The Need for a New Generation of Substructure Searching Software

Paul R. Raithby[a,b] and Robin Taylor[b,*]

[a]*Department of Chemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK*[2]

[b]*Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK*[3]

Abstract

Advances in synthetic chemistry mean that the molecules now synthesized include increasingly complex entities with mechanical bonds or extensive frameworks. For these complex molecular and supramolecular species, single-crystal X-ray crystallography has proved to be the optimal technique for determining full three-dimensional structures in the solid state. These structures are curated and placed in structural databases, the most comprehensive of which (for organic and metallo-organic structures) is the Cambridge Structural Database (CSD). A question of increasing importance is how users can search such databases effectively for these structures. In this *Opinion* we highlight some of the classes of complex molecules and supramolecules and the challenges associated with searching for them. We develop the idea of substructure searches that involve topological searches as well as searches for molecular fragments, and propose significant enhancements to substructure-search programs that are both achievable and highly beneficial for both the database user community and the broader chemistry community.

1. Introduction

Over the last four decades there have been ground-breaking advances in synthetic chemistry that have moved the emphasis from the molecule to interlinked molecules and supramolecules, with the additional control over material properties and functions that this brings (Cram, 1988; Lehn, 1988; Pedersen, 1988). While it can be argued that the chemistry of the covalent bond is now relatively well understood, the control of the formation of molecules and supramolecules assembled through the formation of mechanical linkages and intermolecular interactions remains a challenge. Through their ingenuity synthetic chemists have assembled more and more complex materials containing an increasing number of diverse building blocks (Ward and Raithby, 2013) including the formation of molecular machines (Feringa, 2017; Sauvage, 2017; Stoddart, 2017; Sluysmans and Stoddart, 2019). The characterisation of these increasingly complex systems has also become more challenging. Fortunately, with the advances in instrumentation and computing power the complexity of crystal structures that can now be solved and refined to atomic resolution using modern single-crystal X-ray crystallographic techniques has grown enormously and the line between molecular and macromolecular crystallography has become blurred (Helliwell, 2017). As well as the natural driver for understanding the structures of biologically relevant macromolecules and their relationship to the development of pharmaceuticals, there has been an equally significant push in the functional materials arena because of the relationship between structure and properties (Roy, Reddy and Hazra, 2018; Dai et al., 2020).

Many classes of complex molecular and supramolecular materials have now been studied. Among the first complex systems to be studied were the catenanes (Hamilton et al., 1998) and rotaxanes (Bravo et al., 1998), which can be described as mechanically interlocked molecules (MiMs). Subsequently, these classes of complexes have been extended to include, amongst others, double helicates (Hasenknopf et al., 1996), unimolecular cages (Zhang, Ronson and Nitschke, 2018; Percastegui, Ronson and Nitschke, 2020) and molecular knots (van Dongen et al., 2014; Danon et al.,

[2] P. R. Raithby: e-mail: p.r.raithby@bath.ac.uk; orcid.org/0000-0002-2944-0662
[3] R. Taylor: e-mail: robin@justmagnolia.co.uk; orcid.org/0000-0002-0391-2609

2017; Fielden, Leigh and Woltering, 2017). In addition, over the last three decades the area of coordination polymers (Hoskins and Robson, 1990; Batten and Robson, 1998) and three dimensional metal-organic frameworks (MOFs) (Furukawa et al., 2013) has come to prominence. These types of 3D structures have subsequently expanded to include zeolitic imidazolate frameworks (ZIFs) (Pimentel et al., 2014) and covalent organic frameworks (COFs) (Ding and Wang, 2013; Lee and Cooper, 2020).

There are two aspects to the structures of these complex molecular and supramolecular species: firstly, the connectivities of the component chemical units; and secondly the topology of the complete entity. For example, in the catenane structure NIFLAP[1] (Hamilton et al., 1998) (Fig. 1) the units comprising the structure are dinaphtho crown rings with three hexadiyne linkers while the topology is described as two interlocking rings, a [2]catenane. More complex catenanes and other species such as rotaxanes, cage complexes and molecular knots exhibit a range of different topologies, such as the [6]catenane metal-peptide capsule ROYSAC (Fig. 2a) (Sawada et al., 2019) and the supramolecular pseudo-rotaxane ADOMOW (Fig. 2b) (Miljanić et al., 2007).

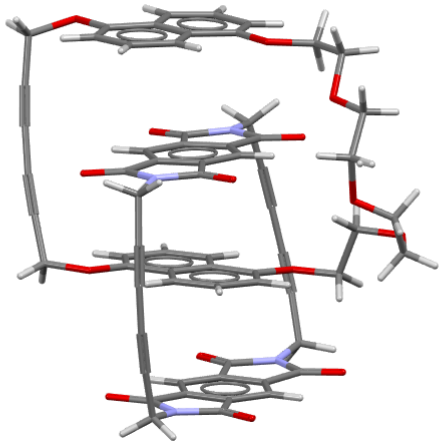
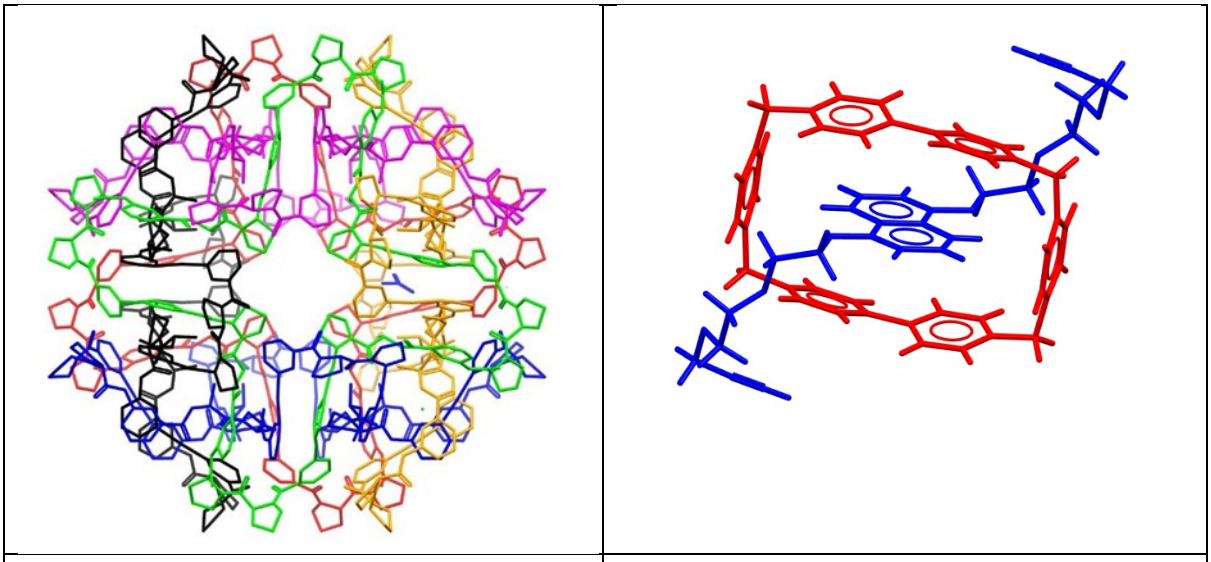


Figure 1. The [2]catenane NIFLAP

When considering MOFs there are again two levels of complexity, viz., identifying the chemical composition of the MOF and then the framework pattern that it adopts. Many MOFs are simply named after the laboratory in which they were prepared and identified; for example, MOF-5 (sometimes called IRMOF-1) which is $Zn_4O(BCD)_3$ (where $(BCD)^{2-}$ is 1,4-benzodicarboxylate) discovered by Yaghi *et. al*. (Rosi et al., 2003), or NOTT-112 with the chemical formula ($[Cu_3(L)(H_2O)_3)]\cdot 8DMSO\cdot 15DMF\cdot 3H_2O$) (L = 1,3,5-*tris*(3′,5′-dicarboxy[1,1′-biphenyl]-4-yl)benzene), characterised by Schröder and Champness at the University of Nottingham (Yan et al., 2009). Many of these names are in common usage but are not particularly helpful for describing the structures of the MOFs. The second requirement is a nomenclature that defines the "nodes" and "linker" groups or "struts" of the MOF as well as the connectivity between them. Collectively they define a framework ("net") topology.

[1] Here and elsewhere we identify structures by their refcodes in the Cambridge Structural Database (CSD) (Groom et al., 2016).



(a) ROYSAC | (b) ADOMOW

Figure 2. Examples of complex catenanes and rotaxanes. (a) ROYSAC a [6]catenane metal-peptide capsule and (b) ADOMOW a pseudo-rotaxane

Because of the complexity of framework structures and the challenges involved in searching for them, a comprehensive analysis of all the structures in the CSD has been carried out and a MOF subset identified. This subset, now containing over 100,000 entries, can be downloaded as a stand-alone database and searched using elements of the CSD software suite (Moghadam et al., 2017). This resource has been shown to be very helpful in enabling the efficient exploration of MOFs in the CSD (Li et al., 2020) which has led to computational advances in the prediction of the properties of MOFs (Moghadam et al., 2020; Sarkisov et al., 2020). Software is also available for identifying the network topology of MOFs (see Section 2.2).

Nevertheless, identifying network structures with particular features remains challenging. Furthermore, searches within the databases are focussed on molecular components and it is not yet possible to search a database for generic structural types. For example, we are unaware of a search program that would allow us to find, with good response time, all MOFs with linkers of the form -OC(=O)-R-C(=O)-O- where R is any organic moiety and the length of the linker must lie in a specified distance range. Nor is it possible to constrain substructure searches so that they only find hits with a particular topology, e.g. [3]catenanes, or to search for molecules with a knot crossing involving two specified substructures, or to perform highly generic searches, e.g. any molecule with a knot. Finally, there are many molecules that are so huge that it is effectively impossible in a sketching tool to draw substructures that will find those molecules without large numbers of extraneous hits. The aim of this *Opinion* is to discuss how these types of searches might be provided and how they should be presented to users; in other words, what a new generation of substructure searching programs might look like.

2. Algorithmic Considerations

2.1 Knots and Links

In mathematical knot theory (Adams, 2004) a knot can only occur in a closed loop, otherwise it could be undone. In dealing with chemical structures, of course, a knot in an acyclic chain could be made to satisfy this requirement by a virtual bond connecting the appropriate terminal atoms. The starting point for knot identification is a 2D projection of the knot showing at each crossing point which part

of the strand is on top. Examples are given in Fig. 3, which shows all the different types of prime knots[2] with ≤ 8 crossings.[3] They are labelled using the Alexander-Briggs notation [later extended by Rolfsen (Rolfsen, 1976; Scharein, 1998)] which simply organises the knots by their crossing number. This number is followed by a subscripted index, the sole function of which is to differentiate between different knots with the same crossing number. The "unknot" – no knot at all – is denoted 0_1. The trefoil knot is 3_1.

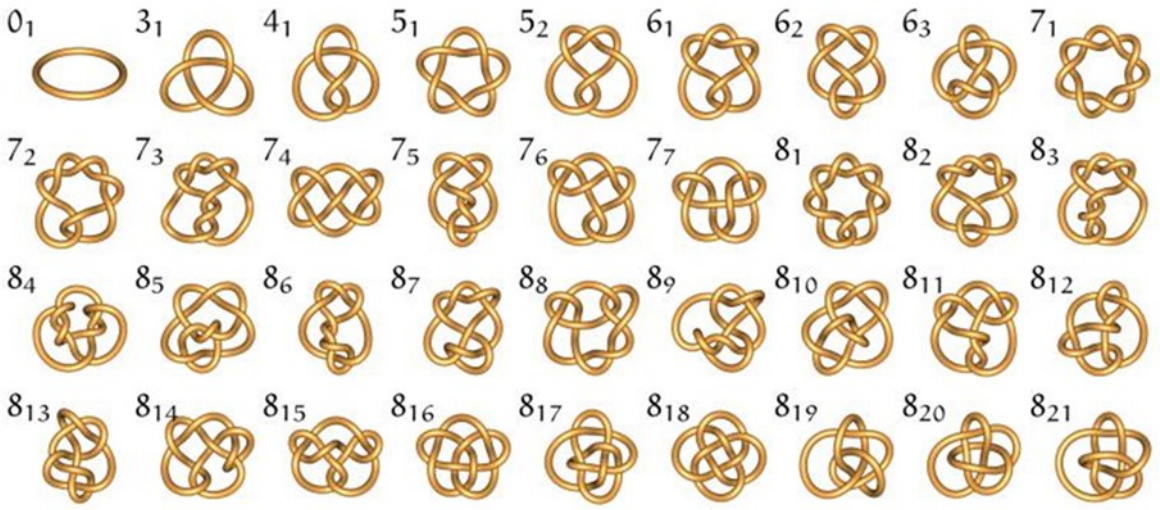


Figure 3. Prime knots with ≤ 8 crossings.

Unfortunately, a 2D projection can contain crossings that are not part of the knot, but due to geometrical factors that can be removed (in chemistry-speak, by changing the conformation). They are called nugatory crossings[4] (Hoste, Thistlethwaite and Weeks, 1998) and will be familiar to everyone; we have all had to untangle an electric lead that at first sight looked as if it was knotted but turned out not to be. In mathematics, the three types of adjustments that can be made to knot diagrams to remove nugatory crossings, thereby producing the minimal knot diagram, are called Reidemeister moves[5] (Reidemeister, 1927). They were used by Leigh et al. (Leigh, Lemonnier and Woltering, 2018) to demonstrate that a molecule thought to have a knot with 16 crossings in fact had only 8. Unfortunately, the number of Reidemeister moves to remove all nugatory crossings can sometimes be very large, the pathway may involve temporary increases in the number of crossings, and there may be no clear indication of when the minimal knot diagram is reached.

Hence the importance of knot polynomials that are invariant to Reidemeister moves (Adams, 2004). These are expressions derived from a knot diagram that will be the same whether or not the diagram is minimal. The Jones polynomial (Jones, 1997) can distinguish between any pair of knots provided the number of crossings in their minimal diagrams is ≤ 9. Above that its discriminatory ability is not guaranteed. It can also distinguish between the enantiomers of chiral knots in some cases.

---

[2] Knots, like integers, are prime or composite.

[3] This and Fig. 4 were taken with permission from the website of the very useful knot-drawing program Knotplot (Scharein, 1998) and its accompanying database of more than 3,000 knots and links [see also (Rawdon, Millett and Stasiak, 2015)].

[4] Nugatory crossings are also called reducible crossings or removable crossings. These can be removed by simply twisting the knot and are not a requirement for the definition of a given knot (Weisstein, 2013)

[5] A Reidemeister move is the conformational change (twisting) of the knot that is required to remove a nugatory crossing. Diagrams showing Reidemeister moves can be found in Reidemeister's original publication (Reidemeister, 1927).

A comparison of different knot polynomials is included in a description of the Python package Topoly, which is designed to find and categorise self-entangled proteins (Dabrowski-Tumanski et al., 2020). The authors have programmed several lesser-known polynomials in addition to the standard ones. Their software can be used to identify and categorise knots and other motifs such as slip knots and lassos. Perego and Potestio (Perego and Potestio, 2019) have reviewed other relevant contributions by the protein community. One is knot localization, i.e., working out where knots are in the polymer chain. Another is closure. This is the linking of the ends of a protein chain (or that part of it containing a knot) to form a closed loop, thereby satisfying the mathematical requirement for a knot. The link must not cross any of the existing features in a 2D projection as this might alter the topology. An alternative approach is to use knotoid theory (Turaev, 2012), which has been developed for analysing "knots" in open chains.

Knot theory and invariants can also be applied to links (catenanes). Link notation is similar to Alexander-Briggs knot notation but with an additional superscript to denote the number of components. The Hopf link, corresponding to a [2]catenane, is 2^2_1. Example links are in Fig. 4. Only a small number of these have been engineered into molecules as yet, but it is clear from a recent, fascinating review that the field is progressing rapidly and increasingly complex molecular links will be made (Gao et al., 2020). The well-known program ToposPro (Blatov, Shevchenko and Proserpio, 2014) (https://topospro.com) is capable of recognising some of these links, including the Hopf link (Fig. 4, extreme left) and the interesting Borromean link (Fig. 4, extreme right). The latter is the simplest of the Brunnian links, where the removal of one component leaves the remainder unlinked. ToposPro is also capable of identifying other interesting molecular topologies such as the single-twist Möbius strip.

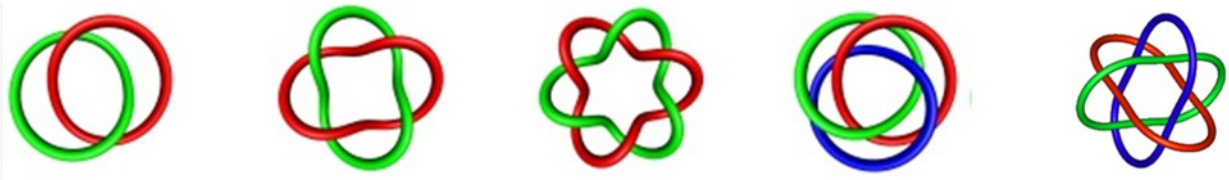


Figure 4. Example link topologies.

We note that many other molecular topologies have not been mentioned in this brief discussion - ravels (knots in which >2 strands can meet at a point), barrels, rotaxanes, braids, braided rotaxanes, weaves, etc.

2.2 MOFs and Other Framework Structures

A considerable amount of work has been done on computational approaches to MOF analysis. Databases of MOF geometries and topologies have been established (Alexandrov, Shevchenko and Blatov, 2019). One of the most useful resources to help with the assignment of the structure of the frameworks and nets is the Reticular Chemistry Structure Resource (RCSR) database (O'Keeffe et al., 2008). It contains many types of net and the symbols used to describe them. The database remains an active resource (http://rcsr.anu.edu.au) and can be used to identify the type of net observed in specific framework structures. While this resource has not been uniformly accepted, it has been generally supported by the structural chemistry community (Batten et al., 2013) and the concept has been developed further in recent years (Ohrstrom, 2015; Barthel et al., 2018). Attempts towards the assignment of MOF identifiers using automated cheminformatics algorithms have also been made recently (Bucior et al., 2019). Another highly valuable tool in the analysis of network structures is ToposPro (Blatov, Shevchenko and Proserpio, 2014). It will undertake a wide range of topological analyses, and in the context of the current discussion, allows for Crystal Information Files (CIFs) containing potential MOF structures to be simply read into the program suite and the parameters of

the network identified. The program can also be used to classify intermolecular hydrogen-bonding networks and analyse entangled networks (i.e., threading or catenation of different nets).

There are some standardisation issues (Bonneau et al., 2018). Four different naming systems exist for network topologies (although one is specific to zeolites) and unification seems desirable. A more difficult matter is the standardisation of node assignments. Networks are typically defined by nodes that represent the positions of only a small proportion of the atoms in a network. Usually, nodes will be assigned to metal atoms and representative atoms of bridging ligands, but there is no universal standard. Indeed, a rigid standard may be counterproductive. The ultimate purpose of analysing network topologies is to give insight into the nature of structures, and the best node assignment for achieving this may not always be the same.

2.3 Search Speed

Searching for structures in a chemical or crystallographic database should ideally be fast enough to be interactive. The principal reason is that one search very often suggests another, either because unwanted hits are found - indicating that the query needs to be modified - or because interesting hits spark off new ideas for searches. Substructure searching involves matching a subgraph onto a graph. This is an NP-complete problem, meaning that time requirements rise exponentially with the size of the system. Extensive work has been done to find good algorithms (Ehrlich and Rarey, 2012) including casting the problem as an SQL query (Golovin and Henrick, 2009). Nevertheless, search times are made acceptable only by pre-screening of database entries using bit strings that code for the presence or absence of substructural features (Leach and Gillet, 2007). Usually, this vastly reduces the number of subgraph-graph matches that have to be performed. The bit strings may be inverted, leading to a new set of bit strings, each of which identifies the entries containing a particular substructural feature (Agrafiotis et al., 2011). Substructure searching is still an active field and probably must remain so to deal with increasingly large databases containing highly complex molecules.

The problem becomes much more difficult when substructures include nonbonded contacts: for example, a search for extended hydrogen-bonding motifs involving specified molecular substructures. Chisholm and Motherwell (Chisholm and Motherwell, 2004) devised an algorithm based on a combination of substructure searching and searching for nonbonded contacts between substructures, organised as a depth-first search with backtracking. It is the basis of the motif-searching functionality in the program Mercury (Macrae et al., 2008) but searches are relatively slow compared with those for queries without nonbonded contacts. Searching for hydrogen-bonding motifs could be speeded up by using bit strings to screen on the presence or absence of hydrogen bonds between specific functional groups, or by using fingerprints derived from the connectivity of extended hydrogen-bonding networks. However, this will require defining in advance what is and is not counted as a hydrogen bond, an issue on which users may disagree. While the IUPAC definition of a hydrogen bond ("*The hydrogen bond is an attractive interaction between a hydrogen atom from a molecule or a molecular fragment X–H in which X is more electronegative than H, and an atom or a group of atoms in the same or a different molecule, in which there is evidence of bond formation*") may be helpful it remains a controversial issue and, in a searching context, users might reasonably wish to tailor the definition to the problem in hand. Users may also wish to search for networks involving other nonbonded interactions such as halogen bonds and aromatic stacking. A more elaborate screening system would therefore be based on bit strings that capture the shortest two or three contacts (after correction for van der Waals radii) formed by all atoms in a structure, together with the binned distances of those contacts.

Searching for molecules with particular topological features – e.g., a 6_1 knot or Borromean rings - will clearly require the topologies of database entries to be evaluated and stored in the database. For very large, knotted molecules there are likely to be many nugatory crossings in the

knot diagram, which can make knot invariant calculation extremely demanding; this is known to be a problem for proteins (Perego and Potestio, 2019). It also seems to us that generating suitable knot diagrams may sometimes be difficult for very complex molecules containing many rings and peripheral groups irrelevant to the knot.

A search for MOFs that have linkers with lengths falling in a specified distance range and of a particular generic form (e.g., -OC(=O)-R-C(=O)-O-) could be speeded up by screening with bit strings in which each bit signifies the presence or absence of a metal-metal linker whose length falls in a particular distance bin and whose terminal atoms belong to particular functional groups. The use of reduced graphs, in which groups of atoms are represented by single nodes, would facilitate searching for generic frameworks. It would also reduce subgraph-graph matching times. A hierarchy of reduced graphs for any given MOF may be useful because, as we noted earlier, there is no universal way of defining the nodes and struts of a MOF.

3. User-Interface Considerations

Presentation of advanced substructure searching to the user will not be easy for two reasons: (a) substructures for finding huge molecules can be extremely difficult to draw; (b) topology also needs to be defined. Some essential features for a user interface follow from this.

Firstly, the drawing of some substructure queries will require a 3D viewer. As evidence, we note that the CSD has many entries that have no two-dimensional diagrams because they would be incomprehensible. Drawing facilities such as the ability to place atoms or substructures on template objects such as cubes and octahedra will be helpful. 2D to 3D converters are useful as they allow substructures to be typed as SMILES or SMARTS strings and then converted to 3D. The ability to copy and paste substructures will be necessary as so many large synthetic molecules are oligomeric. For the same reason, drawing queries as reduced graphs, where each point represents a user-defined multi-atom substructure, is likely to ease complex query definition considerably. It is similar to using single letters to represent amino acids in proteins, though more complicated because amino acids have only two points of attachment and the overall chain is linear. Mapping from reduced graphs to complete graphs will need information not only about the underlying substructure that each node represents but also the atoms involved in each node-node bond. Nevertheless, it seems to us a very promising avenue to explore, breaking a complex drawing problem into a set of smaller ones: the reduced graph, and the substructure that each different type of node represents. Of course, large, well-indexed and user-extensible libraries of both 2D and 3D substructure templates will also aid creation of complex queries.

There will still remain substructures that cannot be drawn in any reasonable amount of time. For example, imagine trying to construct a query to find the sulfido-silver cluster BIVMUP, with the formula $C_{578}H_{868}Ag_{320}P_{24}S_{190}$, shown in Fig. 5 (Anson et al., 2008). Even when rotated in 3D it is hard to understand this structure. A compromise query might be useful; for example, a search for molecules containing all of the different sulfur and silver coordination geometries in this cluster and with the total atom count of the cluster in a user-defined range.

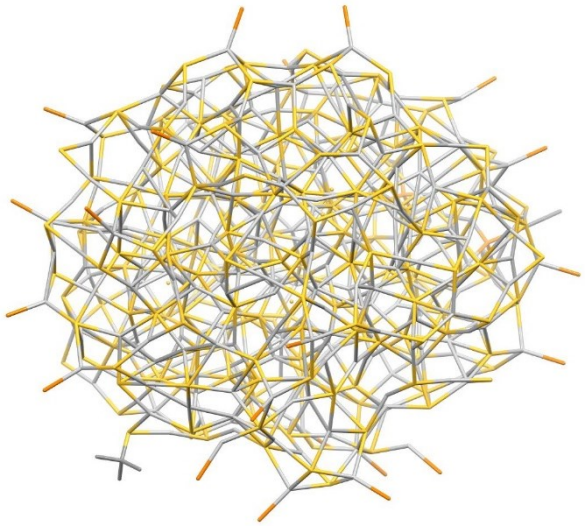


Figure 5. The molecular structure of the sulfido-silver cluster complex BIVMUP (Anson et al., 2008)

Editing existing database entries to create queries is a well-known and highly useful technique. It will be necessary, however, to assist users to understand very complex structures. In the above example, automatic identification and display of the different sulfur and silver geometries would save a lot of time. Cartoon displays of complex molecules can make the underlying topologies much easier to see and would be very useful if the user's objective were to find other molecules with the same topology. It was pointed out earlier that misidentification of knots is easily possible so an algorithm for assigning and reporting knot types on request is needed. This and other topology-assignment algorithms are essential anyway so that databases can be pre-processed and topologies of entries stored.

Given such a database, users would be able to search for topologies as well as, or even instead of, searching for the presence of chemical substructures. It seems to us necessary to separate the definitions of chemical substructure(s) and topology. There are three obvious ways in which the desired topology could be specified by the user: (a) use of standard notations such as Alexander-Briggs or RSCR net notation; (b) selection from a set of templates such as those shown in Figs. 3 and 4; (c) use of an existing database entry with the desired topology. Alternatively, a topology definition could be constructed by the user from simple components, e.g., a [3]rotaxane from three circles and a line. (That, of course, would require the underlying software to detect the topology from the drawing.) Links between the chemical substructures and the topology may be desired. For example, three macrocycle substructures might be drawn and a [3]catenane topology selected, but the user might then wish to indicate which of the macrocycles is to be in the central position. As another example, the user might select a given type of knot topology and then wish to indicate whereabouts in the substructure the knot should be located.

4. Conclusions

Our purpose in writing this paper was to draw attention to a growing problem. Substructure searching programs, which are essential information tools for chemists and crystallographers, are becoming unable to deal effectively with a significant proportion of new molecules and supramolecules. Further advances in synthetic chemistry and increasing interest in advanced materials will only make this problem more pressing. Fortunately, a great deal of relevant theoretical analysis and algorithm and program development has been done, providing a foundation for the development of a new generation of substructure searching software.

We have focussed exclusively on substructure searching but acknowledge that there are many other existing and potential types of searches. These include similarity searching, pharmacophore searching, searching for voids, shape searching, and searching for molecules with particular properties or uses, e.g., gas separation, molecular motors. As molecules become larger and their uses more sophisticated, these types of searches will become increasingly necessary. Nevertheless, substructure searching will remain the most fundamental of chemical information tools because it retrieves molecules and intermolecular networks based on chemical connectivity, which is of central importance in chemistry.

We must stress the realities of software development, especially when the problem domain is research oriented. User requirements can almost never be fully understood up-front, either by software developers or by potential users. Even if they were, they would change. Unforeseen implementation problems are to be expected. There will doubtless remain many difficult challenges, not least in the areas of topology[4] and searching for extended networks involving nonbonded interactions. User interfaces must be designed with particular care. Despite all this, a significant enhancement of substructure-search programs is both achievable and highly desirable.

Acknowledgments

We are grateful to Rob Scharein, the author of KnotPlot, for allowing us to reproduce plots from the website devoted to that excellent program. We thank Professor Andrew Burrows, Dr Dan Pantos and Dr Jason Cole for valuable discussions during the preparation of this article. RT and PRR thank the Cambridge Crystallographic Data Centre for Emeritus Research Fellowships.

References

Adams, C.C., 2004. *The Knot Book. An Elementary Introduction to the Mathematical Theory of Knots.* Providence: American Mathematical Society.

Agrafiotis, D.K., Lobanov, V.S., Shemanarev, M., Rassokhin, D.N., Izrailev, S., Jaeger, E.P., Alex, S. and Farnum, M., 2011. Efficient Substructure Searching of Large Chemical Libraries: The ABCD Chemical Cartridge. *Journal of Chemical Information and Modeling,* 51(12), pp. 3113-3130.

Alexandrov, E.V., Shevchenko, A.P. and Blatov, V.A., 2019. Topological Databases: Why Do We Need Them for Design of Coordination Polymers? *Crystal Growth & Design,* 19(5), pp. 2604-2614.

Anson, C.E., Eichhöfer, A., Issac, I., Fenske, D., Fuhr, O., Sevillano, P., Persau, C., Stalke, D. and Zhang, J., 2008. Synthesis and Crystal Structures of the Ligand-Stabilized Silver Chalcogenide Clusters [Ag154Se77(dppxy)18], [Ag320(StBu)60S130(dppp)12], [Ag352S128(StC5H11)96], and [Ag490S188(StC5H11)114]. *Angewandte Chemie International Edition,* 47(7), pp. 1326-1331.

Barthel, S., Alexandrov, E.V., Proserpio, D.M. and Smit, B., 2018. Distinguishing Metal–Organic Frameworks. *Crystal Growth & Design,* 18(3), pp. 1738-1747.

Batten, S.R., Champness, N.R., Chen, X.M., Garcia-Martinez, J., Kitagawa, S., Ohrstrom, L., O'Keeffe, M., Suh, M.P. and Reedijk, J., 2013. Terminology of metal-organic frameworks and coordination polymers (IUPAC Recommendations 2013). *Pure and Applied Chemistry,* 85(8), pp. 1715-1724.

Batten, S.R. and Robson, R., 1998. Interpenetrating nets: Ordered, periodic entanglement. *Angewandte Chemie-International Edition,* 37(11), pp. 1460-1494.

Blatov, V.A., Shevchenko, A.P. and Proserpio, D.M., 2014. Applied Topological Analysis of Crystal Structures with the Program Package ToposPro. *Crystal Growth & Design,* 14(7), pp. 3576-3586.

[4] For example, shapes that are topologically equivalent from a mathematical viewpoint but are inequivalent in chemistry because of the constraints of chemical bonding.

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. and Ohrstrom, L., 2018. Deconstruction of Crystalline Networks into Underlying Nets: Relevance for Terminology Guidelines and Crystallographic Databases. *Crystal Growth & Design,* 18(6), pp. 3411-3418.
Bravo, J.A., Raymo, F.M., Stoddart, J.F., White, A.J.P. and Williams, D.J., 1998. High Yielding Template-Directed Syntheses of [2]Rotaxanes. *European Journal of Organic Chemistry,* 1998(11), pp. 2565-2571.
Bucior, B.J., Rosen, A.S., Haranczyk, M., Yao, Z., Ziebel, M.E., Farha, O.K., Hupp, J.T., Siepmann, J.I., Aspuru-Guzik, A. and Snurr, R.Q., 2019. Identification Schemes for Metal–Organic Frameworks To Enable Rapid Search and Cheminformatics Analysis. *Crystal Growth & Design,* 19(11), pp. 6682-6697.
Chisholm, J.A. and Motherwell, S., 2004. A new algorithm for performing three-dimensional searches of the Cambridge Structural Database. *Journal of Applied Crystallography,* 37, pp. 331-334.
Cram, D.J., 1988. The Design of Molecular Hosts, Guests, and Their Complexes (Nobel Lecture). *Angewandte Chemie International Edition in English,* 27(8), pp. 1009-1020.
Dabrowski-Tumanski, P., Rubach, P., Niemyska, W., Gren, B.A. and Sulkowska, J.I., 2020. Topoly: Python package to analyze topology of polymers. *Briefings in Bioinformatics*.
Dai, D., Liu, Q., Hu, R., Wei, X., Ding, G., Xu, B., Xu, T., Zhang, J., Xu, Y. and Zhang, H., 2020. Method construction of structure-property relationships from data by machine learning assisted mining for materials design applications. *Materials & Design,* 196, p. 109194.
Danon, J.J., Kruger, A., Leigh, D.A., Lemonnier, J.F., Stephens, A.J., Vitorica-Yrezabal, I.J. and Woltering, S.L., 2017. Braiding a molecular knot with eight crossings. *Science,* 355(6321), pp. 159-+.
Ding, S.Y. and Wang, W., 2013. Covalent organic frameworks (COFs): from design to applications. *Chemical Society Reviews,* 42(2), pp. 548-568.
Ehrlich, H.C. and Rarey, M., 2012. Systematic benchmark of substructure search in molecular graphs - From Ullmann to VF2. *Journal of Cheminformatics,* 4.
Feringa, B.L., 2017. The Art of Building Small: From Molecular Switches to Motors (Nobel Lecture). *Angewandte Chemie-International Edition,* 56(37), pp. 11059-11078.
Fielden, S.D.P., Leigh, D.A. and Woltering, S.L., 2017. Molecular Knots. *Angewandte Chemie-International Edition,* 56(37), pp. 11166-11194.
Furukawa, H., Cordova, K.E., O'Keeffe, M. and Yaghi, O.M., 2013. The Chemistry and Applications of Metal-Organic Frameworks. *Science,* 341(6149), pp. 974-+.
Gao, W.-X., Feng, H.-J., Guo, B.-B., Lu, Y. and Jin, G.-X., 2020. Coordination-Directed Construction of Molecular Links. *Chemical Reviews,* 120(13), pp. 6288-6325.
Golovin, A. and Henrick, K., 2009. Chemical Substructure Search in SQL. *Journal of Chemical Information and Modeling,* 49(1), pp. 22-27.
Groom, C.R., Bruno, I.J., Lightfoot, M.P. and Ward, S.C., 2016. The Cambridge Structural Database. *Acta Crystallographica Section B-Structural Science Crystal Engineering and Materials,* 72, pp. 171-179.
Hamilton, D.G., Feeder, N., Prodi, L., Teat, S.J., Clegg, W. and Sanders, J.K.M., 1998. Tandem Hetero-Catenation: Templating and Self-Assembly in the Mutual Closure of Two Different Interlocking Rings. *Journal of the American Chemical Society,* 120(5), pp. 1096-1097.
Hasenknopf, B., Lehn, J.-M., Kneisel, B.O., Baum, G. and Fenske, D., 1996. Self-Assembly of a Circular Double Helicate. *Angewandte Chemie International Edition in English,* 35(16), pp. 1838-1840.
Helliwell, J.R., 2017. New developments in crystallography: exploring its technology, methods and scope in the molecular biosciences. *Bioscience reports,* 37(4), p. BSR20170204.
Hoskins, B.F. and Robson, R., 1990. DESIGN AND CONSTRUCTION OF A NEW CLASS OF SCAFFOLDING-LIKE MATERIALS COMPRISING INFINITE POLYMERIC FRAMEWORKS OF 3-D-LINKED MOLECULAR RODS - A REAPPRAISAL OF THE ZN(CN)2 AND CD(CN)2 STRUCTURES AND THE SYNTHESIS AND STRUCTURE OF THE DIAMOND-RELATED FRAMEWORKS N(CH3)4 CUIZNII(CN)4 AND CUI 4,4',4'',4'''-TETRACYANOTETRAPHENYLMETHANE BF4.XC6H5NO2. *Journal of the American Chemical Society,* 112(4), pp. 1546-1554.

Hoste, J., Thistlethwaite, M. and Weeks, J., 1998. The first 1,701,936 knots. *Mathematical Intelligencer,* 20(4), pp. 33-48.
Jones, V.F.R., 1997. A Polynomial Invariant for Knots via non Neumann Algebras. In: M.A.I.D. Atiyah, ed. *Fields Medallists' Lectures 1997.* Singapore: World Scientific, pp. 488-458.
Leach, A.R. and Gillet, V.J., 2007. *An Introduction to Chemoinformatics.* Netherlands: Springer.
Lee, J.S.M. and Cooper, A.I., 2020. Advances in Conjugated Microporous Polymers. *Chemical Reviews,* 120(4), pp. 2171-2214.
Lehn, J.M., 1988. SUPRAMOLECULAR CHEMISTRY - SCOPE AND PERSPECTIVES MOLECULES, SUPERMOLECULES, AND MOLECULAR DEVICES. *Angewandte Chemie-International Edition,* 27(1), pp. 89-112.
Leigh, D.A., Lemonnier, J.F. and Woltering, S.L., 2018. Comment on "Coordination-Driven Self-Assembly of a Molecular Knot Comprising Sixteen Crossings". *Angewandte Chemie-International Edition,* 57(38), pp. 12212-12214.
Li, A., Bueno-Perez, R., Wiggin, S. and Fairen-Jimenez, D., 2020. Enabling efficient exploration of metal–organic frameworks in the Cambridge Structural Database. *CrystEngComm,* 22(43), pp. 7152-7161.
Macrae, C.F., Bruno, I.J., Chisholm, J.A., Edgington, P.R., McCabe, P., Pidcock, E., Rodriguez-Monge, L., Taylor, R., van de Streek, J. and Wood, P.A., 2008. Mercury CSD 2.0 - new features for the visualization and investigation of crystal structures. *Journal of Applied Crystallography,* 41, pp. 466-470.
Miljanić, O.Š., Dichtel, W.R., Khan, S.I., Mortezaei, S., Heath, J.R. and Stoddart, J.F., 2007. Structural and Co-conformational Effects of Alkyne-Derived Subunits in Charged Donor−Acceptor [2]Catenanes. *Journal of the American Chemical Society,* 129(26), pp. 8236-8246.
Moghadam, P.Z., Li, A., Liu, X.W., Bueno-Perez, R., Wang, S.D., Wiggin, S.B., Wood, P.A. and Fairen-Jimenez, D., 2020. Targeted classification of metal-organic frameworks in the Cambridge structural database (CSD). *Chemical Science,* 11(32), pp. 8373-8387.
Moghadam, P.Z., Li, A., Wiggin, S.B., Tao, A., Maloney, A.G.P., Wood, P.A., Ward, S.C. and Fairen-Jimenez, D., 2017. Development of a Cambridge Structural Database Subset: A Collection of Metal-Organic Frameworks for Past, Present, and Future. *Chemistry of Materials,* 29(7), pp. 2618-2625.
O'Keeffe, M., Peskov, M.A., Ramsden, S.J. and Yaghi, O.M., 2008. The Reticular Chemistry Structure Resource (RCSR) Database of, and Symbols for, Crystal Nets. *Accounts of Chemical Research,* 41(12), pp. 1782-1789.
Ohrstrom, L., 2015. Let's Talk about MOFs-Topology and Terminology of Metal-Organic Frameworks and Why We Need Them. *Crystals,* 5(1), pp. 154-162.
Pedersen, C.J., 1988. THE DISCOVERY OF CROWN ETHERS (NOBEL LECTURE). *Angewandte Chemie-International Edition,* 27(8), pp. 1021-1027.
Percastegui, E.G., Ronson, T.K. and Nitschke, J.R., 2020. Design and Applications of Water-Soluble Coordination Cages. *Chemical Reviews,* 120(24), pp. 13480-13544.
Perego, C. and Potestio, R., 2019. Computational methods in the study of self-entangled proteins: a critical appraisal. *Journal of Physics-Condensed Matter,* 31(44).
Pimentel, B.R., Parulkar, A., Zhou, E.K., Brunelli, N.A. and Lively, R.P., 2014. Zeolitic Imidazolate Frameworks: Next-Generation Materials for Energy-Efficient Gas Separations. *Chemsuschem,* 7(12), pp. 3202-3240.
Rawdon, E.J., Millett, K.C. and Stasiak, A., 2015. Subknots in ideal knots, random knots and knotted proteins. *Scientific Reports,* 5(1), p. 8928.
Reidemeister, K., 1927. *Abh. Math. Semin. Univ. Hamburg,* 5, pp. 24-32.
Rolfsen, D., 1976. *Knots and Links.* USA: AMS Chelsea Publishing.
Rosi, N.L., Eckert, J., Eddaoudi, M., Vodak, D.T., Kim, J., O'Keeffe, M. and Yaghi, O.M., 2003. Hydrogen Storage in Microporous Metal-Organic Frameworks. *Science,* 300, p. 1127.

Roy, B., Reddy, M.C. and Hazra, P., 2018. Developing the structure–property relationship to design solid state multi-stimuli responsive materials and their potential applications in different fields. *Chemical Science,* 9(14), pp. 3592-3606.
Sarkisov, L., Bueno-Perez, R., Sutharson, M. and Fairen-Jimenez, D., 2020. Materials Informatics with PoreBlazer v4.0 and the CSD MOF Database. *Chemistry of Materials,* 32(23), pp. 9849-9867.
Sauvage, J.P., 2017. From Chemical Topology to Molecular Machines (Nobel Lecture). *Angewandte Chemie-International Edition,* 56(37), pp. 11080-11093.
Sawada, T., Inomata, Y., Shimokawa, K. and Fujita, M., 2019. A metal–peptide capsule by multiple ring threading. *Nature Communications,* 10(1), p. 5687.
Scharein, R.G., 1998. *Knotplot* [Online]. Available from: *https://knotplot.com/* [Accessed.
Sluysmans, D. and Stoddart, J.F., 2019. The Burgeoning of Mechanically Interlocked Molecules in Chemistry. *Trends in Chemistry,* 1(2), pp. 185-197.
Stoddart, J.F., 2017. Mechanically Interlocked Molecules (MIMs)-Molecular Shuttles, Switches, and Machines (Nobel Lecture). *Angewandte Chemie-International Edition,* 56(37), pp. 11094-11125.
Turaev, V., 2012. KNOTOIDS. *Osaka Journal of Mathematics,* 49(1), pp. 195-223.
van Dongen, S.F.M., Cantekin, S., Elemans, J., Rowan, A.E. and Nolte, R.J.M., 2014. Functional interlocked systems. *Chemical Society Reviews,* 43(1), pp. 99-122.
Ward, M.D. and Raithby, P.R., 2013. Functional behaviour from controlled self-assembly: challenges and prospects. *Chemical Society Reviews,* 42(4), pp. 1619-1636.
Weisstein, E., 2013. *"Reduced Knot Diagram." From MathWorld--A Wolfram Web Resource* [Online]. Available from: *https://mathworld.wolfram.com/ReducedKnotDiagram.html* [Accessed 22/07/2021].
Yan, Y., Lin, X., Yang, S.H., Blake, A.J., Dailly, A., Champness, N.R., Hubberstey, P. and Schroder, M., 2009. Exceptionally high H-2 storage by a metal-organic polyhedral framework. *Chemical Communications,* (9), pp. 1025-1027.
Zhang, D.W., Ronson, T.K. and Nitschke, J.R., 2018. Functional Capsules via Subcomponent Self-Assembly. *Accounts of Chemical Research,* 51(10), pp. 2423-2436.