**Classification and use of topological data**

By V. A. Blatov, D. M. Proserpio and R. M. Hanson

**1. Introduction**

Raw crystallographic data, which are obtained from experiment or theoretical modeling, contain no information on the crystal structure connectivity. However, this information, which should describe all kinds of interatomic bonding, is crucial for many tasks of crystal chemistry and materials science, such as classification, comparative analysis, modeling of crystal growth and assembling, as well as search for correlations between crystal composition, structure and properties. Similar to molecular graphs, which are widespread in organic chemistry, crystal structure can be represented as a three-periodic graph, which is usually called *net* (Delgado-Friedrichs & O’Keeffe, 2005; Delgado-Friedrichs *et al*., 2005). Unlike molecular graphs, nets are infinite and require special methods of description. This description becomes more and more common in routine crystal structure determinations and usually follows the standard crystallographic description. Many topological descriptors were proposed for nets, and all of them can be determined from the initial crystallographic data with rigorous algorithms. The software, where these algorithms are implemented (Delgado-Friedrichs & O’Keeffe, 2003; Blatov *et al*., 2014), makes topological analysis available for a broad crystallographic community and has already led to developing special databases (O'Keeffe *et al*., 2008; Blatov *et al*., 2014) that collect the topological parameters of crystal structures. This requires an extension of the CIF format to make it capable to store the full topological information, which describes the crystal architecture at different levels of its organization.

Although the core CIF dictionary contains some topological information in the CHEMICAL\_CONN category, this information is limited to representation of separate organic molecules and does not cover the whole topological pattern of a crystal. Some existing CIF data items in the GEOM\_BOND and GEOM\_CONTACT categories can describe the crystal structure connectivity, but they have been never used for this purpose to be intended for geometrical description. Infinite systems of valence bonds in inorganic, metal-organic or covalent organic frameworks cannot be described by the core CIF data items. Furthermore there are no rules to describe a plethora of non-valence intra- or intermolecular interactions (H bonds, halogen and other specific bonds, van der Waals interactions). Moreover, heterodesmic structures that contain polyatomic building units, such as molecules, metal complex groups, clusters, *etc*., can be simplified to underlying nets, which bear the general topological motif of these units, and these nets should also be formalized in an appropriate format. The current version of the topoCIF dictionary meets all these requirements and is intended to make the exchange of topological information on crystal structures easy for crystallographer-theorists, structural chemists, software and database developers.

**2. Dictionary design considerations**

The topoCIF dictionary contains categories for the description of connectivity and topological properties of periodic nets. Since the corresponding terminology is still under development, only conventional terms and descriptors defined by Delgado-Friedrichs & O’Keeffe (2005) and Blatov *et al*. (2010) are included into the categories; the TOPOL\_REPRES\_ENTANGL category has been left empty until the terminology on entanglements gains consistency. The categories can be divided into two groups: (i) descriptors of crystal structure connectivity and (ii) descriptors of topological properties of a periodic net. The descriptors of the first group can be computed directly from the initial crystallographic data; the descriptors of the second group are determined resting upon the connectivity information. Thus the topoCIF dictionary enables one to generate two new levels of the crystal structure description: just as a connected net without any further detailing, and as a topologically characterized object.

**3. Classification of data definitions**

**3.1. Underlying net description**

*Underlying net* is the net of centroids of structural units; the edges of the net represent the links between the units. Underlying net describes the topology of a particular representation of a crystal structure; the representation is *complete* if the underlying net coincides with the initial structure, and partial in the opposite case. For example, diamond and SiO2 (cristobalite) can be represented by the same underlying net **dia**, however for diamond this net is complete, but for cristobalite it is partial since the oxygen atoms are considered as bridges (links) between silicon atoms. All topological data in the TOPOL category describe one of possible (complete or partial) topological representations of the crystal structure given in the form of a periodic net. The nodes of this net either coincide with the atoms or correspond to the centers of building units, if the initial crystal structure was properly simplified (Alexandrov *et al*. 2011; Bonneau *et al.* 2018; Shevchenko & Blatov, 2021). This means that in general the nodes of the underlying net and the crystal topology as a whole should be considered separately of the atoms. Thus, a special category, TOPOL\_REPRES, is introduced, which describes a particular topological representation. The category TOPOL\_REPRES\_NODE specifies the nodes of the underlying net and contains the following items, which describe the node site, composition and topology:

**node site**

* \_topol\_repres\_node.atom\_label

→ \_atom\_site\_label

\_topol\_repres\_node.label

\_topol\_repres\_node.fract\_x

\_topol\_repres\_node.fract\_y

\_topol\_repres\_node.fract\_z

\_topol\_repres\_node.symmetry\_multiplicity

\_topol\_repres\_node.Wyckoff\_symbol

**node composition**

\_topol\_repres\_node.chemical\_formula\_iupac

\_topol\_repres\_node.chemical\_formula\_moiety

\_topol\_repres\_node.chemical\_formula\_sum

\_topol\_repres\_node.structural\_formula\_InChI

\_topol\_repres\_node.structural\_formula\_SMILES

**node topology**

\_topol\_repres\_node.coordination\_sequence

\_topol\_repres\_node.point\_symbol

\_topol\_repres\_node.extended\_point\_symbol

\_topol\_repres\_node.vertex\_symbol

*The bullet (•) indicates a category key. The arrow (→) is a reference to a parent data item.*

If the \_topol\_repres\_node.atom\_label item was applied, the underlying net node labelled in the \_topol\_repres\_node.label item coincides with the atom specified by the \_atom\_site\_label item. In particular, if the representation is complete, all underlying net nodes coincide with atoms; in this case, the labels \_topol\_repres\_node.label, \_topol\_repres\_node.atom\_label and \_atom\_site\_label can be the same and the items \_topol\_repres\_node.fract\_x, \_topol\_repres\_node.fract\_y, \_topol\_repres\_node.fract\_z, \_topol\_repres\_node.symmetry\_multiplicity, \_topol\_repres\_node.Wyckoff\_symbol should not be applied as the corresponding information can be extracted from the corresponding items of the ATOM\_SITE category.

The connectivity of the nodes is described in the TOPOL\_LINK category and some of the items are similar to those from GEOM\_BOND or GEOM\_CONTACT categories:

* \_topol\_link.node\_label\_1

→ \_topol\_repres\_node.label

* \_topol\_link.node\_label\_2

→ \_topol\_repres\_node.label

* \_topol\_link.site\_symmetry\_symop\_1

→ \_space\_group\_symop.id

* \_topol\_link.site\_symmetry\_symop\_2

→ \_space\_group\_symop.id

\_topol\_link.site\_symmetry\_translation\_1

\_topol\_link.site\_symmetry\_translation\_2

\_topol\_link.distance

*The bullet (•) indicates a category key. The arrow (→) is a reference to a parent data item.*

However, the description of symmetry operations is split into two parts: translational (\_topol\_link.site\_symmetry\_translation\_\*) and non-translational (\_topol\_link.site\_symmetry\_symop\_\*). This enables one to specify any number of translations that can be important for some interpenetrating arrays, where the connected nodes can belong to distant unit cells. There are additional items for description of other link properties:

\_topol\_link.id

\_topol\_link.type

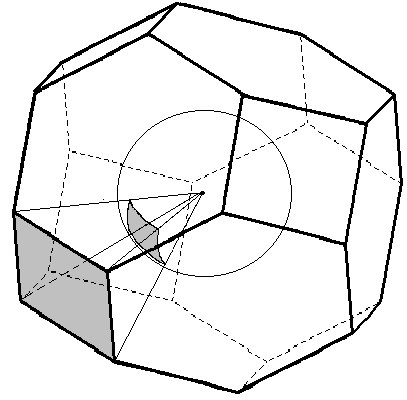
\_topol\_link.order

\_topol\_link.multiplicity

\_topol\_link.special\_details

\_topol\_link.Voronoi\_solidangle

The topol\_link.Voronoi\_solidangle item designates the solid angle of an interatomic contact A-X, which corresponds to a common face of the Voronoi polyhedra of A and X atoms (Blatov, 2004; Fig. 1).



**Figure 1:** The solid angle of the shaded Voronoi polyhedron face in the body-centered cubic lattice is equal to the shaded segment of the unit sphere being cut off by the pyramid with the Voronoi polyhedron atom A at the vertex and the face in the base. The shaded face is shared between Voronoi polyhedra of A and X atoms and corresponds to the bond A-X.

Since edges of the underlying net can correspond not only to bonds between atoms, but also to complex polyatomic links (bridging atoms, ligands, clusters) the category TOPOL\_REPRES\_EDGE is introduced, whose items describe the composition of an edge specified by the \_topol\_repres\_edge.id item:

* \_topol\_repres\_edge.id

→ \_topol\_link.id

\_topol\_repres\_edge.chemical\_formula\_iupac

\_topol\_repres\_edge.chemical\_formula\_moiety

\_topol\_repres\_edge.chemical\_formula\_sum

*The bullet (•) indicates a category key. The arrow (→) is a reference to a parent data item.*

**3.2. Underlying net topological properties**

The topological properties of the underlying net, which is described in the \_TOPOL\_REPRES\_NODE, TOPOL\_LINK and \_TOPOL\_REPRES\_EDGE categories, are collected in TOPOL\_REPRES, TOPOL\_REPRES\_ENTANGL, TOPOL\_REPRES\_TILING, TOPOL\_REPRES\_TILING\_FACES, TOPOL\_REPRES\_TILING\_TILES and TOPOL\_REPRES\_OCCURRENCE categories.

TOPOL\_REPRES

\_topol\_repres.overall\_topology

\_topol\_repres.overall\_topology\_EPINET

\_topol\_repres.overall\_topology\_RCSR

\_topol\_repres.overall\_topology\_SP

\_topol\_repres.overall\_topology\_TOPOS

\_topol\_repres.period

\_topol\_repres.td10

\_topol\_repres.total\_point\_symbol

\_topol\_repres.genus

This set of items collects information on the connectivity of an underlying net. The overall topology of the net can be described in accordance with four known nomenclatures, or given in an arbitrary format. The followingnomenclatures are used to designate the overall topologies of periodic nets:

* RCSR lower-case bold three-letter symbols, see O'Keeffe *et al*. (2008) and http://rcsr.net/ for details. Example: **dia**  (see http://rcsr.net/nets/dia)
* EPINET *sqc*XXXXX symbols, see http://epinet.anu.edu.au/ for details.
* symbols *k*/*m*/*fn* for sphere packings as proposed by Koch *et al*. (2006).
* Zeolite capital three-letter symbols, see http://www.iza-structure.org/databases/ for details.
* TOPOS symbols N*D*n (Aman *et al*., 2014), where N is a sequence of degrees (coordination numbers) of all independent nodes; *D* is one of the letters *C*, *L*, or *T* designating the dimensionality of the net (*C* – chain, *L* – layer, *T* – three-periodic); n enumerates non-isomorphic nets with a given N*D* sequence. For instance, the symbol 3,3,4*T*3 denotes the 3rd (by the order) three-periodic trinodal net with two 3-coordinated and one 4-coordinated independent nodes. For finite (molecular) graphs the symbols N*M*k-n are used, where k is the number of vertices (atoms) in the graph.
* Subnet **s**-d-*G*-n symbols (Blatov, 2007; Blatov & Proserpio 2009), where **s** is a conventional name of the initial net, d is a set of ascending integers equal to degrees of all inequivalent nodes in the subnet, *G* is the space group for the most symmetrical embedding of the subnet, n is optional and enumerates non-isomorphic subnets with a given **s**-d-*G* sequence. Examples: **scu**-3,6-*P*42/*mnm*-2 (is a 3,6-c net derived from 4,8-c **scu**); **acs**-4-*Pbcn* (is a 4-c net derived from 6-c **acs**)
* Subnet transformation symbols **s**/*G*→*S*1→…→*S*n;BS (Blatov, 2007), where **s** is a conventional name of the initial net, *G* is the space group of the initial net, *S*1, …, *S*n is the sequence of group-subgroup transformations to obtain the symmetry of the resulting subnet, *S*n, BS is the set of numbers of non-equivalent edges to be retained in the subnet. For instance, the notation fny/P 63/m c m->P 63 2 2 (0,0,1/4);Bond sets: 2,3,4,5 means that the subnet is derived from the RCSR net **fny** by decreasing its space-group symmetry from *P*63/*mcm* to *P*6322 with shifting the origin by (0, 0, 1/4) vector and breaking all non-equivalent edges in the resulting net except the edges No 2, 3, 4 and 5.

TOPOL\_REPRES\_ENTANGL contains no items because the descriptors of entanglements in crystal structures have not been consistent yet. However, the importance of entanglements in the topological description is undoubtful , and this category should be developed in future versions of the dictionary.

TOPOL\_REPRES\_TILING

\_topol\_repres\_tiling.vertices

\_topol\_repres\_tiling.edges

\_topol\_repres\_tiling.faces

\_topol\_repres\_tiling.tiles

\_topol\_repres\_tiling.signature

\_topol\_repres\_tiling.Dsize

\_topol\_repres\_tiling.dual

TOPOL\_REPRES\_TILING\_FACES

\_topol\_repres\_tiling\_faces.tile\_id

\_topol\_repres\_tiling\_faces.size

\_topol\_repres\_tiling\_tile.count

TOPOL\_REPRES\_TILING\_TILES

\_topol\_repres\_tiling\_tile.id

\_topol\_repres\_tiling\_tile.count

These three categories describe a tiling, which is admitted by the underlying net. As a rule, natural tiling (Blatov *et al*., 2007) is described, since it contains minimal cages, which can be glued yielding all other larger cages including infinite channels.

TOPOL\_REPRES\_OCCURRENCE

\_topol\_repres\_occurrence.total

\_topol\_repres\_occurrence.id

* \_topol\_repres\_occurrence.citation\_id

→ \_citation.id

*The bullet (•) indicates a category key. The arrow (→) is a reference to a parent data item.*

The TOPOL\_REPRES\_OCCURRENCE category stores information on the occurrences of a particular topology in other crystal structures and can be used together with the \_topol\_repres.overall\_topology\_\* items for developing topological databases. Reference or collection codes of crystal structure determinations can be used as \_topol\_repres\_occurrence.id values, for example, Reference Codes from the Cambridge Structural Database or Collection Codes from the Inorganic Crystal Structure Database.

**3.3. Development of the supporting software and databases**

The CIF topological data can be read and generated by the *ToposPro* program package (Blatov *et al*., 2014) and online service *TopCryst* (https://topcryst.com). Jmol software [Jmol] can also read and visualize the nets specified in the topoCIF format. The topological databases Reticular Chemistry Structure Resource (RCSR, http://rcsr.net/) and ToposPro Collections (TTD and TTO, https://topospro.com) use the topoCIF topological descriptors and, hence, one can use the topoCIF format for exporting or importing information to or from these databases. Thus, the crystallographic information on old and new structure determinations can be supplemented with the connectivity data, stored and shared.

**References**

Alexandrov, E. V., Blatov, V. A., Kochetkov, A. V. & Proserpio, D. M. (2011). *Underlying nets in three-periodic coordination polymers: topology, taxonomy and prediction from a computer-aided analysis of the Cambridge Structural Database. CrystEngComm*, **13**, 3947-3958.

Aman, F., Asiri, A. M., Siddiqui, W. A., Arshad, M. N., Ashraf, A., Zakharov, N. S. & Blatov, V. A. (2014). *Multilevel topological description of molecular packings in 1,2-benzothiazines. CrystEngComm*, **16**, 1963-1970.

Blatov, V.A. (2004). *Voronoi-Dirichlet polyhedra in crystal chemistry: theory and applications. Cryst. Rev.* **10**, 249-318.

Blatov, V. A. (2007). *Topological relations between three-dimensional periodic nets. I. Uninodal nets. Acta Cryst.* **A63**, 329–343.

Blatov, V. A., Delgado-Friedrichs, O., O’Keeffe, M. & Proserpio, D. M. (2007). *Three-periodic nets and tilings: natural tilings for nets. Acta Cryst*. **A63**, 418-425.

Blatov, V.A., O’Keeffe, M. & Proserpio, D. M. (2010). *Vertex-, face-, point-, Schläfli-, and Delaney-symbols in nets, polyhedra and tilings: recommended terminology*. *CrystEngComm*, **12**, 44-48.

Blatov, V. A., & Proserpio, D. M. (2009). *Topological relations between three-dimensional periodic nets. I. Binodal nets. Acta Cryst.* **A65**, 202-212.

Blatov, V.A., Shevchenko, A.P. & Proserpio, D.M. (2014). *Applied topological analysis of crystal structures with the program package ToposPro. Cryst. Growth Des.* **14**, 3576-3586.

Bonneau, C., O’Keeffe, M., Proserpio, D. M., Blatov, V. A., Batten, S. R., Bourne, S. A., Lah, M. S., Eon, J.-G., Hyde, S. T., Wiggin, S. B. & Öhrström, L. (2018) *Deconstruction of Crystalline Networks into Underlying Nets: Relevance for Terminology Guidelines and Crystallographic Databases. Cryst. Growth Des.*, **18**, 3411–3418

Delgado-Friedrichs, O. & O’Keeffe, M. (2003). *Identification of and symmetry computation for crystal nets. Acta Cryst.* **A59**, 351-360.

Delgado-Friedrichs, O. & O’Keeffe, M. (2005). *Crystal nets as graphs: terminology and definitions. J. Solid State Chem*. **178**, 2480-2485.

Delgado-Friedrichs, O., Foster, M. D., O'Keeffe, M., Proserpio, D. M., Treacy, M. M. J. & Yaghi, O. M. (2005). *What do we know about nets. J. Solid State Chem*. **178**, 2533-2554.

O'Keeffe, M., Peskov, M. A., Ramsden, S. J. & Yaghi, O. M. (2008). *The reticular chemistry structure resource (RCSR) database of, and symbols for, crystal nets. Acc. Chem. Res.* **41**, 1782-1789.

Koch, E., Fischer, W. & Sowa, H. (2006). *Interpenetration of homogeneous sphere packings and of two-periodic layers of spheres. Acta Cryst.* **A62**, 152-167.

Shevchenko, A.P. & Blatov, V.A. (2021). *Simplify to understand: how to elucidate crystal structures? Struct. Chem.* **32**, 507-519.