



## A to Z of polymorphs related by proton transfer†

Cite this: *CrystEngComm*, 2023, 25, 2845

Amy Woods-Ryan, <sup>ab</sup> Cheryl L. Doherty <sup>a</sup> and Aurora J. Cruz-Cabeza <sup>\*b</sup>

The occurrence of tautomeric polymorphism in the Cambridge Structural Database (CSD) was established to be very rare in a previous study by A. J. Cruz-Cabeza and C. R. Groom (*CrystEngComm*, 2011, **13**, 93). A decade has now elapsed and the CSD has seen a significant increase in its total number of crystal structures, useful CSD subsets have been introduced and the CSD Python API has been developed to allow for complex data mining. Given this, we wanted to revisit tautomeric polymorphs in the CSD alongside other polymorphs related by proton transfer and compare these results with those from an in-house pharmaceutical database in order to assess their prevalence and significance for pharmaceuticals. From A (amine–imine tautomeric polymorphs) to Z (zwitterionic polymorphs), here we study different types of polymorphs related by proton-transfer in the CSD, the CSD drug subset (DrugCSD), the single component drug subset of the CSD (SDrugCSD), and the GSK small molecule crystal structure database (GSD). First, we assess the potential of compounds to exist as tautomers. Whilst 51% of compounds in the CSD are capable of tautomerism, this number increases to 73% and 70% for the SDrugCSD and the GSD respectively. Tautomerism potential is, thus, more prevalent in pharmaceuticals than in common organic compounds in the CSD. Second, in mining the CSD we identify a total of 95 families of polymorphs related by proton transfer which can then be classified into six different categories depending on the type of proton transfer observed and the ionisation of species involved. The most common of such category is that of tautomeric polymorphs followed by zwitterionic polymorphs. The rarest type of proton transfer polymorphs is that of multi-zwitterionic polymorphs where two different zwitterions of the same compound are found in two different crystal structures. Overall, 3% of polymorphic compositions in the DrugCSD are found to be related by proton transfer which, although not very common, is of relevance to pharmaceuticals and drug development due to the potential impact on physical properties. Specific examples of each of the categories are discussed with calculations of lattice energies presented and consideration of  $\Delta pK_a$  values and likelihood of proton transfer and ionisation.

Received 7th March 2023,  
Accepted 25th April 2023

DOI: 10.1039/d3ce00216k

rsc.li/crystengcomm

## Introduction

Molecules with labile protons can exist in different chemical states related by the transfer of a proton. The population for each of these states is dictated by chemical equilibrium and thus the environment in which the species exist. The transfer of a proton can occur within a compound or between compounds in the gas-phase, solution or the solid state, and it may or may not be also accompanied by the generation of charge or the re-arrangement of bonds. The most common types of molecular species related by proton transfer are prototropic tautomers and ampholytic compounds able to

exist in an overall neutral state without or with separation of charges.

Prototropic tautomerism is the interconversion of isomers of a compound *via* the movement of a proton in combination with the rearrangement of double bonds within the molecule.<sup>1</sup> Examples of prototropic tautomerism include functional tautomerism (involving a change in functional groups, *i.e.* keto–enol and enamine–imine tautomerism) and annular tautomerism<sup>2</sup> (involving prototropic tautomerism in heterocyclic ring systems) amongst others. Compounds containing an acidic and a basic group (ampholytes) may be able to exist as a neutral molecule with no separation of charges or as a neutral molecule with localised charges (zwitterion). Zwitterions are also referred to as inner salts.

In solution, tautomers exist in equilibrium and their populations are determined by the relative stability of their molecular structures which can vary as a function of temperature, solvent and pH.<sup>3–5</sup> In most cases, these factors lead to an equilibrium which strongly favours a single tautomer. Similarly, ionisable molecules of various types can

<sup>a</sup> Medicine Development and Supply, GlaxoSmithKline, GSK Medicines Research Centre, Gunnels Wood Road, Stevenage, Hertfordshire, UK

<sup>b</sup> Department of Chemistry, Durham University, South Road, Durham, UK.

E-mail: [aurora.j.cruz-cabeza@durham.ac.uk](mailto:aurora.j.cruz-cabeza@durham.ac.uk)

† Electronic supplementary information (ESI) available. See DOI: <https://doi.org/10.1039/d3ce00216k>



exist as non-ionised, protonated, deprotonated or even zwitterionic, with their speciation in solution impacted by the same factors and thus their interconversion can be considered a subset of tautomerism with proton transfer but no re-arrangement of bonds. The specific molecular species present in solution impact the physicochemical properties of compounds such as reactivity,  $pK_a$ , and even biological activity (which has implications for drug development).<sup>6–8</sup>

In the solid state, it is generally considered that the tautomeric form present in a crystal is fixed under a specific set of conditions. Unlike in solution, a dynamic equilibrium between molecular species often does not exist in the solid state. Instead, the intermolecular interactions found in the crystal can shift the tautomeric state of the compound for tautomeric states differing by up to 35 kJ mol<sup>-1</sup> in tautomeric energy (non-ionised tautomers in this case).<sup>9</sup> Different tautomers can be observed within the same crystal (co-crystallised) or within different polymorphic structures.<sup>10</sup> A rich example of tautomeric and polymorphic diversity is that of 2-thiobarbituric acid (Fig. 1) which can exist in the solid state in six polymorphs, namely its pure enol form (II), its pure keto form (I, III, V and VI) and its enol:keto form (IV).<sup>11</sup> The specific nomenclature of the various types of these polymorphs can become complex,<sup>12</sup> but they all sit under the umbrella of polymorphs related by proton transfer.

As with other types of polymorphism, polymorphs related by proton transfer can have different solid state properties. From a pharmaceutical perspective, the primary concern with this is the impact of polymorphic form on properties such as solubility, dissolution rate and bioavailability of the drug product. Further to these, the polymorphic form's morphology, bulk density and chemical stability can impact drug product manufacturability.<sup>13</sup> Several studies in the literature<sup>14–18</sup> have highlighted some difficulties associated with the control and isolation of polymorphs related by proton transfer, with many of them crystallising concomitantly. Consequently, developing

solid forms of pharmaceuticals able to tautomerise may be complex since a pure single phase is usually desired to ensure consistent quality and performance of a drug product.

In this context, the main motivation for the present work was to mine available crystallographic data to establish how common tautomerism potential is in general chemical compounds and in pharmaceuticals, and to quantify and report the occurrence of complex polymorphism related by various types of proton transfer. Whilst there has been a previous investigation on tautomerism in the CSD,<sup>9</sup> over a decade has elapsed and since the CSD has more than doubled in size to 1 million structures.<sup>19</sup> Further to this, here we look at more broad cases of polymorphism related by proton transfer including tautomeric polymorphs, zwitterionic polymorphs and other more complex systems such as salts related by single *versus* multi-proton transfer.

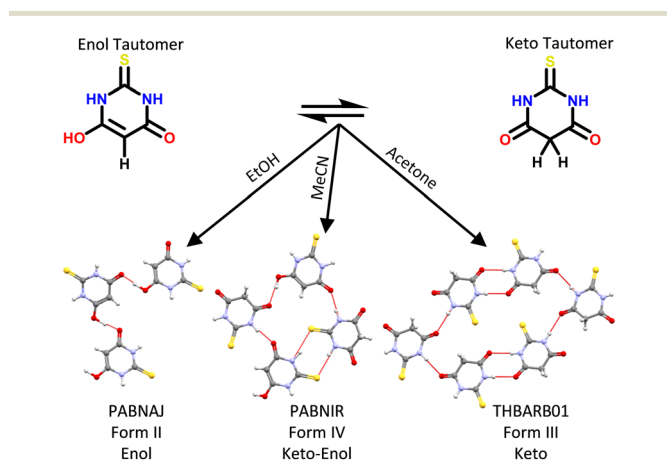
## Overview of tautomerism and the solid state

### Techniques for tautomer characterisation

Historically, hydrogen atoms (H-atoms) have been notoriously difficult to locate in crystal structures from direct measurement of conventional X-ray diffraction due to their weak scattering. As a result, H-atom positions are often implied by the geometries of other heavier atoms, provided there is no disorder in the structure. So whilst it is possible to reliably obtain H-atom positions from good X-ray diffraction data, challenges still exist for the study of tautomeric systems and mis-determinations are not uncommon if the quality of diffraction data is not optimal.<sup>20</sup>

Crystal structure determination is usually conducted at sub-ambient temperatures to minimise thermal effects. However, a change in temperature itself can cause proton migration within a crystal and impact the relevance of the structure to the room temperature form. The thermal migration of protons within a crystal structure has been documented in salt-co-crystal pairs and tautomeric molecules.<sup>18,21,91</sup> The application of both temperature and pressure has also been shown to cause proton transfer in squaric acid:bipyridine adducts together also with a change of polymorphic form.<sup>22</sup> Additionally, light-induced keto-enol transformations have been observed within the solid state, indicating that exposure to light could be important when handling and collecting structural data on potentially photo-sensitive compounds.<sup>23</sup>

For absolute confidence in crystal structure determination of complex systems, especially for molecules capable of tautomerism or materials for which proton positions are critical, orthogonal techniques such as neutron diffraction, solid state nuclear magnetic resonance (ssNMR) and infrared spectroscopy can be used to re-confirm assignment of proton positions.<sup>22,24–27</sup> Emerging techniques such as near-edge X-ray absorption fine structure spectroscopy (NEXAFS), in combination with density functional theory (DFT), have also



**Fig. 1** Three 2-thiobarbituric acid polymorphs containing different tautomeric forms. Single crystals were crystallised by evaporation from hot saturated solutions in the indicated solvents. Hydrogen bond motifs are shown in red.



been used to confirm exact H-atom positions in salt-cocrystal systems.<sup>28</sup>

DFT can be used to calculate the relative stability of tautomers and to predict the effect of different solvents on their equilibrium. For instance, the computed relative stability of sulfasalazine tautomers in DMF and water has been used to rationalise the preference for specific tautomeric forms observed in these solvents.<sup>29</sup> However, modelling can be challenging due to the sensitivity of tautomer energies to model chemistries and basis sets used. Perhaps the most significant errors in the modelling arise from the difficulty of accounting for electron delocalisation<sup>30</sup> with DFT. Such errors can be overcome and accuracy improved with the aid of higher order Hartree-Fock methods which comes at a high computational cost.<sup>31</sup>

### Prevalence of tautomerism in chemical databases (prior work)

The potential for tautomerism in various compounds from various databases has been assessed previously. For this, specific definitions of rules for tautomerisation reactions need to be derived and applied, and the types of rules used will undoubtedly impact the resulting statistics. The Chemical Structure Database (of the National Cancer Institute Computer-Aided Drug Design Group) is a collection of 103.5 million synthesised small molecular structures, 68% of which have been reported to have potential for tautomerism.<sup>32</sup> Using different tautomerisation rules and different databases, Martin<sup>8</sup> reported that 26% of 1791 marketed drugs are tautomeric and Cruz-Cabeza and Groom<sup>9</sup> that 10% of molecules in the Cambridge Structural Database (CSD) have the potential for tautomerism. The range of tautomeric potential reported from these various databases, using various tautomerisation rules, spans from just 10% to 68%.

Perhaps a more unbiased way of looking at tautomerism prevalence is by direct observation of tautomers in the crystal structures rather than by predicting tautomerism potential based on tautomerisation rules applied on chemical compounds. Even if tautomerism potential is predicted, the energy of the various tautomers will dictate whether they can be observed experimentally. Unsurprisingly, calculated relative stabilities of tautomers mirror the frequency of occurrence of tautomers in the CSD, with high-energy tautomers often not observed in the solid state. Thus, whilst potential for tautomerism may be high, only a small fraction of compounds with tautomerism potential exist in various tautomeric forms in the solid state (0.5%)<sup>9</sup> and this is dictated by the tautomer energy as well as the intermolecular interactions in the crystal. Based upon those observations, a general rule was proposed by Cruz-Cabeza and Groom: 'for different tautomers to be observed in the solid state, their relative energy must not exceed that of a strong hydrogen bond in an organic crystal'.<sup>9</sup>

From the point of view of ionisation, the prevalence of zwitterionic polymorphs and salt-cocrystal pairs in the CSD has

also been investigated in previous works.<sup>14,15,33–36</sup> A search of the CSD in 2010 found only four single component molecules (clonixin, norfloxacin, anthranilic acid, torasemide) with both neutral non-ionised and zwitterionic forms.<sup>15</sup> The majority of these were molecules of pharmaceutical interest.

### Relative stabilities of tautomers

Using computational chemistry, the relative stability of tautomers and ionised species can be estimated and used as a predictor for their experimental observation.

For neutral species related by proton transfer, the most commonly observed tautomers are typically also the most stable, except for where the energy difference between them is small (<5 kJ mol<sup>-1</sup>). In those cases, intermolecular interactions in the solid state can shift the tautomeric outcome.<sup>2</sup>

There are exceptions, however, which have reported the observation of very high-energy tautomers in the solid state. Such is the case of the enol-tautomer of barbituric acid which, despite being highly metastable (53.7 kJ mol<sup>-1</sup>), is found in the most stable overall polymorphic form. This stable polymorph of barbituric acid was notoriously difficult to discover and was only produced relatively recently by ball-mill grinding.<sup>37,38</sup> The ability of some molecules to tautomerise can be used to our advantage and a specific tautomer targeted and crystallised. For example, Epa *et al.* used supramolecular cocrystal design to selectively crystallise and isolate the high-energy tautomer of 1-deazapurine.<sup>39</sup> This ability to deliberately stabilise a metastable tautomer is important and provides the opportunity to isolate novel solid state forms containing different tautomers and displaying different physical properties.

For neutral vs. charged species related by proton transfer, energy differences between molecular species can be much larger (in the order of hundreds of kJ mol<sup>-1</sup> in some cases). Some general trends can be assumed. For example, in the gas-phase, a zwitterion or charged pair of molecules will always have a much higher energy than their non-ionised counterparts, due to the unfavourable separation of charges. This relative stability can often change in solution or the solid state due to the stabilisation of species brought about by coulombic interactions in charged systems. Many of the common amino acids exist as zwitterions in solution and the solid state but as neutral in the gas phase (*i.e.* glycine<sup>40</sup>).

## Methods

### Datasets

Refcode lists (entire CSD, CSD drug subset & CSD single component drug subset) were extracted from the Cambridge Structural Database (CSD) version 5.42 using ConQuest.

All datasets were created using the 'Best Hydrogens List' subset which removes duplicates and redeterminations. Additionally, the entire CSD dataset was further refined to only allow for compounds containing any combination of a subset of atom types (H, D, C, N, O, F, P, S, Cl, Br, I) and only include organic, non-polymeric structures with 3D co-



ordinates determined and no errors. Application of these filters resulted in 293 984 entries for the entire CSD (CSD), 7787 entries for the CSD drug subset (DrugCSD) and 729 entries for the CSD single component drug subset (SDrugCSD). Additionally, the GSK structural database (GSD) was assessed with duplicates and redeterminations removed and contained 1820 entries. The GSK database of small-molecule crystal structures contains X-ray crystallography data obtained over the past 40 years by GSK and legacy companies. The structures are not limited to marketed drugs and the GSD contains molecules from all phases of development, including non-API molecules such as intermediates or impurities. Conversely, the DrugCSD consists of small-molecule crystal structures containing only approved drug molecules. The SDrugCSD contains crystal structures of pure neat drug compounds, thus not multi-component systems.

### Tautomer generation

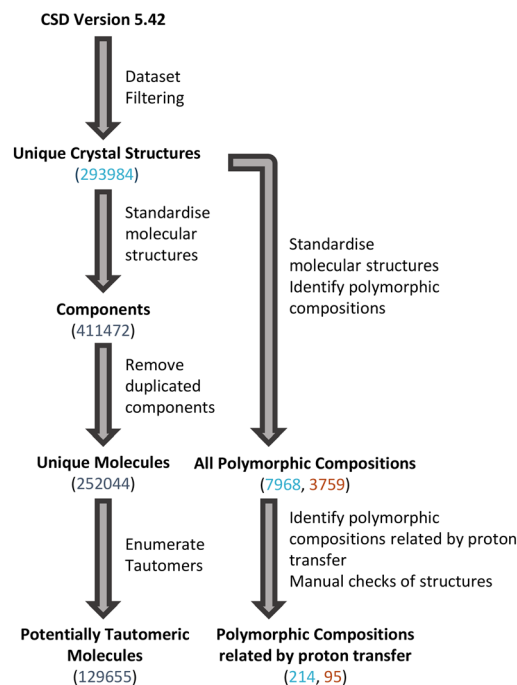
The CSD Python API<sup>41</sup> was utilised to loop through each crystal entry of each dataset and separate the structure into individual molecular components. The resultant molecules were standardised (bond types assigned and H-atoms positions added if missing). Tautomers for each unique molecule were generated using the 'Tautomer Enumerator' function of the RDKit<sup>42</sup> module in Python. The tautomer enumeration rules are based upon those described by Sitzmann *et al.*<sup>32</sup> A maximum of 50 tautomers were enumerated for each molecule. Enumeration was conducted twice for each dataset to assess differences in the tautomers generated if bond and sp<sup>3</sup> stereochemistry was allowed to be lost or not during enumeration. A molecule was deemed to be potentially tautomeric if more than one tautomer was enumerated during the tautomer generation step.

### Identification of polymorphs with protonation diversity

A scripted approach comparing canonical SMILES, SMILES, standard InChI and non-standard InChI (FixedH) strings was used to identify 'same' molecules where multiple tautomers had been observed in the solid state. The use of these identifiers enabled determination of multiple types of tautomerism including forms where charges are in distinct positions in different structures, multi-component systems that can exist as salts and cocrystals, zwitterionic polymorphs, and tautomeric polymorphs.

### Processing of polymorphic structures

The scripted nature of searches in large databases can result in some incorrectly identified structures beyond the potential for structures with mis-determined H-atom positions. Therefore, additional manual checks were conducted to identify and eliminate any groups of structures which were not real sets of polymorphs related by proton transfer. These checks (performed in Mercury<sup>43</sup>) included visualisation of structures, Mogul geometry checks, crystal packing similarity



**Fig. 2** Process for identifying tautomeric molecules and polymorphs related by proton transfer. Numbers quoted are from the 'entire CSD' dataset (number of unique molecules in black with number of crystal structures in blue and number of polymorphic families in red).

and simple metadata checks. The overall process used to identify tautomeric molecules and polymorphic compositions related by proton transfer is shown in Fig. 2.

### Generation of calculated pK<sub>a</sub>

Aqueous pK<sub>a</sub> values were calculated and visualised using JChem for Excel.<sup>44</sup> The pK<sub>a</sub> for the strongest acid and strongest protonated base for relevant molecules in crystal structures of interest were recorded and the ΔpK<sub>a</sub> was calculated as in eqn (1).

$$\Delta pK_a = pK_a[\text{protonated base}] - pK_a[\text{acid}] \quad (1)$$

In some cases, additional pK<sub>a</sub> and ΔpK<sub>a</sub> values were calculated (*e.g.*, self ΔpK<sub>a</sub> and ΔpK<sub>a</sub> for second ionisations as described by Cruz-Cabeza *et al.*<sup>45</sup>).

### DFT calculations

CSD crystal structures of interest were imported into Materials Studio.<sup>46</sup> For each polymorphic pair, structures obtained at the same temperature were chosen where possible. Supercells were prepared when initial unit cell lengths were much less than 10 Å in any dimension. Each crystal structure was geometry optimised using CASTEP<sup>47</sup> with the Perdew–Burke–Ernzerhof<sup>48</sup> gradient-corrected functional (GGA PBE) with the Tkatchenko and Scheffler (TS)<sup>49</sup> scheme for dispersion correction and OTFG ultrasoft pseudopotentials. Geometry optimisations were performed using a plane wave basic cut-off energy of 600 eV.



The Brillouin zone was sampled using a  $k$ -point spacing of approximately  $0.05 \text{ \AA}^{-1}$  and a SCF tolerance of  $1.0 \times 10^{-6}$  eV per atom was used. Full cell optimisations were performed.

Separately, molecules were geometry optimised in the gas phase using finer settings and starting from molecular geometries taken from the crystal structures. Each molecule was placed in a box with the dimensions ensuring approximately  $\geq 12.5 \text{ \AA}$  free space surrounding the molecule in all directions. Geometry optimisations were performed using the same procedure as above.

After crystal and gas-phase geometry optimisations, single-point energy calculations were performed on the gas phase molecules and crystal structures with application of the many-body dispersion correction scheme (MBD\*)<sup>50</sup> and with a plane wave basic cut-off energy of 630 eV. The Brillouin zone was sampled using a  $k$ -point spacing of  $0.07 \text{ \AA}^{-1}$  and a SCF tolerance of  $5.0 \times 10^{-7}$  eV per atom. Geometry optimisation and single-point energy calculation settings were adjusted as required for good convergence.

### Lattice energy calculations

Lattice energies ( $E_{\text{Latt}}$ ) were calculated according to the eqn (2) and (3) for single component and multi-component systems, respectively.

$$E_{\text{Latt}} = (E_{\text{cell}}/Z) - E_{\text{gas}} \quad (2)$$

$$E_{\text{Latt}} = (E_{\text{cell}}/Z) - (E_{\text{gas(A)}} + E_{\text{gas(B)}}) \quad (3)$$

where  $E_{\text{cell}}$  is the total calculated energy of the unit cell,  $Z$  is the number of formula units per unit cell, and  $E_{\text{gas}}$  is the calculated single-point energy for the lowest energy tautomer of a molecule in the gas phase. For multi-component

systems,  $E_{\text{gas(A)}}$  and  $E_{\text{gas(B)}}$  are the gas phase energies for each unique molecule (molecule A, molecule B) present in the lattice.

## Results and discussion

### Tautomerism potential of molecules in the CSD

Based on a broad set of tautomerisation rules (as defined in RDKit), we found that around 52% of molecules in the CSD (129655) have the potential for tautomerisation. Interestingly, this proportion increases slightly for the DrugCSD (53%, 1446 molecules) and significantly for the SDrugCSD (73%, 387 molecules). The SDrugCSD statistics are well in line with the GSD where 70% of compounds were found to have tautomerisation potential (1467 molecules). The statistics are summarised graphically in Fig. 3. Varying the RDKit tautomer enumeration rules (to either keep or lose stereochemistry during enumeration) did not significantly impact the statistics.

The lower incidences of molecules able to tautomerise found in the CSD and DrugCSD datasets may be due to the inclusion of components such as solvents, counter-ions, and co-formers, whereas the SDrugCSD only contains neat drug molecules. Previous analyses of the CSD, DrugCSD and other pharmaceutical databases have shown that drug-like molecules are significantly larger (higher molecular weight) than the organics in the CSD, thus increasing the probability of a drug molecule containing a tautomerisable functional group.<sup>51</sup> Also of note is the fact that the GSD has a much larger proportion of 'free drug' structures (58.07%) than the DrugCSD (19.53%).<sup>52</sup> These results indicate a much higher fraction of molecules in the CSD are capable of tautomerism than initially identified by Cruz-Cabeza and Groom (10%).<sup>9</sup> Similarly, these numbers are much higher than those

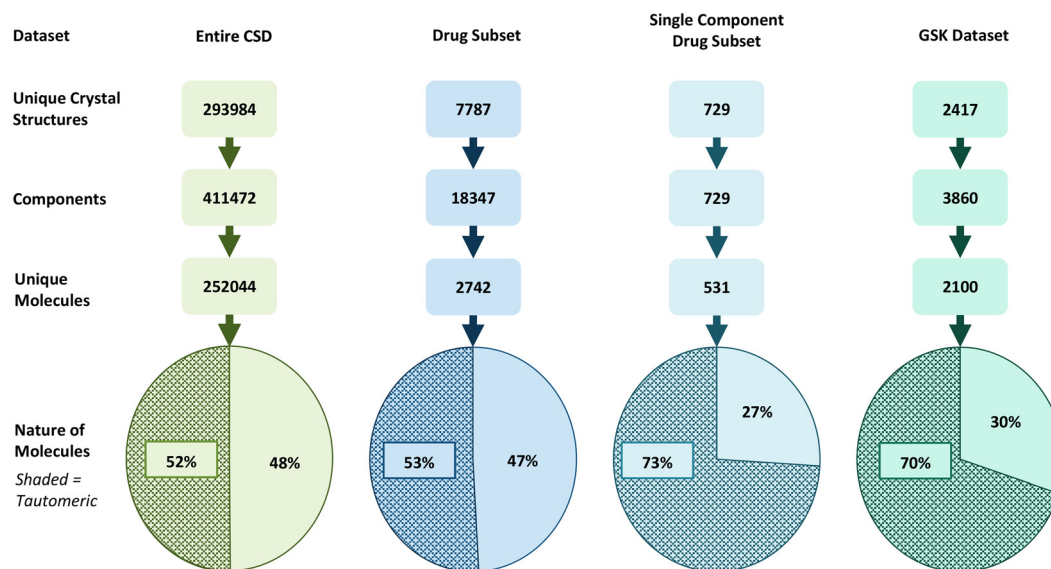


Fig. 3 Datasets of crystal structures studied here split into crystal entries, components, and unique molecules able (shaded) and unable (non-shaded) to tautomerise.



reported in marketed drugs (26%).<sup>8</sup> One reason for this discrepancy is likely to be due to the differences in the tautomer enumeration rules. For example, Cruz-Cabeza and Groom did not allow keto–enol tautomerism, which is considered in the transform rules within RDKit.<sup>32</sup>

Rather than a comparison of absolute numbers across different methods, the value of current analysis lies in the comparison of the different datasets with the same method. This comparison clearly points towards tautomerism potential being significantly more prevalent in drug compounds than in other small molecules in the CSD.

### Ampholyte potential of molecules in the CSD

Another important set of species related by proton transfer are ampholytes which may potentially exist as non-ionised and/or zwitterionic. It was found that 32% of the SDrugCSD structures contain a compound identified to have both an acidic and a basic group and therefore a  $\Delta pK_a^{\text{self}}$  could be calculated.<sup>45</sup> Only 10% of the CSD and 14% of the DrugCSD molecules (heaviest component) were found to have ampholytic potential. Of those compounds able to be ampholytic, 11%, 32% and 26% are found as zwitterions in the CSD, DrugCSD and SDrugCSD respectively with the remaining existing in the non-ionised state (Fig. 4). These statistics highlight the higher potential for self-proton transfer in drug compounds, with the crystal composition also influencing it.<sup>45</sup>

### Classification of polymorphs related by proton transfer

A well accepted definition of crystal polymorphism is that a polymorph is ‘a solid crystalline phase resulting from different arrangements of molecules in the solid state...two polymorphs will be different in crystal structure but identical in the liquid and vapour states’.<sup>53</sup> Where interconversion of molecular species is rapid in the liquid, crystals containing different tautomers or charged species of a compound (or compounds) will lead to identical population of species in

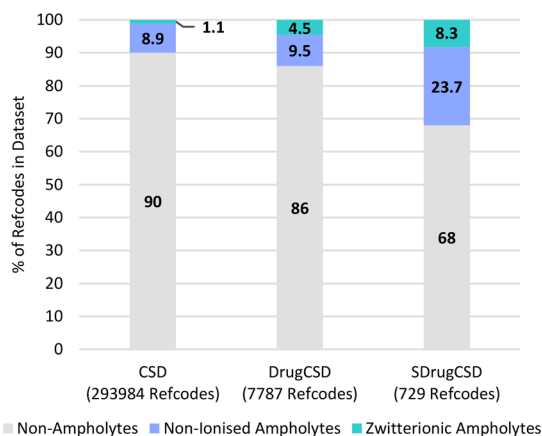


Fig. 4 Ionisation behaviour of heaviest component in the various crystal structures within the three CSD subsets studied here.

the liquid and therefore can be classified as polymorphs. Tautomeric polymorphism of this type has been referred to as desmotropy elsewhere.<sup>12,16,54–59</sup>

Polymorphs related by proton transfer may exist with very different or with nearly identical crystal packings (the main difference being the position of a H-atom). If a change in hydrogen position is accompanied with a discontinuity in the heat capacity at some specific conditions of temperature or pressure, the crystal structures shall be considered as polymorphs or phases. This type of high structural similarity polymorphism may indeed be challenging to identify, especially from X-ray data alone.<sup>60</sup>

In our search of the CSD we have identified three main umbrellas of polymorphs related by proton transfer branching into six distinct categories (Fig. 5). The three umbrellas relate to the ionisation state of the constituent components. Proton transfer may occur generating sets of polymorphs that are non-ionised or have mixed or multiple ionisations.

Under the first umbrella (no ionisation), we find tautomeric polymorphs where multiple tautomers have been isolated individually (and/or jointly) in different crystal structures. These are observed in single as well as multicomponent systems. Under the second umbrella (non-ionised and ionised pairs), we find zwitterionic polymorphs (where some structures are non-ionised whilst others are zwitterionic) as well as salt–cocrystal pairs (structures exist both with and without proton transfer between the components). Finally, under the third umbrella (multiple ionisation), we find three categories of polymorphs namely multi-zwitterionic polymorphs (where the polymorphs are all zwitterionic but the local charges are distributed in different atoms), multi-component systems with multiple proton transfers (*i.e.* monovalent and divalent salts) and structures with different protonation positions (*i.e.* salts where a proton can transfer to/from multiple different functional groups). To

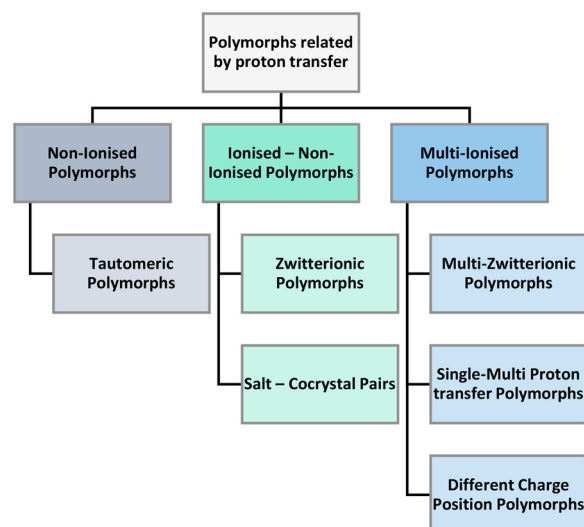


Fig. 5 Classification of polymorphs related by proton transfer.



**Table 1** Nomenclature and classification of polymorphs related by proton transfer. Key: 'A' and 'a' represent two tautomers of a compound. 'B' represents a second component, and 'N' represents additional neutral component(s). '+' and '-' represent positive and negative charges, respectively

Polymorphic type	Composition and speciation of polymorphic pairs
Tautomeric polymorphs	[A], [a] [A:a], [a:a] [A], [A:a] [A:N], [a:N] [A:a:N], [a:N] [A:N], [A:a:N]
Zwitterionic polymorphs	[A], [A <sup>±</sup> ] [A:N], [A <sup>±</sup> :N]
Salt-cocrystal pairs	[A:B], [A <sup>±</sup> :B <sup>∓</sup> ] [A:B:N], [A <sup>±</sup> :B <sup>∓</sup> :N]
Multi-zwitterionic polymorphs	[A <sup>±</sup> ], [a <sup>±</sup> ] [A <sup>±</sup> :N], [a <sup>±</sup> :N]
Single-multi proton transfer polymorphs	[A <sup>±</sup> :B <sup>∓</sup> ], [A <sup>2+</sup> :B <sup>2-</sup> ] [A <sup>±</sup> :B <sup>∓</sup> :N], [A <sup>2+</sup> :B <sup>2-</sup> :N]
Different charge position polymorphs	[A <sup>±</sup> :B <sup>∓</sup> ], [a <sup>±</sup> :B <sup>∓</sup> ] [A <sup>-</sup> :B <sup>+</sup> ], [a <sup>-</sup> :B <sup>+</sup> ] [A <sup>+</sup> :B <sup>-</sup> :N], [a <sup>+</sup> :B <sup>-</sup> :N] [A <sup>-</sup> :B <sup>+</sup> :N], [a <sup>-</sup> :B <sup>+</sup> :N]

aid the understanding of these polymorphs, examples of pairs of each category and their composition are given in Table 1 where the wealth of compositions and ionisation states in these systems can be appreciated.

### CSD statistics on polymorphs related by proton transfer

In our CSD analysis we identified 3759 polymorphic families, only 2.5% of which correspond to polymorphs related by proton transfer (94 families). The other CSD subsets also show similar incidences of polymorphism related by proton transfer with only 2.9% in the DrugCSD and 2.4% in the SDrugCSD (Table 2). Our analysis, however, did not identify any polymorphs related by proton transfer in the GSK database.

All polymorphic structures were analysed manually (to check for errors), and the final 95 families identified were all genuine polymorphs related by proton transfer. We expect, however, that this number is an under-representation of incidences in the CSD since some proton transfer polymorphs may have been eliminated by the redetermination analysis

**Table 2** Polymorphs related by proton transfer in each dataset

	All polymorphic families	Polymorphic families related by proton transfer <sup>a</sup>
CSD	3759	2.5% (94)
DrugCSD	374	2.9% (11)
SDrugCSD	125	2.4% (3)

<sup>a</sup> 95 unique sets of polymorphs with protonation diversity. The CSD dataset is restricted to organics only. This restriction did not apply to the DrugCSD.

algorithm which is used for the generation of the CSD subsets.<sup>61</sup>

Additionally, in some examples, proton positions have been reliably determined by means other than crystallography (*e.g.* IR spectroscopy)<sup>62</sup> or not determined at all. We note that some of these cases cannot feasibly be accounted for by our comparison algorithms of X-ray data only. For instance, two high profile GSK drugs (albendazole<sup>57</sup> and ranitidine hydrochloride<sup>63</sup>) are known to exhibit tautomeric polymorphism and a GSK development compound GSK251<sup>64</sup> exhibits zwitterionic polymorphism, but these were not identified in the GSD for these reasons. Furthermore, we know that not all polymorphs have their structures determined and deposited in the CSD and protons can be difficult to locate, further impacting the confidence in these statistics. Nonetheless, polymorphs related by proton transfer appear to occur frequently enough for them to be considered important. The distribution of polymorphic families across the different types of polymorphs related by proton transfer is given in Fig. 6 where the analysis is given per family. Of the 95 polymorphic systems (214 refcodes), 94 contained only two tautomers whilst one polymorphic family contained three tautomers<sup>16</sup> (two observed in our dataset). Tautomeric polymorphs are most common (51) followed by zwitterionic polymorphs (24) and multi-zwitterionic polymorphs (2) are the rarest. Only 9 salt-cocrystal pairs were found. The different polymorphic systems are further discussed in the sections below.

### Tautomeric polymorphs

The observed tautomeric compositions of the 51 tautomeric polymorphic systems have been summarised in Fig. 7. The majority (43) are pure systems where two tautomers of a compound (one 'a' and the other 'A') were crystallised individually in different polymorphs ([a] and [A]), with an additional 3 systems where two tautomers have been isolated individually as well as a mixture ([a], [A] and [a:A]). In 5 systems, two tautomers had been observed with one of them only seen together with the first ([a] and [a:A]). Only two of these polymorphic systems are of a hydrate or solvate and none were cocrystals with another molecule.

The number of tautomeric polymorphs in the CSD has tripled since 2010 when only 16 such pairs were identified. Additionally, there now exists polymorphic families containing three or more tautomeric forms. For example, 3-methyl-4-(4-methylbenzoyl)-1-phenyl-pyrazo-5-one has now been reported to have three tautomers in the solid state ([a], [a:A] and [α] – where a, A and α are three different tautomers of a compound). Two of these tautomers were found in our CSD analysis within two different refcode families (VIMPAJ<sup>65</sup> and ZUMXIQ<sup>16</sup>).

The crystallisation solvent itself can influence the tautomeric state in a single direction which may partially explain why so few hydrates or solvates form tautomeric polymorphs.<sup>66</sup> However, no conclusions shall be drawn on this given that hydrates and solvates are generally less



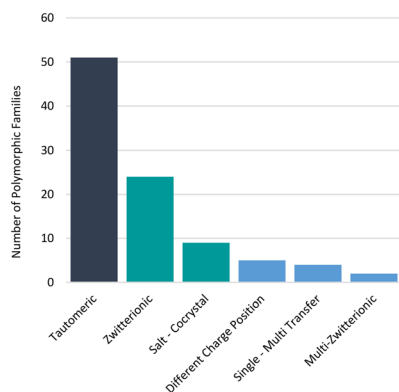


Fig. 6 Distribution of polymorphs related by proton transfer.

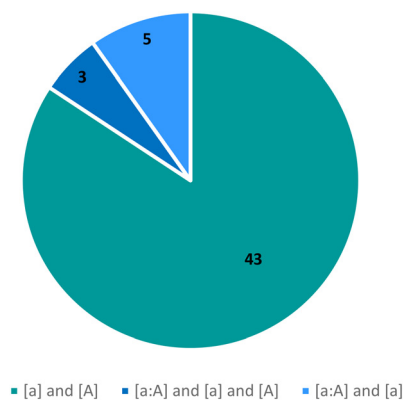


Fig. 7 The distribution of tautomers in tautomeric polymorphs.

commonly observed in the CSD<sup>52,67</sup> and the overall numbers of polymorphs related by proton transfer are small.

### Zwitterionic polymorphs

A total of 24 compounds were found to exhibit zwitterionic polymorphism. The majority are neat anhydrate structures (21, [A] and [A<sup>±</sup>]), none were hydrated and 3 were multi-component systems (cocrystals of piroxicam, [A<sup>±</sup>:N] and [A:N]).

For these compounds, the  $\Delta pK_a^{\text{self}}$  is an important parameter to determine the likelihood of a compound to be able to exist as non-ionised and as a zwitterion in the solid state. The equivalence  $\Delta pK_a^{\text{self}}$  point, where molecules containing both acidic and basic  $pK_a$  have a 50% probability of existing as either the zwitterion or non-ionised molecule, has been reported to be 4.1.<sup>45</sup> The  $\Delta pK_a^{\text{self}}$  scale can be classified into domains which describe the likelihood of zwitterion formation. In zone 1 ( $\Delta pK_a^{\text{self}} < 0.9$ ), > 99% of molecules crystallise as their non-ionised tautomers, in zone 2 ( $0.9 \leq \Delta pK_a^{\text{self}} \leq 7.4$ ) both forms are possible and in zone 3 ( $\Delta pK_a^{\text{self}} > 7.4$ ) > 99% of molecules crystallise as zwitterions. The probability of observing a zwitterion in zone 2 is described by eqn (4).

$$P_{\text{obs}}(\text{zwitterion, \%}) = 15\Delta pK_a^{\text{self}} - 12 \quad (4)$$

Given the prior knowledge and data, we would in principle expect molecules exhibiting zwitterionic polymorphism to sit in zone 2 of the  $\Delta pK_a^{\text{self}}$  scale. Interestingly, however, the majority of the zwitterionic polymorphs in our dataset had a  $\Delta pK_a^{\text{self}}$  within zone 1 instead. A further analysis of these structures reveals that most of these instances correspond to compounds where the acid and base groups sit very close to each other resulting in an intra-molecular rather than an intermolecular proton transfer (see Fig. 8 and 9). Because these groups interact, the  $\Delta pK_a^{\text{self}}$  prediction may not be as accurate as the actual experimental value, an observation which has been highlighted before.<sup>45,68,69</sup> Additionally, different tautomers of a compound will have different  $pK_a$  values, and therefore the choice of tautomer in  $pK_a$  calculations may be important.

Our results suggest that when the acid-base proton transfer for zwitterion formation occurs intermolecularly the majority of zwitterionic polymorphs have  $\Delta pK_a^{\text{self}}$  values in zone 2 with a significant proportion also in zone 1. However, when the proton transfer for zwitterion formation occurs intramolecularly, most zwitterionic polymorphs have  $\Delta pK_a^{\text{self}}$  within zone 1. No molecules exhibiting zwitterionic polymorphism had  $\Delta pK_a^{\text{self}} > 4$ , indicating that when there is a large positive  $\Delta pK_a^{\text{self}}$ , zwitterionic polymorphism does not occur, presumably due to the very large driving force to form solids with the zwitterions only. In this zone, normal polymorphism of zwitterionic compounds prevails.

Most of the structures exhibiting zwitterionic polymorphism and intra-molecular proton transfer were Schiff bases with general formula Ar-CH=N-R (frequently used in co-ordination chemistry). *Ortho*-Hydroxy derivatives of Schiff bases can exist in enol-imine, keto-enamine or zwitterionic forms depending on the location of the hydrogen in the O···H···N bond, as shown in Fig. 10.<sup>70</sup> These molecules were observed in the enol, zwitterionic and keto forms in our datasets, but the  $pK_a$  values were calculated for the enol-imine forms rather than the keto-enamine forms. Studies of the tautomerism in Schiff bases have indicated that the presence of peripheral groups enable stabilisation of metastable tautomers in the solid state.<sup>55</sup> This appears to be a much bigger factor enabling zwitterionic polymorphism in these compounds rather than the  $pK_a$  values.

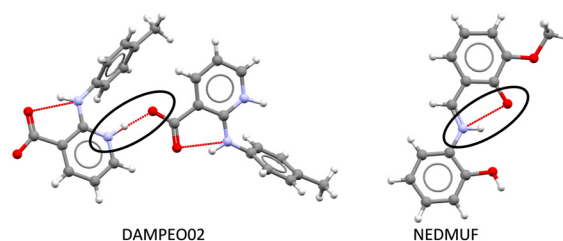


Fig. 8 Zwitterion formation via intermolecular proton transfer in DAMPEO02 (left) and via intramolecular proton transfer in NEDMUF (right).



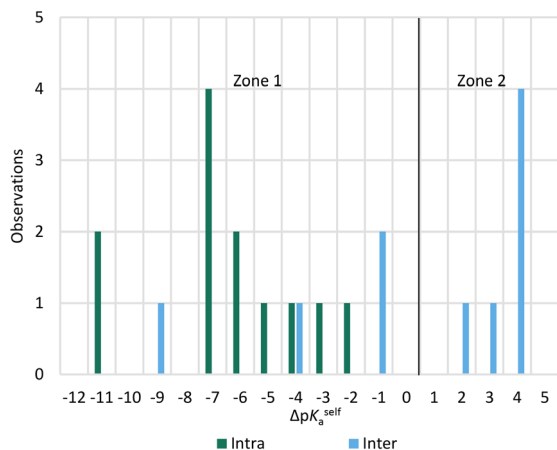


Fig. 9  $\Delta pK_a$  for molecules exhibiting zwitterionic polymorphism, split by the nature of the proton transfer in the crystal (intra or intermolecular).

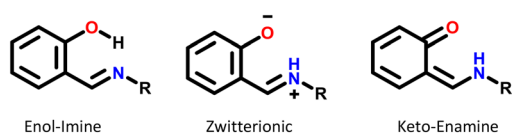


Fig. 10 Possible tautomerism in Schiff bases.

In summary, whilst there are no reliable trends to draw hard conclusions from these observations, there does appear to be a slight tendency for molecules exhibiting zwitterionic polymorphism to sit within (or close to) zone 2 when proton transfer is intermolecular and zone 1 when it is intramolecular. A  $\Delta pK_a^{\text{self}}$  of  $\sim 4$  appears to be an upper limit for observing zwitterionic polymorphism.

### Salt-cocrystal polymorphic pairs

Only 9 salt-cocrystal polymorphic pairs were found in the CSD, all of which were anhydrous forms ( $[A:B]$  and  $[A^+B^-]$ ). Cases where the salt and cocrystal have different stoichiometries were not considered to be polymorphs (e.g. TODNIM and WAFCOX, AFICEZ and AFICID) and are thus not included in this number. For the salt-cocrystal polymorphic pairs, the most acidic  $pK_a$  for the proton donating molecule, and most basic  $pK_a$  for the proton accepting molecule in each pair was calculated and used to determine the  $\Delta pK_a$ . A recent analysis of the CSD has shown that for non-solvated multicomponent systems, the point of 50% probability for salt *versus* cocrystal formation occurs where  $\Delta pK_a$  is  $\sim 1.4$  whereas for hydrates, this point is at  $\Delta pK_a \sim -0.5$ .<sup>45</sup> The distribution of  $\Delta pK_a$  for these non-solvated salt-cocrystal polymorphic pairs is consistent with that finding, with 7 of the 9 pairs having a  $\Delta pK_a$  within one unit of 1.4 (range 0.43 to 2.21). Thus, the key question when presented with a multicomponent system with a small  $\Delta pK_a$

should therefore not be whether a salt or cocrystal can form, but whether both can form.

### Multi-zwitterionic polymorphs

Only two compounds exhibiting multi-zwitterionic polymorphism were found ( $[A^{\pm}]$  and  $[a^{\pm}]$ ), making this the rarest type of proton transfer polymorphism in our search. These compounds were cinchomeronic acid (often used as a ligand) and *N*-2-hydroxyethylpiperazine-*N*-2-ethane sulfonic acid (commonly known as HEPES and used as a buffering agent). Their molecular structures are shown in Fig. 11.

Cinchomeronic acid has 8 crystal structures in its refcode family CINMER<sup>71-73</sup> with only two unique polymorphic forms (CINMER02 and CINMER04). All crystal structures of the most stable form I (four structures, which includes CINMER02) show their entries with proton transfer from the 3-substituted carboxylic acid. For form II, however, the proton position has been a matter of debate with two structures showing proton transfer from the 4-substituted carboxylic acid (CINMER04, CINMER05) and two from the 3-substituted carboxylic acid (CINMER01, CINMER03). Orthogonal approaches had to be used to confirm the proton positions for forms I (3-substituted) and II (4-substituted) respectively and established that there is no temperature-induced proton migration for the two forms.<sup>73</sup>

These disagreements in the literature and the requirement for large amounts of orthogonal data highlight how difficult it can be to accurately describe the protonation states of these molecules in the solid state and how 'incorrect' determinations can be hidden in the CSD with seemingly little warning or comment. The calculated  $pK_a$  for the two carboxylic acid groups are 3.69 for 3-substituted and 5.16 for the 4-substituted. It is intriguing that a zwitterion forms at all due to the weakly basic nature of the pyridine (calculated  $pK_a$  of the protonated base being 0.95) and the reasonably negative  $\Delta pK_a$  values of between  $-4.21$  and  $-2.74$ . The Gibbs free energy for proton transfer in aqueous state at 298 K was calculated according to eqn (5).<sup>45</sup>

$$\Delta G_{\text{ionisation,water}}^{\circ} = -5.7\Delta pK_a \quad (5)$$

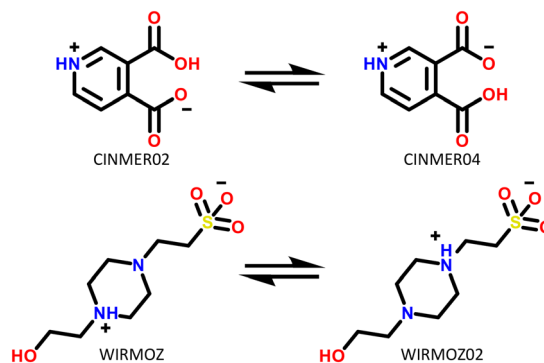


Fig. 11 Molecular structures of compounds with multi-zwitterionic polymorphic pairs, and their associated refcodes.



For both zwitterionic structures,  $\Delta G_{\text{ionisation,water}}^{\circ}$  is positive, indicating that in solution, zwitterion formation will not be spontaneous or favoured. The value is slightly larger for CINMER04, indicating that this zwitterion is the least favourable of the two in water, and corresponds with the lower  $\Delta pK_a$ . In the gas state, the zwitterionic forms are also unfavourable, however in the solid state we only see zwitterionic forms.

HEPES has 5 crystal structures in its refcode family WIRMOZ<sup>20,74,75</sup> with only two unique polymorphic forms. The two piperazine nitrogen atoms have quite different calculated  $pK_a$  values (1.62 for the N protonated in WIRMOZ, 7.34 for the N protonated in WIRMOZ02). The sulfonic acid is a reasonably strong acid with a calculated  $pK_a$  of  $-1.34$ . Therefore, the  $\Delta pK_a$  is either 2.96 or 8.68 respectively, so a zwitterion might be expected to form in either case (with lower probability for the lower  $\Delta pK_a$  as in WIRMOZ).  $\Delta G_{\text{ionisation,water}}^{\circ}$  is negative for both possible zwitterions of HEPES, meaning this proton transfer is also favourable in aqueous solution. It is lower by 32.6 kJ mol<sup>-1</sup> for the zwitterionic molecule corresponding to WIRMOZ04 making this the more stable zwitterion in solution.

### Single-multi proton transfer polymorphs

Four systems were found to have single-multi proton transfer polymorphism ( $[A^+ : B^-]$  and  $[A^{2+} : B^{2-}]$ ) all of which consist of a symmetrical di-base with a symmetrical di-acid (Table 3). The first  $\Delta pK_a^I$  (the strongest base and strongest acid) and second  $\Delta pK_a^{II}$  (the second strongest base and second strongest acid) in each pair were calculated for the four pairs of polymorphs identified. In all our systems here, the  $\Delta pK_a^I$  is always greater than 1, thus salt formation is predicted as likely for all systems.<sup>45</sup> For  $\Delta pK_a^{II}$ , the equivalence point (50% probability) for making a monovalent *versus* a di-valent salt is  $-4.2$ . All the  $\Delta pK_a^{II}$  data for our systems lie around that value, indicating the good predictability of the  $\Delta pK_a$  rule for these systems.

Finally, the proton positions in the crystal structures for HAZFAP06 (with its HAZFAP01 pair) are not unambiguously

determined by X-ray crystallography due to the high temperature of the data collection. Supporting neutron diffraction and modelling data provides evidence that a second proton transfer has occurred in HAZFAP06 compared to HAZFAP01. The second proton transfer and resultant change in form is thermally induced, reversible and also results in a colour change.<sup>22</sup> This form change can also be induced by applying an electric field and has been studied using synchrotron X-ray diffraction.<sup>76</sup> With increased use of orthogonal techniques such as neutron diffraction or ssNMR, in combination with computational modelling, more polymorphs related by proton transfer may be found where previously proton positions may have been ambiguous.







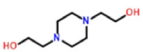

### Polymorphs with charge at different positions

All 5 systems exhibiting polymorphism with charge located on different positions were multicomponent systems (e.g.  $[A^+ : B^-]$  and  $[a^+ : b^-]$ ). An example of this polymorphism is found in sodium dihydrogen citrate (NAHCIT<sup>77</sup> and NAHCIT01<sup>78</sup>), where the citric carboxylic acid deprotonation is different in the two polymorphs. A particularly interesting example of this type of polymorphism is lamivudine hydrochloride (LASZAI<sup>79</sup> and LASZAI02<sup>80</sup>) because the protonation at a different location in the lamivudine molecule is also related by tautomerism. Both tautomers of lamivudine (Table 4) are very weakly basic with calculated  $pK_a$  of 0.21 and  $-0.05$ . However, HCl is a reasonably strong acid (experimental  $pK_a = -5.9$ )<sup>81</sup> and thus the  $\Delta pK_a$  is sufficiently large that salt formation is still expected to occur for either tautomer ( $\Delta pK_a$  being 6.1 and 5.9 respectively).

### Energy analysis

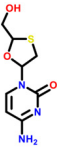
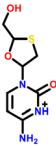
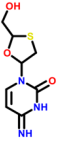
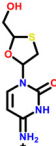
The lattice energies for several of these proton transfer polymorphs were calculated to assess how differences in lattice energies compare to those of 'usual' polymorphs (Table 5). In addition to this, the gas-phase energy difference of the corresponding molecular species were computed. Reassuringly, despite the very large gas-phase energy differences of the

**Table 3** Polymorphic systems with single-multi proton transfer polymorphs

Monovalent refcode	Divalent refcode	Basic component	Acidic component	$\Delta pK_a^I$	$\Delta pK_a^{II}$
HAZFAP01 <sup>72</sup>	HAZFAP06 <sup>22</sup>			4.2	-2.5
UWUJUS <sup>73</sup>	MEJYIM <sup>74</sup>			1.3	-5.5
UWUKAZ <sup>73</sup>	UWUKAZ02 <sup>75</sup>			1.5	-5.3
JYHAD <sup>76</sup>	ZELD0M <sup>77</sup>			6.7	-5.0



**Table 4** Lamivudine tautomers and calculated  $pK_a$  and  $\Delta pK_a$  for corresponding HCl salts

Refcode	Structure	Protonated structure	$pK_a$	$\Delta pK_a$
LASZAI			0.2	6.1
LASZAI02			-0.1	5.9

molecular species involved in some cases, all polymorphic pairs studied differ in overall lattice energies by somewhere between 1 to 7  $\text{kJ mol}^{-1}$ , typical of polymorphs.<sup>83,84</sup>

The amide and imide tautomeric polymorphs of sulfasalazine (QIJZOY03<sup>85</sup> and KIJB0X,<sup>86</sup> respectively) were computed to differ by almost 7  $\text{kJ mol}^{-1}$  in lattice energy. The literature was searched for experimental evidence of the stability ranking these polymorphs, but nothing conclusive was found to have been reported. Commercially available batches of sulfasalazine appear to be consistent with form I which is computed to be the most stable form despite containing the higher-energy tautomer.<sup>87</sup> Form II, is computed to be metastable and this agrees with the unusual crystallisation conditions used for its isolation from supercritical  $\text{CO}_2$ .<sup>88</sup>

The simple compound *m*-aminobenzoic acid was identified as forming zwitterionic polymorphs with two structures identified in our searches (AMBNZA<sup>89</sup> being the non-ionised form II, and AMBNZA02<sup>34</sup> being the zwitterionic form IV). There is a large difference in energy in the gas phase for non-ionised *versus* zwitterionic *m*-aminobenzoic acid species (240.5  $\text{kJ mol}^{-1}$ ) as expected since there is no stabilisation of

charges in the gas-phase. However, the  $\Delta E_{\text{Latt}}$  is 5.1  $\text{kJ mol}^{-1}$ , and the relative stability correlates with available experimental data.<sup>34,62,90</sup>

Four pairs of salt-cocrystal polymorphs have been previously assessed, with differences in formation energy of between  $\sim 8.5$  to 11.7  $\text{kJ mol}^{-1}$  reported in the literature.<sup>91</sup>

Because there were only two examples, the lattice energies of both pairs of multi-zwitterionic polymorphs were assessed. For the CINMER pair, the non-ionised molecule had the lowest energy by  $\sim 100$   $\text{kJ mol}^{-1}$ , relative to either of the two zwitterionic molecules. The difference in energy between the two zwitterions in the gas phase was just 4.3  $\text{kJ mol}^{-1}$ , comparable to the energy difference between many of the non-ionised tautomers assessed.  $\Delta E_{\text{Latt}}$  of 4  $\text{kJ mol}^{-1}$  was calculated with CINMER02 being the more stable form. This stability ranking of the two forms correlates with experimental relative stability data in the literature.<sup>72</sup> For the WIRMOZ polymorphs, the calculated  $\Delta E_{\text{Latt}}$  was 2.2  $\text{kJ mol}^{-1}$ .

For single-multi proton transfer polymorphs, the polymorphic pair from the refcodes MEJYIM and UWUJUS were assessed, with the component molecules assumed to be most stable in the non-ionised state in the gas phase. The structure with two proton transfers, MEJYIM, was calculated to be more stable by 1.1  $\text{kJ mol}^{-1}$ .

Lamivudine hydrochloride (LASZAI) exhibits charge position polymorphism, seemingly as a consequence of tautomerism. The gas phase energies of the neutral and protonated tautomers were calculated. Interestingly, the amine form is more stable in the gas phase by 11.9  $\text{kJ mol}^{-1}$  for the non-ionised molecule, whereas the imine is more stable by 9.9  $\text{kJ mol}^{-1}$  when the molecule is ionised. In the solid state, the polymorph containing the protonated amine was more stable by 5.8  $\text{kJ mol}^{-1}$ .

## Conclusions

By assessing the CSD, its drug subsets and the GSD, we have found that tautomerism is not just a serious consideration in

**Table 5** Gas phase and lattice energies for some polymorphs related by proton transfer

Type of polymorphism	Refcode, form	Species	Relative $E_{\text{Gas}}$ ( $\text{kJ mol}^{-1}$ )	Relative $E_{\text{Lattice}}$ ( $\text{kJ mol}^{-1}$ )
Tautomeric	KIJB0X, form I	Imide tautomer	+22.9	0.0
	QIJZOY03, form II	Amide tautomer	0.0	+6.9
Zwitterionic	AMBNZA02, form IV	Zwitterion	+240.5	0.0
	AMBNZA, form II	Non-ionised	0.0	+5.1
Multi-zwitterionic	CINMER02, form I	Zwitterion <i>para</i> $\text{COO}^-$	0.0	0.0
	CINMER04, form II	Zwitterion <i>meta</i> $\text{COO}^-$	+4.3	+4.0
Multi-zwitterionic	WIRMOZ, form I <sup>a</sup>	Zwitterion OH side	+79.1	0.0
	WIRMOZ04, form II <sup>a</sup>	Zwitterion $\text{SO}_3$ side	0.0	+2.2
Single-multi proton transfer	MEJYIM, form I <sup>a</sup>	Double	+1101.0 <sup>c</sup>	0.0
	UWUJUS, form II <sup>a</sup>	Single	0.0	+1.1
Different charge position	LASZAI, form I	Amine tautomer: non-ionised, ionised <sup>b</sup>	0.0, +9.9	0.0
	LASZAI02, form II	Imine tautomer: non-ionised, ionised <sup>b</sup>	+11.9, 0.0	+5.8

<sup>a</sup> Form names were not found in the literature. <sup>b</sup> Relative energies for LASZAI gas phase molecules are comparisons of the amine and imine tautomers per specified ionisation state. <sup>c</sup> DMol3 (ref. 82) with GGA PBE functional and TS DFT-D correction used to calculate gas phase energy of charged species only.



the pharmaceutical industry for medicinal chemists, but for solid state scientists alike. Drug-like molecules have an increased tendency to contain functional groups capable of tautomerism, as well as to contain both acidic and basic functional groups, thus enabling zwitterion formation. Evidence suggests that both tautomer equilibrium and speciation in solution may have implications for the solid state outcome and polymorphic control in crystallisations.

Polymorphs related by proton transfer represent approximately 3% of polymorphs for drug-like molecules: that is potentially 1 in every 33 drug candidates. Of the 6 categories of these polymorphs found in the CSD, tautomeric polymorphs are by far the most common (54%) followed by zwitterionic polymorphs (25%). Salt-cocrystal pairs make up only 9% of the polymorphs related by proton transfer but a relatively high proportion of cocrystal polymorphs; there were only 145 polymorphic cocrystals identified in a 2015 search of the CSD,<sup>83</sup> hence roughly 6% of polymorphic cocrystals may exist as salt-cocrystal pairs. Despite the slower uptake of cocrystals in the industry due to regulatory ambiguity,<sup>92</sup> it is significant that these many are salt-cocrystal pairs thus reinforcing the need to intentionally assess whether both can form, especially when the  $\Delta pK_a$  rule indicates a similar probability of forming a salt or cocrystal ( $\Delta pK_a$  is close to 1).

Furthermore, we have noted that some literature examples were not identified in our searches highlighting the limitations of our method. The limitations may be due to inconsistent quality of structures deposited in the CSD (when they exist), restrictions applied to our datasets or the scripted comparison algorithm, which treats individual molecules in a bulk way and may not account for unique scenarios. Importantly, we know that the true number of polymorphs related by proton transfer is likely to be greater than that reported here and when studying individual systems experimentally proton positions should be investigated by multiple orthogonal techniques.

Finally, despite the sometimes large differences in gas-phase energies of different tautomers or ionisation states of a molecule, we have shown the overall lattice energy of these polymorphs is small as expected from 'typical' polymorphs. The existence of proton transfer polymorphs demonstrates the power of the crystal lattice to stabilise metastable tautomers or other ionised molecular species and the ability of a molecule to tautomerise or ionise should present opportunities for crystal engineering. Given that these polymorphs could have substantially different properties given the changes in both chemical structure and solid state packing, for a pharmaceutical company with a large portfolio, this type of polymorphism is common enough to be a problem or an opportunity (or both) when designing and developing solid forms.

## Author contributions

All authors have contributed to the manuscript. AWR is the primary author and conducted the searches, coding,

calculations, and analysis. CLD has contributed to helpful discussions. AJCC has designed the project and contributed to ideas and discussions. AJCC is corresponding author.

## Conflicts of interest

A. Woods-Ryan and C. L. Doherty are employed by and hold shares in GSK.

## Acknowledgements

The authors would like to thank Leen Kalash, Nick Henley, Luca Russo, Ian Rosbottom, Jack Ryan and Daniel Woods for their valuable discussions, advice and support and GSK for funding.

## Notes and references

- P. J. Taylor, G. van der Zwan and L. Antonov, in *Tautomerism: Methods and Theories*, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2013, pp. 1–24.
- A. J. Cruz-Cabeza, A. Schreyer and W. R. Pitt, *J. Comput.-Aided Mol. Des.*, 2010, **24**, 575–586.
- J. Powling and H. J. Bernstein, *J. Am. Chem. Soc.*, 1951, **73**, 4353–4356.
- R. M. Izatt, J. L. Oscarson, S. E. Gillespie, H. Grimsrud, J. A. R. Renuncio and C. Pando, *Biophys. J.*, 1992, **61**, 1394–1401.
- E. Evangelio, C. Rodriguez-Blanco, Y. Coppel, D. N. Hendrickson, J. P. Sutter, J. Campo and D. Ruiz-Molina, *Solid State Sci.*, 2009, **11**, 793–800.
- F. Mesiti, A. Maruca, V. Silva, R. Rocca, C. Fernandes, F. Remião, E. Uriarte, S. Alcaro, A. Gaspar and F. Borges, *Eur. J. Med. Chem.*, 2021, **213**, 113183.
- A. R. Katritzky, C. Dennis Hall, B. E. D. M. El-Gendy and B. Draghici, *J. Comput.-Aided Mol. Des.*, 2010, **24**, 475–484.
- Y. C. Martin, *J. Comput.-Aided Mol. Des.*, 2009, **23**, 693–704.
- A. J. Cruz-Cabeza and C. R. Groom, *CrystEngComm*, 2011, **13**, 93–98.
- T. L. Threlfall, *Analyst*, 1995, **120**, 2435.
- M. R. Chierotti, L. Ferrero, N. Garino, R. Gobetto, L. Pellegrino, D. Braga, F. Grepioni and L. Maini, *Chem. – Eur. J.*, 2010, **16**, 4347–4358.
- J. Elguero, *Cryst. Growth Des.*, 2011, **11**, 4731–4738.
- Polymorphism in the Pharmaceutical Industry*, ed. R. Hilfiker, Wiley-VCH Verlag GmbH & Co. KGaA, 2006.
- S. S. Kumar and A. Nangia, *Cryst. Growth Des.*, 2014, **14**, 1865–1881.
- N. K. Nath, S. S. Kumar and A. Nangia, *Cryst. Growth Des.*, 2011, **11**, 4594–4605.
- V. B. Kurteva, B. L. Shivachev, R. P. Nikolova, S. D. Simova, L. M. Antonov, L. A. Lubenov and M. A. Petrova, *RSC Adv.*, 2015, **5**, 73859–73867.
- C. Wales, L. H. Thomas and C. C. Wilson, *CrystEngComm*, 2012, **14**, 7264–7274.
- G. R. Desiraju, *J. Chem. Soc., Perkin Trans. 2*, 1983, 1025.
- CSD Statistics, <https://www.ccdc.cam.ac.uk/CCDCStats/Stats>, (accessed 17 October 2021).



- 20 M. Woińska, S. Grabowsky, P. M. Dominiak, K. Woźniak and D. Jayatilaka, *Sci. Adv.*, 2016, **2**, 1–8.
- 21 T. Steiner, I. Majerz and C. C. Wilson, *Angew. Chem., Int. Ed.*, 2001, **40**, 2651–2654.
- 22 D. M. S. Martins, D. S. Middlemiss, C. R. Pulham, C. C. Wilson, M. T. Weller, P. F. Henry, N. Shankland, K. Shankland, W. G. Marshall, R. M. Ibberson, K. Knight, S. Moggach, M. Brunelli and C. A. Morrison, *J. Am. Chem. Soc.*, 2009, **131**, 3884–3893.
- 23 A. G. Lvov, A. V. Yadykov, K. A. Lyssenko, F. W. Heinemann, V. Z. Shirinian and M. M. Khusniyarov, *Org. Lett.*, 2020, **22**, 604–609.
- 24 J. S. Stevens, S. J. Byard, C. C. Seaton, G. Sadiq, R. J. Davey and S. L. M. Schroeder, *Phys. Chem. Chem. Phys.*, 2014, **16**, 1150–1160.
- 25 J. S. Stevens, S. J. Byard, C. A. Muryn and S. L. M. Schroeder, *J. Phys. Chem. B*, 2010, **114**, 13961–13969.
- 26 K. Wolnica, G. Szklarz, M. Dulski, M. Wojtyniak, M. Tarnacka, E. Kaminska, R. Wrzalik, K. Kaminski and M. Paluch, *Colloids Surf., B*, 2019, **182**, 110319.
- 27 P. D. S. Carvalho, L. F. Diniz, G. T. S. T. Da Silva, N. D. Coutinho, P. G. Dos Santos, V. H. Carvalho-Silva, C. Ribeiro and J. Ellena, *Cryst. Growth Des.*, 2021, **21**, 1122–1135.
- 28 P. T. Edwards, L. K. Saunders, D. C. Grinter, P. Ferrer, G. Held, E. J. Shotton and S. L. M. Schroeder, *J. Phys. Chem. A*, 2022, **126**, 2889–2898.
- 29 A. S. Aljaber and A. D. Bani-Yaseen, *J. Mol. Graphics Modell.*, 2019, **86**, 160–169.
- 30 K. R. Bryenton, A. A. Adeleke, S. G. Dale and E. R. Johnson, *WIREs Comput. Mol. Sci.*, 2022, e1631.
- 31 I. Ciofini, C. Adamo and H. Chermette, *Chem. Phys.*, 2005, **309**, 67–76.
- 32 M. Sitzmann, W. D. Ihlenfeldt and M. C. Nicklaus, *J. Comput.-Aided Mol. Des.*, 2010, **24**, 521–551.
- 33 L. Orola, M. V. Veidis, I. Sarcevic, A. Actins, S. Belyakov and A. Platonenko, *Int. J. Pharm.*, 2012, **432**, 50–56.
- 34 P. A. Williams, C. E. Hughes, G. K. Lim, B. M. Kariuki and K. D. M. Harris, *Cryst. Growth Des.*, 2012, **12**, 3104–3113.
- 35 E. Simone and Z. K. Nagy, *CrystEngComm*, 2015, **17**, 6538–6547.
- 36 P. W. Carter and M. D. Ward, *J. Am. Chem. Soc.*, 1994, **116**, 769–770.
- 37 M. U. Schmidt, J. Brüning, J. Glinnemann, M. W. Hützler, P. Mörschel, S. N. Ivashevskaya, J. Van De Streek, D. Braga, L. Maini, M. R. Chierotti and R. Gobetto, *Angew. Chem., Int. Ed.*, 2011, **50**, 7924–7926.
- 38 M. R. Chierotti, R. Gobetto, L. Pellegrino, L. Milone and P. Venturello, *Cryst. Growth Des.*, 2008, **8**, 1454–1457.
- 39 K. Epa, C. B. Aakeröy, J. Desper, S. Rayat, K. L. Chandra and A. J. Cruz-Cabeza, *Chem. Commun.*, 2013, **49**, 7929.
- 40 Z. Liu, L. Zhong, P. Ying, Z. Feng and C. Li, *Biophys. Chem.*, 2008, **132**, 18–22.
- 41 C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**, 171–179.
- 42 G. Landrum, *RDKit: Open-source cheminformatics, Version 2021.03.1*, <https://www.rdkit.org/>.
- 43 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. Crystallogr.*, 2020, **53**, 226–235.
- 44 JChem for Office (Excel) Version 18.22.4.11, ChemAxon, 2018, <https://www.chemaxon.com>.
- 45 A. J. Cruz-Cabeza, M. Lusi, H. P. Wheatcroft and A. D. Bond, *Faraday Discuss.*, 2022, **235**, 446–466.
- 46 BIOVIA, *Dassault Systèmes, Materials Studio, 20.1.0.2728*, San Diego: Dassault Systèmes, 2020.
- 47 S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. I. J. Probert, K. Refson and M. C. Payne, *Z. Kristallogr.*, 2005, **220**, 567–570.
- 48 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.
- 49 A. Tkatchenko and M. Scheffler, *Phys. Rev. Lett.*, 2009, **102**, 073005.
- 50 A. Ambrosetti, A. M. Reilly, R. A. Distasio and A. Tkatchenko, *J. Chem. Phys.*, 2014, **140**, 18A508.
- 51 M. J. Bryant, S. N. Black, H. Blade, R. Docherty, A. G. P. Maloney and S. C. Taylor, *J. Pharm. Sci.*, 2019, **108**, 1655–1662.
- 52 L. N. Kalash, J. C. Cole, R. C. B. Copley, C. M. Edge, A. A. Moldovan, G. Sadiq and C. L. Doherty, *CrystEngComm*, 2021, **23**, 5430–5442.
- 53 J. Halebian and W. McCrone, *J. Pharm. Sci.*, 1969, **58**, 911–929.
- 54 M. Bauer, R. K. Harris, R. C. Rao, D. C. Apperley and C. A. Rodger, *J. Chem. Soc., Perkin Trans. 2*, 1998, 475–482.
- 55 M. Rubčić, K. Užarević, I. Halasz, N. Bregović, M. Mališ, I. Dilović, Z. Kokan, R. S. Stein, R. E. Dinnebier and V. Tomišić, *Chem. – Eur. J.*, 2012, **18**, 5620–5631.
- 56 C. Foces-Foces, A. L. Llamas-Saiz, R. M. Claramunt, C. López and J. Elguero, *J. Chem. Soc., Chem. Commun.*, 1994, 1143–1145.
- 57 A. Bongioanni, B. S. Araújo, Y. S. de Oliveira, M. R. Longhi, A. Ayala and C. Garnerero, *AAPS PharmSciTech*, 2018, **19**, 1468–1476.
- 58 A. M. Araya-Sibaja, M. Urgellés, F. Vásquez-Castro, F. Vargas-Huertas, J. R. Vega-Baudrit, T. Guillén-Girón, M. Navarro-Hoyos and S. L. Cuffini, *RSC Adv.*, 2019, **9**, 5244–5250.
- 59 A. M. Araya-Sibaja, C. E. Maduro de Campos, C. Fandaruff, J. R. Vega-Baudrit, T. Guillén-Girón, M. Navarro-Hoyos and S. L. Cuffini, *J. Pharm. Anal.*, 2019, **9**, 339–346.
- 60 P. Sacchi, M. Lusi, A. J. Cruz-Cabeza, E. Nauha and J. Bernstein, *CrystEngComm*, 2020, **22**, 7170–7185.
- 61 A. J. Cruz-Cabeza and J. Bernstein, *Chem. Rev.*, 2014, **114**, 2170–2191.
- 62 L. Gopal, C. I. Jose and A. B. Biswas, *Spectrochim. Acta, Part A*, 1967, **23**, 513–518.
- 63 M. Mirmehrabi, S. Rohani, K. S. K. Murthy and B. Radatus, *J. Cryst. Growth*, 2004, **260**, 517–526.
- 64 K. Down, A. Amour, N. A. Anderson, N. Barton, S. Campos, E. P. Cannons, C. Clissold, M. A. Convery, J. J. Coward, K. Doyle, B. Duempelfeld, C. D. Edwards, M. D. Goldsmith, J. Krause, D. N. Mallett, G. A. McGonagle, V. K. Patel, J.



- Rowedder, P. Rowland, A. Sharpe, S. Sriskantharajah, D. A. Thomas, D. W. Thomson, S. Uddin, J. N. Hamblin and E. M. Hessel, *J. Med. Chem.*, 2021, **64**, 13780–13792.
- 65 P. N. Remya, C. H. Suresh and M. L. P. Reddy, *Polyhedron*, 2007, **26**, 5016–5022.
- 66 L. Qi, Y. Jin, H. Li, Y. Dong and C. Xie, *Trans. Tianjin Univ.*, 2020, **26**, 458–469.
- 67 J. E. Werner and J. A. Swift, *CrystEngComm*, 2021, **23**, 1555–1565.
- 68 C. Liao and M. C. Nicklaus, *J. Chem. Inf. Model.*, 2009, **49**, 2801–2812.
- 69 A. J. Cruz-Cabeza, *CrystEngComm*, 2012, **14**, 6362.
- 70 P. M. Dominiak, E. Grech, G. Barr, S. Teat, P. Mallinson and K. Woźniak, *Chem. – Eur. J.*, 2003, **9**, 963–970.
- 71 D. Ang, G. B. Deacon, P. C. Junk and D. R. Turner, *Polyhedron*, 2007, **26**, 385–391.
- 72 D. Braga, L. Maini, C. Fagnano, P. Taddei, M. R. Chierotti and R. Gobetto, *Chem. – Eur. J.*, 2007, **13**, 1222–1230.
- 73 I. R. Evans, J. A. K. Howard, J. S. O. Evans, S. R. Postlethwaite and M. R. Johnson, *CrystEngComm*, 2008, **10**, 1404–1409.
- 74 F. Gao, C. Yin, P. Yang and G. Xue, *Acta Crystallogr., Sect. E: Struct. Rep. Online*, 2004, **60**, o1328–o1329.
- 75 P. Śledzied, R. Kamiński, M. Chruszcz, M. D. Zimmerman, W. Minor and K. Woniak, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2010, **66**, 482–492.
- 76 L. K. Saunders, H. H.-M. Yeung, M. R. Warren, P. Smith, S. Gurney, S. F. Dodsworth, I. J. Vitorica-Yrezabal, A. Wilcox, P. V. Hathaway, G. Preece, P. Roberts, S. A. Barnett and D. R. Allan, *J. Appl. Crystallogr.*, 2021, **54**, 1349–1359.
- 77 J. P. Glusker, D. Van Der Helm, W. E. Love, M. Dornberg, J. A. Minkin, C. K. Johnson and A. L. Patterson, *Acta Crystallogr.*, 1965, **19**, 561–572.
- 78 A. Rammohan and J. A. Kaduk, *Acta Crystallogr., Sect. E: Crystallogr. Commun.*, 2016, **72**, 854–857.
- 79 J. Ellena, N. Papparidis and F. T. Martins, *CrystEngComm*, 2012, **14**, 2373.
- 80 J. C. Tenorio Clavijo, F. F. Guimarães, J. Ellena and F. T. Martins, *CrystEngComm*, 2015, **17**, 5187–5194.
- 81 A. Trummal, L. Lipping, I. Kaljurand, I. A. Koppel and I. Leito, *J. Phys. Chem. A*, 2016, **120**, 3663–3669.
- 82 B. Delley, *J. Chem. Phys.*, 1990, **92**, 508–517.
- 83 A. J. Cruz-Cabeza, S. M. Reutzel-Edens and J. Bernstein, *Chem. Soc. Rev.*, 2015, **44**, 8619–8635.
- 84 J. Nyman and G. M. Day, *CrystEngComm*, 2015, **17**, 5154–5165.
- 85 L. A. Filip, M. R. Caira, S. I. Fărcaș and M. T. Bojiță, *Acta Crystallogr., Sect. C: Cryst. Struct. Commun.*, 2001, **57**, 435–436.
- 86 A. J. Blake, X. Lin, M. Schröder, C. Wilson and R.-X. Yuan, *Acta Crystallogr., Sect. C: Cryst. Struct. Commun.*, 2004, **60**, o226–o228.
- 87 S. S. Chourasiya, D. R. Patel, C. M. Nagaraja, A. K. Chakraborti and P. V. Bharatam, *New J. Chem.*, 2017, **41**, 8118–8129.
- 88 W. Y. Wu and C. S. Su, *J. Cryst. Growth*, 2017, **460**, 59–66.
- 89 J. Voogd, B. H. M. Verzijl and A. J. M. Duisenberg, *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.*, 1980, **36**, 2805–2806.
- 90 M. Svärd, F. L. Nordström, T. Jasnobulka and Å. C. Rasmuson, *Cryst. Growth Des.*, 2010, **10**, 195–204.
- 91 C. L. Jones, J. M. Skelton, S. C. Parker, P. R. Raithby, A. Walsh, C. C. Wilson and L. H. Thomas, *CrystEngComm*, 2019, **21**, 1626–1634.
- 92 O. N. Kavanagh, D. M. Croker, G. M. Walker and M. J. Zaworotko, *Drug Discovery Today*, 2019, **24**, 796–804.

