Crystal structure prediction by combining graph network and Bayesian optimization

Guanjian Cheng
Soochow University

Xin-Gao Gong
Fudan university

Wan-Jian Yin (wjyin@suda.edu.cn)
Soochow University

Keywords: crystal structure prediction, graph network, GN-BOSS, DFT-free

DOI: https://doi.org/10.21203/rs.3.rs-814684/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Crystal structure prediction by combining graph network and Bayesian optimization

Guanjian Cheng\textsuperscript{1}, Xin-Gao Gong\textsuperscript{2,3}, Wan-Jian Yin\textsuperscript{1,4,5,*}

\textsuperscript{1}College of Energy, Soochow Institute for Energy and Materials InnovationS (SIEMIS), and Jiangsu Provincial Key Laboratory for Advanced Carbon Materials and Wearable Energy Technologies, Soochow University, Suzhou 215006, China
\textsuperscript{2}Key Laboratory for Computational Physical Sciences (MOE), State Key Laboratory of Surface Physics, Department of Physics, Fudan University, Shanghai 200433, China
\textsuperscript{3}Shanghai Qi Zhi Institute, Shanghai, 200232, China
\textsuperscript{4}Light Industry Institute of Electrochemical Power Sources, Soochow University, Suzhou 215006, China
\textsuperscript{5}Key Lab of Advanced Optical Manufacturing Technologies of Jiangsu Province & Key Lab of Modern Optical Technologies of Education Ministry of China, Soochow University, Suzhou 215006, China
\textsuperscript{*wjyin@suda.edu.cn}

Abstract
We developed a density functional theory (DFT)-free approach for crystal structure prediction, in which a graph network (GN) is adopted to establish a correlation model between the crystal structure and formation enthalpies, and Bayesian optimization (BO) is used to accelerate the search for crystal structure with optimal formation enthalpy. The approach of combining GN and BO for crystal structure searching (GN-BOSS) can predict crystal structures at given chemical compositions with and without additional constraints on cell shapes and lattice symmetries. The applicability and efficiency of the GN-BOSS approach is then verified by solving the classical Ph-vV challenge. The approach can accurately predict the crystal structures with a computational cost that is three orders of magnitude less than that required for DFT-based approaches. The GN-BOSS approach may open new avenues for data-driven crystal structural predictions without using expensive DFT calculations.
Predicting crystal structure at a given chemical composition prior to experimental synthesis has attracted a long-lasting interest in condensed matter science. Earlier attempts based on empirical rules provide qualitative description of structures, for example, Pauling’s five rules for ionic crystals\(^1\), Goldschmidt’s tolerance factor for perovskite formability\(^2\), and dimensional descriptors to classify the zinc-blend (ZB)/wurtzite (WZ) and rock-salt (RS) structures for binary semiconductor compounds\(^3,4\). Owing to reliable energy calculation via density functional theory (DFT), the current state-of-the-art approaches for crystal structure prediction (CSP) mainly combine DFT calculations with structural searching algorithms such as (quasi-) random search\(^5,6\), simulated annealing\(^7\), genetic algorithm\(^8–11\), particle-swarm optimization (PSO)\(^12,13\), and differential evolutionary process\(^14\). These approaches extensively explore the structural candidates via searching algorithms, adopting DFT-calculated energy as a stability metric. The necessary DFT calculations involve the evaluation of numerous structural candidates in the process of structure searching and are thus time-consuming.

Advancement of machine learning (ML) in materials science has recently focused on its applications in predicting materials properties such as formation enthalpies (\(\Delta H\))\(^15,16\), Gibbs free energies\(^17\), bandgaps\(^18,19\), even wave function and electron density\(^20\), X-ray absorption spectra\(^21\), and phase transitions\(^22\). The accuracy of this approach is close to that of quantum mechanics calculations, but the computational costs are orders of magnitude lower. In addition to the influence of compositional atoms, the influence of their spatial arrangement, i.e., crystal structure, on materials properties has recently been analyzed via structural characterizing approaches such as the Wyckoff-species matrix-based method\(^23\), Voronoi tessellation method\(^24\), and graph network\(^18,19\). A crystal \(G\) can be represented by a vector \(\{C_i\}_{i=1,N}, \{R_i\}_{i=1,N}, L\), where \(\{C_i\}\) and \(\{R_i\}\) are elemental features and coordinates of the \(i\)th atom, \(N\) is the total number of atoms in a periodic cell, and \(L\) is the vector (\(a, b, c, \alpha, \beta, \gamma\)) defining the cell shape. In these approaches, crystal structures are transformed to a physically meaningful and algorithm-readable data format, such as a symmetry-invariant matrix\(^23\), bond configurations\(^24\), or crystal graphs\(^18\), enabling the establishment of a correlation model between a crystal and its formation enthalpy as follows:

\[
\Delta H = f(G(\{C_i\}, \{R_i\}, L)).
\]  

(1)

In principle, CSP can be efficiently performed using equation (1) by optimizing \(\{R_i\}, L\) to minimize \(\Delta H\) at a given \(\{C_i\}\). This approach substitutes DFT calculations with the ML model; therefore, it has the potential to significantly accelerate the CSP.
Despite this potential advantage, the practical approach of DFT-free CSP still has challenges. First, the ML model should have a sensitive response to the crystal structure; therefore, the fixed-structure model and symmetry-invariant model, which have a constraint on the crystal structures, are inapplicable or limited in determining the ground state crystal structure that may have arbitrary cell shape and atomic coordinates. Second, the high accuracy of DFT calculations benefit from the systematic cancellation of errors relative to the experiment, and the claimed DFT-level accuracies of the ML models are obtained from training data composed of stable crystal structures. The extension of ML models to structural searching is questionable because most structural candidates in the searching process are metastable or unstable, and their relative energies are crucial in determining the ground state structure. Finally, an appropriate optimization algorithm compatible with the black-box ML model is required.

In this study, we constructed an improved graph network to establish the ML model between ~320,000 crystal compounds and their DFT-calculated formation enthalpies, with mean absolute errors (MAE) as low as 26.9 meV/atom. This model was then combined with Bayesian optimization (BO) for CSP. The combined approach of the graph network-based Bayesian optimization structural searching (GN-BOSS) was then applied to successfully predict the crystal structures of 24 octet binary compounds, i.e., I-VII (I=Li, Na, K, Rb, Cs; VII=F, Cl) and II-VI (II=Be, Mg, Ca, Sr, Ba, Zn, Cd; VI=O, S), as well as C, Si, and CsPbI$_3$. The comparative studies show that the DFT-free GN-BOSS approach can predict crystal structures from scratch with extremely low computational cost. This study may open a new avenue for data-driven crystal structural prediction without using the expensive DFT calculations during structural searching.

First, a GN model was constructed to establish the correlation between the crystals and their formation enthalpies $\Delta H$. In GN, crystal structures are considered as a crystal graph $G$, numerically represented using atomic and structural features. We initially used 23 atomic features $\{C_i\}$ and 10 structural features. The structural features contain four types of attributes of Voronoi polyhedrons derived from $(\{R_i\}, L)$. The advantage of using GN to describe crystals is that it has no constraints on cell shape and lattice symmetry. In addition, it has well-developed update operations, concatenation, and aggregation for $G$, which help map an arbitrary crystal graph $G$ to $\Delta H$ [equation (1)]. Because the DFT-calculated $\Delta H$ depends on the setup parameters, we also considered three key setup parameters as atomic features in $\{C_i\}$ during GN model training. The details of the features (Tables S1–S3) and GN architecture are provided in Supplementary Materials.

GN was updated based on the DFT-calculated $\Delta H$ from the offline Open Quantum Materials
Database (OQMD). Data cleaning was performed to exclude data with incomplete information and restrictions: 

1. the number of atoms in the unit cell (less than 50),
2. PBE as exchange-correlation functional, and
3. kinetic energy cutoff (520 eV), making data as reliable and comparable as possible.

Accordingly, we obtained more than 320,000 data points, including ~40,000 experimentally known ones and ~280,000 hypothetical ones, covering 85 elements, 7 lattice systems, and 167 space groups. All data were divided into three parts: training set (50%), validation set (12.5%), and test set (37.5%). The best model was selected to minimize the errors between the GN-predicted and DFT-calculated $\Delta H$.

Previous studies have shown that the network architecture and update functions have a significant impact on the model performance. Here, we incorporated the MEGNet architecture (Fig. S1) using the atomic features and structural features of Voronoi polyhedrons. The MEGNet architecture was adopted because it is a universal framework for crystals and molecules, facilitating the future extension of the current approach to molecular structure prediction. In particular, our GN models can reach a MAE of 26.9 meV/atom when the portion of training data is 50% (Fig. 2), which is less than that of previous models such as CGCNN (39 meV, MP database), iCGCNN (30.5 meV, OQMD database), and original MEGNet (28 meV, MP database). We observed a significant decrease in MAE as the amount of training data increased, and the values rapidly converged below 50 meV when the portion of training data reached as low as 25%, showing the generality of the current GN model.

In the framework of GN, CSP is an optimization problem to identify the crystal graph $G(\{R_i\}, L)$ at a given chemical composition $\{C_i\}$ to minimize $\Delta H$ according to GN-constructed equation (1). The enumeration of all possible structures is a long-standing challenge in CSP. Here, we applied BO via a Gaussian mixture model based on tree of Parzen estimators (TPE) to explore the structural space. BO is a selection-type algorithm that is compatible with the black-box ML model, demonstrating a great capability to identify the global minimum, and has been recently combined with DFT calculations for CSP in fixed crystal systems. Compared to the normal BO algorithm based on the Gaussian process, which performs better in low-dimensional space (number of features less than 20), the TPE-based Gaussian mixture model demonstrated higher efficiency in high-dimensional space. Here, we combined the GN model with a BO algorithm as the GN-BOSS approach, as shown in Fig. 1, to solve the minimization problem, i.e., $\min_{\{C_i\}} f(G(\{C_i\}, \{R_i\}, L)).$ Since the PSO algorithm is successfully applied in DFT-based CSP [12], we also combined GN with PSO for a comparative study.

First, for a given number of atoms in a cell, $n$ initial structures $G_k(\{R_i\}, L)_{k=1,n}$ of random cell
shape and atomic positions were generated with or without crystal symmetry constraints (among 230 space groups excluding P1), and their corresponding structural features were obtained by conducting a Voronoi analysis\(^{24}\). Accordingly, \(\Delta H\)'s were predicted using our GN model (equation [1]). Next, the TPE-based Gaussian mixture model and acquisition functions were used to suggest a potential global minimum at structural spaces. Subsequently, its true formation energy was obtained using our GN model and iteratively fed back into the model to refine the BO model to suggest the \((n+2)th\) structure. For practical implementation, we added an additional constraint to avoid the generation of unreasonable structures with extremely small or large volumes. In particular, the surrogate function in the BO model is a type of approximation to the black-box function in equation (1), based on the limited dataset of \((G, \Delta H)\).

The GN-BOSS approach was then applied to solving the typical Ph-vV challenge\(^{3,4,38}\), i.e. identifying the crystal structures of octet binary compounds, including 10 I-VII (I = Li, Na, K, Rb, Cs; VII = F, Cl) compounds and 14 II-VI compounds (II = Be, Mg, Ca, Sr, Ba, Zn, Cd; VI = O, S). There are more than 300 types of prototype structures for AB compounds\(^{31}\); two representatives of these are tetrahedral-coordination ZB/WZ and octahedral-coordination RS structures. Predicting ZB/WZ and RS structures proves the ability of CSP from ionic to covalent systems\(^{38}\). Notably, in contrast to conventional approaches discerning two fixed structures of ZB/WZ and RS, GN-BO predicts their structures starting from an arbitrary cell shape. The GN model is trained based on the data, excluding the 24 compounds mentioned above. Here, we considered four-atom cell CaS, for example, to show how GN-BO works. First, 200 structures with random \(G\left(\{R_{i}\}, L\right)_{i=1,4}\) were generated, and then BO was iteratively performed to choose a low-\(\Delta H\) structure, with their GN-predicted \(\Delta H\) and typical structures shown in Fig. 3.

To show the performance of BO, random searching (RAS) and the PSO algorithm are also shown in Fig. 3 for comparison. To verify the extension of the GN model to the structural candidates in the searching procedure, we pick up structures every 50 RAS steps and use DFT to calculate their total energies. The DFT-calculated total energies follow a trend similar to the GN-predicted \(\Delta H\) (Fig. S2), ensuring the applicability of GN to select a low-\(\Delta H\) structure among numerous structural candidates.

GN-BO shows superior optimization performance than that of GN-PSO or GN-RAS. Most of the BO-selected structures are located in low-\(\Delta H\) configurations (exploitation); however, some of them are in high-\(\Delta H\) configurations (exploration). It can be clearly seen that GN-BO can find the ground state structures while GN-PSO and GN-RAS cannot. For GN-BO, the initial random structures do not make sense due to the existence of a Ca-Ca wrong bond and isolated unbonded S atoms, as shown in
Fig. 3(b). Next, the formation enthalpy decreased rapidly, and the reasonable structures without wrong bonds appeared as in Fig. 3(c). After 3000 steps, the four-coordination structure [Fig. 3(d)] was found, and after 5000 steps, the ground-state RS structure [Fig. 3(e)] was successfully discovered. Because the accuracy of the current GN model is mildly inferior to state-of-the-art DFT calculations (Fig. S2), the final structure in Fig. 3(e) still has some distortions. Additional BO steps may help to minimize these distortions (Fig. S3). However, predicting a perfect RS structure remains a current challenge because the GN model is still inferior to accurately perceive small coordinate perturbation to total energies. We further used DFT calculations to relax the GN-BO predicted structures and obtain the precise crystal structures. To verify that the ground state crystal structures were derived using GN-BO rather than DFT relaxations, we adopted 18 prototype structures of AB compounds, which frequently appeared in the Inorganic Crystal Structure Database as initial structures of CaS, and only three of them were relaxed to RS structure (Fig. S4). These results further confirm that the ground state structure is derived using the GN-BO approach, not by DFT relaxation. This indicates that DFT relaxations help to find the precise location of a local minimum in structural space \( \{\{R_i\}, L\}_{i=1}^{4} \) but cannot jump from one local minimum to another. In contrast, the GN-BO approach is able to find the basin around the global minimum, and DFT calculation help to identify the exact location of this minimum at that basin. To overcome the undesirable local minimums from GN-BO, crystal symmetry constraints were implemented and found to enable CSP within a large cell (up to 10 atoms) with a high efficiency (shown as GN-BO-SYM in Fig. 3).

The approaches of GN-BO, GN-PSO, GN-BO-SYM, and DFT-PSO-SYM [12] were then applied to structural searching for 23 other binary compounds. The DFT-PSO-SYM approach has no problem predicting experimental structures. The corresponding \( \Delta H \) evolution of the GN-RAS, GN-BO, GN-PSO, GN-BO-SYM approaches are shown in Figs. S5–S10. It can be observed that GN-BO and GN-BO-SYM can reproduce most of the experimental structures predicted by DFT-PSO-SYM, while GN-PSO failed to predict 19 of the 23 compounds within 10,000 steps. Notably, GN-BO and GN-BO-SYM finished those tasks with a computational cost about two or three orders of magnitude faster than DFT-PSO-SYM (Fig. 4). Finally, we adopted GN-BO-SYM and DFT-PSO-SYM to predict C (8 atoms), Si (8 atoms), and CsPbI\(_3\) (5 and 10 atoms) (Supplementary Materials) and found that GN-BO-SYM can also reproduce most of the results of DFT-PSO-SYM (Figs. S11-S14). An exceptional case is 10-atom CsPbI\(_3\) (Fig. S15). GN-BO-SYM predicted a cubic perovskite phase (\( Pm\bar{3}m \)) while DFT-PSO-SYM predicted a non-perovskite phase (\( Pnma \), which has an energy about 7 meV/atom lower than that of
the \textit{Pm\textbar m} phase. Such a small difference may not be discerned by the current GN model. The accuracy of the GN-BOSS approach depends on the reliability of the GN model. Since the development of the GN model is currently in progress\cite{40,41}, it is believed that the accuracy of GN-BOSS and similar ML-based CSP approaches would have a great prospect of improvement in the future.

In conclusion, we combined graph network and Bayesian optimization as a GN-BOSS approach for crystal structure prediction without performing DFT calculations. GN-BOSS was then applied to successfully predict the crystal structures of typical I-VII (I=Li, Na, K, Rb, Cs; VII=F, Cl) and II-VI (II=Be, Mg, Ca, Sr, Ba, Zn, Cd; VI=O, S) compounds. The comparative studies show that GN-BOSS, although with less accuracy than DFT results, can predict crystal structures from scratch without DFT. Meanwhile, the limitations of the current GN-BOSS approach are also apparent, such as its strong dependence on a reliable database. In terms of methodology, there are several directions that deserve further development, including crystal structure characterization, structural searching, and algorithm parallelization, to predict more complicated and unknown structures in a more efficient way. Current work may open a new avenue for data-driven crystal structural prediction without using the expensive DFT calculations during structural searching.

The work was supported by National Key Research and Development Program of China (grant Nos. 2020YFB1506400), National Natural Science Foundation of China (grant Nos. 11974257), Jiangsu Distinguished Young Talent Funding (grant No. BK20200003), Yunnan Provincial Key S&T Program (grant No. 202002AB080001-1), the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). DFT calculations were carried out at the National Supercomputer Center in Tianjin [TianHe-I(A)].

Methods

\textbf{Data collection and split.} All the data set for GN model training and evaluation are from the offline Open Quantum Materials Database (OQMD)\textsuperscript{31} of version 1.3, including crystal structure and DFT-calculated $\Delta H$. Data cleaning was performed to exclude data with incomplete information and restrictions: \textit{i}) the number of atoms in the unit cell (less than 50), \textit{ii}) PBE as exchange-correlation functional, and \textit{iii}) kinetic energy cutoff (520 eV), making data as reliable and comparable as possible. Accordingly, we obtained more than 320,000 data points has been obtained, including ~40,000 experimentally known ones and ~280,000 hypothetical ones, covering 85 elements, 7 lattice systems, and 167 space groups. All data were divided into three parts: training set (50%), validation set (12.5%), and test set (37.5%).
**Features.** As shown in Fig. 5, a crystal structure is considered as a crystal graph $G$, numerically represented using atomic and structural features [18,19]. A crystal graph $G$ can be represented by a vector $\{C_i\}_{i=1..N}$, $\{R_i\}_{i=1..N}$, $L$, where $\{C_i\}$ and $\{R_i\}$ are elemental features and coordinates of the $i$th atom, $N$ is the total number of atoms in a periodic cell, and $L$ is the vector $(a, b, c, \alpha, \beta, \gamma)$ defining the cell shape. We initially used 23 atomic features $\{C_i\}$ and 4 structural features $\{A_{ij}\}$. The structural features contain four types of attributes of Voronoi polyhedrons [28,29] derived from $\{R_i, L\}$. The details of the features (Tables S1–S3) are provided in Supplementary Materials.

**GN model.** In this article, we used MEGNet [19] as our GN model architecture and improved it (Fig. 6). The MEGNet architecture was adopted because it is a universal framework for crystals and molecules, facilitating the future extension of the current approach to molecular structure prediction.

**Bayesian optimization.** Here, we applied BO via a Gaussian mixture model based on tree of Parzen estimators (TPE) [34] to explore the structural space. BO is a selection-type algorithm that is compatible with the black-box ML model, demonstrating a great capability to identify the global minimum [35,36], and has been recently combined with DFT calculations for CSP in fixed crystal systems [37]. Compared to the normal BO algorithm based on the Gaussian process, which performs better in low-dimensional space (number of features less than 20), the TPE-based Gaussian mixture model demonstrated higher efficiency in high-dimensional space [34].

**Code availability**
The code for GN-BOSS approach is available on [http://www.comates.group/links?software=gn_boss](http://www.comates.group/links?software=gn_boss)
All GN-based results reported in this work can be reproduced by this code. The DFT-based results are produced by CALYPSO code.

**Reference**
1. Pauling, L. THE PRINCIPLES DETERMINING THE STRUCTURE OF COMPLEX IONIC CRYSTALS. *J. Am. Chem. Soc.* **51**, 1010–1026 (1929).
2. Goldschmidt, V. M. Die Gesetze der Krystallochemie. *Naturwissenschaften* **14**, 477–485 (1926).
3. Van Vechten, J. A. Quantum Dielectric Theory of Electronegativity in Covalent Systems. I. Electronic Dielectric Constant. *Phys. Rev.* **182**, 891–905 (1969).
4. PHILLIPS, J. C. Ionicity of the Chemical Bond in Crystals. *Rev. Mod. Phys.* **42**, 317–356 (1970). 8
5. Pickard, C. J. & Needs, R. J. Ab initio random structure searching. *J. Phys. Condens. Matter* **23**, 053201 (2011).

6. He, C. et al. Complex Low Energy Tetrahedral Polymorphs of Group IV Elements from First Principles. *Phys. Rev. Lett.* **121**, 175701 (2018).

7. Pannetier, J., Bassas-Alsina, J., Rodríguez-Carvajal, J. & Caignaert, V. Prediction of crystal structures from crystal chemistry rules by simulated annealing. *Nature* **346**, 343–345 (1990).

8. Deaven, D. M. & Ho, K. M. Molecular Geometry Optimization with a Genetic Algorithm. *Phys. Rev. Lett.* **75**, 288–291 (1995).

9. Glass, C. W., Oganov, A. R. & Hansen, N. USPEX—Evolutionary crystal structure prediction. *Comput. Phys. Commun.* **175**, 713–720 (2006).

10. Trimarchi, G. & Zunger, A. Global space-group optimization problem: Finding the stablest crystal structure without constraints. *Phys. Rev. B* **75**, 104113 (2007).

11. Zhao, X. et al. Exploring the Structural Complexity of Intermetallic Compounds by an Adaptive Genetic Algorithm. *Phys. Rev. Lett.* **112**, 045502 (2014).

12. Wang, Y., Lv, J., Zhu, L. & Ma, Y. Crystal structure prediction via particle-swarm optimization. *Phys. Rev. B* **82**, 094116 (2010).

13. Peng, F. et al. Hydrogen Clathrate Structures in Rare Earth Hydrides at High Pressures: Possible Route to Room-Temperature Superconductivity. *Phys. Rev. Lett.* **119**, 107001 (2017).

14. Zhang, Y.-Y., Gao, W., Chen, S., Xiang, H. & Gong, X.-G. Inverse design of materials by multi-objective differential evolution. *Comput. Mater. Sci.* **98**, 51–55 (2015).

15. Faber, F. A., Lindmaa, A., von Lilienfeld, O. A. & Armiento, R. Machine Learning Energies of 2 Million Elpasolite (\(AB\)\(C\)\(2\)\(D\)\(6\)) Crystals. *Phys. Rev. Lett.* **117**, 135502 (2016).

16. Li, Z., Xu, Q., Sun, Q., Hou, Z. & Yin, W.-J. Thermodynamic Stability Landscape of Halide Double Perovskites via High-Throughput Computing and Machine Learning. *Adv. Funct. Mater.* **29**, 1807280 (2019).

17. Bartel, C. J. et al. Physical descriptor for the Gibbs energy of inorganic crystalline solids and temperature-dependent materials chemistry. *Nat. Commun.* **9**, 4168 (2018).

18. Xie, T. & Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **120**, 145301 (2018).

19. Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **31**, 3564–3572 (2019).

20. Tsubaki, M. & Mizoguchi, T. Quantum Deep Field: Data-Driven Wave Function, Electron Density Generation, and Atomization Energy Prediction and Extrapolation with Machine Learning. *Phys. Rev. Lett.* **125**, 206401 (2020).

21. Carbone, M. R., Topsakal, M., Lu, D. & Yoo, S. Machine-Learning X-Ray Absorption Spectra to Quantitative Accuracy. *Phys. Rev. Lett.* **124**, 156401 (2020).

22. Liu, Y.-H. & van Nieuwenburg, E. P. L. Discriminative Cooperative Networks for Detecting Phase Transitions. *Phys. Rev. Lett.* **120**, 176401 (2018).

23. Jain, A. & Bligaard, T. Atomic-position independent descriptor for machine learning of material properties. *Phys. Rev. B* **98**, 214112 (2018).

24. Ward, L. et al. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Phys. Rev. B* **96**, 024104 (2017).
25. Ryan, K., Lengyel, J. & Shatruk, M. Crystal Structure Prediction via Deep Learning. *J. Am. Chem. Soc.* **140**, 10158–10168 (2018).

26. Oliynyk, A. O., Adutwum, L. A., Harynuk, J. J. & Mar, A. Classifying Crystal Structures of Binary Compounds AB through Cluster Resolution Feature Selection and Support Vector Machine Analysis. *Chem. Mater.* **28**, 6672–6681 (2016).

27. Bartel, C. J. *et al.* A critical examination of compound stability predictions from machine-learned formation energies. *Npj Comput. Mater.* **6**, 97 (2020).

28. Aurenhammer, F. Voronoi diagrams - a survey of a fundamental geometric data structure. *ACM Comput. Surv.* **23**, 345–405 (1991).

29. Rycroft, C. H. VORO++: A three-dimensional Voronoi cell library in C++. *Chaos Interdiscip. J. Nonlinear Sci.* **19**, 041111 (2009).

30. Battaglia, P. W. *et al.* Relational inductive biases, deep learning, and graph networks. *ArXiv180601261 Cs Stat* (2018).

31. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **65**, 1501–1509 (2013).

32. Seko, A., Hayashi, H., Nakayama, K., Takahashi, A. & Tanaka, I. Representation of compounds for machine-learning prediction of physical properties. *Phys. Rev. B* **95**, 144110 (2017).

33. Park, C. W. & Wolverton, C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Phys. Rev. Mater.* **4**, 063801 (2020).

34. Bergstra, J., Yamins, D. & Cox, D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. in *International Conference on Machine Learning* 115–123 (PMLR, 2013).

35. Bergstra, J. S., Bardenet, R., Bengio, Y. & Kégl, B. Algorithms for Hyper-Parameter Optimization. in *Advances in Neural Information Processing Systems 24* (eds. Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F. & Weinberger, K. Q.) 2546–2554 (Curran Associates, Inc., 2011).

36. Seko, A. *et al.* Prediction of Low-Thermal-Conductivity Compounds with First-Principles Anharmonic Lattice-Dynamics Calculations and Bayesian Optimization. *Phys. Rev. Lett.* **115**, 205901 (2015).

37. Yamashita, T. *et al.* Crystal structure prediction accelerated by Bayesian optimization. *Phys. Rev. Mater.* **2**, 013803 (2018).

38. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).

39. Bergerhoff, G., Hundt, R., Sievers, R. & Brown, I. D. The inorganic crystal structure data base. *J. Chem. Inf. Model.* **23**, 66–69 (1983).
**Fig. 1 | Flowchart of GN-BOSS approach.** For clarity, in the text, GN-BO and GN-BO-SYM indicate GN-BOSS approach without and with crystal symmetry constraints.
Fig. 2 | GN-predicted and DFT-calculated formation enthalpies for ∼320,000 crystals in OQMD. The mean absolute error as a function of percentage of training crystals.
Fig. 3 | (a) Process of GN-RAS, GN-BO, GN-PSO, and GN-BO-SYM approaches, to search the crystal structure of CaS. Absolute values of $\Delta H$ for four approaches have been shifted for clarity. (b–e) Typical structures in the process of GN-BO. (f–i) Lowest energy structures in the process of GN-RAS, GN-PSO, GN-BO-SYM, and DFT-PSO-SYM, respectively, within 6,000 steps.
Fig. 4 | The computational cost of GN-BO-SYM, GN-BO and DFT-PSO-SYM to find the right crystal structures. The number of atoms of the 24 binary compounds predicted by GN-BO-SYM, GN-BO and DFT-PSO-SYM are 8, 4, and 8 in a simulation cell; and the numbers of atoms of C, Si, CsPbI₃ are indicated in bracket.
Fig. 5 | Construction of the crystal graph.
Fig. 6 | The GN model architecture.
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SI2021.08.15.docx