**Classification and use of topological data**

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

**1. Introduction**

Raw crystallographic data obtained from experiment or theoretical modeling contain no information on the crystal structure connectivity. However, this information, which might describe all sorts 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 among 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 a *net* (Delgado-Friedrichs & O’Keeffe, 2005; Delgado-Friedrichs *et al*., 2005). Unlike molecular graphs, nets are infinite and require special methods of description. Such descriptions have become increasingly common in routine crystal structure determinations and usually follow the conventional crystallographic description, which includes atomic coordinates, space group, unit cell dimensions, and other geometrical information, but does not concern the structure connectivity. Many topological descriptors have been proposed for nets; all of these descriptors are determined from the initial crystallographic data using algorithms. Software implementing these algorithms (Delgado-Friedrichs & O’Keeffe, 2003; Blatov *et al*., 2014) has made topological analysis available for the general crystallographic community and has already led to the development of special databases (O'Keeffe *et al*., 2008; Blatov *et al*., 2014) that collect the topological parameters of crystal structures. This work could be facilitated by an extension of the CIF format that would make it possible to store topological information along with a crystal structure and to describe the crystal architecture at different levels of organization.

==============================

Bob-Vlad conversation 4/25/21 -- remove at leisure

SO, WE MEET AGAIN! :)

What would you like me to look at, if anything, Vlad?

I will try to finish this version, then we will see what to do netx

OK. I am available and have two days pretty much free tomorrow and Wednesday.

Great; anyway we will be able to ask for some more days if needed

Ah, good. And I would like to get some better test files. All I have are old ones that only have TOPOL\_LINK, not NODE. Do we have those?

Not yet because ToposPro produces the old cif version, without

Sure. Maybe we could mock some up. I can do that. I'd like to have a CIF file that has both an actual structure and a TOPOL section. That should be no problem, right? I'm thinking that a bridging edge with multiple atoms would be good to have.

Ok, this is easy to do I suppose.

What do you think of:

1) My section 4? OK? Too much?

2) The idea of allowing for the bridging atoms to be identified -- tied to atoms in the CIF -- not just their MF?

Section 4 is a good idea - it is very helpful. I will read it carefully and let you know if anything should be changed. Concerning 2) - we should extend the format for this?

It would require, say, allowing for nodes to be identified not only for link ends but also for edge groups.

Ah -- which reminds me -- nodes can be centroids. So I propose allowing a node to describe a set of atoms, not just a single atom reference. Maybe:

TOPOL\_NODE\_ATOM\_LABELS (plural)

The problem is that we should discuss all changes at the IUCr Commissions - I am not sure that even our decision to remove “REPRES” will be immediately accepted. If they are going to issue a new IT G volume soon perhaps we have no time to make big changes.

Argument would be before is better than after ;)

We could ask them. But I think we will develop the format anyway.

This is James?

Yes

Let's see if we can agree on a small set of improvements. Something like the addition of a label -- that could just be a "version 2.0" idea; something like harmonizing the labels by removing REPRES, that's more fundamental -- and in my opinion more important to get right the first time.

I was pretty happy that the net construction from just the LINK data was (relatively) easy. That part looks excellent to me. OK -- I will leave you to your work! Fun communicating this way! Feel free to delete all this!

Thanks - we will be in touch; I hope to finish these corrections today or tomorrow; in the meantime you could think about extending the format for this idea. Plural labels is really very fruitful

OK. Looks like you did some with the dictionary? Was I right about the KEY business? (The bullets in this document)

Yes you could edit the .dic file

I will put it in a Google Doc so that we can comment.

Perfect. Should I remove this dialog to proceed editing? Or we can leave it for Davide.

========================

Although the core CIF dictionary contains some topological information in the CHEMICAL\_CONN, GEOM\_BOND, and GEOM\_CONTACT categories, this information is limited and does not cover all topological aspects of a crystal. The core CIF dictionary does not have the ability to describe a plethora of non-valence intra- or intermolecular interactions (hydrogen bonds, halogen and other specific bonds, van der Waals interactions, *etc*.). 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 intends to meet all these issues and requirements to make the exchange of the topological information 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. The TOPOL\_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 based on that connectivity information. Thus, the topoCIF dictionary enables one to generate two new levels of the crystal structure description: a connected net without any further detailing and a topological description of that connected net.

**3. Classification of data definitions**

**3.1. Underlying net description**

The *underlying net* is the net of centroids of structural units. The edges of the net represent the links between these units. The underlying net describes the topology of a particular representation of a crystal structure. This representation is said to be *complete* if the underlying net coincides with the initial structure, and partial if not. For example, diamond and SiO2 (cristobalite) can be represented by the same underlying net **dia**. For diamond, this net is complete, because all the bonding we would normally associate with the diamond structure are represented. But for cristobalite, **dia** is only a partial net, since it only represents the silicon atoms. The oxygen atoms can be considered to be bridges (links) between silicon atoms, but those links are represented just by edges of the **dia** net.

TOPOL\_NODE, TOPOL\_LINK, and TOPOL\_EDGE

All topological data in the TOPOL category describe one of several 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*; in the latter case the initial crystal structure should be properly simplified (Alexandrov *et al*. 2011; Bonneau *et al.* 2018; Shevchenko & Blatov, 2021). These three categories describe the connectivity of the underlying net.

TOPOL\_NODE

In general, the nodes of the underlying net and the crystal topology as a whole should be considered independently of the atoms. The category TOPOL\_NODE specifies the nodes of the underlying net and contains the following items, which describe the node site, composition and topology. In this document, bullet (∙) indicates a category key, and arrow (→) indicates a reference to a parent data item having that same value:

**node site**

* \_topol\_node.label

\_topol\_node.atom\_label

→ \_atom\_site\_label

\_topol\_node.fract\_x

\_topol\_node.fract\_y

\_topol\_node.fract\_z

\_topol\_node.symmetry\_multiplicity

\_topol\_node.Wyckoff\_symbol

**node composition**

\_topol\_node.chemical\_formula\_iupac

\_topol\_node.chemical\_formula\_moiety

\_topol\_node.chemical\_formula\_sum

\_topol\_node.structural\_formula\_InChI

\_topol\_node.structural\_formula\_SMILES

**node topology**

\_topol\_node.coordination\_sequence

\_topol\_node.point\_symbol

\_topol\_node.extended\_point\_symbol

\_topol\_node.vertex\_symbol

If the \_topol\_node.atom\_label item is present, the underlying net node labelled in the \_topol\_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, and labels \_topol\_node.label can coincide with \_topol\_node.atom\_label ~~can both refer to the same \_atom\_site\_label value~~, and items \_topol\_node.fract\_x, \_topol\_node.fract\_y, \_topol\_node.fract\_z, \_topol\_node.symmetry\_multiplicity, and \_topol\_node.Wyckoff\_symbol should not be present, as that information can be extracted from the corresponding items of the ATOM\_SITE category.

TOPOL\_LINK

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

\_topol\_link.node\_label\_1

→ \_topol\_node.label

\_topol\_link.node\_label\_2

→ \_topol\_node.label

\_topol\_link.distance

However, link symmetry operations are expressed differently from GEOM\_BOND or GEOM\_CONTACT. They are 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. Thus, we have:

\_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

There are several additional items for description of other link properties, including:

* \_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.

TOPOL\_EDGE

When a link of the underlying net corresponds to a briding atom or a complex polyatomic bridge (a ligand, or cluster, for example), the category TOPOL\_EDGE describes the chemical composition of that link, identified by the \_topol\_edge.id item:

* \_topol\_edge.id

→ \_topol\_link.id

\_topol\_edge.chemical\_formula\_iupac

\_topol\_edge.chemical\_formula\_moiety

\_topol\_edge.chemical\_formula\_sum

**3.2. Underlying net topological properties**

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

TOPOL

The following items describe the connectivity of an underlying net:

\_topol.overall\_topology

\_topol.overall\_topology\_EPINET

\_topol.overall\_topology\_RCSR

\_topol.overall\_topology\_SP

\_topol.overall\_topology\_TOPOS

\_topol.period

\_topol.td10

\_topol.total\_point\_symbol

\_topol.genus

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\_ENTANGL

This category contains no items, because the descriptors of entanglements in crystal structures have not been standardized yet. However, the importance of entanglements in the topological description is acknowledged , and this category is expected to be developed in future versions of the dictionary.

TOPOL\_TILING, TOPOL\_TILING\_FACES, and TOPOL\_TILING\_TILES

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 combined to provide all other larger cages, including infinite channels.

Items in these categories include:

\_topol\_tiling.vertices

\_topol\_tiling.edges

\_topol\_tiling.faces

\_topol\_tiling.tiles

\_topol\_tiling.signature

\_topol\_tiling.Dsize

\_topol\_tiling.dual

\_topol\_tiling\_faces.tile\_id

\_topol\_tiling\_faces.size

\_topol\_tiling\_tile.count

\_topol\_tiling\_tile.id

\_topol\_tiling\_tile.count

TOPOL\_OCCURRENCE

The TOPOL\_OCCURRENCE category stores information on the occurrences of a particular topology in other crystal structures and can be used together with the \_topol.overall\_topology\_\* items for developing topological databases. Items include:

\_topol\_occurrence.total

\_topol\_occurrence.id

* \_topol\_occurrence.citation\_id

→ \_citation.id

Reference or collection codes of crystal structure determinations can be used as \_topol\_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 (https://jmol.sourceforge.net) 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.

**4. Net Reconstruction from CIF Data**

The periodic net described by a topoCIF file is fully defined by its set of TOPOL\_NODE and TOPOL\_LINK data. The method of constructing a net for visualization or analysis from this data is not trivial, since generally a CIF file only provides locations for the asymmetric unit. Below we describe a sketch of the corresponding algorithm as implemented in Jmol. Net construction involves cataloging all symmetry operators, asymmetric unit nodes, and links, applying symmetry operators to the links to determine all possible link directions from a given symmetry-related position in the unit cell, and then matching those possibilities to a given set of nodes:

1) A list of symmetry operators **listSymOps** is created based on SPACE\_GROUP\_SYMOP data.

2) A list of nodes **listNodes** is created based on TOPOL\_NODE data, and the list is expanded using specified symmetry operators and lattice translations to fill as many unit cells as desired. This provides the full list of nodes we wish to connect.

3) A list of links **listLinks** is created based on TOPOL\_LINK data.

4) For each link, a list of symmetry-transformed primitive normalized edges is created. First, the link’s specified rotation and translation operations are applied to each of the two associated nodes’ fractional coordinates, and the pair of transformed points, p1 and p2, are normalized such that point p1’s fractional coordinates are all within the interval [0,1). Then, all symmetry operators are applied to these node positions, and if the distance between the two resultant Cartesian coordinates is within tolerance (0.01 Å for Jmol) of TOPOL\_LINK.DISTANCE, the primitive is saved; otherwise, it is discarded, and an error is reported. This ensures that all possible link positions and orientations are identified.

5) All relevant net connections associated with the nodes of **listNodes** are then calculated for each link. First, two lists of nodes are created: all nodes matching the first node reference, **listNode1**, and all those matching the second reference, **listNode2**. (In some cases, such as diamond, these lists are the same.) Two nested loops then iterate over all nodes in **ListNode1** and **ListNode2**. For each pair of nodes – node1 and node2 – an initial check is made that the pair satisfies the condition that their Cartesian distance matches the link within tolerances. The job of the first loop is to ascertain which of the cached primitives for our link match node1. We take the node’s associated asymmetric unit point symmetry operator-transformed location and unitize it into the [0,1) interval for each of its three fractional coordinates. If that unitized point is the primitive’s unitized node1, we mark that operator as appropriate. Otherwise, we skip it. In a second loop, which is within the node2 scanning loop, we scan through the symmetry operators, checking to see if the primitive’s p1-p2 vector matches our symmetry-transformed node1-node2 vector. If and only if this match is made and the connection does not already exist, we add a new connection. Note that we check here to make sure that we aren’t declaring a connection multiple times. When a new connection is found, we also register the nodes that are connected. This allows us to remove unconnected nodes later, if desired.

**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.