Topological Methods and Tools for the Analysis of Big Crystallographic Data

Vladislav A. Blatov[1][0000-0002-4048-7218]

[1] Samara State Technical University, Samara, 443100, Russian Federation
blatov@topospro.com

Abstract. We briefly overview mathematical models, methods and computer tools for the topological description, analysis and classification of crystal structures. We present *ToposPro*, a program package for comprehensive geometrical and topological analysis of periodic architectures of any composition and complexity. *ToposPro* is designed for the automated analysis of big crystallographic data, which are either collected in the world-wide electronic databases or generated by theoretical methods. *ToposPro* was used to create a system of topological databases, which are available both in the local version and as a number of interactive web-services integrated into the *TopCryst* system. All described topological tools are considered as applied to solving typical tasks of crystal chemistry and materials science.

Keywords: Topological Analysis, Crystal Structure, Big Data, Software, Web-Service.

1 Introduction

Experimental data on crystal structures form one of the biggest and well-organized sets of information in natural sciences. At present, about two million records are collected in the world-wide crystallographic databases such as Cambridge Structural database (CSD), Inorganic Crystal Structure Database (ICSD), Pearson's Crystal Database (PCD), Crystallography Open Database (COD). This dataset definitely contains a lot of correlations 'chemical composition – chemical structure', but has not yet been explored in detail. For a long time, the main reason for that was the absence of automated tools for processing such huge and extremely diverse information with the same universal algorithms. Another important problem was that the initial crystallographic data contained no knowledge about chemical bonding, which is crucial for the usage of these data to solve the tasks of crystal chemistry and materials science.

In this paper, we present a number of computer tools for the description, analysis and classification of crystal structures of any chemical composition and complexity. An important feature of all mathematical models and algorithms behind these tools is that they are based on the topological representation of a crystal structure, where the overall connectivity between the atoms is restored hence providing naturally chemical treatment of the experimental crystallographic data. Such approach enables one to solve the problem of processing the crystallographic datasets mentioned above.

2 Models and algorithms

2.1 Voronoi partition

The initial experimental crystallographic data provide only geometrical information about positions of atoms in the crystal space and symmetry operations between them. To transform these data into a chemical object we need to establish interatomic contacts. For this purpose we represent the atoms by their *Voronoi polyhedra*, which confine parts of the space belonging to particular atoms (Fig. 1). The Voronoi polyhedra form a partition of the crystal space (*Voronoi partition*) thus assigning volume and shape to the atoms. The faces of the Voronoi polyhedra correspond to possible interatomic contacts; to treat them as chemical bonds or weaker interactions additional chemical and geometrical criteria can be applied. As a result, a rigorous algorithm *Domains* was developed [1], which enables one to determine chemical bonding in crystal structures of any nature, from metals to proteins.

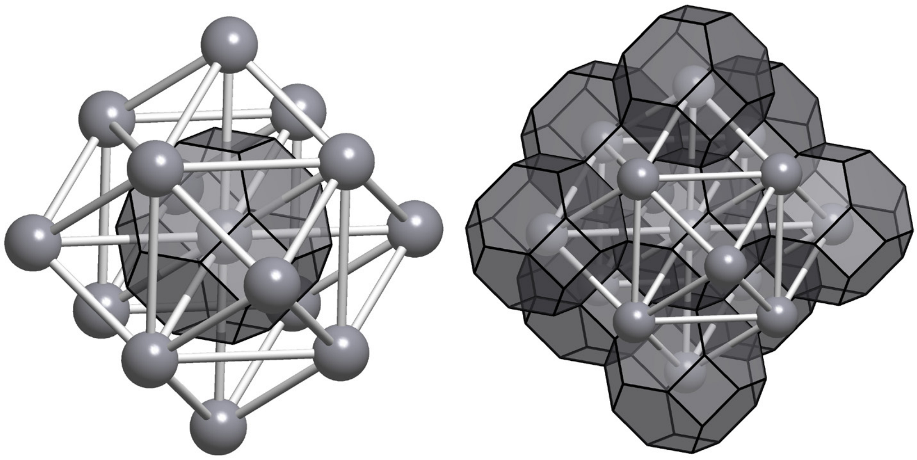

Fig. 1. Body-centered cubic packing: (left) a fragment of atomic net with the Voronoi polyhedron of an atom; each of 14 faces of the polyhedron corresponds to an interatomic contact; (right) a part of the Voronoi partition; the net of the vertices and edges of the Voronoi polyhedra represents the Voronoi net.

However, Voronoi partition describes not only atoms and their interactions, but also the free space in the crystal. Indeed, the vertices of the Voronoi polyhedra are geometrically the most distant points from the surrounding atoms and hence can be considered as centers of voids. Similarly, the edges of the Voronoi polyhedra mimic channels between the voids. This approach is especially important when studying absorption, catalytic, or diffusion properties of the crystal [2].

2.2 Atomic net

After determining interatomic interactions the crystal structure is represented as an infinite periodic graph (*atomic net*), which nodes and edges correspond to atoms and interactions between the atoms (Fig. 1). This is the *complete* topological representation of the crystal structure as it provides the comprehensive information about the structure connectivity. The topology of the net is completely described by its *labeled quotient graph* [3], which can be obtained by wrapping the net to the unit cell using the periodic conditions and keeping the information about bonding of the atoms inside and outside the unit cell. This approach enables one to transform the infinite net to a finite object and hence store the topological information in a database. The topology of the labeled quotient graph and the corresponding net can be characterized by a set of topological indices (invariants), which can be used for matching periodic nets and uniting the nets with the same topology into the same *topological type*. As a result, a rigorous topological classification of crystals structures becomes possible [4].

2.3 Underlying net

The complete atomic net can then be represented at several levels of detailing by *simplification* algorithms [5]. Such algorithms enable one to separate atomic groups (building units), which assemble the whole crystal structure. Both chemical and topological criteria are used for this purpose and as a result the atomic net is transformed into a net of the centroids of the building units connected according to the connections of the atoms of these units in the atomic net (Fig. 2). This simplified net is called *underlying* [4] as it encodes the method of constructing the crystal from atomic units and hence underlies the chemical model of the crystal. Since building units can be chosen in different ways depending on the treatment of the chemical nature of the crystalline substance, several underlying nets can match the same crystal structure. This is a powerful tool for determining structural correlations between chemical substances of different composition, crystal structure parameters and bonding [6].

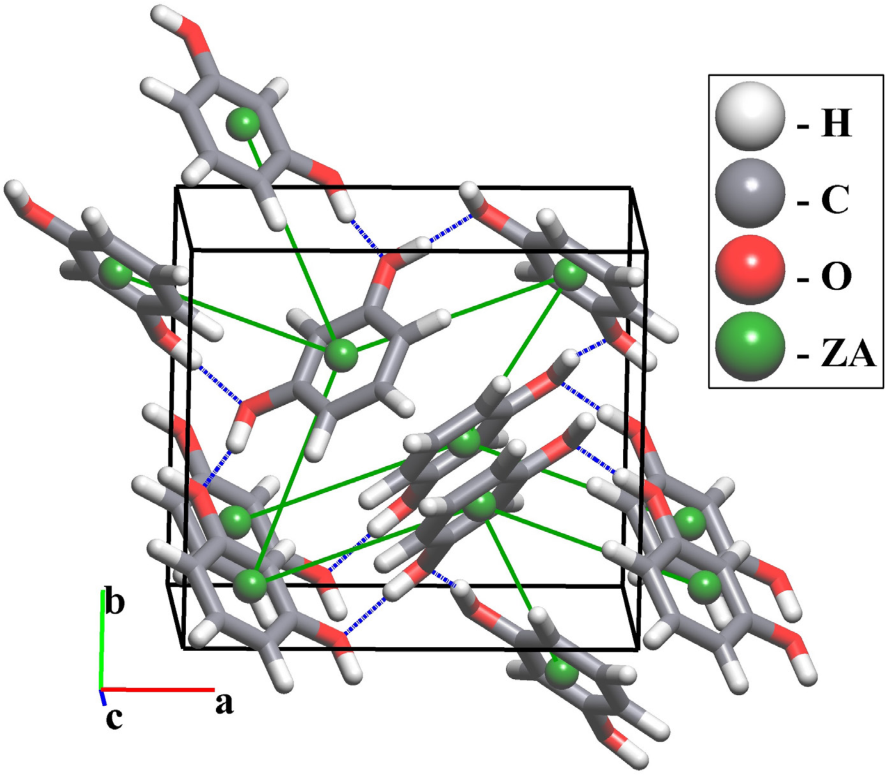

Fig. 2. Crystal structure of resorcinol. The underlying net is shown by green balls ZA (centroids of the resorcinol molecules) and green edges, which mimic the H-bonding between the resorcinol molecules (H-bonds are shown by dotted blue lines).

2.4 Voronoi net

The net of vertices and edges of all atomic Voronoi polyhedra is called *Voronoi net* (Fig. 1). This kind of periodic net describes the topology of the free space between the atoms and can be analyzed by the same methods and tools as the atomic net. This allowed us to talk about *dual crystal chemistry* [2]., which inverts the consideration of a crystal from material objects (atoms and bonds as clouds of electron density) to virtual objects (voids and channels), which however admit a similar vision.

2.5 Natural tiling

Another method of the analysis of the crystal free space is based on the tiling model. A *tile* is a generalized polyhedron, which vertices and edges are vertices and edges of the atomic or underlying net. Unlike a convex Voronoi polyhedron, the tile faces are not necessarily flat and vertices can be bivalent. However, similar to Voronoi polyhedra tiles form a face-to-face partition (*tiling*) of the crystal space, but the centers of tiles are centers of cages, not atoms (Fig.3). In general, the number of possible tilings, which can be constructed for a given net, is infinite but we have proposed e set of rigorous conditions [7] that enabled us to algorithmize the construction of a unique *natural tiling* with the smallest possible tiles, the so-called *natural tiles*. The natural tiles have clear physical meaning of the smallest cages in the crystal structure while their faces mimic the sections of the channels between the cages. The net of the tile centers and edges between the centers of adjacent tiles is called *dual net* (Fig. 3); similar to the Voronoi net it indicates the topology of the system of cages and channels between them. Thus the Voronoi net and natural tiling models supplement each other altogether characterizing the free (porous) space of the crystal.

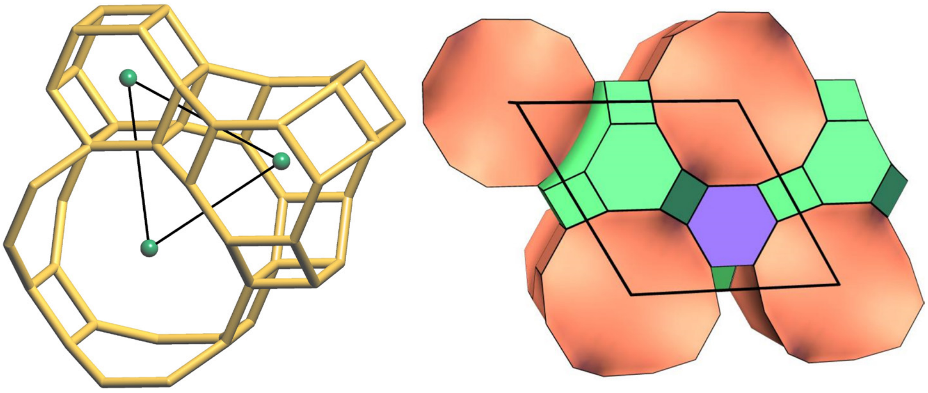

Fig. 3. Silicon framework in the crystal structure of the zeolite gmelinite (GME): (left) three different natural tiles and the corresponding fragment of the dual net (green balls); (right) a fragment of the natural tiling.

3 Software

We have implemented all mathematical models and algorithms mentioned above into the free program package *ToposPro* [4, 8], which currently (May 2023) has more than 6,800 registered users from 102 countries. *ToposPro* is designed for the Windows© operation systems and allows the users to create their own crystallographic databases in a unique binary format. An important advantage of the *ToposPro* format compared with other crystallographic databases is that it enables one to store the complete information on periodic nets, *i.e.* about topological properties of crystal structures. *ToposPro* includes a database management system, which is supplied with a user-friendly interface and integrates a number of applied programs for the comprehensive geometrical and topological analysis of crystal structures. Each applied program is designed for processing the crystallographic data in a batch mode that enables the user to work with large samples of information on hundreds of thousands of crystalline substances. At present, the researchers over the world use *ToposPro* for solving many tasks of crystal chemistry and materials science and for various classes of crystalline substances [9].

4 Databases

We have used *ToposPro* to extract topological information from the world-wide crystallographic databases and to develop the first *topological* databases. These databases form a set of the *ToposPro* topological collections [4], which contain millions of records with the geometrical and topological parameters characterizing crystal structures at different levels of their organization. The following structural units are characterized: atoms, molecules, ligands, nanoclusters, natural tiles, and periodic nets. For all these objects their occurrences in crystal structures are given, and the classification into topological types is provided. The collections are available at the *ToposPro* website and are distributed in a demo version together with *ToposPro*.

5 Web-services

We have recently started the implementation of the *ToposPro* tools and databases into free online web-services. All actual *ToposPro* collections are available *via* the *TopCryst* service [10, 11]. The users can upload their crystallographic data as a CIF file and get a detailed topological analysis of the crystal structure in a fully automated manner. *TopCryst* provides the information on all possible structure representations and for each representation it gives a list of building units together with the corresponding underlying net, which is related to one of more than 800,000 topological types from the *ToposPro* collection. The occurrences of all revealed topological objects are output and the links to the website of the CSD and ICSD are provided to let the user see the initial crystallographic information.

6 Conclusion

The models, algorithms, methods and computer tools described above form the first automated system of the topological analysis of crystal structures. This approach is naturally chemical as it introduces the concept of chemical interaction into purely geometrical crystallographic model. Long experience of the applications of this approach to various classes of crystal structures and to different crystallochemical tasks has proved its effectiveness. The system is in progress: the general trend for its further development is the transformation of the topological databases into the knowledge bases by the application of machine-learning methods [6] and the creation of artificial-intelligence systems for materials science.

The work was supported by the Russian Science Foundation (grant No. 22-13-00062).

References

1. Blatov, V. A.: A Method for Topological Analysis of Rod Packings. Struct. Chem. 27 (6), 1605–1611 (2016).

2. Blatov, V. A., Shevchenko, A.P.: Analysis of voids in crystal structures: the methods of 'dual' crystal chemistry. Acta Cryst.A59(1), 34-44 (2003).
3. Chung, S.J., Hahn, Th., Klee, W.E.: Nomenclature and generation of three-periodic nets: the vector method. Acta Cryst. A40(1), 42-50 (1984).
4. Blatov, V. A., Alexandrov, E. V., Shevchenko, A. P.: Topology: ToposPro. In: Comprehensive Coordination Chemistry III; Elsevier, Oxford (2021).
5. Shevchenko, A.P., Blatov, V.A.: Simplify to understand: how to elucidate crystal structures? Struct. Chem. 32(2), 507-519 (2021).
6. Shevchenko, A.P., Smolkov, M.I., Wang, J., Blatov, V.A.: Mining knowledge from crystal structures: oxidation states of oxygen-coordinated metal atoms in ionic and coordination compounds. J. Chem. Inf. Model. 62(10), 2332–2340 (2022).
7. Blatov, V. A., Delgado-Friedrichs, O., O'Keeffe, M., Proserpio, D. M.: Three-periodic nets and tilings: natural tilings for nets. Acta Cryst. A63(5), 418–425 (2007).
8. ToposPro Homepage, https://topospro.com, last accessed 2023/05/17.
9. Blatov, V. A., Shevchenko, A. P., Proserpio, D. M.: Applied topological analysis of crystal structures with the program package *ToposPro*. Cryst. Growth Des. 14(7), 3576–3586 (2014).
10. Shevchenko, A. P., Shabalin, A. A., Karpukhin, I. Y., Blatov, V. A.: Topological representations of crystal structures: generation, analysis and implementation in the *TopCryst* system. Sci. Technol. Adv. Mater. Methods 2(1), 250–265 (2022).
11. TopCryst Homepage, https://topcryst.com, last accessed 2023/05/17.