Struct Chem (2017) 28:39–44
DOI 10.1007/s11224-016-0853-3

REVIEW ARTICLE

Crystal structures as periodic graphs: the topological genome and graph databases

Olaf Delgado-Friedrichs[1] · Stephen T. Hyde[1] · Michael O'Keeffe[2] · Omar M. Yaghi[3,4]

Received: 11 July 2016 / Accepted: 13 September 2016 / Published online: 23 September 2016

Abstract We call attention to methods of enumerating periodic structures and to the databases that contain them. These provide information essential to the systematic design of crystalline materials. The underlying topology is uniquely specified and identifiable from the *Systre key* which is thus the topological genome.

Keywords Nets · Tilings · Databases · Materials design

The last 20 years have seen a profound shift in the way the structures of crystalline materials are described. The change is from crystal structures as sets of points (sphere packings etc.) to crystal structures as sets of *connected* points: graphs. The graph describing the underlying topology of a crystal structure is usually termed a *net* which is a special kind of graph—*connected* (all vertices linked by a path of edges) and *simple* (no edges beginning and ending on the same vertex, or multiple edges) [1].

Parallel to that change in the way of regarding crystal structures was the development of *reticular chemistry* which is concerned with the designed synthesis of materials of predetermined structure by linking molecular modules (secondary building units, SBUs) into periodic frameworks [2]. However, as a distinguished crystal chemist observed some years ago, "the synthesis of new structures requires not only chemical skill but also some knowledge of the principal topological possibilities" [3]. This requirement is equally important to the development of theoretical databases of potentially useful materials. The topic of analysis of crystal structures in terms of underlying nets is large and has been reviewed elsewhere [1]. In this paper, we describe our work, which we believe to be unique, in which those principal topological possibilities have been discovered, characterized, and recorded in searchable databases.

The *systematic* discovery of new nets has been made possible by the development of combinatorial tiling theory. In particular, the generalized Schläfli symbol—the Delaney-Dress symbol (D-symbol)—can be used to determine all periodic tilings of a space (e.g., three-dimensional Euclidean space) of a given kind. The 1-skeleton of a periodic tiling (the set of vertices and edges) is in fact a periodic net *carried* by the tiling—just the topological structure that is sought. In an early application to enumeration of potential zeolite structures, the D-symbol was identified as an example of an "inorganic gene," a concept earlier introduced by Mackay [4]. However, as a given three-dimensional net can have many tilings (or none at all because of self entanglement), we consider the D-symbol rather as a seed or embryo and describe here a topological genome that provides the information that uniquely and completely specifies a periodic net.

A route to uniquely characterizing a periodic net came from what is called an *equilibrium placement.* In that, the vertices are assigned *barycentric* coordinates in which each vertex has coordinates that are the average of those of its neighbors. It is easy to show that if one vertex is arbitrarily

✉ Olaf Delgado-Friedrichs
olaf.delgado@gmail.com

[1] Department of Applied Mathematics, Research School of Physics and Engineering, The Australian National University, Canberra, ACT 2601, Australia

[2] School of Molecular Sciences, Arizona State University, Tempe, AZ 85287, USA

[3] Department of Chemistry, University of California, Berkeley, CA 94720, USA

[4] Kavli Energy Nanosciences Institute, Berkeley, CA 94720, USA

assigned coordinates (say 0, 0, 0) and a translation lattice given, the rest are then uniquely determined [5]. A key result is that for nets in which no two vertices have the same barycentric coordinates (nets without *collisions*), the automorphism group of the abstract graph is isomorphic with a crystallographic space group (an automorphism of a graph is a permutation of vertices that preserves the edge-vertex connectivity) [5, 6]. Virtually all the nets of practical interest in materials crystal chemistry are such crystallographic nets. For them a unique identifier (the *Systre key*) can be determined so that it can be definitely stated whether or not two nets are the same [7]. Thus for this class of structures, the notorious graph isomorphism problem is solved.

The *Systre key* is the net genome. It contains *all* topological information about the net such as automorphism group and topological invariants like point symbols, cycle structure, and coordination sequences. Importantly, it is also unique for a given net.

We illustrate these points by considering the structure of an inorganic material, faujasite, also known as zeolite X and zeolite Y, which is one of the most valuable inorganic materials due to the central role it plays in the petrochemical industry. The structure is an aluminosilicate framework with tetrahedral atoms (Al and Si) linked into a 4-coordinated net by –O– links. The framework structure is actually a special type known as a simple tiling by simple polyhedra. A simple polyhedron is one with only 3-coordinated vertices. In a simple tiling, two such polyhedra meet at a face (the tiling is *face-to-face*), three at an edge, and four at a vertex. The D-symbol tells us how these polyhedra are arranged in a tiling. Such simple tilings are ubiquitous in nature, for example as foams, biological structures, and the grain pattern of polycrystalline materials.

But we can go deeper. Each tiling has a dual structure in which the vertices of the dual are inside the original tiles and linked by new edges to new vertices inside adjacent tiles. Noting that the dual of a dual is the original tiling completes the specification. The dual of a simple tiling is a tiling by tetrahedra.

An important mathematical result is that there are exactly nine combinatorial types of tiling by one tetrahedron (i.e., nine isohedral tilings by tetrahedra) [8]. It follows at once that there are just nine uninodal (one kind of vertex) simple tilings. Interestingly, seven of these are known zeolite structures including that of faujasite. The other two contain 3-rings which are unfavorable for aluminosilicate zeolite structures.

In practice, in deriving a D-symbol, a tile is divided into tetrahedral *chambers* whose vertices are one each of the vertices of the tile and the centers of edges, of faces, and of the tile itself. Figure 1 shows how the tetrahedron of the faujasite dual structure is divided into 24 chambers. A D-symbol is then a labeled graph which contains a node for each kind (w.r.t. to symmetry) of chamber, colored edges describing where chambers share faces with each other, and finally numbers indicating how many of these chambers meet around a given common edge [9]. The latter determine such properties of the tiling as the sizes of its two-dimensional faces, the degrees of vertices within individual tiles, and so on. Hence the name generalized Schläfli symbol.

The figure also shows a textual representation of the D-symbol which completely specifies the combinatorial structure of the tiling and its dual, and hence implicitly the net of the faujasite structure. D-symbols can be read by the program 3dt (available at gavrog.org) which, among many other things, can export the net topology.

The faujasite structure as drawn by 3dt is shown in Fig. 2 as a tiling of space by polyhedra—hexagonal prisms, truncated octahedra, and larger polyhedra (*faujasite cages*). But also shown in the figure is a second description of the structure—a tiling by polygons of a two-dimensional periodic surface. In this description, to which we return below, the structure is a $4^3.6$ tiling of a surface known as the *D* surface.

Figure 3 shows the Systre key for the faujasite net. This is in fact what is known as a *labeled quotient graph* [10]. The first number indicates that the structure is 3-periodic. Thereafter, each set of five numbers defines an edge of the repeat unit. Thus, the last five say that the vertex 44 is joined to vertex 48 in the unit cell displaced by 0, 0, 1 from the origin. An essential point is that there are an impossibly large number of different labelings of the quotient graph for a typical net. Thus for the faujasite net, there are 48 vertices in the repeat unit, and thus $48! \approx 10^{61}$ vertex numberings, and for these in principle an infinite number of choices of coordinate bases. However, by use of the equilibrium placement, combined with graph traversing techniques, Systre is able to reduce these numbers dramatically. The key observation here is that, once a coordinate system is fixed, the neighbors of any given vertex can be sorted lexicographically by their assigned coordinate values, which in the absence of collisions allows one to construct a unique numbering inductively. For a graph with E edges in the repeat unit, at most E^3 different coordinate systems and hence numberings have to be considered. In practice, the actual number is usually much smaller, e.g., 1152 (vs. $96^3 \approx 10^6$) in the case of faujasite. Of all the potential representations based on these numbering, the lexicographically smallest is chosen as the canonical key [11]. Systre is available as part of the Gavrog package (gavrog.org).

We turn now to enumeration of nets. Tilings, and hence the nets they carry, can be systematically enumerated, and

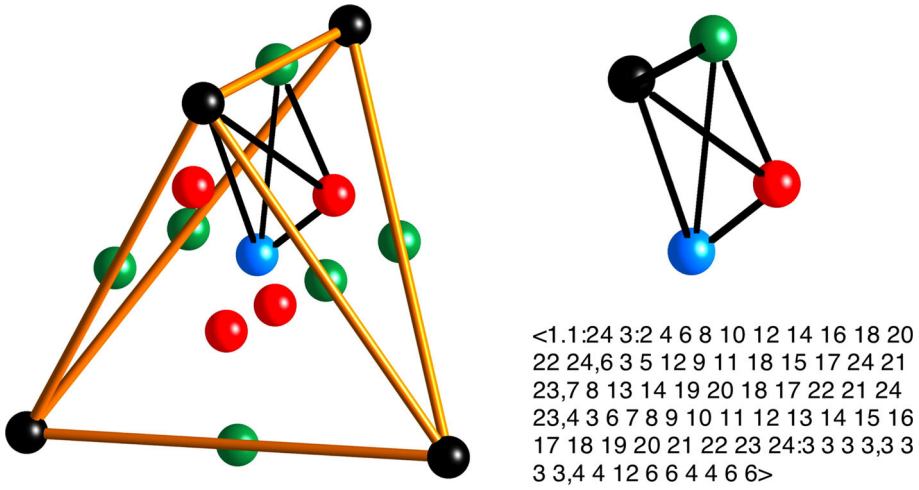

Fig. 1 *Left* a space-filling tetrahedron whose tiling is the dual of the faujasite tiling. *Black balls* are at the tetrahedron vertices. *Green*, *red,* and *blue balls* are at the centers of, respectively, edges, faces, and the tile. *Top right* one is the tetrahedra chamber of the tile. Below that is the D-symbol which specifies how the chambers are assembled in the tiling (Color figure online)

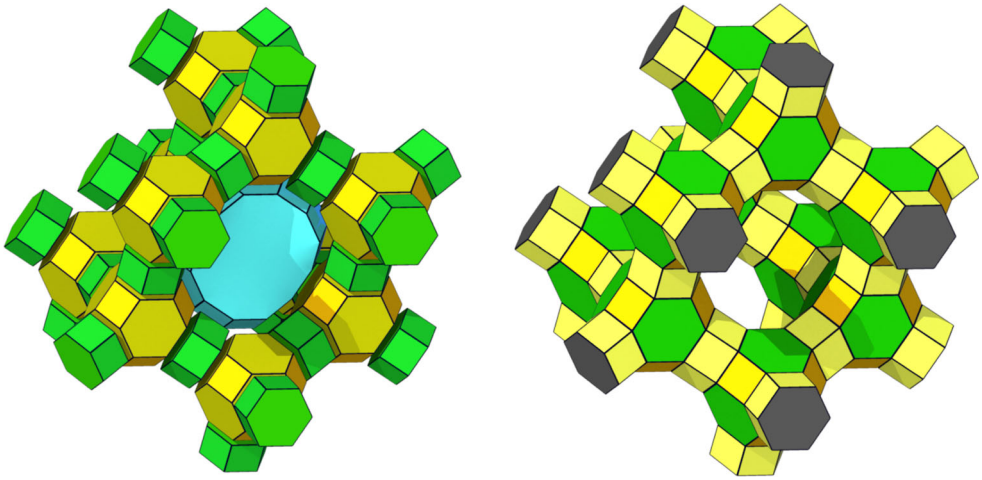

Fig. 2 Faujasite structure. *Left* as a tiling of Euclidean space by three-dimensional tiles (*yellow*, *green,* and *blue*). *Right* as a two-dimensional tiling of a periodic surface by quadrangles (*yellow*) and hexagons (*green*) (Color figure online)

3 1 2 0 0 0 1 3 0 0 0 1 4 0 0 0 1 5 0 0 0 2 6 0 0 0 2 7 0 0 0
2 8 0 0 0 3 7 0 0 0 3 9 0 0 0 3 10 0 0 0 4 8 0 0 0 4 11 0 0
0 4 12 0 0 0 5 10 0 0 0 5 13 0 0 0 5 14 0 0 0 6 15 0 0 0 6
16 0 0 0 6 17 0 0 0 7 18 0 0 0 7 19 0 0 0 8 17 0 0 0 8 20 0
0 0 9 19 0 0 0 9 21 0 0 0 9 22 0 0 0 10 23 0 0 0 10 24 0 0
0 11 20 0 0 0 11 21 0 0 0 11 25 0 0 0 12 25 0 0 0 12 26 0
0 0 12 27 0 0 0 13 23 0 0 0 13 28 0 0 0 13 29 0 0 0 14 27
0 0 0 14 29 0 0 0 14 30 0 0 0 15 31 0 0 0 15 32 0 0 0 15
33 0 0 0 16 32 0 0 0 16 34 0 0 0 16 35 0 0 0 17 35 0 0 0
17 36 0 0 0 18 33 0 0 0 18 37 0 0 0 18 38 0 0 0 19 38 0 0
0 19 39 0 0 0 20 39 0 0 0 20 40 0 0 0 21 39 0 0 0 21 41 0
0 0 22 31 1 0 0 22 32 1 0 0 22 41 0 0 0 23 42 0 0 0 23 43
0 0 0 24 32 1 0 0 24 34 1 0 0 24 43 0 0 0 25 42 0 1 0 25
44 0 0 0 26 31 0 0 1 26 33 0 0 1 26 42 0 1 0 27 33 0 0 1
27 37 0 0 1 28 31 0 −1 1 28 36 0 −1 1 28 41 −1 −1 1 29 36
0 −1 1 29 40 0 −1 1 30 34 1 0 0 30 37 0 0 1 30 40 0 −1 1
34 45 0 0 0 35 46 0 0 0 35 47 0 0 0 36 46 0 0 0 37 47 1 0
−1 38 47 1 0 −1 38 48 0 0 0 39 45 1 1 −1 40 45 1 1 −1 41
46 1 0 0 42 48 0 −1 1 43 45 1 0 0 43 48 0 −1 1 44 46 1 0
0 44 47 1 0 0 44 48 0 0 1

Fig. 3 Systre key for the net of the faujasite structure

this has been done, for example, for simple tilings and quasi-simple tilings (duals of tilings by generalized tetrahedra). These all provide 4-coordinated nets of special interest to the design and synthesis of materials like zeolites and zeolitic imidazolate frameworks (ZIFs) [12, 13].

Of particular interest was the enumeration of tilings with one kind of face (face transitive)—the duals of these have nets with one kind of edge (edge transitive) [14]. Such nets are by far the most important for design and synthesis of materials such as metal–organic frameworks (MOFs) [2, 15].

But there is another way of generating nets from tilings which is, perhaps, even more powerful. We remarked above that the faujasite net could be considered as a $4^3.6$ tiling of a periodic surface. But note that a tiling of the plane by quadrangles is 4^4. A tiling formed replacing a quadrangle by a hexagon requires excess area compared with the flat plane. In fact, the $4^3.6$ pattern tiles the hyperbolic plane (a surface of negative curvature). One representation of this tiling is shown in Fig. 4.

The geometry and topology of the faujasite structure can be constructed in three-dimensional space by mapping that hyperbolic tiling onto a sponge-like Euclidean crystalline surface of negative curvature [16]. Such surfaces are well

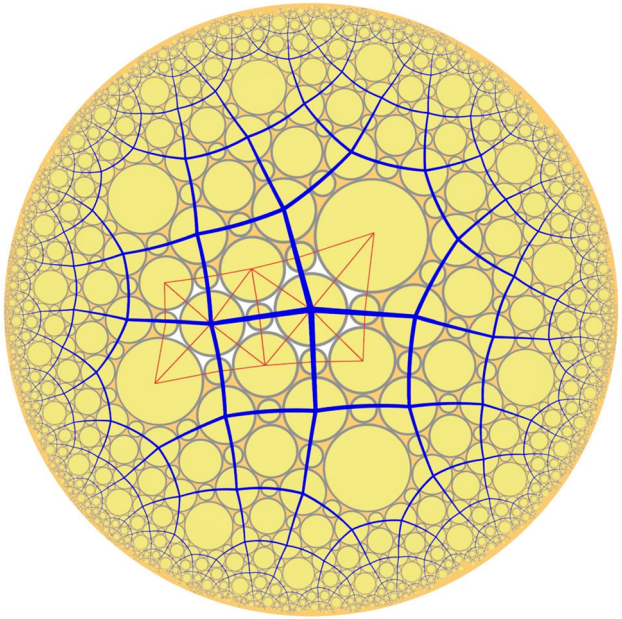

Fig. 4 Hyperbolic tiling $4^3.6$ shown on a Poincaré disk, from EPINET (http://epinet.anu.edu.au/hqc576)

known in materials physics and chemistry, particularly the three cubic surfaces known as *P*, *D*, and *G (gyroid)*. Both the *D* and *G* surfaces, for example, are formed in the bicontinuous cubic phases of liquid crystals [17], and the G structure underlies the structure of the MCM-48 mesoporous silicate material [18]. Two-dimensional tilings of the hyperbolic plane—that are readily enumerable—can be used to construct three-dimensional nets systematically [19, 20]. The EPINET project ("Euclidean Patterns Induced by Non-Euclidean Tilings" [21]) systematically generates crystallographic nets in this way. The variety of nets that can be produced is incredibly rich. An initial enumeration derived from the most symmetric hyperbolic tilings of limited complexity has already produced some 14,000 3-periodic nets; many more are in the pipeline.

A second database, much smaller (about 2600 nets), is the RCSR (Reticular Chemistry Structure Resource), a collection of the nets, known so far, likely to be of special interest to the design and syntheses of materials [22]. The criteria for inclusion include topological simplicity (small numbers of kinds of vertices and edges) and/or special interest in the theory of periodic nets and tilings. Many of the RCSR nets also occur in EPINET (indeed often come from there). Although primarily a database of periodic graphs, they are given as optimal embeddings with space group, unit cell, and coordinates for vertices and edges. The entry for faujasite is shown as Fig. 5. Each RCSR net has a symbol (lower case bold letters) such as **fau** for the faujasite net and **dia** for the net of the diamond structure. Note that **fau** is also identified as sqc13159, its EPINET identifier.

There are also other specialized databases of nets suitable for silicate zeolites [23, 24]. These are restricted to nets with embeddings in which every vertex is in tetrahedral coordination. The program package ToposPro has a large database of nets including all RCSR and EPINET nets [25]. This can be searched for occurrences of nets in reported crystal structures. This program can also be used to generate tilings for nets in a form suitable as inputs for 3dt.

We have presented D-symbols and labeled quotient graphs (in the form of Systre keys) as examples of Alan Mackay's general concept of an inorganic gene [26] (see also [27–30])—a mechanism that allows us to encode a spatial structure in the form of a lineal or one-dimensional sequence such that we can translate back and forth between the structure and its encoding, similarly to how a biological gene encodes a protein in the form of a piece of DNA strand and conversely, at least in principle, the gene could be recovered from the protein. In contrast to Mackay's original proposal, which uses bond lengths, angles, and torsion for the encoding [26], both the D-symbols proposed later [4] and the labeled quotient graphs we focus on here are purely topological and do not contain any explicit measurements.

It is customary and convenient to use invariants such as vertex symbols and coordination sequences for structure recognition, but as these do not contain the complete information necessary for reconstructing a structure, they can never provide positive proof that two candidate structures are indeed identical. By contrast, a topological gene contains all the information necessary, but unlike its biological counterpart, the same abstract gene can have a large number of concrete, one-dimensional representations. As we have demonstrated above, however, this difficulty can be overcome and we can efficiently compute a "canonical" genetic sequence, the Systre key, which uniquely and completely describes a crystallographic net and hence combines many of the advantages of traditional invariants and an inorganic gene.

The Systre key is a unique textual representation of a net such that two nets are the same if and only if their keys are identical. Once this representation is computed, a simple text comparison is all that is needed to identify a given net. Besides the Systre software itself, this property has been used in the EPINET project to remove duplicates from the enumeration, as well as in ToposPro, where the Systre key serves as a final tie-breaker when identifying nets. But we believe that there is even greater potential. Well-established hashing techniques can be used to condense the Systre key—which can be quite large for complicated nets—into a fixed-sized "fingerprint," e.g., a 40 character sequence of letters and digits using the popular SHA1 algorithm [31]. Such a fingerprint can then be used for

Fig. 5 Faujasite page from the RCSR database (http://rcsr.anu.edu.au/nets/fau)

fau

RCSR reference: http://rcsr.net/nets/fau

names: faujasite, 4/4/c17, sqc13519
key words: zeolite net, simple tiling, good
references: J. Solid State Chem. 178, 2533-2554 (2005)

embed type	space group	volume	density	genus	td10
1a	Fd-3m	505.6330	0.3797	49	579

a	b	c	alpha	beta	gamma
7.9667	7.9667	7.9667	90.0	90.0	90.0

vertices: 1

vertex	cn	x	y	z	symbolic	Wyckoff	symmetry	order
V1	4	0.0362	0.1250	0.3025	x, y, z	192 i	1	1

vertex	cs_1	cs_2	cs_3	cs_4	cs_5	cs_6	cs_7	cs_8	cs_9	cs_{10}	cum_{10}	vertex symbol
V1	4	9	16	25	37	53	73	96	120	145	579	4.4.4.6.6.12

edges: 4

edge	x	y	z	symbolic	Wyckoff	symmetry
E1	0.0806	0.0806	0.3025	x, x, z	96 g	m
E2	0.0806	0.1694	0.3025	x, 1/4-x, z	96 g	m
E3	0.0000	0.1612	0.3388	0, y, 1/2-y	96 h	2
E4	0.9919	0.1250	0.2581	x, y, 1/4-x	96 g	m

tiling:

tiling	dual	vertices	edges	faces	tiles	D-symbol
2[4^6.6^2] + [4^6.6^8] + [4^18.6^4.12^4]	-	1	4	5	3	24

example in URLs, and we envision a "meta-search engine" for structures in which a net can be entered and searched for in a variety of participating structure databases via the Systre key and its fingerprint. We further hope that Mackay's work as well as our own contributions may inspire future research into "genes" and canonical representations in chemistry and material science.

Acknowledgments Work supported by the US National Science Foundation, Grant No. DMR 1104798.

References

1. O'Keeffe M, Yaghi OM (2012) Chem Rev 112(2):675–702
2. Yaghi OM, O'Keeffe M, Ockwig NW, Chae HK, Eddaoudi M, Kim J (2003) Nature 423:705–714
3. Brunner GO (1981) Zeit Kristallogr 156(1–4):295–304
4. Delgado-Friedrichs O, Dress AWM, Huson DH, Klinowski J, Mackay AL (1999) Nature 400:644–647
5. Delgado-Friedrichs O (2005) Discrete Comput Geom 33:67–81
6. Moriera de Oliveira M Jr, Eon JG (2014) Acta Crystallogr A 70(3):217–228
7. Delgado-Friedrichs O, O'Keeffe M (2003) Acta Crystallogr A 59(4):351–360
8. Delgado-Friedrichs O, Huson DH (1999) Discrete Comput Geom 21(2):299–315
9. Delgado-Friedrichs O (2003) Theor Comput Sci 303(2):431–445
10. Chung SJ, Hahn T, Klee WE (1984) Acta Crystallogr A 40(1):42–50
11. Delgado-Friedrichs O, O'Keeffe M (2005) Acta Crystallogr A 61(3):358–362

12. Delgado-Friedrichs O, Huson DH (2000) Discrete Comput Geom 24(2–3):279–292
13. Foster MD, Delgado-Friedrichs O, Bell RG, Almeida Paz FA, Klinowski J (2003) Angew Chem 115(33):4026–4029
14. Delgado-Friedrichs O, O'Keeffe M (2007) Acta Crystallogr A 63(4):344–347
15. Delgado-Friedrichs O, O'Keeffe M, Yaghi OM (2007) Phys Chem Chem Phys 9:1035–1043
16. Andersson S, Hyde ST, von Schnering HG (1984) Zeit Kristallogr 168(1–4):1–18
17. Hyde ST (1996) Curr Opin Solid State Mater Sci 1(5):653–662
18. Carlsson A, Kaneda M, Sakamoto Y, Terasaki O, Ryoo R, Joo SH (1999) J Electron Microsc 48(6):795–798
19. Hyde ST, Delgado-Friedrichs O, Ramsden SJ, Robins V (2006) Solid State Sci 8(7):740–752
20. Ramsden SJ, Robins V, Hyde ST (2009) Acta Crystallogr A 65(2):81–108
21. Euclidean Patterns in Non-Euclidean Tilings (2009) Australian National University, Canberra. http://epinet.anu.edu.au. Accessed 9 Jul 2016
22. O'Keeffe M, Peskov MA, Ramsden SJ, Yaghi OM (2008) Acc Chem Res 41(12):1782–1789
23. Treacy MMJ, Rivin I, Balkovsky E, Randall KH, Foster MD (2004) Microporous Mesoporous Mats 74(1):121–132
24. Pophale R, Cheeseman PA, Deem MW (2011) Phys Chem Chem Phys 13(27):12407–12412
25. Blatov VA, Shevchenko AP, Proserpio DM (2014) Cryst Growth Des 14(7):3576–3586
26. Mackay AL (1995) J Mol Struct (Theochem) 336(2):293–303
27. Mackay AL (1986) Comput Math Appl 12(1–2):21–37 (reprinted, Mackay AL (1986) In: Hargittai I (ed) Symmetry: unifying human understanding. Pergamon Press, New York, pp 21–37)
28. Mackay AL, Klinowski J (1986) Comput Math Appl 12(3–4):803–824 (reprinted, Mackay AL, Klinowski J (1986) In: Hargittai I (ed) Symmetry: unifying human understanding. Pergamon Press, New York, pp 803–824)
29. Shevchenko VY, Krivovichev SV (2008) Struct Chem 19(4):571–577
30. Rajan K (2015) Annu Rev Mater Res 45:153–169
31. Eastlake D 3rd, Jones P (2001) US secure hash algorithm 1 (SHA1) (No. RFC 3174)