those constraints such as the contact map and lattice parameters, which may be biased by known structures. However, this issue can be addressed to certain extent by the DFT based first principle calculation of the formation energy or even thermodynamic stability phonon calculation to verify if the predicted structure is stable or not. Heuristic rules of crystal structures such as Pauling's rules can also be used to correct some errors in contact map prediction. On the other hand, the ab initio approaches have the advantages of unbiased free-energy guided global search in the configuration space, which is however usually subject to getting stuck in local optima and failure to find the true stable structures, especially for relatively large systems. Considering that the currently known 200,000 crystal structures in the ICSD database can be classified into about 9000 structure prototypes, it means that there is a large similarity among the crystal structures that can be exploited to do CSP. Actually, the advantages and disadvantages of our knowledge-rich approach CMCrystal compared to the ab initio approaches [59] are analogous to what has been discussed in the context of the protein structure prediction field, where the prior approaches have the dominating role[60]. While there is no coevolutionary information can be exploited in CSP, the physical and chemical rules that govern atom interactions allow predicting the contact with good performance.

Another factor that should be considered for CMCrystal algorithm is how well the required constraint information can be predicted. In a recent work on space group prediction[31], the top three space group prediction performance has reached a range of 0.81 to 0.98 in terms of R^2 scores based on different crystal systems. For cubic structures, this performance is 0.96 on average. For the lattice parameter prediction, it has achieved a performance of around 0.82 while our in-house algorithm has reached a R^2 of 0.97 for cubic structures and a R^2 range of 0.77 to 0.89 for other five crystal systems. Based on current performance and ongoing efforts in this area, we believe that our proposed algorithm can achieve good performance for quite some types of crystal materials.

While CMCrystal has been shown to be able to reconstruct atomic configurations from the contact map in this study, the quality depends on how accurate is the predicted contact map, which is an unsolved problem to be studied. However, considering the strong consistency of the bond lengths among atom pairs of different species in different compounds, the contact maps of crystal materials are expected be predictable. Actually, one of our ongoing works have focused on the development of deep neural network models for contact map prediction and has achieved very promising results.

5 Conclusion

We formulate a crystal structure prediction/reconstruction problem based on its space group symmetry and the atom contact map, and applied a series of state-of-the-art global optimization algorithms to solve the problem. Our experiments show that global optimization algorithms can reconstruct the crystal structure for some materials by optimizing the placement of the atoms using the contact map as the objective given only their space group and 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 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. However, we found that using the contact map alone is in general not enough to guide the search for the true structure precisely and additional geometric and physical constraints may be needed such as pairwise distance information to further improve the reconstruction quality, which is under our investigation. Another potential improvement is to conduct more extensive parameter tuning for the optimization algorithms used here for different structures as here we mostly use the default parameters for the algorithms.

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., E.S., R.D., Y.L.; investigation, J.H., W.Y., R.D., E.S., and S.L.; resources, J.H.; data curation, J.H. and W.Y.; writing–original draft preparation, J.H., R.D., Y.L, W.Y. and X.L.; writing–review and editing, J.H and R.D.; visualization, J.H., R.D., 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] Artem R Oganov. *Modern methods of crystal structure prediction*. John Wiley & Sons, 2011.

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

[11] Felipe Oviedo, Zekun Ren, Shijing Sun, Charles Settens, Zhe Liu, Noor Titan Putri Hartono, Savitha Ramasamy, Brian L DeCost, Siyu IP Tian, Giuseppe Romano, et al. Fast and interpretable classification of small x-ray diffraction datasets using data augmentation and deep neural networks. *npj Computational Materials*, 5(1):1–9, 2019.

[12] Yoshihiko Ozaki, Yuta Suzuki, Takafumi Hawai, Kotaro Saito, Masaki Onishi, and Kanta Ono. Automated crystal structure analysis based on blackbox optimisation. *npj Computational Materials*, 6(1):1–7, 2020.

[13] 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.

[14] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018.

[15] Artem R Oganov and Colin W Glass. Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. *The Journal of chemical physics*, 124(24):244704, 2006.

[16] Yanchao Wang, Jian Lv, Li Zhu, Shaohua Lu, Ketao Yin, Quan Li, Hui Wang, Lijun Zhang, and Yanming Ma. Materials discovery via calypso methodology. *Journal of Physics: Condensed Matter*, 27(20):203203, 2015.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] J Christian Schön. How can databases assist with the prediction of chemical compounds? *Zeitschrift für anorganische und allgemeine Chemie*, 640(14):2717–2726, 2014.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

[35] 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.

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

[37] 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.

[38] 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.

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

[40] 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.

[41] 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.

[42] 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.

[43] Kenneth Price, Rainer M Storn, and Jouni A Lampinen. *Differential evolution: a practical approach to global optimization*. Springer Science & Business Media, 2006.

[44] Riccardo Poli, James Kennedy, and Tim Blackwell. Particle swarm optimization. *Swarm intelligence*, 1(1):33–57, 2007.

[45] Jonas Mockus. Application of bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization*, 4(4):347–365, 1994.

[46] Daniel James Lizotte. *Practical bayesian optimization*. University of Alberta, 2008.

[47] Peter I Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

[48] Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Mostofa Patwary, Mr Prabhat, and Ryan Adams. Scalable bayesian optimization using deep neural networks. In *International conference on machine learning*, pages 2171–2180, 2015.

[49] Ky Khac Vu, Claudia d'Ambrosio, Youssef Hamadi, and Leo Liberti. Surrogate-based methods for black-box optimization. *International Transactions in Operational Research*, 24(3):393–424, 2017.

[50] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[51] Turab Lookman, Prasanna V Balachandran, Dezhen Xue, and Ruihao Yuan. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials*, 5(1):1–17, 2019.

[52] Nikolaus Hansen. The cma evolution strategy: a comparing review. In *Towards a new evolutionary computation*, pages 75–102. Springer, 2006.

[53] Nikolaus Hansen, Anne Auger, Raymond Ros, Steffen Finck, and Petr Pošík. Comparing results of 31 algorithms from the black-box optimization benchmarking bbob-2009. In *Proceedings of the 12th annual conference companion on Genetic and evolutionary computation*, pages 1689–1696, 2010.

[54] Peter Bayer and Michael Finkel. Evolutionary algorithms for the optimization of advective control of contaminated aquifer zones. *Water Resources Research*, 40(6), 2004.

[55] L Damp. Gonzalez., lf: Optimisation of the nose of a hypersonic vehicle using dsmc simulation and evolutionary optimisation. In *5th AIAA ASSC Space Conference*, 2005.

[56] Changying Li and Paul H Heinemann. A comparative study of three evolutionary algorithms for surface acoustic wave sensor wavelength selection. *Sensors and actuators B: Chemical*, 125(1):311–320, 2007.

[57] Christoph Waibel, Thomas Wortmann, Ralph Evins, and Jan Carmeliet. Building energy optimization: An extensive benchmark of global search algorithms. *Energy and Buildings*, 187:218–240, 2019.

[58] Alberto Costa and Giacomo Nannicini. Rbfopt: an open-source library for black-box optimization with costly function evaluations. *Mathematical Programming Computation*, 10(4):597–629, 2018.

[59] Jooyoung Lee, Peter L Freddolino, and Yang Zhang. Ab initio protein structure prediction. In *From protein structure to function with bioinformatics*, pages 3–35. Springer, 2017.

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