DISTANCE MATRIX BASED PREDICTION OF CRYSTAL STRUCTURES 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, Rongzhi Dong
School of Mechanical Engineering
Guizhou University
Guiyang China 550050

Dilanga
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC 29201

August 18, 2021

ABSTRACT

Crystal structure prediction has been frequently used in discovery of new materials. Global optimization methods such as genetic algorithms (GA) and particle swarm optimization (PSO) have been combined with first principle free energy calculations to predict crystal structures given composition or only a chemical system. However, these approaches usually do not exploit the large amount of implicit rules and constraints of atom configurations embodied in the large number of known crystal structures. Here we propose DMCrystal, an algorithm that exploits geometric constraints such as the atomic contact map of a target crystal structure to predict its structure given its space group information. We develop a novel encoding scheme for GA that incorporates space group symmetry properties of crystal structures to reconstruct the crystal structure given the atomic contact map. We show that atomic interaction information learned from existing materials can be used to improve crystal structure prediction.

***K*eywords** crystal structure prediction · space group · contact map · genetic algorithms · implicit rules

1 Introduction

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]. In the past decade, crystal structure predictions [2, 1, 3] are among the most promising approaches for new materials discovery.

In a standard crystal structure prediction (CSP) problem[4], one has to find a crystal structure with low free energy for a given chemical composition (or a chemical system such as Mg-Mn-O with variable composition) [1]. With the crystal structure of a chemical substance, many physichochemical properties can be predicted reliably and routinely using first-principle calculation or machine learning models [5]. CSP algorithms based on evolutionary algorithms[6] and particle swarm optimization[7] have led to a series of new materials discoveries [1]. However, these global free energy search based algorithms have a major obstacle that limits their successes [1, 8] due to their dependence on the costly DFT calculations of free energies for sampled structures. The scalability of these approaches remains an unsolved issue.

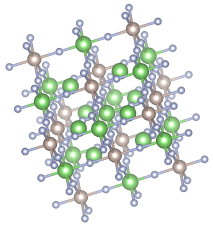
Herein we explore a new knowledge-rich approach for crystal structure prediction, which is inspired by the recent success of deep learning approaches for protein structure prediction (PSP) led by the famous AlphaFold [9] algorithm from Google. In the PSP problem, one has to predict the 3D tertiary structure of a protein given only its amino acid sequence. The latest approach uses deep learning to predict the contact maps[10] or distance matrix[9], which can then be used to reconstruct the full three-dimensional (3D) protein structure with high accuracy[11]. In this paper, we are exploring whether genetic algorithms can be used 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 [12] 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 [13] or be inferred from domain knowledge [14]. In [15], 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.

2 Materials and Methods

2.1 Problem definition:contact map based CSP for structures with high symmetry

```
Lattice type         F
Space group name     F m -3 m
Space group number   225
Setting number       1

Lattice parameters

   a        b        c        alpha    beta     gamma
 7.89265  7.89265  7.89265  90.0000  90.0000  90.0000

Unit-cell volume = 491.663896 Å^3

Structure parameters

                        x          y          z          Occ.     B     Site
   1 Li  Li0          0.25000    0.25000    0.25000    1.000    1.000    8c
   2 Li  Li1          0.00000    0.00000    0.50000    1.000    1.000    4b
   3 Ru  Ru2          0.00000    0.00000    0.00000    1.000    1.000    4a
   4 F   F3           0.00000    0.00000    0.25313    1.000    1.000   24e
================================================================================
```

(a) Cif file of crystal material $Li_{12}Ru_4F_4$



(b) Graph representation of the crystal $Li_{12}Ru_4F_4$

Figure 1: Cif and graph representation of crystal materials

The basic problem of CSP is how to predict its crystal structure given only its composition, e.g. $Li_{12}Ru_4F_4$. A periodic crystal structure (Figure1) 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 of 3.5 Å, the crystal structure can be converted into a graph, which can be represented as an adjacency matrix, or contact map. The contact map 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. 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 if global optimization algorithms such as genetic algorithms (GA) can solve this problem.

For the crystal $Li_{12}Ru_4F_4$ in Figure1, 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 20 atoms in the unit cell, which can be mapped into a 20x20 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 225, the generated crystal structure will have the same contact map matrix. However, structures with high symmetry have a special physical periodic patterns as reflected by their fractional coordinates (Figure1(a)): while the space group 225 has 192 symmetric operations, the symmetry multiplicity of the three atoms of elements K0, K1, Lu2, Cl3 are only 8, 4, 4, 24 respectively. This is due to the special fraction coordinate values such as 0.25, 0, 0.5 that maps multiple transformations into identical positions. However, during the global fraction coordinate search, most arbitrary coordinate values will lead to totally different number of symmetry multiplicity, which are invalid structures for a given space group.

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 [15, 16, 17]. 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 synergistic effect by e.g. reducing the number of DFT free energy calculations.

2.2 Contact map based CSP for high symmetry structures using genetic algorithms

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 contact map of the predicted structure and the contact map of the target crystal structure. In this framework,first, the existing inorganic materials samples in the databases such as ICSD, Materials Project, and OQMD are used to train three prediction models including a space group predictor [13], a lattice constant predictor[18, 17], and a contact map predictor. And then given these information a global optimization algorithm such as the genetic algorithm will be used 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 the genetic algorithm optimization to different problem instances to evaluate the performance.

2.3 Genetic algorithms

Genetic algorithms[19] 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 and has been used in several crystal structure prediction algorithms [2, 20], but mainly for free energy minimization. The main hyper-parameters include the population size, crossover and mutation rates. Here we apply the binary encoded GA as the global optimization procedure for crystal structure reconstruction.

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]. However, for crystal structures with high symmetry, the number of symmetric operations are not equal to the symmetry multiplicity of some of the independent atoms. This is due to the special fraction coordinate values including 8 types of them: 1/6,1/4,1/3,1/2,2/3,3/4,5/6,0,-3/8,-1/4, -1/8. Atom coordindates with these values will lead to degenerate atom positions, leading to different dimensions of the contact map. We also find that some degeneracy is caused by the double, triple, or quadratic relationships of x,y/x,z/y,z relationships. The existence of such degeneracy will makes the structure derived from varying coordinates to be invalid for a given space group with a given composition/formula. There are two approaches to address this issue, either by using aftermath penalty to the objective function, which is inefficient or by using special encoding to ensure that the generated structures are valid structures of the given space group.

encoding of special fractional coordinates

Figure 2: Encoding for high symmetry structure search.

2.4 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 as 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:

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

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

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 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,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 3. For GA and DE, we set the population size to 100 and the number of generations to 1000 with mutation probability of xx. For PSO, the number of particles is set as 100. For CMA-ES, we set the population size to be 300 and 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.

3.3 Results

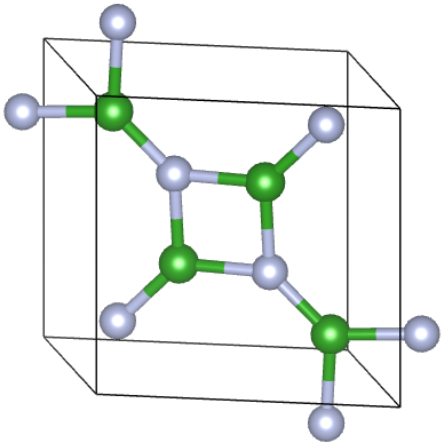
3.3.1 Successful contact map based crystal structure predictions

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).

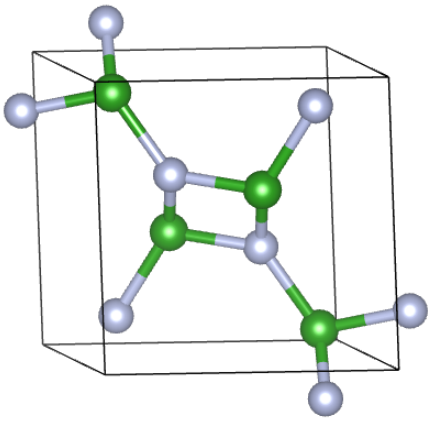
Figure3 shows three sets of predicted and target crystal structures of B_4N_4, Bi_4Se_4, and Co_4As_8. For both B_4N_4 and Bi_4Se_4 (Figure3(a)-(d)), the contact map accuracy reaches 100% and the predicted structures are very close to the target structures. The RMSD of B_4N_4 is 0.07 which is smaller 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.

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

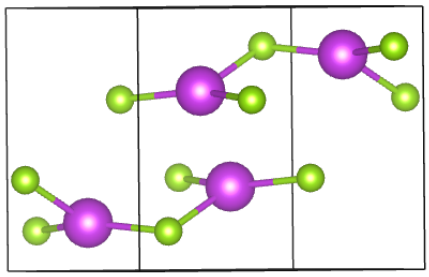
Target material	**contact map accuracy**	**RMSD**	**MAE**
Ag_4S_2	1.000	0.320	0.233
Bi_4Se_4	1.000	0.124	0.097
B_4N_4	1.000	0.070	0.054
Pb_4O_4	1.000	0.246	0.196
S_4N_4	1.000	0.156	0.137
Te_4O_8	1.000	0.379	0.266
W_4N_8	1.000	0.368	0.214
Cd_4P_8	1.000	0.320	0.204
Co_4As_8	0.923	0.197	0.149
Bi_8Se_4	0.889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



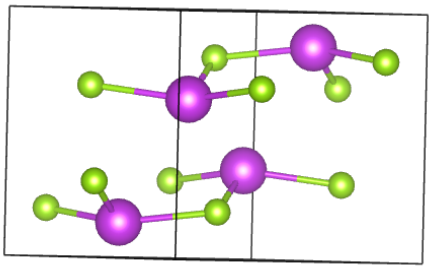
(a) Target structure B_4N_4



(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

Figure 3: Examples of predicted versus target crystal structures.

4 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 are able to reconstruct the crystal structure for some materials by optimizing the placement of the atoms using the contact map as the objective for certain inorganic materials 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 free energy 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.

5 Contribution

Conceptualization, J.H.; methodology, J.H. and W.Y.; software, W.Y. and J.H; validation, W.Y, R.D., Y.L. and J.H.; investigation, J.H. and W.Y.; resources, J.H.; data curation, J.H. W.Y.; writing–original draft preparation, J.H., R.D., Y.L,

and W.Y.; writing–review and editing, J.H; visualization, W.Y, R.D., and J.H; supervision, J.H.; funding acquisition, J.H. and S.L.

6 Acknowledgement

Research reported in this work was supported in part by NSF under grant and 1940099 and 1905775 and by NSF SC EPSCoR Program under award number (NSF Award OIA-1655740 and GEAR-CRP 19-GC02). The views, perspective, and content do not necessarily represent the official views of the SC EPSCoR Program nor those of the NSF. This work was also partially supported.

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

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

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

[20] 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.