SCIENCE CHINA
Chemistry

•COMMUNICATIONS•

August 2019 Vol.62 No.8: 982–986
https://doi.org/10.1007/s11426-019-9502-5

Identify crystal structures by a new paradigm based on graph theory for building materials big data

Mouyi Weng[†], Zhi Wang[†], Guoyu Qian[†], Yaokun Ye, Zhefeng Chen, Xin Chen, Shisheng Zheng & Feng Pan[*]

School of Advanced Materials, Peking University, Shenzhen Graduate School, Shenzhen 518055, China

Received March 19, 2019; accepted May 14, 2019; published online June 12, 2019

Material identification technique is crucial to the development of structure chemistry and materials genome project. Current methods are promising candidates to identify structures effectively, but have limited ability to deal with all structures accurately and automatically in the big materials database because different material resources and various measurement errors lead to variation of bond length and bond angle. To address this issue, we propose a new paradigm based on graph theory (GT scheme) to improve the efficiency and accuracy of material identification, which focuses on processing the "topological relationship" rather than the value of bond length and bond angle among different structures. By using this method, automatic deduplication for big materials database is achieved for the first time, which identifies 626,772 unique structures from 865,458 original structures. Moreover, the graph theory scheme has been modified to solve some advanced problems such as identifying highly distorted structures, distinguishing structures with strong similarity and classifying complex crystal structures in materials big data.

structures identification, graph theory, big data, topological relationship, materials database

Citation: Weng M, Wang Z, Qian G, Ye Y, Chen Z, Chen X, Zheng S, Pan F. Identify crystal structures by a new paradigm based on graph theory for building materials big data. *Sci China Chem*, 2019, 62: 982–986, https://doi.org/10.1007/s11426-019-9502-5

The materials genome project (MGP) and development of structure chemistry are to pursue methods to construct new materials by using big data [1–8]. Nowadays several hundreds of thousands of inorganic crystal structures have been collected in material databases [9–13], but some of materials are equivalent to each other. Data deduplication, a key technique based on material identification, can eliminate the impact of repeated data during high-throughput calculations.

Research in recent decades has yielded a variety of methods to identify structures [14–17]. One open source code, XtalComp [15], has an ability to demonstrate a standard unit cell with atom positions by transforming unit cell several times. Besides, pair distribution functions and spherical harmonics functions have been used to compare crystal structures. A method named bond characterization matrix (BCM) has been employed in Crystal Structure AnaLYsis by Particle Swarm Optimization (CALYPSO) code [14], and the difference of two BCMs indicates the degree of difference between two crystal structures.

However, current schemes tend to set multiple tolerances as thresholds of value of bond length and bond angle in order to deal with all structures, and they are not reliable to distinguish isomorphic crystal structures because different material resources and various measurement errors lead to deviations of bond lengths and bond angles. Thus, there is a great need for exploring reliable and an artificial intelligence method to identify structures accurately and automatically in materials big data.

Herein, we propose a new paradigm based on graph theory (GT scheme) to improve the efficiency and accuracy of material identification. By simplify crystal structures into a

†These authors contributed equally to this work.
*Corresponding author (email: panfeng@pkusz.edu.cn)

graph only with the information of "topological connection", rather than the value of bond length and bond angle, we can identify isomorphic structures in a big database directly without setting thresholds for the complex functions. We do only need to set the threshold based on Pauling's rules to determine two atoms are "connected" or not. That means we could compare thousands of types of crystal structures with a unique set of parameters, which is given to the realizations of automatic high-throughput materials deduplication.

During data deduplication in material databases, we set the following requirements for the algorithm: (1) the algorithm should distinguish different sets of unit cell. By using different unit cell selections, atom structures can be much different (Figure S1, Supporting Information online). The algorithm should have the ability to distinguish and compare them. (2) The algorithm can tolerate the negligible deviation between experimental data and computed results of atoms and lattice constant, and induce the identical structures into one group. For instance, Table S1 (Supporting Information online) shows the different experimental and calculated lattice constants of spinel Co_3O_4. Although the results have minor differences, the structures still come from the same source. (3) Using continuous functions to represent the differences between structures should be avoided. As mentioned above, instead of using automatic technique, previous studies need to analyze the results of continuous functions manually to judge whether the structures are equivalent or not.

In GT scheme, we first simplify structures into a graph, which only consists of vertices and edges, in which atoms are simplified as vertices and adjacent atoms with the actual distance less than maximum bond length that are "connected" with edges. Actually, the bond length depends on element types, and it is the only place where we set thresholds in this method.

We next define the "distance" between two atoms. Here, the distance is not the actual distance in real space, but the minimum number of steps which can connect two different vertices in the graph. To take spinel Co_3O_4 as an example, multiple atoms with different "distances" (distance 1, 2, 3, 4) from central oxygen atom are shown in Figure 1(a), respectively.

Based on graph theory, if the topological relationship of the simplified graphs which come from different chemical compounds are completely the same, we can suppose that these crystal structures are equivalent. On the other hand, we have not found any two different structures that can have the same graph so far, as long as we take enough "distance" (shown in Figure 1) into consideration, which presents the soundness and completeness of GT scheme.

A graphical representation of spinel Co_3O_4 structures has been shown in Figure 1(a), which originates from the conventional Co_3O_4 unit cell (Figure 1(b)). All the topological relationship of the unit cell has been involved in this simplified graph in distance 4, and the periodicity of unit cell is considered. To simplify the expression, we are seeking a way to describe spinel Co_3O_4 within a "shorter distance."

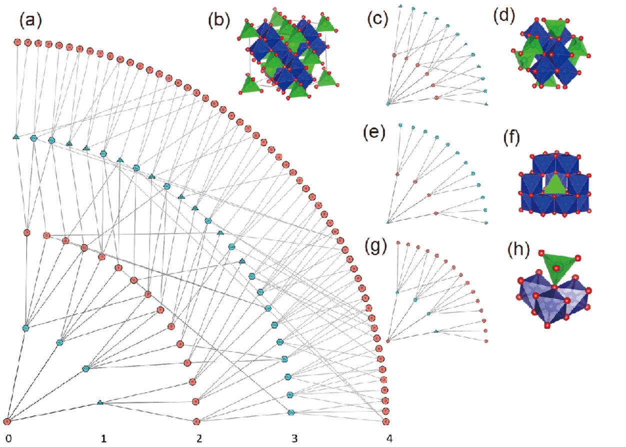


Figure 1 (a) The simplified graph of spinel Co_3O_4 in "distance" 4. The red circles represent oxygen atoms; the blue triangles represent the central cobalt atoms located in tetrahedrons and the blue hexagons represent the central cobalt atoms located in octahedrons; the numbers in bottom indicate the "distance" from the central oxygen atom. (b) Schematic representation of crystal structure of spinel Co_3O_4. The red dots represent oxygen atoms; the cobalt atoms are located in the blue octahedron and green tetrahedrons, respectively. (c) The simplified graph of spinel Co_3O_4 in distance 2 with an octahedral cobalt in center, and its crystal structure (d). (e) The simplified graph of spinel Co_3O_4 in distance 2 with an octahedral cobalt in center, and its crystal structure (f). The graph representation with three layers of spinel Co_3O_4 with a tetrahedral cobalt in center. (g) The simplified graph of spinel Co_3O_4 in distance 2 with an oxygen in center, and its crystal structure (h) (color online).

Since there are in total 56 atoms in spinel Co_3O_4 unit cell [18], we convert 56 separated graphs in distance 2 by making each atom as the central vertex. Most of them are the same, and we obtain three different types of graphs after deduplication. Two of them are centered on cobalt atoms, and one of them is centered on oxygen atom. We call them spinel-Co-1 (Figure1(c, d)), spinel-Co-2 (Figure 1(e, f)) and spinel-O (Figure 1 (g, h)).

Thus, for spinel Co_3O_4 structure, we can describe it by using the combination of Figure 1(c, e, g) in distance 2. It should be noted that the portion of these three types in distance 2 are 2/7, 1/7 and 4/7, respectively.

For most crystal structures, comparing the graphs consisting of each central atom with its connected atoms (within distance 1) seems sufficient to make a judgment on isomorphism in materials, but sometimes there are some exceptions.

For instance, layered $NaNiO_2$ is widely used in sodium ion batteries (SIBs) [19]. In O3 phase and P3 phase of different layered $NaNiO_2$, the central sodium ions are six-coordinated with octahedral configuration and triangular prismatic configuration, respectively (Figure 2(a, b)). That means if we only compare the O3 phase with P3 phase of layered $NaNiO_2$

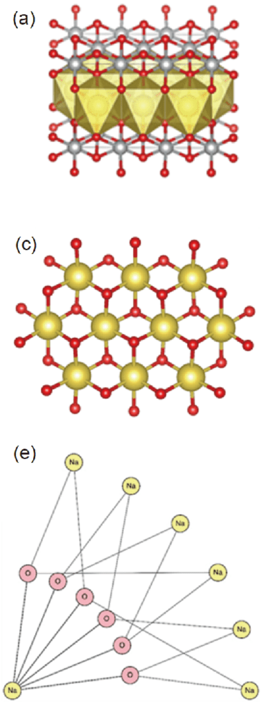
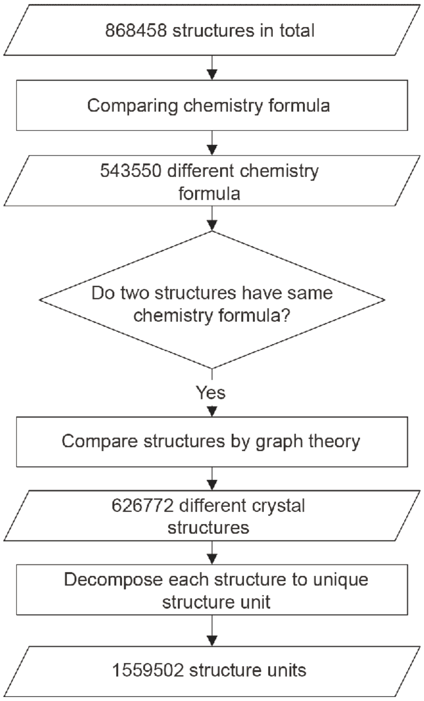


Figure 2 (a) The crystal structures of (a) O3 phase and (b) P3 phase of layered $NaNiO_2$ with central sodium ions located in octahedrons and triangular prisms. (c) The simplified graph of O3 phase of layered $NaNiO_2$ in distance 2 with an octahedral sodium ion in center, and its crystal structure (e). (d) The simplified graph of P3 phase of layered $NaNiO_2$ in distance 2 with an octahedral sodium ion in center, and its crystal structure (f) (color online).


868458 structures in total
Comparing chemistry formula
543550 different chemistry formula
Do two structures have same chemistry formula?
Yes
Compare structures by graph theory
626772 different crystal structures
Decompose each structure to unique structure unit
1559502 structure units

Scheme 1 The workflow and the result of comparing structures in database.

within distance 1, the simplify graphs are isomorphic. To ensure the accuracy of calculation results, we need to consider enough "distance" (usually 3 or 4), as shown in Figure 2(c–f).

Moreover, compared with the traditional structure chemistry methods, the GT scheme can classify distorted structures which remain the original topological connection into the same type of structure much more easily, which enhances the efficiency and reliability of material identification. Structural distortion usually leads to the increased tolerance, and it is difficult for the traditional structure chemistry methods to identify structures. For GT expression, although the number of different atoms in a unit cell is different, the graph around the central atoms is not changed. Also, the portion of each type of graphs remains. As long as there is no generated or missing chemical bonds during the structural distortion process, the distorted structures can be easily classified.

Benefit from the distinctive advantage of GT scheme, automatic deduplication for big materials database can be realized for the first time. We collect and integrate all inorganic crystal structure data to have 865,458 original structures in total [9–11,13].

We deduplicate all structures within 3 steps as shown in Scheme 1. Since different chemistry formula cannot be the same structure, we do only judge the structures which have the same chemistry formula are equivalent or not by GT scheme. The calculation process is paralleled into 80 threads on eight Intel® Xeon® CPU Processor E5-2640 v4 CPUs, and it takes us 96 h to obtain the final result (626,772 unique structures). To speed up the topological information processing, we also develop a Pruning Algorithm, which is described in the Supporting Information online.

Once we get the topological information inside 626,772 crystal structures, we are wondering that what kinds of structural units are the most "popular" in inorganic crystal structures. The structural unit is a secondary structure regardless of its geometrical structure. For example, 1 sodium atom connected with 6 atoms of oxygen located in either an octahedron or a prism is isomorphic. We have excavated 1,559,502 structural units totally from the 626,772 structures when we set the distance as 1 (Scheme 1). Another code (Pseudo code) has been written to describe the geometric structure of each structural unit and calculate the occurrence frequency of each structural unit. For instance, in the case of Mn-O coordination, six-coordinate, penta-coordinate, and tetra-coordinate structural units are the most common, with 39,890, 3,799, and 2,730 times appearing in our inorganic material database, respectively. Some statistical results have been selected in the Supporting Information online, and further analysis would be done in the near future.

The GT scheme can be applied in "higher dimensions". If we can regard a structural unit as a "pack", the calculation efficiency would be significantly enhanced in some cases.

Taking the SiC crystal as an example, SiC has hundreds of phases, and each phase has different stacking mode. Some of them are too similar to be distinguished by the original GT scheme, as the graphs of 2H phase and 4H phase are still isomorphic until distance 3 (Figure 3(a–c)). Particularly, in most SiC crystals, each silicon atom with four adjacent carbon atoms forms a tetrahedral structural unit (Figure 3 (d)). If we set this Si–C tetrahedron manually as a vertex in graph, and set the shared carbon atom between two tetrahedrons as the edge, we can find the difference between the two new graphs obviously (Figure 3(e, f)). We can find two quadrangles share a side in Figure 3(f), but not in Figure 3(e). Thus, the two structures are different. Benefited from the usage of pack, the identification process can be accelerated significantly, and more efforts can be done from manually to automatically in order to improve the overall operating speed of GT scheme.

In particular, the GT scheme provides access to identify and classify the complex crystal structures from materials big data. For example, the spinel structures represent a class of minerals with general formulation AB_2X_4, where the X anions are arranged in a cubic close-packed lattice and the cations A and B occupy some or all of the octahedral and tetrahedral sites in the lattice. The $Li_{11}Ti_4Fe_9O_{32}$ belongs to spinel structure [9], but it is difficult by traditional structure chemistry methods to detect it as the severe distortion and displacement in its unit cell. The lattice constants of $Li_{11}Ti_4Fe_9O_{32}$ are a=8.34 Å, b=8.36 Å, c=8.50 Å, $\alpha=\beta=\gamma$=89.7°, while lattice constants of the typical spinel Co_3O_4 structure are $a=b=c$=8.15 Å, $\alpha=\beta=\gamma$=90°. That means the unit cell of $Li_{11}Ti_4Fe_9O_{32}$ suffers from severe distortion after atom substitutions, which may bring a problem to judge the actual crystal structure by traditional methods as the deviation between experimental data and theoretical results cannot be ignored. When using GT scheme, we only need to export the topological information of $Li_{11}Ti_4Fe_9O_{32}$. Without considering the types of element, if there are the isomorphic graphs mentioned as Figure 1(c, e, g), and the portion are 2/7, 1/7 and 4/7, respectively, we can classify the $Li_{11}Ti_4Fe_9O_{32}$ to spinel structure (Figure 4).

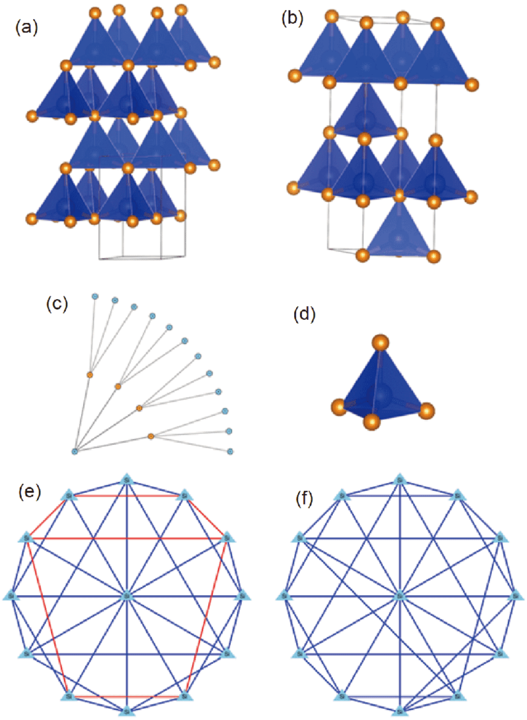


Figure 3 The crystal structures of (a) 4H SiC and (b) 2H SiC. (c) The same simplified graph of 4H SiC and 2H SiC in distance 2. (d) A structural unit of tetrahedral SiC. The modified graph of (e) 4H SiC and (f) 2H SiC exported on the basis of tetrahedral SiC structural units. All silicon atoms are shown in blue dots, while all carbon atoms are shown in brown dots (color online).

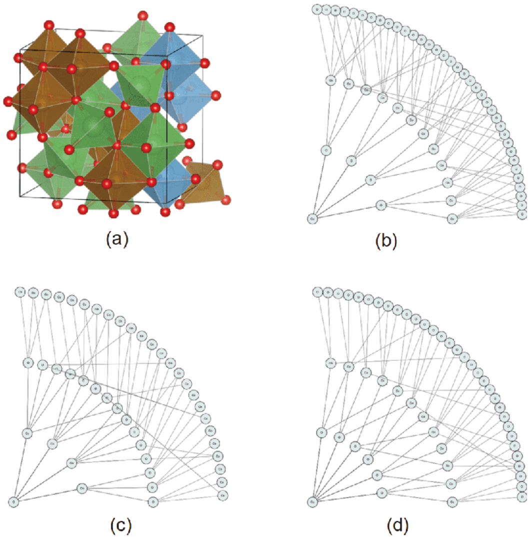


Figure 4 (a) Schematic representation of crystal structure of $Li_{11}Ti_4Fe_9O_{32}$; (b–d) the simplified subgraphs of spinel structures, which are isomorphic to Figure 1(c, e, g), respectively (color online).

The novel GT scheme enhances the efficiency and accuracy of material identification. By simplifying crystal structures into a graph, we only need to judge whether the topological information is isomorphic or not, instead of setting elusive thresholds for the complex functions. Thus, 626,772 unique structures have been filtered successfully from materials big data by high-throughput screening. Furthermore, the modified GT scheme has demonstrated the potential ability to address complex issues in material identification field. For perspective, inspired by the simplified principle from graph theory, this scheme may benefit the exploration of new material design and structural evolution.

Acknowledgements The authors thank Dr. Lin-Wang Wang from Lawrence Berkeley National Laboratory and Dr. Wenfei Fan from the University of Edinburgh for their helpful discussions. This work was supported by the National Key R&D Program of China (2016YFB0700600), the National Natural Science Foundation of China (21603007, 51672012), Soft Science Research Project of Guangdong Province (2017B030301013), and New

Energy Materials Genome Preparation & Test Key-Laboratory Project of Shenzhen (ZDSYS201707281026184).

Conflict of interest The authors declare that they have no conflict of interest.

Supporting information The supporting information is available online at http://chem.scichina.com and http://link.springer.com/journal/11426. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

1 Hautier G, Fischer CC, Jain A, Mueller T, Ceder G. *Chem Mater*, 2010, 22: 3762–3767

2 Hautier G, Fischer C, Ehrlacher V, Jain A, Ceder G. *Inorg Chem*, 2011, 50: 656–663

3 Hautier G, Jain A, Ong SP, Kang B, Moore C, Doe R, Ceder G. *Chem Mater*, 2011, 23: 3495–3508

4 Jain A, Hautier G, Moore CJ, Ping Ong S, Fischer CC, Mueller T, Persson KA, Ceder G. *Comput Mater Sci*, 2011, 50: 2295–2310

5 Mueller T, Hautier G, Jain A, Ceder G. *Chem Mater*, 2011, 23: 3854–3862

6 Wu Y, Lazic P, Hautier G, Persson K, Ceder G. *Energy Environ Sci*, 2013, 6: 157–168

7 Yang L, Ceder G. *Phys Rev B*, 2013, 88: 224107

8 Raccuglia P, Elbert KC, Adler PDF, Falk C, Wenny MB, Mollo A, Zeller M, Friedler SA, Schrier J, Norquist AJ. *Nature*, 2016, 533: 73–76

9 Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G, Persson KA. *APL Mater*, 2013, 1: 011002

10 Crystallography Open Database. http://www.crystallography.net/cod/index.php

11 Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C. *JOM*, 2013, 65: 1501–1509

12 Downs RT, Hall-Wallace M. *Am Mineral,* 2003, 88: 247–250

13 Kirklin S, Saal JE, Meredig B, Thompson A, Doak JW, Aykol M, Rühl S, Wolverton C. *npj Comput Mater*, 2015, 1: 15010

14 Wang Y, Lv J, Zhu L, Ma Y. *Phys Rev B*, 2010, 82: 094116

15 Lonie DC, Zurek E. *Comput Phys Commun*, 2012, 183: 690–697

16 Sadeghi A, Ghasemi SA, Schaefer B, Mohr S, Lill MA, Goedecker S. *J Chem Phys*, 2013, 139: 184118

17 Zhu L, Amsler M, Fuhrer T, Schaefer B, Faraji S, Rostami S, Ghasemi SA, Sadeghi A, Grauzinyte M, Wolverton C, Goedecker S. *J Chem Phys*, 2016, 144: 034203

18 Liu X, Prewitt C. *Phys Chem Miner*, 1990, 17: 168–172

19 Han MH, Gonzalo E, Singh G, Rojo T. *Energy Environ Sci*, 2015, 8: 81–102