

## PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a publisher's version.

For additional information about this publication click this link.

<http://hdl.handle.net/2066/204870>

Please be advised that this information was generated on 2019-12-31 and may be subject to change.

# Cocrystals in the Cambridge Structural Database: a network approach

Jan-Joris Devogelaer, Hugo Meekes, Elias Vlieg and René de Gelder\*

Radboud University, Heyendaalseweg 135, 6525AJ Nijmegen, The Netherlands. \*Correspondence e-mail: r.degelder@science.ru.nl

Received 23 January 2019

Accepted 5 April 2019

Edited by A. Nangia, CSIR–National Chemical Laboratory, India

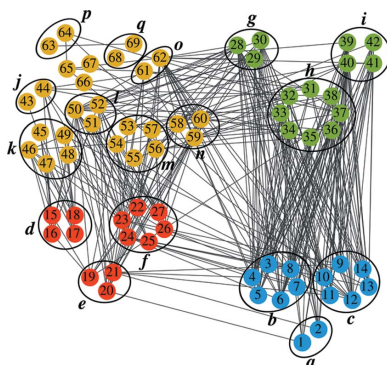
**Keywords:** cocrystallization; networks; data mining; Cambridge Structural Database; cocrystal prediction; knowledge-based approach.

To obtain a better understanding of which coformers to combine for the successful formation of a cocrystal, techniques from data mining and network science are used to analyze the data contained in the Cambridge Structural Database (CSD). A network of coformers is constructed based on cocrystal entries present in the CSD and its properties are analyzed. From this network, clusters of coformers with a similar tendency to form cocrystals are extracted. The popularity of the coformers in the CSD is unevenly distributed: a small group of coformers is responsible for most of the cocrystals, hence resulting in an inherently biased data set. The coformers in the network are found to behave primarily in a bipartite manner, demonstrating the importance of combining complementary coformers for successful cocrystallization. Based on our analysis, it is demonstrated that the CSD coformer network is a promising source of information for knowledge-based cocrystal prediction.

## 1. Introduction

The opportunity to alter several physico-chemical properties of high-value chemicals, such as pharmaceuticals (Berry & Steed, 2017) and agrochemicals (Nauha & Nissinen, 2011), without changing their molecular structure and function, has promoted the use of multi-component crystals (or systems) as a formulation tool. Multi-component systems, such as salts, solvates and cocrystals, are crystalline aggregates containing multiple ionic and/or neutral species in the crystal lattice (Grothe *et al.*, 2016). For a molecule of interest, a variety of multi-component solid forms can be prepared, each characterized by a distinct set of properties including solubility, bioavailability, hydration stability, and mechanical, optical and thermal properties. Additionally, the crystallization behavior of chiral molecules is influenced when using multi-component systems, possibly resulting in the formation of chiral conglomerates (*i.e.* a physical mixture of separate enantiomer crystals), enabling their efficient separation using crystallization-based techniques (Lorenz & Seidel-Morgenstern, 2014).

Having knowledge of the solid-state landscape of the molecule, not only in terms of polymorphism but also in terms of the available multi-component forms, is therefore crucial during the design and optimization of the final product and its production route. The types of multi-component systems a molecule can form is strongly influenced by its molecular structure. For instance, the lack of ionizable functional groups generally precludes the molecule from forming salts, leaving only solvate formation or cocrystallization as feasible options. Yet, whereas the pairing of complementary ions for the



formation of salts is rather straightforward, the design of solvates, and in particular cocrystals, using weak (directional) non-covalent interactions remains challenging. Nevertheless, the number of additional components (or *coformers*) is much larger than the available solvents or counterions (Almarsson & Zaworotko, 2004), making cocrystallization an attractive formulation tool.

There are several strategies to design a new cocrystal. A well-known approach uses supramolecular synthons (Desiraju, 1995) (*i.e.* a variety of common intermolecular interactions) to rationalize the feasibility of cocrystal formation. In general, one aims to match complementary hydrogen bond motifs,  $\pi$ - $\pi$  interactions, ion- $\pi$  interactions, halogen bonds or even van der Waals interactions between the coformers to predict the formation of a cocrystal. A distinction is generally made between *homosynthons*, using self-complementary functional groups such as carboxylic acids or amides, and *heterosynthons*, where the moieties of different functional groups are combined (*e.g.* combining a carboxylic acid with an amide group). Although this strategy has been quite successful and conforms with general, chemical insights, the synthon-based approach is based on an *a posteriori* understanding of crystal structures and relies on isolated structural attributes. The method does not account for more complex factors beyond functional group matching, such as issues with packing, or experimental difficulties (*e.g.* difference in solubility). Additionally, a recent study (Taylor & Day, 2018) has demonstrated that just the presence of hydrogen and halogen bonds alone is not necessarily a good descriptor for successful cocrystallization, stressing the importance of including more subtle effects in the design process.

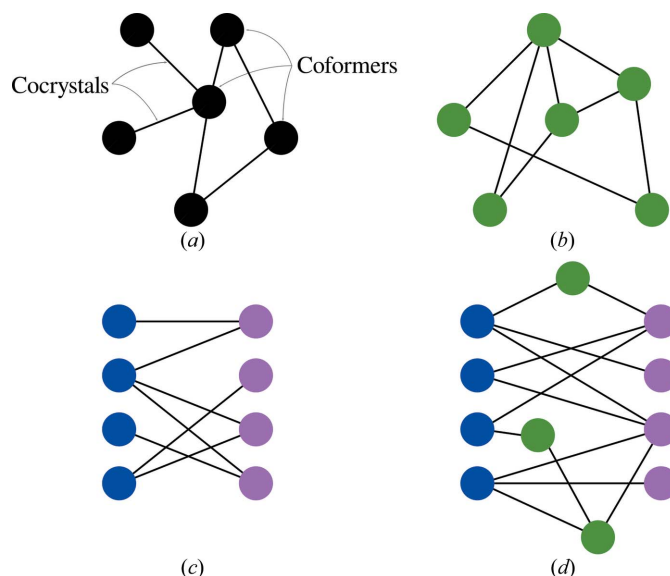
Because the experimental determination of cocrystals is time and labor intensive, various computational tools have been developed to understand and predict cocrystallization. These methods include the use of molecular modelling (Taylor & Day, 2018; Issa *et al.*, 2009; Karamertzanis *et al.*, 2009), the analysis and application of molecular descriptors (Fabian, 2009; Wicker *et al.*, 2017), the use of hydrogen bond propensity calculations (Delori *et al.*, 2013) and molecular electrostatic potential surfaces (Greco *et al.*, 2014). Again, a possible drawback of these tools is their focus on isolated molecular features and dependence on too general or simplified rules for cocrystallization.

A valuable addition to the set of tools would therefore be a more comprehensive (or *holistic*) method that looks beyond the isolated structural properties of coformers and implicitly includes the decisive but subtle factors for successful cocrystallization. In this article, we present a knowledge-based approach that attempts to do this by studying cocrystallization in the form of a network with the theoretical tools provided through *network science*. Network science is a growing field that has originated from graph theory and has found many applications in diverse research areas. By converting a complex problem into a network, a set of new characteristics of the system can be revealed that can improve the understanding and use of its underlying structure and dynamics.

The Cambridge Structural Database (CSD; Groom *et al.*, 2016) is the most extensive crystallographic database and currently contains about a million small molecule entries, including a large number of cocrystal structures. By identifying the relations between the coformers found in these cocrystals, a network can be constructed, which can then be analyzed. The goal of this network analysis is to provide a set of empirical, data-driven insights about cocrystallization that can later be applied in an enhanced design strategy.

## 2. Methods

A network is essentially a collection of *nodes* and *edges* (or connections) between these nodes. The binary cocrystals (*i.e.* containing two distinct coformers) from the Cambridge Structural Database (CSD; version 5.39, November 2017 + two updates) were used to build up the network, drawing them as the edges and their coformers as the nodes (as illustrated in Fig. 1*a*). By converting the database's cocrystal entries into a network, an enormous amount of relational information is deduced that is normally not accessible with the CSD's software [*e.g.* *ConQuest* (Bruno *et al.*, 2002), *Mercury* (Macrae *et al.*, 2008)]. The network was subsequently studied using a set of common network analysis techniques to acquire a better understanding of its structure. These tools, as described below, include clustering, analyzing the network's *degree* distribution and determining to which network type it belongs. The extraction of cocrystal data, construction of the network and further analyses were all performed with scripts written in Python (version 2.7.15) in conjunction with the CSD's Python API.



**Figure 1**  
(*a*) Example of a network, consisting of nodes (coformers) and edges (cocrystals). (*b*) A monopartite network, characterized by a single set of nodes, and edges between any of the nodes. (*c*) A bipartite network with two distinct sets of nodes, and edges only between these sets. (*d*) A mixture network, having the properties of both networks (*b*) and (*c*).

## 2.1. Construction of the network

The CSD was scanned for entries that contain two distinct chemical entities, are organic, non-ionic, error-free, and have their three-dimensional coordinates determined (including disordered structures). From these entries, the binary cocrystals were discriminated from solvates, or structures crystallized with a gas molecule, using a custom *classifier* algorithm (see Appendix A). The algorithm also removes erroneous entries<sup>1</sup> and effectively handles difficulties arising from chiral entries, adding cocrystals for only one representative enantiomer. The process resulted in a set of binary cocrystals, formed by a set of unique coformers.

The set of cocrystals was then transformed into an undirected, unweighted network  $G(N, E)$ , consisting of nodes  $N$  (coformers) and edges  $E$  (cocrystals). In fact, an *adjacency matrix*  $\mathbf{A} \in \mathbb{R}^{|N| \times |N|}$  is constructed, of which the row and column indices correspond to the nodes (coformers), and for which the elements are set to 1 for every known edge (cocrystal) between these nodes (Fig. 2). The adjacency matrix is a symmetric matrix that serves as the mathematical basis of the network and permits the study of its properties.

Our philosophy behind the construction of the network was to solely map the relations originating from cocrystals, hence without including polymorphism, stoichiometry, structural information or (physico-)chemical properties. Nevertheless, the resulting network is informative enough to study cocrystallization from a theoretical point of view: our results show that structural and chemical properties can be recovered using the correct tools from network science.

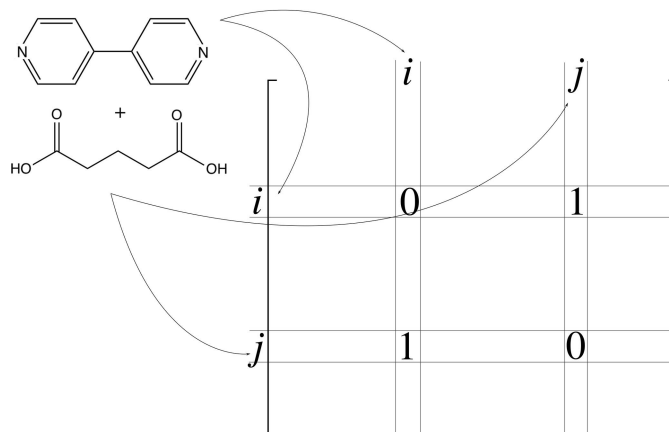
## 2.2. Clustering

The extent to which the structure of the network can reflect some of the generally accepted principles of cocrystallization was studied by clustering the coformers. Clusters are mutually exclusive groups of nodes that are related through some measure of topological similarity, and are expected to demonstrate a specific function within the network. In the case of coformers, it is envisaged that clusters will emerge that are responsible for different cocrystallization mechanisms (*e.g.* hydrogen bond acceptors). The proposed similarity, also known as the Jaccard similarity coefficient (Jaccard, 1912), between two coformers  $i$  and  $j$  is defined as:

$$s_{i,j} = \frac{|n_i \cap n_j|}{|n_i \cup n_j|}, \quad (1)$$

with  $n_i$  and  $n_j$  the sets of neighbors of coformers  $i$  and  $j$ , respectively. The neighbors of a coformer are defined as the set of all the coformers it forms cocrystals with, or mathematically  $n_i = \{j \in N | \mathbf{A}_{(i,j)} = 1\}$ , with  $\mathbf{A}$  the adjacency matrix and  $N$  the set of nodes of the network. The similarity measure

<sup>1</sup>For some entries, the three-dimensional data, and more specifically the connectivities between its atoms, is poorly determined. As a result, the distinct coformers cannot be extracted from the entry, and are therefore discarded from the data set. Also, an additional check of the coformer's neutrality is performed, removing any ionic molecules that may have been incorrectly added to the data set.



**Figure 2**

Addition of the cocrystal entry SOVDIQ to the adjacency matrix. The cocrystal is first split into its coformers (4,4'-bipyridine and glutaric acid), which are then labeled as  $i$  and  $j$ . Next, elements  $\mathbf{A}_{(i,j)}$  and  $\mathbf{A}_{(j,i)}$  of the adjacency matrix are set to 1 for the existing cocrystal. Conversely, coformer combinations for which no cocrystal is known, are set to 0.

in equation (1) is larger for combinations of coformers that have more neighbors in common, and punishes those that cocrystallize with more diverse partners. The similarity was calculated for each pair of coformers and stored in a coformer similarity matrix. This matrix is similar to the adjacency matrix, but instead of containing 0's or 1's, it contains the calculated similarities for each coformer combination ( $s_{i,j} \in [0, 1]$ ).

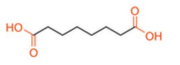
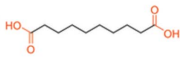



































































A smaller set of  $m$  popular coformers was then clustered using Ward's hierarchical clustering method (Ward, 1963) [as implemented in the SciPy library (Jones *et al.*, 2001)]. The coformer similarity matrix was first transformed into a dissimilarity or distance matrix, containing distances  $d_{i,j} = 1 - s_{i,j}$ . Next, the coformers were placed in  $m$  separate clusters or *singletons* and the cluster pair with the lowest distance is merged into a larger cluster, reducing the number of clusters to  $m - 1$ . The distance matrix was updated for the smaller set of clusters, where the distance to a joined cluster  $p$  is defined as:

$$d(p, q) = \left[ \frac{|q| + |s|}{|q| + |s| + |t|} d(q, s)^2 + \frac{|q| + |t|}{|q| + |s| + |t|} d(q, t)^2 - \frac{|q|}{|q| + |s| + |t|} d(s, t)^2 \right]^{1/2} \quad (2)$$

with  $p$  the cluster that is formed by joining clusters  $s$  and  $t$ , and  $q$  one of the remaining clusters. This agglomerative process was repeated, recording the distances at which clusters were merged, and was terminated when a single cluster, containing all the coformers, was obtained. In contrast to coformers, the distance between clusters can exceed a value of 1: for a remaining cluster  $q$  that is relatively dissimilar to clusters  $s$  and  $t$ , the first two terms under the square root in equation (2) can be large compared to the last term, resulting in a cluster distance larger than 1. Cluster merges at such a distance are, however, only expected in the final stages of the procedure, where rather distant clusters are eventually combined.

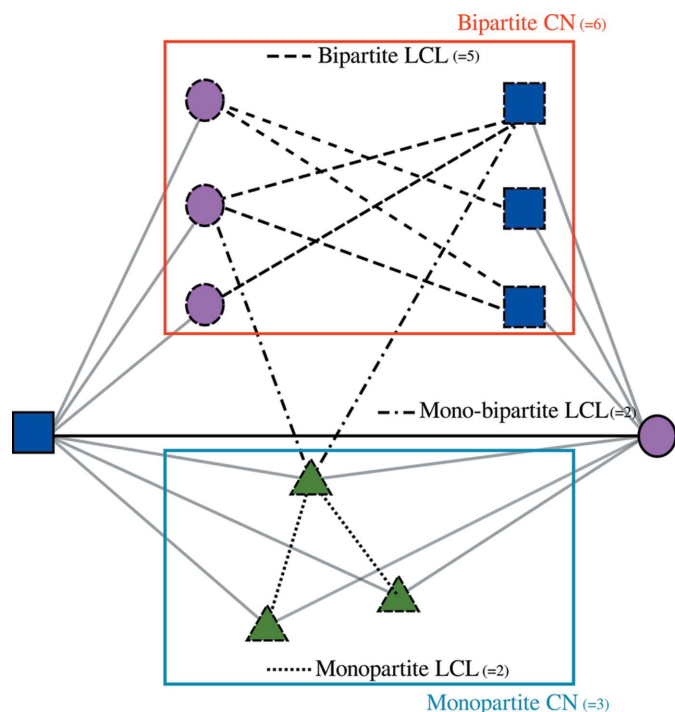
**Table 1**  
Summary of the clusters of coformers in Fig. 5.

The coformers are grouped per cluster and labeled by an index that corresponds to the endpoints in the dendrogram going from left to right (e.g. coformers **3** and **4** are the leftmost and second leftmost endpoints of cluster **b**). Additionally, the coformers are provided with their number of distinct, observed cocrystals in the CSD (i.e. coformer degree *k*) between parentheses.

cluster	a	b	c	d
	<p>1(34) </p> <p>2(41) </p>	<p>3(81)  6(51) </p> <p>4(105)  7(34) </p> <p>5(76)  8(46) </p>	<p>9(65)  12(60) </p> <p>10(54)  13(31) </p> <p>11(66)  14(44) </p>	<p>15(52)  17(36) </p> <p>16(60)  18(32) </p>
cluster	e	f	g	h
	<p>19(36) </p> <p>20(61) </p> <p>21(72) </p>	<p>22(47)  25(34) </p> <p>23(57)  26(34) </p> <p>24(73)  27(94) </p>	<p>28(288) </p> <p>29(167) </p> <p>30(107) </p>	<p>31(34)  35(37) </p> <p>32(32)  36(31) </p> <p>33(46)  37(31) </p> <p>34(53)  38(39) </p>
cluster	i	j	k	l
	<p>39(75)  41(70) </p> <p>40(112)  42(79) </p>	<p>43(42) </p> <p>44(152) </p>	<p>45(58)  47(74) </p> <p>46(64)  48(41) </p> <p>49(38) </p>	<p>50(177) </p> <p>51(34) </p> <p>52(40) </p>
cluster	m	n	o	p
	<p>53(35)  55(50) </p> <p>54(56)  56(37) </p> <p>57(33) </p>	<p>58(58) </p> <p>59(135) </p> <p>60(45) </p>	<p>61(57) </p> <p>62(107) </p>	<p>63(43) </p> <p>64(75) </p>
cluster	Not clustered	q		
	<p>65(106) </p> <p>66(44)  67(66) </p>	<p>68(48) </p> <p>69(71) </p>		

The clustering process was graphically represented using a tree-like *dendrogram*. The dendrogram has the separate coformers (or singleton clusters) as its endpoints and schematically shows the relative (dis)similarity of coformers or clusters of coformers using the distance *d* at which they were

merged. In the case of a merge between two coformers (singletons), this distance is simply equal to  $d_{i,j} = 1 - s_{i,j}$ , and for multi-coformer clusters, the distance is given by equation (2). Therefore, the smaller the distance at which two clusters are merged, the more neighbors are shared among its



**Figure 3**  
Example of a subnetwork encountered upon the inspection of a cocrystal, containing two types of common neighbors (CN) and three types of local community links (LCLs).

members. By cutting the dendrogram at a carefully chosen distance  $d$ , a set of clusters was obtained that was analyzed further.

### 2.3. Degree distribution and power-law model

A characteristic property of a network is the distribution of its nodes' connectivities, or *degrees*. The degree  $k$  of a node is defined as the number of neighbors it has ( $k_i = |n_i|$ ), or here, the number of distinct cocrystals known for a given coformer. The degree distribution is usually presented as the fraction of nodes  $p(k)$  with degree  $k$  as a function of the degree  $k$ .

Because the shape of this distribution for the coformer network is right-skewed, the data was transformed to a log log plot, where it is found to demonstrate quasi-linear behavior. Consequently, a power-law model in the form of:

$$p(k) = Ck^{-\alpha} \quad (3)$$

was fitted to the data. Here,  $\alpha$  is the exponent of the power-law model, which was estimated from the distribution data using a maximum-likelihood estimator (MLE), and  $C$  is a constant. The estimation protocol for  $\alpha$  and  $C$  is described in more detail in Appendix B.

### 2.4. Network types

Two main types of networks can be used to represent many real-world problems: monopartite and bipartite. In a monopartite network (Fig. 1b), all nodes belong to one single group and may be connected to any other node through edges. This is similar to popular social media platforms, where an associa-

tion between any two users (or nodes) is possible. On the other hand, bipartite networks (Fig. 1c) consist of two distinct, non-overlapping groups of nodes with connections only between nodes of different groups. Examples of bipartite networks are a co-authorship network, consisting of author-article relationships, and a consumer-product network. A third type of network consists of a mixture of a mono- and bipartite network. Similar to a bipartite network, still two types of nodes can be identified; however, some nodes may form edges to both sets instead of only one, breaking the constraint for pure bipartition (Fig. 1d). In principle, the mixture network can be seen as a general way to describe the type of a network, with mono- and bipartite networks as its limiting cases (Chang & Tang, 2014). An example of such a mixture network is a network of shareholders: while there are two sets of nodes, owners and corporations, some corporations may also act as shareholders and have shares in other corporations, leading to a mixture network.

Having knowledge of the network's type is crucial when trying to understand its structure and when trying to develop strategies to use the network's information. For instance, link prediction algorithms<sup>2</sup> require the knowledge of the network's type to produce relevant new edge suggestions. For mixture networks, it may therefore be interesting to analyze them in terms of their limiting cases. For example, if the network appears to be mostly bipartite (with only a few monopartite nodes), the use of bipartite link prediction algorithms can be justified for the mixture network.

Whereas a network of binary organic salts can be interpreted as a purely bipartite network with two ion sets (cat- and anions), there is no such straightforward grouping for cocrystals. A certain degree of complementarity has been observed between coformers, such as in hydrogen-bonding or  $\pi$ -electron systems, suggesting that the network is bipartite. However, it is sometimes impossible to unambiguously define the nature (or role) of coformers in such a framework. For example, isonicotinamide (Table 1, coformer 40) has the structural features of both a hydrogen bond donor and acceptor. Besides, since most coformers form crystalline structures with themselves, it comes as no surprise that cocrystals exist that combine structurally analogous molecules [e.g. cocrystal NEHJER (Eddleston *et al.*, 2012) consisting of theophylline and caffeine (coformers 41 and 42, respectively)].

Therefore, instead of hypothesizing to which (limiting) type the coformer network belongs, it was assumed to be of a mixed type and was consequently quantified in terms of its mono- and bipartiteness. To that end, each cocrystal present in the network was consecutively investigated by mapping out the direct periphery of its nodes (*i.e.* paths of length 2 and 3, involving single and pairs of nodes, respectively). These small subnetworks were characterized using different formulations of the *common neighbors* (CN) and *local community links*

<sup>2</sup> These algorithms attempt to find missing edges within a network based on its structural properties. Hereby, it is assumed that some topological measure (*e.g.* the degree) is related to the likelihood of forming edges between nodes, and hence node combinations with higher values are presumed to exist.

(LCL) (Fig. 3) that were introduced by Cannistraci *et al.* (2013) and Daminelli *et al.* (2015):

Monopartite CN: the number of (first) common neighbors, equivalent to  $|n_i \cap n_j|$ .

Bipartite CN: the number of first neighbors connected to each other (excluding monopartite common neighbors).

Monopartite LCL: the number of links between the monopartite common neighbors.

Bipartite LCL: the number of links between bipartite common neighbors.

Monopartite-bipartite LCL: the number of links between mono- and bipartite common neighbors.

The calculation of these metrics using the adjacency matrix is straightforward. By mapping these for each cocrystal in the network, conclusions can be drawn regarding the overall coformer network type.

### 3. Results and discussion

#### 3.1. Network construction

A set of 9222 cocrystals, formed by 7188 unique coformers, was successfully extracted from CSD using the *classifier* algorithm (see Appendix A). The cocrystals were subsequently transformed into a network of coformers (or adjacency matrix), permitting the analysis of its properties and characteristics. The subnetwork formed by the coformers with 30 or more unique cocrystals in the CSD (and the cocrystals

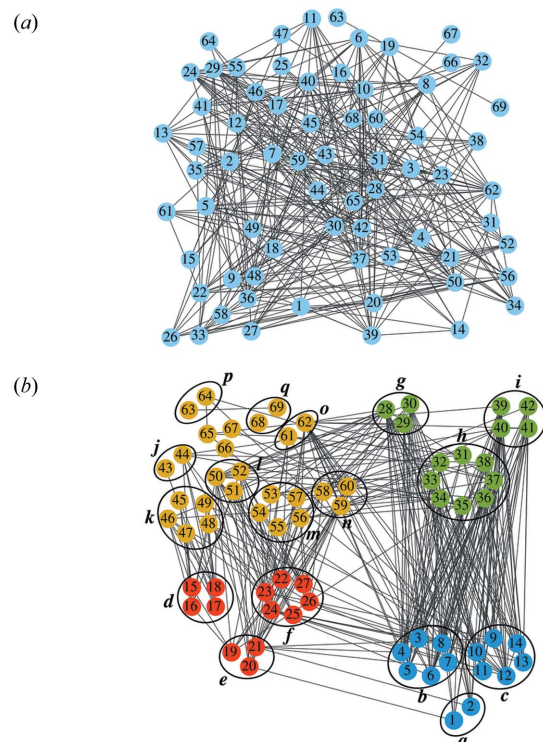
between them) is shown in Fig. 4a. A graphical representation of the total network, or even the subnetwork presented here, is rather uninformative. Using the techniques discussed in Sections 3.2, 3.3 and 3.4, quantitative statements about the structure and type of the network can be made, resulting in a deeper understanding about how coformers relate to each other and how new cocrystals could be predicted.

#### 3.2. Coformer clusters

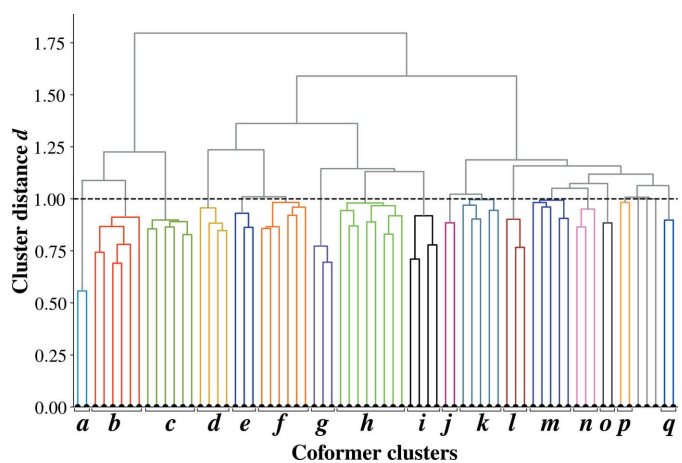
It is common to design cocrystals using supramolecular synthons (Berry & Steed, 2017), where structural motifs are combined that are known to play an important role in the formation and stabilization of the cocrystal. Consequently, coformers are often labeled with a specific function; for example, carboxylic acid containing molecules are classified as hydrogen bond donors and acceptors, and can be combined with themselves (homosynthon) or a different hydrogen bond donor and/or acceptor (heterosynthon). To investigate whether such a grouping of coformers, based on molecular features, can be retrieved from the network, clusters of coformers are sought. Because clusters bring together molecules which have coformers in common, hence without taking any chemical features into account, they are anticipated to reveal purely functionally related coformers.

The clustering is performed for the subset of 69 highly-connected coformers introduced in Section 3.1. Using the data from the complete adjacency matrix, a  $69 \times 69$  similarity matrix is computed, which is subsequently clustered using the abovementioned agglomerative procedure. The dendrogram resulting from such a clustering is shown in Fig. 5, from which a set of relevant clusters is extracted by taking the groups that merged below a distance of 1. This ensures that subsets of the most related coformers are found and prevents completely dissimilar coformers ( $d_{ij} = 1$ ) from being clustered.

In general, the molecular structures of the coformers found in the same clusters in Fig. 5 are similar. For example in Table 1, cluster **b** consists entirely of small aliphatic dicar-



**Figure 4**  
The part of the network showing the 69 distinct coformers with more than 30 neighbors, and the cocrystals formed between them. The numbers and letters each correspond to the coformers and clusters shown in Table 1. (a) Random placement of the nodes. (b) Placement of the nodes according to the clusters and grouped in a hierarchical way.



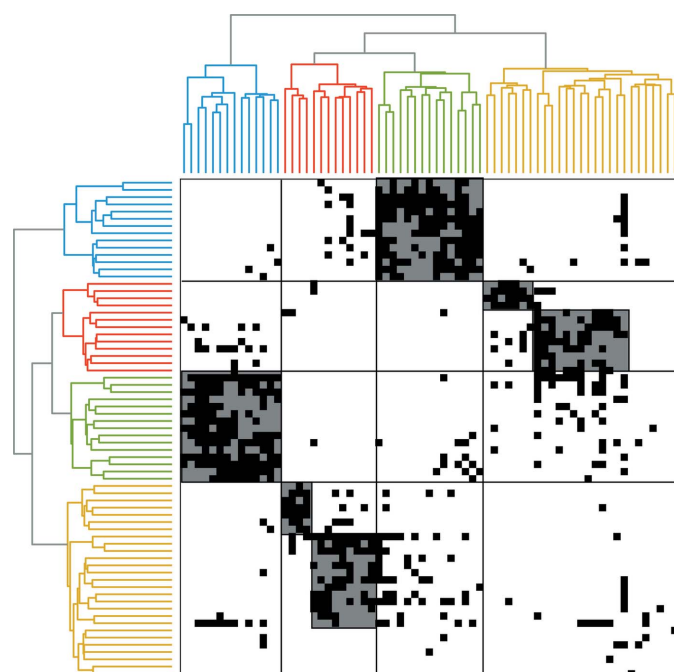
**Figure 5**  
The dendrogram resulting from clustering the set of 69 popular coformers (for which  $k > 30$ ). The clusters are labeled by letters, and the structures corresponding to these clusters can be found in Table 1.

boxylic acids, and the cofomers of cluster *l* are all sixfold substituted aryl halides. As expected, the structural features that connect the clustered cofomers play a profound role in the formation of cocrystals: groups of hydrogen-bond donors, acceptors, or containing electron rich or deficient  $\pi$ -systems are identified among the clusters. The network approach thus recovers the grouping of cofomers that is often used *a priori* for the design of new cocrystals (with for example specific synthons).

On the other hand, for some clusters, the molecular structures involved can be rather different. For instance in cluster *d*, the aggregation mechanism of the cocrystals formed by these cofomers is mostly face-to-face planar stacking [e.g. CSD entries REQWAM (Rosokha *et al.*, 2006), MURPYR (Damiani *et al.*, 1965), ANTPML01 (Robertson & Stezowski, 1978) and PVVBHJ01 (Banerjee, & Brown, 1985)]. However, tetrathiafulvalene (coformer **15**), a heterocyclic sulfur-containing compound, is structurally dissimilar to the other polycyclic aromatic hydrocarbons in the cluster. Another example is cluster *e*, of which the cofomers all can function as both non-aromatic hydrogen bond acceptors [e.g. CSD entries QUIDON (Sakurai, 1968), COLGUG (Timmons *et al.*, 2008), FEQXIJ (Ghosh *et al.*, 2005)] and electron-pair donors in halogen bonds [e.g. entries BNQBRP (Shipley & Wallwork, 1967), FUYDEK (Catalano *et al.*, 2015), QIHCOZ (Walsh *et al.*, 2001)].

Again, while being functionally similar, 1,4-benzoquinone (coformer **19**) is structurally different from the other cofomers in the same cluster. The two examples above highlight the power of this data-driven approach: it is able to successfully identify functionally similar cofomers, free of any structural prejudices.

The hierarchical structure of the dendrogram in Fig. 5 exists at several scales, and therefore, cutting off the tree at heights different from the one proposed above ( $d = 1$ ) is also assumed to result in meaningful clusters. This is exemplified in Fig. 6, where four distinct clusters are extracted at a distance  $d = 1.3$ . As expected, the larger clusters contain more diverse cofomers, which still exhibit a tendency to cocrystallize with similar cofomers. By reorganizing the small  $69 \times 69$  adjacency matrix in such a way that clustered cofomers are placed side-by-side, blocks of dense interconnections (or cocrystals) can be seen (illustrated in Fig. 6). The most obvious example is the block connecting the green and blue clusters, demonstrating the clear complementarity between cofomers containing carboxylic acid groups and aromatic nitrogen atoms in the formation of hydrogen bonds (see also Fig. 4*b*). Cofomers within the same cluster also rarely form cocrystals with themselves (sparse blocks on the diagonal), hinting that the network is organized primarily in a bipartite way (and hence not monopartite). The subnetwork of popular cofomers is, however, only a very small part of the complete network, and a more in-depth analysis of the network type is presented in Section 3.4.



**Figure 6**

A visual representation of the adjacency matrix from the set of 69 popular cofomers, reorganized using the dendrograms (cut off at  $d = 1.3$ ). The small black squares correspond to existing cocrystals between the cofomers. Areas with a relatively large density of cocrystals are emphasized in grey, and black lines are added as guides to distinguish between the clusters. The clusters, characterized by a color, consist of the following smaller clusters that were determined earlier in Fig. 5. Blue: *a*, *b*, *c*. Red: *d*, *e*, *f*. Green: *g*, *h*, *i*. Yellow: *j*, *k*, *l*, *m*, *n*, *o*, *p*, *q*.

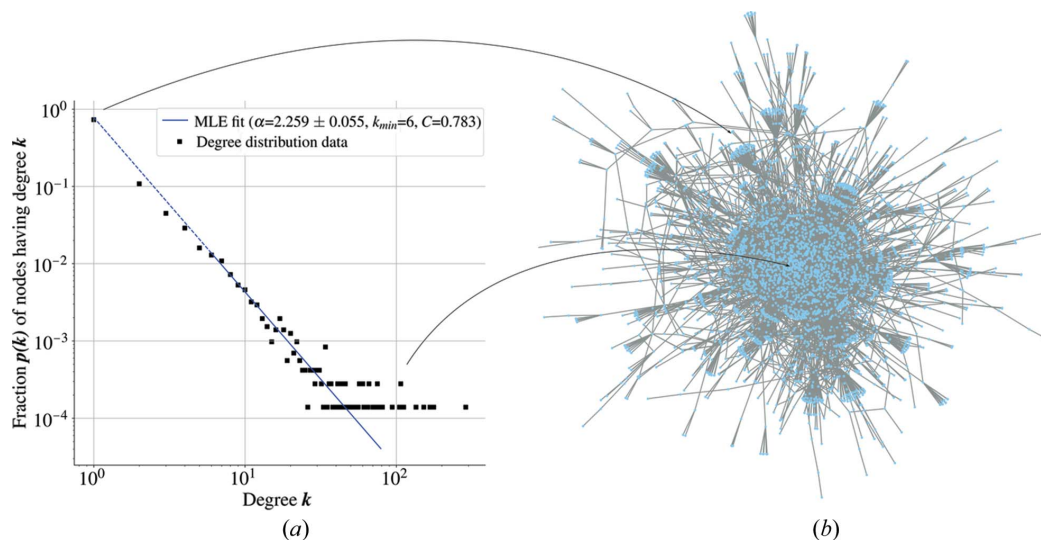
### 3.3. Coformer popularity and bias

A closer look at Fig. 4 and the coformer degrees in Table 1 reveals that some cofomers are significantly more popular than others; for instance, whereas 4,4'-bipyridine (coformer **28**) has cocrystallized with 288 different cofomers, malonic acid (coformer **7**) is found in only 34 distinct cocrystals. In addition, while the network consists of 7188 unique cofomers, only 69 of them appear to have more than 30 cocrystals, implying that the coformer degree is unevenly distributed.

The imbalance of popularity was analyzed using the coformer degree distribution (Fig. 7). Remarkably, a (quasi-)linear relation between  $\log p(k)$  and  $\log k$  is seen, and the distribution was fitted with a power-law model [equation (3)]. Networks that have such a degree distribution are classified as *scale-free*<sup>3</sup> and are characterized by a set of interesting properties (see Appendix B). In the case of the coformer network, this implies that while most cofomers are present in only one or a few cocrystals [small  $k$ , large  $p(k)$ ], a small group exists for which the degree is up to two orders of magnitude larger [large  $k$ , small  $p(k)$ ]. By plotting the cumulative fraction of cocrystals  $W$  as a function the fraction of highest degree cofomers  $p$  (Fig. 8), the imbalance in popularity becomes even clearer: a relatively small group of cofomers (10% of the total number) is found in most of the cocrystals (approximately 70%).

<sup>3</sup> That is, when the degree distribution is fitted with a power-law model and has an exponent  $\alpha$  between 2 and 3.





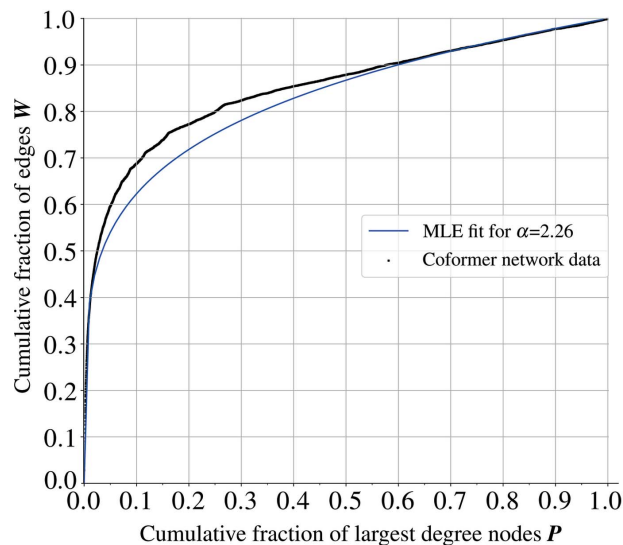
**Figure 7**  
 (a) Distribution of the coformer degrees. The solid blue line is the fitted power-law model for  $k \geq k_{\min}$  and the dashed blue line is an extrapolation of the model over the initial part of the degree range. (b) Largest connected component of the coformer network, containing 83% of the cocrystals and 62% of the cofomers. The arrows highlight the structural differences in (b) associated with the data points in (a). While most of the nodes have  $k = 1$  (and are drawn towards the outside of the network), the central core or *glue* of the network, consisting of a small number of cofomers with a larger  $k$ , is responsible for the coherent structure of the network.

The cocrystal data in the CSD is thus heavily biased: combinations of the same, popular cofomers (large  $k$ ) with relatively unknown cofomers (small  $k$ ) make up for the largest part of the cocrystals entries. Consequently, knowledge-based approaches that use data sets obtained by randomly selecting cocrystals are undoubtedly susceptible to this bias, which may hinder the formulation of general design rules for cocrystallization. On the other hand, as shown in Fig. 8, omitting these highly popular cofomers would drastically reduce the number of cocrystals in the data set, making it impossible to obtain an overall understanding about cocrystallization since only niche cocrystals would be left in the data set.

A plausible explanation for the scale-free topology of the coformer network is that the choice of a second coformer for cocrystallization experiments is frequently biased. For example, new pharmaceuticals are commonly combined with a small group of well-known GRAS<sup>4</sup> cofomers [US Food & Drug Administration (FDA), 2018], such as benzoic acid and nicotinamide (coformer 9 and 39 in Table 1, respectively). This suggests that preferential attachment (Barabási & Albert, 1999; Albert & Barabási, 2002) plays a crucial role in the expansion of the network: whereas highly-connected cofomers are very likely to be used for cocrystal formation, cofomers with smaller connectivity remain relatively unexplored, resulting in a power-law distribution of the degrees. It may thus be worthwhile to consider a broader coformer set when designing new cocrystals, looking beyond the select group of cofomers in the tail of Fig. 7 (or in Table 1). Further, although models based on preferential attachment are presumed to describe the network's evolution fairly well<sup>5</sup>,

they do not coincide with the abovementioned cocrystal design strategies. Therefore, models that take into account the inherent bias of the network should be regarded when choosing a suitable prediction algorithm.

The specific distribution of the degree influences the clustering of the cofomers. For highly-connected cofomers, the denominator of equation (1) is generally large, resulting in relatively low similarities, usually independent of the other coformer. Indeed, as illustrated in Fig. 5, the distance  $d$

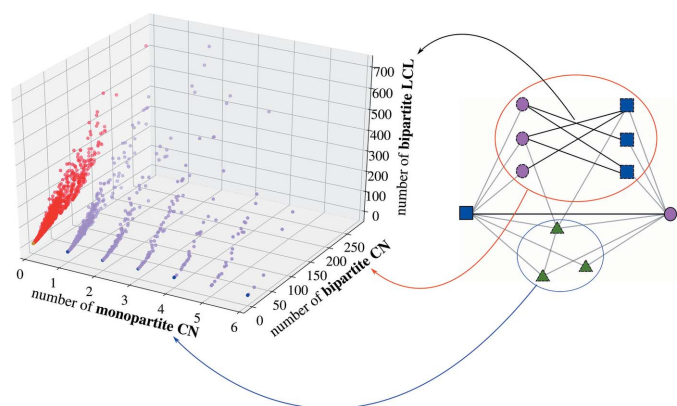


**Figure 8**  
 Cumulative fraction of the edges (or cocrystals)  $W$  plotted against the fraction of nodes (or cofomers) with the highest degrees  $p$ . In practise, a list of nodes with decreasing degrees is constructed, and one records the fraction of edges covered by these nodes while descending through the list. The solid blue line corresponds to a theoretical curve for the power-law model [equation (9)] with  $\alpha = 2.26$ .

<sup>4</sup> Generally recognized as safe.

<sup>5</sup> That is, popular nodes are more likely to form new connections.

between any coformer pair is larger than 0.5 ( $s_{i,j} < 0.5$ ), and closer inspection of the degrees in Table 1 confirms that the difference between the degrees in some clusters (for example



**Figure 9**

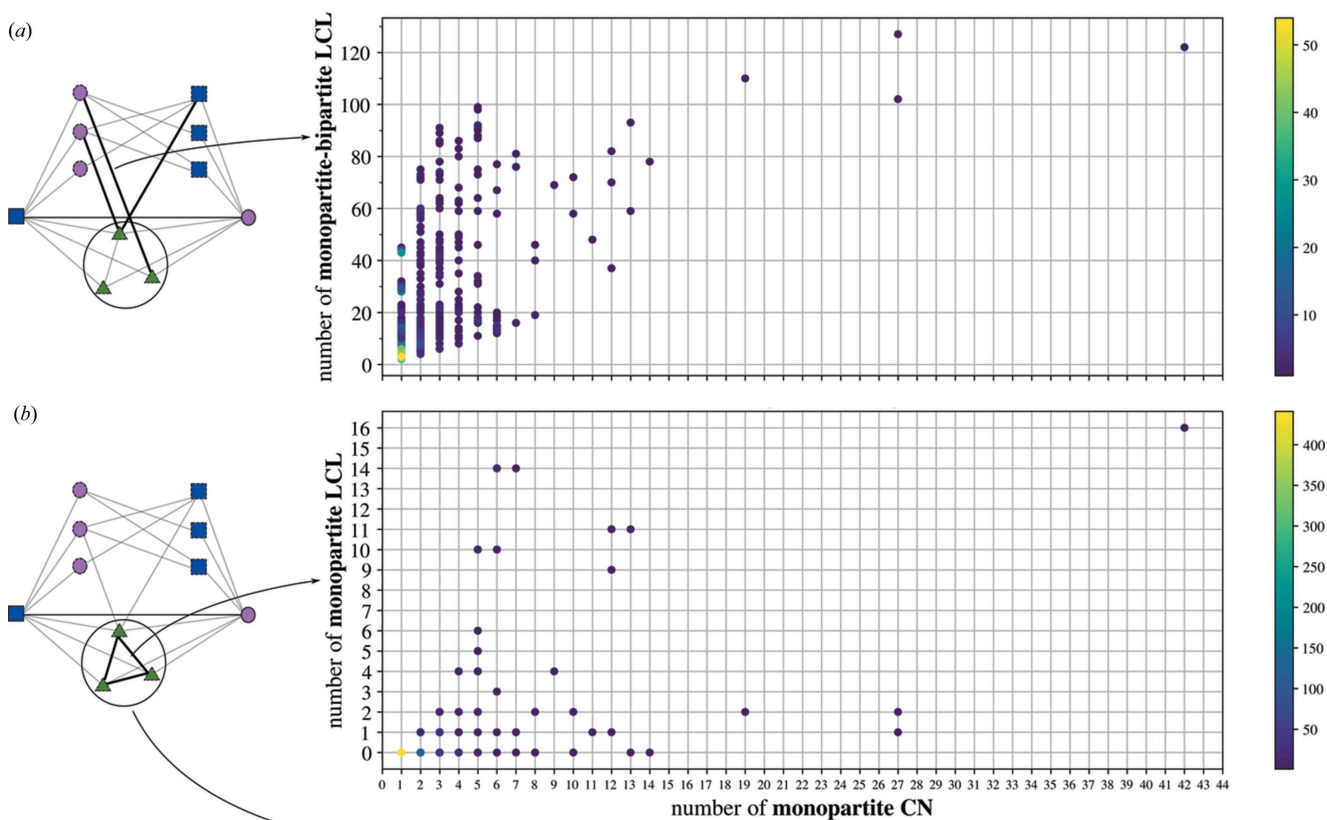
Part of the three-dimensional scatter plot that quantifies the cocrystals in the network according to their number of monopartite CN, bipartite CN and bipartite LCL (low monopartite CN part). For clarification, the small subnetwork of Fig. 3, highlighted for the relevant metrics, is included. Of the 9222 cocrystals, 3572 behave as purely bipartite (red points; monopartite CN = 0, bipartite CN > 0), 72 as purely monopartite (blue points; monopartite CN > 0, bipartite CN = 0), 671 as mixed (purple points; monopartite CN > 0, bipartite CN > 0) and 4907 cannot be characterized by any common neighbors (yellow points; monopartite CN = bipartite CN = 0).

in cluster  $n$ ) can be rather large. Additionally, the coformer network is far from complete, and the clusters that were obtained here are likely to be susceptible to the choices made by researchers in the past few decades when designing cocrystallization experiments. The true set of neighbors (or *profile*) of a coformer may be unattainable, and the actual set of neighbors may be just a reflection of biased experiments, which unavoidably directs the outcome of the clustering procedure. Nevertheless, the obtained clusters still manage to present similar coformers for cocrystallization, and are assumed to improve with the discovery of new cocrystals.

### 3.4. Coformer network type

The type of the coformer network is not *a priori* known, and is therefore assumed to be a mixture of mono- and bipartite. The extent to which the network is similar to either of these two limiting types, is studied by mapping the five metrics introduced above (Section 2.4) for every single cocrystal in the network.

The bipartiteness of the network is analyzed in Fig. 9, where the number of monopartite CN, bipartite CN and bipartite LCL (*i.e.* local community links, edges between bipartite CN) is visualized for each cocrystal present in the network. As a logical consequence of the power-law behavior of the degree distribution, a large part (53%) of the cocrystals cannot be



**Figure 10**

An analysis of the peripheral metrics for the 743 cocrystals with monopartite common neighbors. (a) Number of mono-bipartite LCL and (b) number of monopartite LCL versus the number of monopartite common neighbors (CN). The color of the dots corresponds to a number of cocrystals with those values (using the attached color bars).

characterized by a number of CN (yellow points) due to the limited connectivity of most of the coformers in the network (73% of the coformers with  $k = 1$ , Fig. 7). Of the remaining cocrystals that are interconnected, only a minority of the coformer combinations (2%) is connected exclusively through monopartite common neighbors. On the other hand, approximately 83% behaves purely bipartite (red points) and 16% demonstrates mixed behavior (purple points), with diverse numbers of bipartite CN and LCL, and usually small numbers of monopartite CN ( $\leq 5$ ).

The analysis above therefore suggests that at least the interconnected part of the coformer network is primarily organized in a bipartite manner. Also, as seen in Fig. 9, the number of bipartite CN and LCL is strongly correlated (Pearson correlation coefficient of 0.95), which further supports the claim that this part of the coformer network is predominantly bipartite.

A similar study of the number of monopartite and mono-bipartite LCL for the 743 monopartite and mixed cocrystals (Fig. 10) shows that the number of monopartite common neighbors for most cocrystals rarely exceeds 1. Surprisingly, whenever monopartite neighbors are present, they always form one or several cocrystals with other bipartite common neighbors (via mono-bipartite LCL, Fig. 10*a*), and hardly ever with the other monopartite neighbors (via monopartite LCL, Fig. 10*b*), thus contributing to the bipartiteness of the network.

While an exact bipartition of the entire network in two groups is in principle not possible due to the presence of monopartite noise, it is nevertheless remarkable that a certain level of complementarity is seen for 99% of the interconnected cocrystals. This observation supports the above-mentioned cocrystal design approaches (*e.g.* combining complementary hydrogen bonds or  $\pi$ -electron systems), while being free of any prior hypothesis.

The design of new cocrystals using the network (*e.g.* with link prediction) should therefore be performed in a bipartite way instead of a monopartite way. For example, when two APIs (API 1 and API 2) have several coformers in common ( $U = n_1 \cup n_2$ ), candidates for new cocrystals with API 1 and API 2 should be sought in the sets of non-shared neighbors of API 2 and API 1 ( $n_2 \setminus U$  and  $n_1 \setminus U$ , quadrangular closure), respectively, rather than combining API 1 and API 2 (triangular closure). The exact algorithm for cocrystal prediction based on the network should nonetheless be validated on the data itself, and take into account the other results discussed here. Link prediction applied to cocrystallization will be the topic of a subsequent paper.

#### 4. Conclusions

A network of 7188 coformers was successfully constructed from the information contained in the CSD, making it possible to study cocrystallization using techniques from data mining and network science.

The network is divided in groups or clusters of coformers, which are connected by a common interaction principle (*e.g.* hydrogen bond acceptor). With the addition of new cocrystals

to the database (and consequently to the network), an even more accurate profile of the coformers in terms of bonding will be obtained, leading to better, more refined clusters. Notably, the coformers in these clusters are not necessarily structurally similar, but exhibit an analogous role or function for cocrystal formation. The latter is beneficial when screening for chiral conglomerate cocrystals, since more structural variation is included in the experiments.

The popularity of the coformers in the network is distributed unevenly, and varies approximately over two orders of magnitude. The CSD contains a relatively small subset of highly-popular coformers that is responsible for most of the cocrystals, and hence the data on cocrystallization is inherently biased towards these coformers. Therefore, it is more insightful to choose coformers outside of this small subset when designing new cocrystals and studying cocrystallization in general.

The distribution of the coformer degrees (or connectivities) follows a power law over the largest part of its range, and the network is classified as *scale-free* (see Appendix B). An interesting consequence of the network's specific structure is its lack of an internal scale. Because of the arbitrary fluctuations around it, the average degree is a poor parameter to assess a coformer. A possible reason for the network's scale-freeness is its evolution through preferential attachment, where a select group of coformers is consistently chosen for cocrystallization experiments. While such an evolution can be modelled and is even anticipated to have a good validation performance on the network, its underlying principle (*i.e.* higher connectivity corresponds to higher likelihood of forming a cocrystal) seems unreasonable compared to the cocrystal design strategies proposed in literature.

Even though the coformer network was initially assumed to be of a mixed type, almost all of the interconnected cocrystals in the network are found to behave in a bipartite way. While an exact bipartition of the network (division of the nodes in two groups) is inconsistent due to monopartite noise, there are several clusters (or modules) of coformers in the network that are complementary to each other. This observation may serve as the basic principle to model the coformer network's evolution (with link prediction) and develop an automated, knowledge-based prediction tool.

In conclusion, we have confirmed that the coformer network is a rational representation of cocrystal information, rather than a random assembly of nodes. An automated screening tool based on the structure of the network can thus be justified, provided that the correct model is used. We have developed such a tool, and it is currently being validated using cocrystallization experiments.

#### APPENDIX A Classifier algorithm

##### A1. Splitting

There are three possible types of multi-component crystals that can emerge when inspecting a non-ionic, binary entry

from the CSD. These include cocrystals, solvates and crystals containing a gas molecule. To correctly classify binary entries, an algorithm was written in Python that first converts the structural data of the entries into canonical SMILES strings (Weininger, 1988; Daylight Chemical Information Systems Inc., 2008) [with *OpenBabel* (O'Boyle *et al.*, 2011)] and then splits these strings into their components using standard string manipulations.

SMILES strings are human-readable representations of molecules, or systems containing multiple molecules. When several, distinct molecules are present in the crystal, the SMILES string of the entry is made up of the strings of its constituents, separated by a '.'. The canonicalization of such a SMILES string then results in a unique string for each molecule, promoting the use of canonical SMILES strings as molecular identifiers. In addition, canonical SMILES include the correct absolute configuration of chiral substances (stereogenic centres, *cis-trans* chirality) when computed from three-dimensional data.

## A2. Classification

After splitting the strings of the entries into their constituents, the multi-component crystals are correctly classified as cocrystal, solvate or structure containing a gas molecule by comparing the components to a predefined list of 182 common solvents and 384 common gases. An additional check is performed, confirming the neutrality of the molecule and filtering out erroneous systems coming from faulty three-dimensional coordinates.

**A2.1. Chirality.** When one of the cofomers is chiral, the cocrystal can be either *enantiopure*, because the cocrystal was crystallized from an enantiopure solution or due to the formation of a racemic conglomerate cocrystal, or *racemic*, where the two enantiomers and cofomer are present in the same lattice.<sup>6</sup>

When imposing that due to their configurational difference, enantiomers are different molecules, racemic cocrystals are in principle ternary systems. However, often only one of the enantiomers is present in the asymmetric unit of such a cocrystal, whereas the other one is implied by symmetry operations. In this case, the nature of such a cocrystal is still regarded as binary, and hence racemic compound cocrystals are treated as binary cocrystals (one for each enantiomer). The choice of enantiomer taken up in the asymmetric unit is arbitrary, and thus splitting the cocrystal would result in a system with only one of the enantiomers. Therefore, the same cocrystal but with the counter enantiomer (or exact mirror image) is added to the dataset. In the case where both enantiomers are present in the asymmetric unit of a racemic cocrystal, a binary cocrystal for each enantiomer is added. The deliberate addition of binary cocrystals for racemic systems can also be justified by the observation that enantiopure cocrystals are likely to exist when the racemic compound was successfully cocrystallized (George *et al.*, 2014).

<sup>6</sup> These includes racemic compounds as well as kryptoracemates.

For enantiopure cocrystals, it is very challenging to distinguish a racemic conglomerate cocrystal from a cocrystal that is obtained from a enantiopure mixture.<sup>7</sup> In the case of racemic conglomerates, the enantiomer in the asymmetric unit is again arbitrary, and hence the counter enantiomer should be added to the data set. Because of mirror symmetry, cocrystallization of one of the enantiomers also implies that a cocrystal with the other enantiomer must exist. Therefore, regardless of conglomerate forming behaviour, the counter enantiomer is always deliberately added. While a counter-enantiomer may not always exist in the case of enantiopure cocrystallization, this procedure ensures no indications are missed (so no enantiomers are given too few cocrystals).

The explicit addition of chiral cocrystals, however, falsely increases the popularity (or *degree*, see section 2.3) of the counter cofomer; for example, two edges are drawn for every racemic compound a cofomer cocrystallizes with. Consequently, for every pair, only one representative enantiomer was kept, effectively dealing with the randomness of the asymmetric unit of racemic compound and conglomerate cocrystals, while not overestimating the popularity of the counter cofomer.

## APPENDIX B

### B1. Power-law model fitting

A power-law model in the form  $p(k) = Ck^{-\alpha}$  was fitted to the degree distribution data. Instead of fitting a straight line to the logarithmic data (which is known to result in biased parameter estimations (Goldstein *et al.*, 2004), the exponent of the power-law  $\alpha$  was calculated from the distribution data itself using a maximum-likelihood estimator (MLE) (Newman, 2015):

$$\alpha = 1 + n_{\text{tail}} \left[ \sum_i \ln \frac{k_i}{k_{\text{min}} - \frac{1}{2}} \right], \quad (4)$$

where the summation is performed over data points  $i$  in the tail of the distribution ( $n_{\text{tail}}$  data points) with a degree larger than or equal to the lower bound  $k_{\text{min}}$ , which is the point from where the distribution can be described by a power-law. The exact value of  $k_{\text{min}}$  is usually not *a priori* known, and is therefore estimated from the data by iteratively increasing its value from 1 to 25 and testing where  $\alpha$  reaches a stable value. The simultaneous estimation of  $\alpha$  and  $k_{\text{min}}$  is shown by the blue curve in Fig. 11, where  $\alpha$  reaches a temporarily stable value of 2.26 at  $k_{\text{min}} = 6$ . Hereby, a trade-off is made between the accuracy of  $\alpha$  and the number of observations ( $n_{\text{tail}}$ ) used for its determination, since  $n_{\text{tail}}$  drastically decreases with increasing  $k_{\text{min}}$  (red curve in Fig. 11).

<sup>7</sup> This requires a check of the literature, since the CSD does not provide such information.

The constant  $C$  is determined by summing over both sides of equation (3) in the power-law region ( $k \geq k_{\min}$ ):

$$\sum_{k=k_{\min}}^{\infty} p(k) = C \sum_{k=k_{\min}}^{\infty} k^{-\alpha} \quad (5)$$

$$C = \frac{\sum_{k=k_{\min}}^{\infty} p(k)}{\sum_{k=k_{\min}}^{\infty} k^{-\alpha}} = \frac{\sum_{k=k_{\min}}^{\infty} p(k)}{\zeta(\alpha, k_{\min})} \quad (6)$$

with  $\zeta(\alpha, k_{\min})$  the generalized Riemann zeta function. Using the estimations for  $\alpha$  and  $k_{\min}$ , equation (6) resulted in  $C = 0.783$ .

The goodness of the power-law fit was quantified using the Kolmogorov-Smirnov statistic (or KS-statistic) (Clauset *et al.*, 2009; Press *et al.*, 1992):

$$D = \max_{k \geq k_{\min}} |S(k) - P(k)|, \quad (7)$$

where  $D$  is the largest absolute difference between the observed cumulative degree distribution  $S(k)$  and its power-law fit  $P(k)$ , both in the power-law region ( $k \geq k_{\min}$ ). The cumulative degree distribution is an alternative representation of the degree distribution, where the ordinate axis is transformed to the fraction of nodes  $P(k)$  with a degree  $\geq k$ , or mathematically  $P(k) = \sum_{k'=k}^{\infty} p(k')$ . By approximating this sum as an integral, the corresponding power-law expression for the cumulative degree distribution becomes:

$$P(k) \simeq \frac{C}{\alpha - 1} k^{-(\alpha-1)}. \quad (8)$$

The estimated values of the model parameters for the first 15 lower bound degrees, together with the length of the distribution tail  $n_{\text{tail}}$  and KS statistic for the model fit are summarized in Table 2. The power-law distribution model fits the data reasonably well: the KS statistic for the model fit was

**Table 2**

Summary of the power-law model parameters and corresponding KS statistics for the first 15  $k_{\min}$  values.

The row where  $k_{\min}$  is equal to 6 contains the chosen model parameters.

$k_{\min}$	$\alpha$	$C$	$n_{\text{tail}}$	KS statistic
1	1.94	0.59	7188	0.376
2	2.03	0.44	1941	0.062
3	2.10	0.50	1165	0.027
4	2.18	0.61	843	0.016
5	2.20	0.65	635	0.011
6	2.26	0.78	520	0.011
7	2.28	0.84	427	0.011
8	2.26	0.77	349	0.011
9	2.25	0.76	297	0.011
10	2.26	0.77	259	0.011
11	2.24	0.73	226	0.011
12	2.25	0.75	203	0.011
13	2.24	0.73	182	0.011
14	2.26	0.79	168	0.011
15	2.29	0.90	157	0.011

0.011, which is sufficiently small to confirm the power-law hypothesis for the data with sample size  $n_{\text{tail}} = 520$  (Goldstein *et al.*, 2004).

Networks for which the degree distribution follows a power-law with an exponent  $\alpha \in [2,3]$  are classified as *scale-free*. Scale-free networks are characterized by a peculiar structure: a dense, central core exists, containing only a small fraction of the nodes, but most of the edges (see Figs. 7b and 8), which is surrounded by a large number of unpopular nodes in its periphery. Unlike random networks that are generally characterized by a mean degree and variance, scale-free networks lack such an internal scale: due to its uneven distribution, the expected node degree can be either very small or arbitrarily large<sup>8</sup>, making it a meaningless property. Moreover, assuming a perfect power-law distribution, a theoretical expression can be formulated for the cumulative edge data (Newman, 2015):

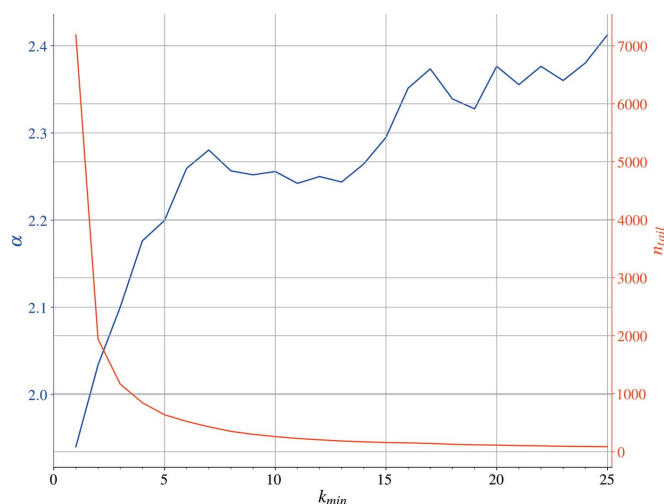
$$W = P^{\frac{\alpha-2}{\alpha-1}}. \quad (9)$$

As was shown in Fig. 8, the theoretical formulation slightly underestimates the network's data. This can be explained by the fact that equation (9) assumes pure power-law behavior over the entire degree range, whereas for the cofomer network, it only holds for degrees larger than the lower bound  $k_{\min} = 6$ .

### Funding information

This research received funding as part of the CORE ITN Project by the European Union's Horizon 2020 Research and Innovation Program under the Marie Skłodowska-Curie Grant Agreement No. 722456 CORE ITN.

<sup>8</sup> Because the second and higher order moments diverge for power-law distributions with a large number of nodes ( $n \rightarrow \infty$ ), the fluctuations around the mean of  $k$  are very large.



**Figure 11**  
Blue line: determination of the lower bound  $k_{\min}$  using the power-law exponent  $\alpha$ . Starting at  $k_{\min} = 6$ ,  $\alpha$  temporarily reaches a stable value, and increases again around  $k_{\min} = 14$ . Red line: number of coformers in the tail of the distribution ( $n_{\text{tail}}$ ) as a function of the lower bound  $k_{\min}$ .

## References

- Albert, R. & Barabási, A.-L. (2002). *Rev. Mod. Phys.* **74**, 47–97.
- Almarsson, O. & Zaworotko, M. J. (2004). *Chem. Commun.* **17**, 1889–1896.
- Banerjee, A. & Brown, C. J. (1985). *Acta Cryst.* **C41**, 82–84.
- Barabási, A.-L. & Albert, R. (1999). *Science*, **286**, 509–512.
- Berry, D. J. & Steed, J. W. (2017). *Adv. Drug Delivery Rev.* **117**, 3–24.
- Bruno, I. J., Cole, J. C., Edgington, P. R., Kessler, M., Macrae, C. F., McCabe, P., Pearson, J. & Taylor, R. (2002). *Acta Cryst.* **B58**, 389–397.
- Cannistraci, C. V., Alanis-Lobato, G. & Ravasi, T. (2013). *Sci. Rep.* **3**, 1613.
- Catalano, L., Pérez-Estrada, S., Terraneo, G., Pilati, T., Resnati, G., Metrangolo, P. & Garcia-Garibay, M. A. (2015). *J. Am. Chem. Soc.* **137**, 15386–15389.
- Chang, C. & Tang, C. (2014). *New J. Phys.* **16**, 093001.
- Clauset, A., Shalizi, C. R. & Newman, M. E. J. (2009). *SIAM Rev.* **51**, 661–703.
- Damiani, A., de Santis, P., Giglio, E., Liquori, A. M., Puliti, R. & Ripamonti, A. (1965). *Acta Cryst.* **19**, 340–348.
- Daminelli, S., Thomas, J. M., Durán, C. & Cannistraci, C. V. (2015). *New J. Phys.* **17**, 113037.
- Daylight Chemical Information Systems Inc. (2008). Smiles - a simplified chemical language. [Online; accessed 1/7/2019]. <http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>
- Delori, A., Galek, P. T. A., Pidcock, E., Patni, M. & Jones, W. (2013). *CrystEngComm*, **15**, 2916–2928.
- Desiraju, G. R. (1995). *Angew. Chem. Int. Ed. Engl.* **34**, 2311–2327.
- Eddleston, M. D., Lloyd, G. O. & Jones, W. (2012). *Chem. Commun.* **48**, 8075–8077.
- Fabian, L. (2009). *Cryst. Growth Des.* **9**, 1436–1443.
- George, F., Tumanov, N., Norberg, B., Robeyns, K., Filinchuk, Y., Wouters, J. & Leysens, T. (2014). *Cryst. Growth Des.* **14**, 2880–2892.
- Ghosh, K., Datta, M., Fröhlich, R. & Ganguly, N. C. (2005). *J. Mol. Struct.* **737**, 201–206.
- Goldstein, M. L., Morris, S. A. & Yen, G. G. (2004). *Eur. Phys. J. B*, **41**, 255–258.
- Greco, T., Hunter, C. A., Gardiner, E. J. & McCabe, J. F. (2014). *Cryst. Growth Des.* **14**, 165–171.
- Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. (2016). *Acta Cryst.* **B72**, 171–179.
- Grothe, E., Meekes, H., Vlieg, E., ter Horst, J. H. & de Gelder, R. (2016). *Cryst. Growth Des.* **16**, 3237–3243.
- Issa, N., Karamertzanis, P. G., Welch, G. W. A. & Price, S. L. (2009). *Cryst. Growth Des.* **9**, 442–453.
- Jaccard, P. (1912). *New Phytol.* **11**, 37–50.
- Jones, E., Oliphant, T., Peterson, P. *et al.* (2001). *SciPy*: Open source scientific tools for Python. [Online; accessed 1/7/2019]. <http://www.scipy.org/>
- Karamertzanis, P. G., Kazantsev, A. V., Issa, N., Welch, G. W., Adjiman, C. S., Pantelides, C. C. & Price, S. L. (2009). *J. Chem. Theor. Comput.* **5**, 1432–1448.
- Lorenz, H. & Seidel-Morgenstern, A. (2014). *Angew. Chem. Int. Ed.* **53**, 1218–1250.
- 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. & Wood, P. A. (2008). *J. Appl. Cryst.* **41**, 466–470.
- Nauha, E. & Nissinen, M. (2011). *J. Mol. Struct.* **1006**, 566–569.
- Newman, M. E. J. (2015). *Networks: An Introduction*. Oxford University Press.
- O’Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T. & Hutchison, G. R. (2011). *J. Cheminform.* **3**, 33.
- Press, H., W., Teukolsky, A., S., Vetterling, T. W. & Flannery, P. B. (1992). *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. Cambridge University Press.
- Robertson, B. E. & Stezowski, J. J. (1978). *Acta Cryst.* **B34**, 3005–3011.
- Rosokha, S. V., Dibrov, S. M., Rosokha, T. Y. & Kochi, J. K. (2006). *Photochem. Photobiol. Sci.* **5**, 914–924.
- Sakurai, T. (1968). *Acta Cryst.* **B24**, 403–412.
- ShIPLEY, G. G. & WALLWORK, S. C. (1967). *Acta Cryst.* **22**, 593.
- Taylor, C. R. & Day, G. M. (2018). *Cryst. Growth Des.* **18**, 892–904.
- Timmons, D. J., Pacheco, M. R., Fricke, K. A. & Sleboznick, C. (2008). *Cryst. Growth Des.* **8**, 2765–2769.
- US Food & Drug Administration (FDA) (2018). Generally Recognized as Safe (GRAS). [Online; accessed 1/8/2019]. <https://www.fda.gov/food/ingredientspackaginglabeling/gras/>
- Walsh, R. B., Padgett, C. W., Metrangolo, P., Resnati, G., Hanks, T. W. & Pennington, W. T. (2001). *Cryst. Growth Des.* **1**, 165–175.
- Ward, J. H. J. (1963). *J. Am. Stat. Assoc.* **58**, 236–244.
- Weininger, D. (1988). *J. Chem. Inf. Comput. Sci.* **28**, 31–36.
- Wicker, J. G. P., Crowley, L. M., Robshaw, O., Little, E. J., Stokes, S. P., Cooper, R. I. & Lawrence, S. E. (2017). *CrystEngComm*, **19**, 5336–5340.