ARTICLE

The extensive solid-form landscape of sulfathiazole: geometrical similarity and interaction energies

David S. Hughes,*[ab] Ann L. Bingham,[b] Michael B. Hursthouse,[b] Terry L. Threlfall[b] and Andrew D. Bond*[a]

A set of 96 crystal structures containing sulfathiazole (SLFZ) is presented, comprising 52 new crystal structures and 39 structures retrieved from the Cambridge Structural Database. The set comprises five polymorphs, 59 co-crystals, 29 salts and three other structures, providing one of the most extensive solid-form landscapes established for a single active pharmaceutical ingredient. The crystal structures are energy-minimised using DFT-D calculations to yield a standardised set. Geometrical comparisons are made using the programs *CrystalCMP*, *COMPACK* and *XPac*, and the results are combined and compared. Consistent conclusions are drawn on full 3-D isostructurailty within the set, identifying a group of 18 isostructural co-crystals, and 11 further isostructural groups of salts or co-crystals comprising two or three structures. Aside from the fully isostructural groups, common 2-D supramolecular constructs (SCs) are restricted to groups of only two or three structures and there are no 2-D SCs that are observed especially frequently. Transferable 1-D SCs are more common, and examples are identified based on hydrogen-bonded and non-hydrogen bonded interactions between SLFZ molecules. Closely-related 1-D SCs comprising translated SLFZ molecules linked by hydrogen bonds are found in one polymorph and almost half of the multi-component set. A comparison of the five SLFZ polymorphs and the 91 multi-component crystal structures identifies several pairwise interactions between SLFZ molecules that are present in one of the polymorphs and at least one multi-component structure. A centrosymmetric $R^2_2(8)$ N—H···N hydrogen-bonded pair occurs in one polymorph and approximately 80% of the co-crystals. Intermolecular interaction energies, calculated using the *PIXEL* method, show that this $R^2_2(8)$ dimer is by far the most stabilising pairwise interaction in any structure. In general, however, there is no straightforward correlation between intermolecular interaction energies of the pairwise motifs in the polymorphs and their frequency of occurrence in the multi-component set. The extensive SLFZ set provides a challenge for systematic geometrical comparison of crystal structures, and some observations are made on the methodology and consistency of the applied programs.

Introduction

Interest in the crystal structures and properties of molecular solids is, broadly speaking, driven by two main demands. On the one hand are the manufacturing and patent aspects associated with the production and marketing of industrially important solid materials, exemplified particularly by pharmaceuticals.[1-9] On the other is the search for designer solids with utilisable properties, such as novel conjugated materials for use in organic solar cells, light-emitting diodes and transistors.[10] Such studies relate to the broader topics of crystal engineering and crystal structure prediction (CSP).[11-14] Research in these areas has

developed to a point where crystal assembly can be designed to a significant extent, especially where intermolecular interactions are at the stronger end of the scale.[15-17] Successes in CSP are also increasing at a steady rate and the current state of the art means that structures with many degrees of freedom and multi-component crystals can often be predicted, using a number of programs.[13] Synergy between CSP and experimental studies has been realised in some cases,[18-20] and the long-term potential for application of machine learning and artificial intelligence to the design of solid forms is clear.[21-24]

In this context, one compound that we have studied extensively is sulfathiazole (SLFZ), a well-known active pharmaceutical ingredient (API; Scheme 1).[25-27] Whilst in search of reliable crystallisation procedures for the SLFZ polymorphs, an extensive solvent-screening exercise was carried out, which yielded over 100 crystalline solvates.[28] This represents one of the most extensive sets of multi-component crystals to be established for a single API,[29] providing a rich opportunity to explore its solid-form landscape. This paper presents crystallographic data for the multi-component SLFZ crystal forms and discusses some initial efforts to analyse the structures. In total, the structure set comprises 91 multi-component structures (52 new structures, 39 retrieved from

the Cambridge Structural Database (CSD)[30]) plus five SLFZ polymorphs. Systematic comparison of such a large set is a significant challenge which might be approached in various ways. The principal focus in this paper is on geometrical similarity, assessed using three generally available programs: *COMPACK*,[31] (as implemented in *Mercury*[32]), *CrystalCMP*[33,34] and *XPac*.[35] Of particular interest is the comparison of results from the three different sources and the challenges that arise while seeking to establish consistent conclusions. The geometrical analysis is accompanied by calculation of intermolecular interaction energies in the polymorphs, with a view towards establishing the extent to which these might be correlated with transferability of pairwise motifs within the structure set. Seaton *et al.* have previously taken a similar approach with a more limited set of SLFZ salts.[36] A complementary analysis of the SLFZ set based on hydrogen-bond topology is planned for a subsequent paper.
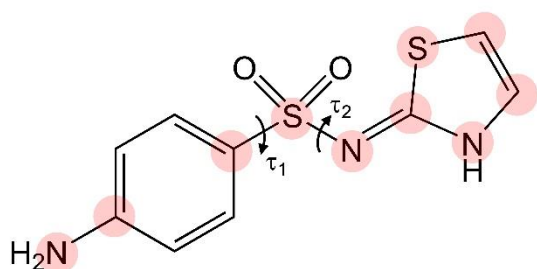
Scheme 1. Sulfathiazole (SLFZ), indicating rotatable torsions (τ_1 and τ_2). Shaded atoms have unique (non-H) connectivity and are used to provide corresponding ordered sets for geometrical comparison. The imino tautomer shown is found exclusively in the structure set. The alternative amino tautomer is not seen.

Experimental Section

X-ray crystallography

52 new crystal structures were obtained from single-crystal X-ray diffraction measurements made on various instruments at the University of Southampton. Experimental and refinement details are provided in the ESI. The available data were in some cases of limited resolution and several structures showed disorder of the solvent molecules. In all disordered cases, except **8**, it was possible with suitable restraints to refine two distinct orientations of the solvent molecule, to give satisfactory coordinate sets for subsequent energy minimisation (see below). For **8**, the location of the solvent molecule was clear, but the electron density appeared to comprise an overlay of several orientations which could not easily be resolved. The most prominent set of peaks in the electron density was modelled with restraints to provide a starting set for energy minimisation, but acceptable *R*-factors could only be produced in the X-ray refinement by application of the *SQUEEZE* algorithm.[37] The structure of **8** is part of a large isostructural group (see Results section) and the solvent molecule is not involved in hydrogen bonding, so uncertainty in its exact position has no significant influence on the subsequent analysis.

Numbering scheme and standardisation of the structures

The literature contains conflicting numbering schemes for the SLFZ polymorphs. Table 1 shows the scheme applied here,[38] with representative CSD refcodes. Throughout this paper, the suffix "**p**" is applied to indicate the polymorphs. The multi-component structures are labelled **1–91**, with **1–59** being co-crystals, **60–88** being salts and **89–91** being other types (described in the Results section).

Table 1. Numbering scheme applied to the SLFZ polymorphs, with crystallographic information and representative CSD refcodes.

	CSD refcode	Space group	Unit-cell parameters (Å, °)			Vol (Å³)	Z / Z′
1p	SUTHAZ16	$P2_1/c$	10.534	12.936	17.191	2230.8	8/2
			90	107.77	90		
2p	SUTHAZ05	$P2_1/n$	10.399	15.132	14.280	2246.6	8/2
			90	91.21	90		
3p	SUTHAZ17	$P2_1/c$	17.448	8.498	15.511	2120.0	8/2
			90	112.81	90		
4p	SUTHAZ18	$P2_1/c$	8.193	8.538	15.437	1077.2	4/1
			90	94.01	90		
5p	SUTHAZ19	$P2_1/n$	10.774	8.467	11.367	1036.5	4/1
			90	91.65	90		

In each structure, a consistent atom numbering scheme is applied to SLFZ, as shown in Fig. 1. Since several structures have $Z' > 1$, a 2-digit code is adopted, where the first number identifies the molecule index and the second is the atom label within the molecule. The molecule has two torsion angles expected to show significant variation amongst the set, denoted τ_1 and τ_2 in Scheme 1. In the crystal, the molecule can exist in two pseudo-chiral conformations, leading to atropisomerism.[39] For consistency in the standardised set, the SLFZ molecule in the asymmetric unit (or the molecule given index 1 in cases with $Z' > 1$) is chosen so that the thiazole ring lies to the left when the molecule is viewed along the bisector of the SO_2 group with the S=O bonds directed toward the viewer (Fig. 1), which corresponds to a negative value for τ_2. This is referred to as the "*R*" (reference) conformation, and the conformation of other molecules is labelled *R* or *S* relative to that reference. The choice of the *R* conformation is arbitrary; the purpose is to describe whether specific molecules have the same or different pseudo-chirality. In some of the non-centrosymmetric structures, experimental absolute structure determination indicated that the *S* conformation is exclusively present in the crystal analysed, while others showed inversion twinning. All structures are converted to the *R* conformation in the standardised set.
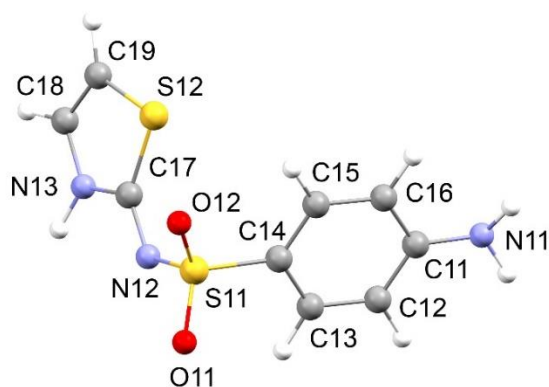
Figure 1. Reference ("*R*") conformation of the SLFZ molecule applied in the standardised structures. The S=O bond vectors are directed towards the viewer, with O12 uppermost and the thiazole ring to the left.

Energy minimisation using DFT-D calculations

Since the crystal structures originate from various sources and have a range of quality indicators, each structure was energy minimised using dispersion-corrected density functional theory (DFT-D) calculations. These optimisation methods have been established to reproduce correct crystal structures.[40] For this work, they provide a "cleaned" data set in which all structures are placed on a common basis. Particular clarification is achieved in the positions of H atoms, which enables a more confident (automated) assessment of hydrogen bonding, to be discussed in a subsequent paper. Energy minimisation is also helpful where the X-ray structures contain poorly-resolved or disordered solvent molecules. All further discussion refers to the set of standardised and energy-minimised structures, which are provided in the ESI. In some cases, it was necessary to reduce the space group symmetry of the X-ray structure to generate complete molecules, or to eliminate disorder of the solvent molecules. These cases are noted in the ESI.

Pairwise intermolecular interaction energies

Pairwise intermolecular interaction energies were calculated for the polymorphs using the *PIXEL* methodology.[41] Similar calculations have been published previously for **1p**, **3p**, **4p** and **5p** by Sovago *et al.*[42] The calculations were applied here to the DFT-D minimised structures, retaining the H atom positions. Tables of interaction energies, including symmetry notation consistent with the rest of the study, are provided in the ESI. Conversion of the *PIXEL* output was implemented through a modified version of the *processPIXEL* utility.[43]

Computer programs

Geometrical comparison of the crystal structures was carried out using *COMPACK*[31] (implemented as the *Crystal Packing Similarity* tool in *Mercury*[32]), *CrystalCMP*[33,34] and *XPac*.[35] To permit comparison of the full solid-form landscape (*i.e.* polymorphs and multi-component forms), the comparisons are applied to the SLFZ molecules only. H atoms are excluded (enabling direct comparison of tautomers and different charge states) and partner molecules are omitted. On the basis of

connectivity, 10 atoms are identified uniquely in each SLFZ molecule (Scheme 1), providing corresponding ordered sets of points for use in *CrystalCMP* and *XPac*. *COMPACK* does not require an ordered set of points to be defined.

Symmetry notation

To compare output between the various programs, the symmetry notation of *PLATON*[44] is adopted. Each SLFZ molecule in the asymmetric unit is designated by the symmetry code 1555_01 (and 1555_02, *etc.*, where *Z′* > 1). The first digit identifies an applied symmetry operator by its position in a specified list, followed by encoding of any further translation along *x*, *y* and *z*, respectively: 555 denotes no additional translation, 655 denotes +1 along x, 545 denotes −1 along *y*, *etc*. Where there is more than one molecule, an index is appended. For example, 3654_02 specifies the second molecule after application of the third symmetry operator in the input operator list, with a further translation of +1 along *x* and −1 along *z*. The notation obviously depends on the defined sequence of symmetry operators, so care was taken to ensure that the sequences in the standardised CIFs are identical to those within *PLATON*. In general, the requirement for such consistency amongst inputs cannot form part of any robust automated methodology, but it is helpful here to achieve consistency between the various programs to be applied.

Results and Discussion

Range of the structure set

The 91 multi-component structures comprise 59 co-crystals (containing neutral SLFZ and co-former molecules) and 29 salts (containing charged SLFZ and partner anions/cations) (see ESI). Three structures fall outside of this straightforward classification: **89** contains both neutral and charged SLFZ molecules, forming [(SLFZ)$_2$]$^-$ units with sparteine cations; **90** contains SLFZ$^+$ with both anionic dinitrobenzoate and neutral dinitrobenzoic acid; **91** contains neutral SLFZ together with ion-separated adamantyl chloride. Amongst the salts, SLFZ$^-$ is roughly twice as common as SLFZ$^+$, and all neutral SLFZ molecules are found as the imino tautomer. To enable a broad survey across the diverse structure set, the structural comparisons throughout this paper are applied to the standardised, DFT-D optimised structures, including only the non-H atoms of the SLFZ molecules.

A previous survey of the conformational characteristics of *N*-substituted arylsulfonamides[45] identified two torsion angles expected to show significant variation, denoted τ_1 and τ_2 in Scheme 1. An analysis of these torsion angles for the SLFZ set (see ESI) shows that τ_2 resembles a Gaussian distribution with mean 78° and standard deviation 8°. Torsion angle τ_1 also resembles a Gaussian with approximate mean 111° and standard deviation 10° but with a residual tail extending to higher values, populated principally by salts. A scatterplot of τ_2 versus τ_1 shows a loose cluster centred around the mean values of τ_1 and τ_2, with the extension to higher τ_1 values seen clearly for the salts (ESI). The polymorphs fall mostly within the bulk

cluster, except for molecule 1 of **2p** (τ_1 = 137.2, τ_2 = –98.6°), which is an outlier due to its high τ_2 value. On this basis, **2p** might be distinguished as a conformational polymorph.[46]

Overview of comparison methodology

All of the applied comparison programs consider clusters of molecules built around a kernel molecule, thereby being independent of choices for the unit cell, space group, *etc.* Clusters of 15 molecules are typically considered to be sufficient to compare structures. The clusters being compared (**A** and **B**) are effectively aligned by overlaying the kernel molecules, then the remaining molecules in **A** and **B** are compared to each other using some geometrical criteria. An inversion-related copy of cluster **B** should also be tested, and separate clusters might be built for independent molecules in structures with $Z' > 1$. The result is a group of molecules in the two structures that are considered to match, with some quantitative measures of the geometrical similarity. The three applied programs differ in the details of their application and the nature of the results reported. Full details of the methodologies are given in the original papers,[31,33-35] but the main differences are summarised as follows.

CrystalCMP produces a single continuous figure-of-merit, PS_{AB}, calculated from the distances and relative orientations of mapped molecules.[33,34] A smaller value of PS_{AB} indicates a greater degree of structural similarity. Clusters **A** and **B** are initially aligned to give the least-squares distance overlay of the atoms in the kernel molecules, then the remaining molecules in **A** and **B** are mapped by identifying the shortest distances between their centroids. A search fragment must be defined to compare molecules, and the PS_{AB} value incorporates an RMSD measure between corresponding atoms.[34] The final PS_{AB} value is based on all (usually 15) mapped molecules.

COMPACK and *XPac* map molecules in a fundamentally different way, by considering local pairwise similarity. A shell is built around the kernel in the reference cluster **A**, comprising molecules "connected" by any intermolecular interatomic distance shorter than the sum of the van der Waals radii plus some tolerance. Each shell molecule in **A** is then compared to each molecule in cluster **B**, and molecules are retained in the growing group of mapped molecules if an **A**↔**B** match is established according to specified distance and angle criteria (which differ between the two programs). The process is continued to second-shell contacts for those molecules retained in the growing group until all molecules in cluster **A** have been visited. Full 3-D isostructurality is established in *COMPACK* if all (usually 15) molecules in cluster **A** are matched, and sub-structure similarity is indicated where only some molecules are matched. *XPac* interprets the established mappings to identify supramolecular constructs (SCs) within structure sets, comprising groups of matched molecules that may be 0-D (isolated) or extend in 1-D, 2-D or 3-D. *XPac* also reports symmetry operators applied to generate each molecule within each SC, which can be helpful to identify them within a large set of structures and to compare with other programs such as *PLATON* or *PIXEL*. *XPac*, like *CrystalCMP*, requires a search fragment to be defined, while *COMPACK* establishes

corresponding atoms automatically by comparing atom types and connectivity. In both *COMPACK* and *XPac*, the requirement for threshold judgements during the mapping of molecules means that results depend on the chosen tolerances.

Application of *CrystalCMP*

The 10 unique non-H atoms identified in Scheme 1 were applied as the search fragment within *CrystalCMP*. Since this fragment omits atoms C12/C13/C15/C16, the comparisons are not affected by the relative rotation of the phenyl ring (torsion angle τ_1). The *CrystalCMP* dendrogram and accompanying similarity matrix for all 96 structures is included in the ESI. Groups of similar structures emerge, including one particularly large group, as shown in Fig. 2. In this group, 15 structures are clustered at $PS_{AB} \leq 5$ (links shaded dark or light green), and 5 further structures (**7**, **33**, **34**, **37**, **41**) are linked to the group with larger PS_{AB} values (links shaded orange). Structures **7** and **37** are quite closely similar to each other (PS_{AB} = 2.8882), but less closely related to the rest of the group. Inspection shows that most of these structures have directly comparable unit cells in space group $P2_1/n$ (see ESI). For **41**, the unit-cell volume is doubled on account of ordering of the solvent molecules (pyrrolidine-1-carbonitrile), but the SLFZ molecules alone are described by the smaller unit cell in $P2_1/n$, as for the rest of the group. Hence, 18 of the structures (excluding **7** and **37**) constitute a 3-D isostructural group, denoted Group 1 in Table 2. Although the isostructurality is clear on visual inspection (Fig. 3), there is considerable variation in the unit-cell parameters amongst the group. The unit-cell volume (halved for **41**) ranges from 1494 (**28**) to 1797 Å^3 (**33**) due to accommodating different solvent molecules. Thus, the SLFZ framework in this group shows considerable geometrical flexibility and *CrystalCMP* is effective to highlight the isostructurality despite these metric differences. All structures in Group 1 are co-crystals, and the co-former molecules are not involved in conventional hydrogen-bonding with SLFZ (*i.e.* they are principally "space filling").

Table 2. 3-D isostructural groups identified amongst the multi-component SLFZ structures. Consistent results are obtained using CrystalCMP, COMPACK and XPac.

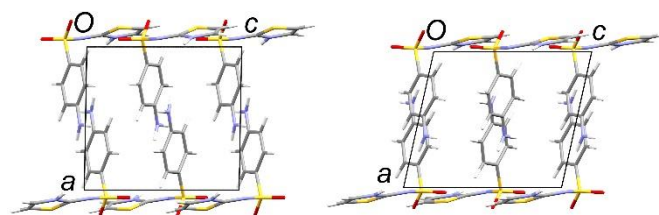|		
Group 1	{ **8, 11, 12, 13, 14, 15, 18, 21, 22, 28, 29, 30, 31, 32, 33, 34, 35, 41** }	Co-crystals
Group 2	{ **7, 37** }	Co-crystals
Group 3	{ **56, 57, 58, 59** }	Co-crystals
Group 4	{ **62, 63, 72** }	Salts
Group 5	{ **16, 17, 19** }	Co-crystals
Group 6	{ **71, 85** }	Salts
Group 7	{ **69, 73** }	Salts
Group 8	{ **25, 47** }	Co-crystals
Group 9	{ **38, 52** }	Co-crystals
Group 10	{ **39, 40** }	Co-crystals
Group 11	{ **77, 79** }	Salts
Group 12	{ **66, 88** }	Salts
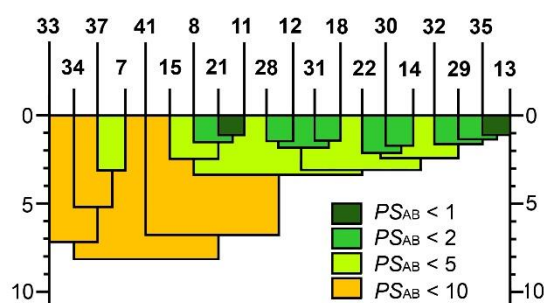
Figure 2. Extract from the *CrystalCMP* dendrogram for the multi-component SLFZ set. These structures (excluding **7** and **37**) constitute a large 3-D isostructural group (Group 1 in Table 2). Structures **7** and **37** form a separate group (Group 2), as discussed in the text.
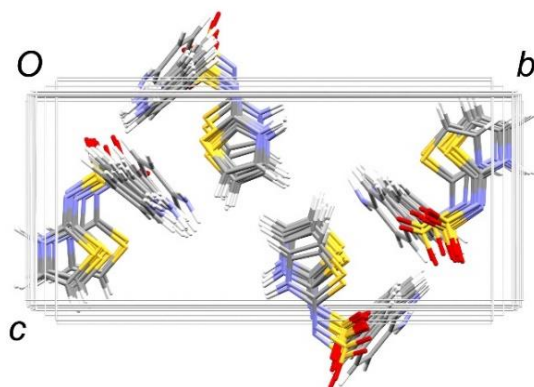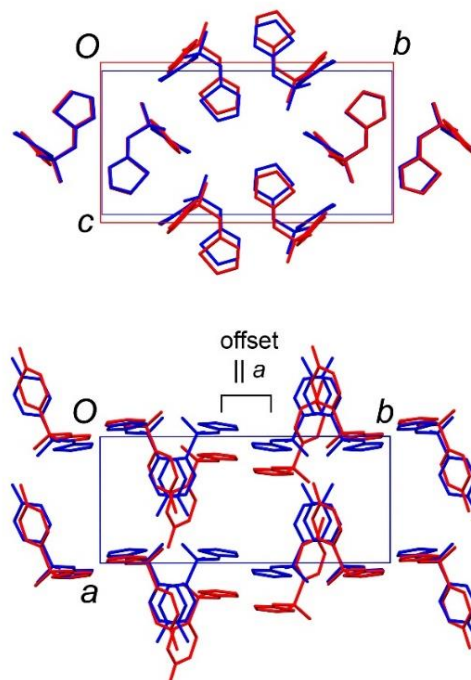

Figure 3. Overlay of the SLFZ molecules in Group 1, viewed down the *a* axis. 17 structures are shown (**41**, having a doubled unit cell, is omitted for clarity). The metric variability in the structures is apparent.

Figure 4. Arrangement of SLFZ molecules in **7** (left) and **37** (right). The structures contain identical 2-D sections in the *bc* planes, although adjacent layers are offset and the unit cell is sheared due to accommodating different solvent molecules (γ-butyrolactone in **7** and pyridazine in **37**).

Figure 5. SLFZ molecules in **37** (blue) compared to Group 1 (red; **33** is shown as a representative example). The structures look essentially identical in projection along the *a* axis (top), but adjacent columns of H-bonded pairs are offset in **33** compared to **37** (bottom) to produce a different overall hydrogen-bonded network. Pyridazine solvent molecules, omitted from these diagrams, accept hydrogen bonds from SLFZ in **37**.

The structures of **7** and **37** are useful to illustrate the sensitivity and potential ambiguity of *CrystalCMP*. Visually, **7** and **37** appear similar as a pair, although the distortion of the unit cell is quite substantial (Fig. 4). The unit-cell parameters are comparable to Group 1, but **7** and **37** are described in space group $P2_1/c$ rather than $P2_1/n$ (for the same unit-cell setting). Comparing **7** or **37** to Group 1, the structures look essentially identical when viewed along the *a* axis (Fig. 5), and they contain consistent columns of hydrogen-bonded pairs running along *a*. However, neighbouring columns are shifted relative to each other along *a*. In Group 1, the relative position of neighbouring columns is established through N11—H···O hydrogen bonds between SLFZ molecules. In **7** and **37**, these are replaced by hydrogen bonds to the solvent molecules (γ-butyrolactone in **7** and pyridazine in **37**), and the SLFZ molecules instead form O···S12 interactions.[47,48] The geometrical difference between the molecular positions is subtle, but the difference in hydrogen bonding is clearly significant, and identifies **7** and **37** as a separate group (Group 2). This conclusion is subsequently supported by results from *COMPACK* and *XPac* (see below).

Visual inspection of the other groups in the *CrystalCMP* dendrogram identifies 3-D isostructurality as listed in Table 2. Isostructural groups exist for both co-crystals and salts, but there are no mixed groups. Some further observations can be made in relation to the methodology. In the dendrogram, **38**, **52** and **50** are linked at $PS_{AB} \approx 12$. These structures resemble Group 1/Group 2 in that **38** and **52** are isostructural, but **50** is subtly different. As for Group 1/Group 2, identical 1-D columns exist in all three structures along the *a* axis, but neighbouring columns in **50** are shifted by ½*a* compared to the other two. Again, this is driven by the occurrence of N11—H···O hydrogen bonds between SLFZ molecules in **50**, which are replaced by N—H···solvent hydrogen bonds in **38** and **52**. Hence, **50** is not included in Group 9. For Group 12, comprising **66** and **88**, the structures are clearly isostructural on visual inspection, but

their similarity measure (PS_{AB} = 14.9780) is significantly larger than some of the cases deemed not to be isostructural. Given the clear visual similarity between **66** and **88**, the value of PS_{AB} is surprisingly high, and could indicate that corresponding molecules may not be appropriately mapped. In general, the geometrical PS_{AB} measure is clearly helpful to identify cases of potential 3-D isostructurality, but it is difficult to select a consistent cut-off value for fully automated grouping of the SLFZ set.

Application of *COMPACK*

Using COMPACK, it was found that distance/angle tolerances of 30%/30° (extended from the default 20%/20°) were required to identify the established cases of 3-D isostructurality in Table 2. Applying a full 96 *vs* 96 comparison with these tolerances followed by automatic grouping reproduces Table 2 exactly, including separation of Groups 1 and 2 (see ESI). Hence, *COMPACK* is marginally more convenient than *CrystalCMP* for identifying 3-D isostructural groups consistent with visual expectations. Within Group 1, however, not all pairwise matches are made at the 15-molecule level. In particular, **33** and **34** fail to match fully with several other structures in the group, indicating that their similarity with the group is close to the upper threshold for acceptance. This is consistent with the relatively larger unit-cell volume of **33** and **34** (see ESI) and also with the fact that both are linked to Group 1 at a higher PS_{AB} value in the *CrystalCMP* dendrogram (Fig. 2). To achieve a complete 15-molecule match between all pairs of structures in Group 1, it is necessary to increase the *COMPACK* tolerances to 45%/40°. However, such a liberal tolerance fails to distinguish between Groups 1 and 2, so it is again difficult to identify one set of *COMPACK* tolerances that would produce all pairwise matches consistent with Table 2.

Although *COMPACK* automatically groups structures having 15-molecule similarity, it is a substantial manual task to distil the information for sub-structure similarity. An example has been published for 50 structures containing carbamazepine.[49] For the SLFZ set, 2234 out of 4560 pairwise comparisons identify some match beyond the kernel molecule, so a fully comprehensive description of the *COMPACK* output is impractical. The discussion below is restricted to a few illustrative examples.

Considering pairwise matches down to the 9-molecule level yields only a few new links between structures in addition to the groups identified in Table 2. An interesting methodological feature emerges, however. Structure **48** shows 13-molecule similarity with Group 3, while **64** matches the same group at the apparently less similar 8-molecule level. Visual inspection shows that both matches actually correspond to the same structural feature, which is a 2-D hydrogen-bonded layer (Fig. 6). For **64** versus **56**, one clearly corresponding molecule within the layer just fails to match at the 30%/30° tolerance level (so the match essentially involves 9 molecules), but the remaining difference in the number of matched molecules is not due to tolerances. Rather, it is a consequence of the relative positions of the common SLFZ layers. In **48**, they are well separated due to inclusion of 18-crown-6 and acetonitrile in the multi-

component structure. As a result, 13 of the 15 SLFZ molecules in the initial cluster built for **48** belong to the common 2-D layer, and the different relative positions of the layers compared to Group 3 is revealed by only 2 mismatched molecules (Fig. 7, top). For **64**, the common SLFZ layers are in direct contact, and only 8 molecules in the initial cluster around the kernel molecule belong to the common 2-D layer. Now the difference between layers is revealed by 6 mismatched molecules in neighbouring layers (Fig. 7, bottom). This example highlights that it is not straightforward to interpret the substructure information generated by *COMPACK*, or even to state immediately that a greater number of matched molecules corresponds to a higher degree of structural similarity. Although it is possible to vary the size of the initial cluster in *COMPACK*, this type of discrepancy will remain in situations where common sub-structure motifs are arranged in significantly different ways. Multi-component structures will be more susceptible to such effects because the target molecules are likely to be dispersed more widely to accommodate the partner molecules.
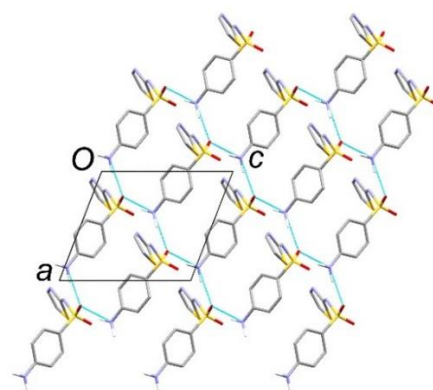

Figure 6. Projection onto the plane of the common 2-D layer of SLFZ molecules identified in the structures of Group 3, **48** and **64** (structure **48** is shown). The dashed lines indicate N—H···O hydrogen bonds.
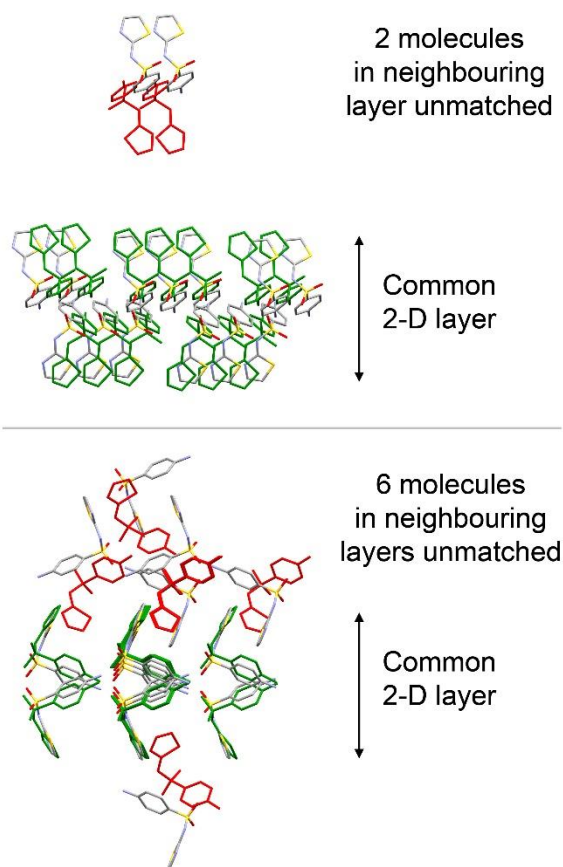
Figure 7. 15-Molecule clusters in **48** (top) and **64** (bottom), both compared to **56**. Matching molecules are coloured green and unmatched molecules are red. Solvent molecules ((18-crown-6)/acetonitrile in **48** and 2-methyl-2-imidazoline in **64**) are omitted.

Amongst the other structures matched at the 9-molecule level or greater in *COMPACK*, a polytypic relationship is identified between polymorphs **3p**, **4p** and **5p** (Table 1), whereby consistent 2-D layers lie in the (100) planes for **3p** and **4p** and in the (10–1) planes for **5p** (Fig. 8). For **4p**, the layers are stacked by translation along the *a* axis (*AAAA* stacking pattern), while in **5p**, every second layer is mirrored perpendicular to the *b* axis (*ABAB* stacking pattern). Polymorph **3p** shows an intermediate *AABB* pattern. The various pairwise matches between **3p/4p/5p** in *COMPACK* range between 9 and 13 matched molecules. This is clearly helpful to draw attention to the similarity between the structures, but manual inspection is still required to extract details of the polytypism.
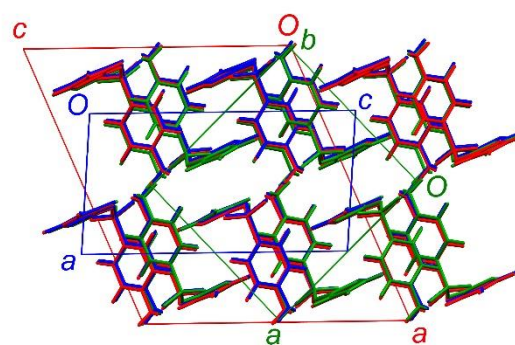

Figure 8. Polytypic relationship between the structures of **3p** (red), **4p** (blue) and **5p** (green). The structures share common 2-D layers (horizontal), but have different stacking sequences.

Application of *XPac*

A full 96 *vs* 96 comparison was carried out using *XPac*, with "high" tolerances (δ_{ang} = 12, δ_{dhd} = 18°). The 3-D SCs identified within the set are consistent with the 3-D isostructural groups listed in Table 2. Most pairwise comparisons within each group yielded a 3-D SC, except for **28** *vs* **33** in Group 1, which appears just to exceed the tolerance limits and returns 2-D similarity. Groups 1 and 2 are distinguished at the applied tolerance level, returning a common 2-D SC in the *ac* planes (a single layer of hydrogen-bonded pairs, running vertically in Fig. 5). Comparison of the isostructural **39** and **40** also returns a 2-D SC, but this is not due to tolerances. Rather, it is a reflection of the cluster building process, similar to that described for *COMPACK*. Structures **39** and **40** show a long *c* axis (~39 Å), which means that the generated clusters do not contain any molecules related by translation along *c*. The problem can be eliminated by increasing the initial cluster size, but again it raises a methodological question of how an initial cluster of molecules might best be defined without manual intervention.

Across the whole structure set, common 2-D SCs are generally restricted to groups of only two or three structures and there are no 2-D SCs that are observed especially frequently. One example links Groups 8 and 10, which contain a common 2-D SC comprising SLFZ molecules linked by N—H···O hydrogen bonds into polar layers (Fig. 9). In **25/47**, the SLFZ molecules in neighbouring layers are linked by N—H···N hydrogen bonds forming an inversion-symmetric $R_2^2(8)$ motif (discussed further below). The structure adopted by **39/40** is more complex, showing alternating polar and non-polar layers (Fig. 9). An $R_2^2(8)$ motif is again found between neighbouring layers, but with C_2 symmetry rather than inversion symmetry.
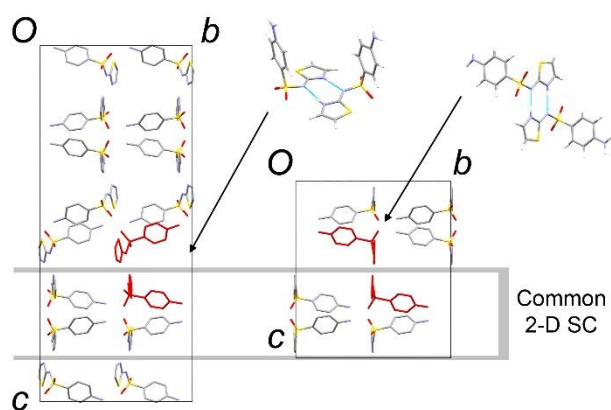
Figure 9. Common 2-D supramolecular construct in the structures of Group 10 (left; **39** shown) and Group 8 (right; **47** shown). The molecules linked by N—H···N $R_2^2(8)$ motifs discussed in the text are highlighted. Solvent molecules (1,4-dioxane in **47** and propionitrile in **39**) are omitted.

Transferable 1-D SCs are more common within the set. For example, the arrangement along the *a* axis of **4p** is built from N—H···O hydrogen bonds between SLFZ molecules related by translation (Fig. 10). *XPac* identifies this 1-D SC in 14 co-crystals, 7 salts and one other structure, totalling *ca* one quarter of the multi-component crystals. An identical arrangement of hydrogen-bonded aminobenzene rings is seen along the *a* axis of the Group 1 structures, plus two other co-crystals (**20**, **37**) and two salts (**70**, **84**), again totalling *ca* one quarter of the multi-component set. Hence, in total, almost one half of the multi-component crystals adopt this hydrogen-bonding arrangement. The two 1-D SCs are geometrically different because the N—H···O hydrogen bonds are formed either by H10 (in **4p**; Fig. 10(a)) or H11 (in Group 1; Fig 10(b)), so that the direction of the translation relative to the SLFZ molecule is different. A closer look at some of the structures reveals the possibility for a subtle change in hydrogen bonding within these 1-D SCs. For example, the N—H10 bond in **2** points clearly at N12 rather than O11 (Fig. 11). In some of the salts (**75**, **83**, **86**), the amino group is protonated, and the NH_3^+ group clearly interacts with both O11 and N12. It is perhaps to be expected that these transferable SCs should be built from hydrogen bonds, but the structure set also contains other 1-D SCs that are not based on hydrogen bonding, *e.g.* see Fig. 12.
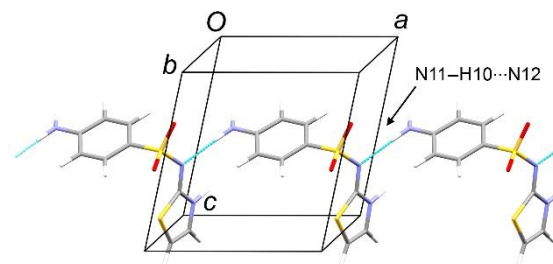
(a)
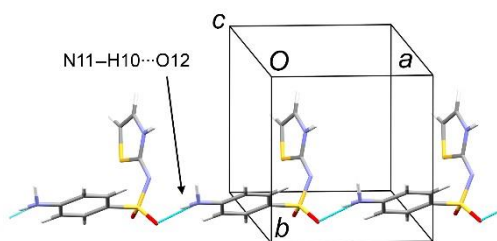
(b)
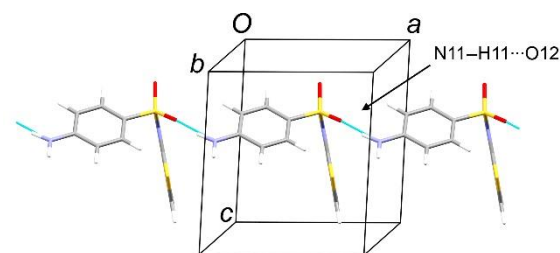
Figure 10. 1-D SC built from N—H···O hydrogen bonds between SLFZ molecules related by translation: (a) along the *a* axis in **4p**; (b) along the *a* axis in Group 1 (**8** is shown).

Figure 11. 1-D SC in **2**, forming N—H···N hydrogen bonds between translated SLFZ molecules.

To summarise the extensive *XPac* output, a Hasse diagram might typically be constructed, showing the relationships between SCs identified in all structures.[35,50,51] A complete diagram for the SLFZ set would be extraordinarily complex, however, and the largely manual task of constructing it is forbidding. Details of the *XPac* comparison between the polymorphs and multi-component structures (5 *vs* 91) are included in the ESI. Further description of the *XPac* output is deferred for a potential additional study.

Pairwise intermolecular interactions in the polymorphs and multi-component structures

A "bottom up" approach to structural similarity, which is effectively implemented in the molecule mapping processes of *COMPACK* and *XPac*, involves local matching of molecular pairs. From a chemical perspective, such an analysis of SLFZ across the structure set should provide insight into the balance between SLFZ-SLFZ and SLFZ-solvent interactions. The information is output directly by *XPac*, which identifies molecular pairs on the basis of their symmetry labels within identified SCs. Alternatively, the "Crystal Packing Feature" search within *Mercury* can be applied to the set, using a given molecular pair extracted from one of the structures. A combination of these two methods identified 15 pairwise SLFZ-SLFZ interactions that occur in the polymorphs and at least one multi-component structure (Table 3). The tolerance-based approach inevitably produces some inconsistency between the results obtained using *XPac* and *Mercury*, but Table 3 provides a fair guide to the relative frequencies of occurrence. The geometrical analysis is augmented by *PIXEL* intermolecular interaction energies calculated for each pairwise motif (see ESI).

Table 3. Frequently occurring pairwise interactions identified within the SLFZ polymorphs and at least one multi-component structure. Results are derived from *XPac* and *Mercury*. The quoted range of intermolecular interaction energies refers to different instances of the same pairwise motif within the polymorphs. A full list, with diagrams, is given in the ESI.

Motif	Found in polymorph	No. of structures	H-bond?	*PIXEL* interaction energy (kJ mol⁻¹)
A	**1p**	42	Y	−147.3 to −136.7
B	**3p**, **4p**	25	Y	−33.2 to −30.1
C	**2p**, **3p**, **4p**	17	N	−50.4 to −35.9
D	**2p**	15	Y	−39.6 to −39.5
E	**3p**, **4p**, **5p**	11	N	−31.9 to −30.1
F	**3p**, **5p**	11	N	+19.5
G	**1p**	9	N	+15.7
H	**3p**, **5p**	8	N	−22.4 to −22.2
I	**2p**, **3p**, **4p**, **5p**	8	N	−48.7 to −36.7

The pairwise motif seen most frequently is the centrosymmetric $R^2_2(8)$ dimer formed by a complementary pair of N—H···N hydrogen bonds. This was also highlighted in the study by Seaton *et al.*[36] The *PIXEL* calculations confirm that this is by far the most stabilising pairwise interaction in any of the polymorphs, and it occurs in roughly half of the multi-component structures, including the large isostructural Group 1. An alternative C_2-symmetric motif with the same $R^2_2(8)$ hydrogen-bonding pattern has been mentioned earlier (Fig. 9). Since the $R^2_2(8)$ motif requires N13 to be protonated, it is seen only in the co-crystals, and in total *ca* 80% of the co-crystals contain either the centrosymmetric or C_2-symmetric $R^2_2(8)$ motif. In the 12 structures where the $R^2_2(8)$ motif is not seen, all but one make an N—H···N/O hydrogen bond to the solvent molecule. The sole exception is **10**, where N13 makes an N—H···O interaction to another SLFZ molecule. Hence, the co-crystals are dominated by the $R^2_2(8)$ motif, but the probability of its formation is reduced where the co-former molecule is able to accept an N—H···N/O hydrogen bond.

In most cases, the pairwise interaction energies (assessed only for the polymorphs) are consistent for a given pair found in different structures, but some instances were identified where a subtle change in geometry has quite a significant effect on the resulting interaction energy. For example, the structures of **2p**, **3p**, **4p** and **5p** contain a common 1-D motif along one lattice direction, comprising two alternating SLFZ-SLFZ pairwise interactions (Fig. 12): (i) a centrosymmetric "closed" dimer involving face-to-face contact between thiazole rings, and (ii) a centrosymmetric pair involving C—H···O interactions between aminobenzene rings. The geometries and interaction energies are consistent in **3p**, **4p** and **5p**, but **2p** shows a subtle geometrical distortion that affects both interactions. For the thiazole-thiazole pair, a greater degree of face-to-face overlap in **2p** gives a larger repulsion term in the *PIXEL* energy and changes the total interaction energy from *ca* −46 to −38 kJ mol⁻¹. For the C—H···O interaction, the change in geometry in **2p** is visually more subtle, but the centroid-centroid distance

decreases by *ca* 0.2 Å and the total interaction energy changes from *ca* −48 to −37 kJ mol⁻¹. This example illustrates that the premise of transferable pairwise motifs, each with a well-defined interaction energy, must be viewed flexibly.
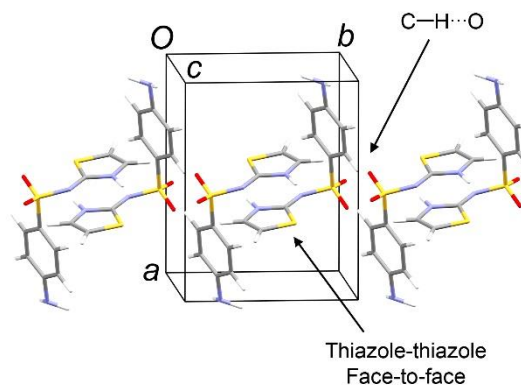
Figure 12. Thiazole-thiazole and C—H···O dimer interactions produce a 1-D supramolecular construct common to the structures of **2p**, **3p**, **4p** and **5p** (**5p** is shown).

Conclusions

The new crystallographic data presented in this paper, combined with existing structures in the CSD, establishes a set of 96 crystal structures containing the active pharmaceutical ingredient (API) sulfathiazole (SLFZ). This is one of the largest groups of crystal structures currently available for any API, providing an unusually broad view of its solid-form landscape. Identifying and describing structural similarity in this extensive set is challenging. This paper has focussed on available programs to assess geometrical similarity: *CrystalCMP*, *COMPACK* and *XPac*. Each program provides valuable results, but they depend on the applied metric measures/tolerances, and it remains difficult and time-consuming to synthesise the output to yield consistent and coherent conclusions, particularly regarding sub-structure similarity. Some aspects of the methodology also seem specifically less suitable for multi-component structures.

For the SLFZ set, some confident conclusions can be drawn. For example, 3-D isostructural groups amongst the multi-component structures are robustly established. Some transferable supramolecular constructs have also been shown, although a comprehensive overview for the whole structure set is still to be addressed. Common pairwise motifs are identified in the polymorphs and multi-component structures, some of which are based on conventional hydrogen bonding, and some of which are not. Although *PIXEL* calculations confirm that frequently occurring pairwise motifs are generally quite strongly stabilising, some less stabilising and even destabilising pairs are transferred, and there are numerous more stabilising interactions in the polymorphs that are not seen in the multi-component structures. Hence, there is no straightforward correlation between interaction energy and transferability of a given pairwise motif between the polymorphs and multi-component structures.

ARTICLE **Journal Name**

There is undoubtedly a great deal more knowledge to be extracted from the SLFZ structure set. A planned subsequent paper will augment this geometrical study with a complementary topological analysis of hydrogen bonding. Many more questions might be considered. For example, can the structure set reveal why SLFZ should be so prolific in forming multi-component crystal forms? Is it a quantifiable function of its shape and/or propensity to form H-bonded networks, or is it simply proportional to the time that has been spent looking? What can be learned about the likelihood of SLFZ forming a multi-component crystal with a given solvent/partner molecule? Are 91 known multi-component structures sufficient to make meaningful conclusions, or do we need more? These types of questions are directly relevant to the practical task of "de-risking" the solid-form selection process, for example in pharmaceutical production. It is hoped that the SLFZ set will be valuable in this and other similar contexts.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Notes and references

1 Y. A. Abramov, ed., *Computational Pharmaceutical Solid State Chemistry*, Wiley, Hoboken, New Jersey, 2016.
2 J. Bernstein, *Polymorphism in Molecular Crystals*, Oxford University Press, Oxford, 2nd edn., 2020.
3 A. D. Bond, *Pharmaceutical Crystallography: A Guide to Structure and Analysis*, Royal Society of Chemistry, Cambridge, 2019.
4 H. G. Brittain, *J. Pharm. Sci.*, 2012, **101**, 464-484.
5 S. R. Byrn, R. R. Pfeiffer and J. G. Stowell, *Solid-State Chemistry of Drugs*, SSCI, Inc., West Lafayette, Indiana, USA, 2nd edn., 1999.
6 S. Datta and D. J. W. Grant, *Nat. Rev. Drug Discovery*, 2004, **3**, 42-57.
7 R. Hilfiker, ed., *Polymorphism: in the Pharmaceutical Industry*, Wiley-VCH, Weinheim, 2006.
8 A. S. Myerson, D. Erdemir and A. Y. Lee, eds., *Handbook of Industrial Crystallization*, Cambridge University Press, Cambridge, 2019.
9 P. H. Stahl and C. G. Wermuth, eds., *Handbook of Pharmaceutical Salts: Properties, Selection and Use*, Wiley-VCH, Weinheim, 2002.
10 H. Bronstein, C. B. Nielsen, B. C. Schroeder and I. McCulloch, *Nat. Rev. Chem.* 2020, **4**, 66-77.
11 C. Bissantz, B. Kuhn and M. Stahl, *J. Med. Chem.* 2010, **53**, 5061-5084.
12 A. K. Nangia and G. R. Desiraju, *Angew. Chem. Int. Ed.* 2019, **58**, 4100-4107.
13 A. M. Reilly, R. I. Cooper, C. S. Adjiman, S. Bhattacharya, A. D. Boese, J. G. Brandenburg, P. J. Bygrave, R. Bylsma, J. E. Campbell, R. Car, D. H. Case, R. Chadha, J. C. Cole, K. Cosburn, H. M. Cuppen, F. Curtis, G. M. Day, R. A. DiStasio, A. Dzyabchenko, B. P. van Eijck, D. M. Elking, J. A. van den Ende, J. C. Facelli, M. B. Ferraro, L. Fusti-Molnar, C. A. Gatsiou, T. S. Gee, R. de Gelder, L. M. Ghiringhelli, H. Goto, S. Grimme, R. Guo, D. W. M. Hofmann, J. Hoja, R. K. Hylton, L. Iuzzolino, W. Jankiewicz, D. T. de Jong, J. Kendrick, N. J. J. de Klerk, H. Y. Ko, L. N. Kuleshova, X. Y. Li, S. Lohani, F. J. J. Leusen, A. M. Lund, J. Lv, Y. M. Ma, N. Marom, A. E. Masunov, P. McCabe, D. P. McMahon, H. Meekes, M. P. Metz, A. J. Misquitta, S. Mohamed, B. Monserrat, R. J. Needs, M. A. Neumann, J. Nyman, S. Obata, H. Oberhofer, A. R. Oganov, A. M. Orendt, G. I. Pagola, C. C. Pantelides, C. J. Pickard, R. Podeszwa, L. S. Price, S. L. Price, A. Pulido, M. G. Read, K. Reuter, E. Schneider, C. Schober, G. P. Shields, P. Singh, I. J. Sugden, K. Szalewicz, C. R. Taylor, A. Tkatchenko, M. E. Tuckerman, F. Vacarro, M. Vasileiadis, A. Vazquez-Mayagoitia, L. Vogt, Y. C. Wang, R. E. Watson, G. A. de Wijs, J. Yang, Q. Zhu and C. R. Groom, *Acta Cryst.* 2016, B**72**, 439-459.
14 P. A. Wood, T. S. G. Olsson, J. C. Cole, S. J. Cottrell, N. Feeder, P. T. A. Galek, C. R. Groom and E. Pidcock, *CrystEngComm*, 2013, **15**, 65-72.
15 R. N. Costa, D. Choquesillo-Lazarte, S. L. Cuffini, E. Pidcock and L. Infantes, *CrystEngComm*, 2020, **22**, 7460-7474.
16 D. Loveland, B. Kailkhura, P. Karande, A. M. Hiszpanski and T. Y. J. Han, *J. Chem. Inf. Model.* 2020, **60**, 6147-6154.
17 G. X. Sun, Y. D. Jin, S. Z. Li, Z. C. Yang, B. M. Shi, C. Chang and Y. A. Abramov, *J. Phys. Chem. Lett.* 2020, **11**, 8832-8838.
18 A. R. Hill, P. Cubillas, J. T. Gebbie-Rayet, M. Trueman, N. de Bruyn, Z. al Harthi, R. J. S. Pooley, M. P. Attfield, V. A. Blatov, D. M. Proserpio, J. D. Gale, D. Akporiaye, B. Arstad and M. W. Anderson, *Chem. Sci.* 2021, **12**, 1126-1146.
19 J. Nyman and S. M. Reutzel-Edens, *Faraday Discussions*, 2018, **211**, 459-476.
20 C. X. Zhao, L. J. Chen, Y. Che, Z. F. Pang, X. F. Wu, Y. X. Lu, H. L. Liu, G. M. Day and A. I. Cooper, *Nature Commun.* 2021, **12**.
21 N. Artrith, K. T. Butler, F. X. Coudert, S. Han, O. Isayev, A. Jain and A. Walsh, *Nature Chem.* 2021, **13**, 505-508.
22 J. J. Devogelaer, H. Meekes, P. Tinnemans, E. Vlieg and R. de Gelder, *Angew. Chem. Int. Ed.* 2020, **59**, 21711-21718.
23 A. P. Frade, P. McCabe and R. I. Cooper, *CrystEngComm*, 2020, **22**, 7186-7192.
24 A. Y. T. Wang, R. J. Murdock, S. K. Kauwe, A. O. Oliynyk, A. Gurlo, J. Brgoch, K. A. Persson and T. D. Sparks, *Chem. Mater.* 2020, **32**, 4954-4965.
25 D. S. Hughes, M. B. Hursthouse, T. Threlfall and S. Tavener, *Acta Cryst.* 1999, C**55**, 1831-1833.
26 M. B. Hursthouse, D. S. Hughes, T. Gelbrich and T. L. Threlfall, *Chem. Central J.* 2015, **9**, 1.
27 T. Gelbrich, D. S. Hughes, M. B. Hursthouse and T. L. Threlfall, *CrystEngComm*, 2008, **10**, 1328-1334.
28 A. L. Bingham, D. S. Hughes, M. B. Hursthouse, R. W. Lancaster, S. Tavener and T. L. Threlfall, *Chem. Commun.* 2001, 603-604.
29 The next largest set of crystallographic data is likely that for the biologically active compound gossypol: B. T. Ibragimov, S. A. Talipov and Z. P. M., *Supramol. Chem.*, 1994, **3**, 147-165.
30 C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, *Acta Cryst.* 2016, B**72**, 171-179.
31 J. A. Chisholm and S. Motherwell, *J. Appl. Cryst.* 2005, **38**, 228-231.

32 C. F. Macrae, I. Sovago, S. J. Cottrell, P. T. A. Galek, P. McCabe, E. Pidcock, M. Platings, G. P. Shields, J. S. Stevens, M. Towler and P. A. Wood, *J. Appl. Cryst.* 2020, **53**, 226-235.

33 J. Rohlicek, E. Skorepova, M. Babor and J. Cejka, *J. Appl. Cryst.* 2016, **49**, 2172-2183.

34 J. Rohlicek, E. Skorepova, M. Babor and J. Cejka, *J. Appl. Cryst.* 2016, **49**, 2172-2183.

35 T. Gelbrich and M. B. Hursthouse, *CrystEngComm*, 2005, **7**, 324-336.

36 C. C. Seaton, R. R. Thomas, E. A. A. Essifaow, E. Nauha, T. Munshi and I. J. Scowen, *CrystEngComm*, 2018, **20**, 3428-3434.

37 A. L. Spek, *Acta Cryst.* 2015, C**71**, 9-18.

38 D. C. Apperley, R. A. Fletton, R. K. Harris, R. W. Lancaster, S. Tavener and T. L. Threlfall, *J. Pharm. Sci.* 1999, **88**, 1275-1280.

39 E. L. Eliel, S. H. Wilen and L. N. Mander, *Stereochemistry of Organic Compounds*, Wiley, New York, 1994.

40 J. van de Streek and M. A. Neumann, *Acta Cryst.* 2010, B**66**, 544-558.

41 A. Gavezzotti, *New J. Chem.*, 2011, **35**, 1360-1368.

42 I. Sovago, M. J. Gutmann, J. G. Hill, H. M. Senn, L. H. Thomas, C. C. Wilson and L. J. Farrugia, *Cryst. Growth Des.* 2014, **14**, 1227-1239.

43 A. D. Bond, *J. Appl. Cryst.* 2014, **47**, 1777-1780.

44 A. L. Spek, *Acta Cryst.* 2009, D**65**, 148-155.

45 A. Kálmán, M. Czugler and G. Argay, *Acta Cryst.* 1981, B**37**, 868-877.

46 A. J. Cruz-Cabeza and J. Bernstein, *Chem. Rev.* 2014, **114**, 2170-2191.

47 R. E. Rosenfield, R. Parthasarathy and J. D. Dunitz, *J. Am. Chem. Soc.*, 1977, **99**, 4860-4862.

48 X. J. Zhang, Z. Gong, J. Li and T. Lu, *J. Chem. Inf. Model.* 2015, **55**, 2138-2153.

49 S. L. Childs, P. A. Wood, N. Rodriguez-Hornedo, L. S. Reddy and K. I. Hardcastle, *Cryst. Growth Des.*, 2009, **9**, 1869-1888.

50 T. Gelbrich and M. B. Hursthouse, *CrystEngComm*, 2006, **8**, 448-460.

51 T. Gelbrich, M. B. Hursthouse and T. L. Threlfall, *Acta Cryst.* 2007, B**63**, 621-632.