Use of crystal structure informatics for defining the conformational space needed for predicting crystal structures of pharmaceutical molecules

SCHOLARONE™
Manuscripts

Use of crystal structure informatics for defining the conformational space needed for predicting crystal structures of pharmaceutical molecules.

Luca Iuzzolino,[†] Anthony M. Reilly,[‡, #] Patrick McCabe,[‡] and Sarah L. Price[†,]*

[†]Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ, England.

[‡]The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England

ABSTRACT: Determining the range of conformations that a flexible pharmaceutical-like molecule could plausibly adopt in a crystal structure is a key to successful crystal structure prediction (CSP) studies. We aim to use conformational information from the crystal structures in the Cambridge Structural Database (CSD) to facilitate this task. The conformations produced by the CSD Conformer Generator are reduced in number by considering the underlying rotamer distributions, an analysis of changes in molecular shape, and a minimal number of molecular *ab initio* calculations. This method is tested for five pharmaceutical-like molecules where an extensive CSP study has already been performed. The CSD informatics-derived set of crystal structure searches generates almost all the low-energy crystal structures previously found, including all experimental structures. The

1

workflow effectively combines information on individual torsion angles and then eliminates the combinations that are too high in energy to be found in the solid state, reducing the resources needed to cover the solid-state conformational space of a molecule. This provides insights into how the low energy solid-state and isolated-molecule conformations are related to the properties of the individual flexible torsion angles.

1. **INTRODUCTION**

Establishing the range of conformations that a potential drug molecule can adopt is not only relevant to drug design. Most drugs are formulated with the active molecule in the solid state,[1, 2] and the physical properties of the crystalline forms of an API can be very dependent on the crystal structure[1-11] with most pharmaceuticals having a range of solid-state structures (polymorphs) containing different conformations of the molecule.[5, 11, 12] Hence, in drug development, the pharmaceutical undergoes a screen for possible solid-forms involving a diverse range of crystallization experiments.[3, 10, 13-16] Since this cannot bring total reassurance that all relevant solid forms of the molecule are known,[3, 7, 10, 13] there is considerable interest in using computational modeling to predict the polymorphs of organic molecules.[1, 3, 10, 13, 17, 18] Such crystal structure prediction (CSP) studies[19] should generate the observed structure and usually other structures that are within the likely energy range of possible polymorphism, which can be considered as putative polymorphs (PPMs). There are examples of using the predicted structure of a PPM to devise an experiment to find it.[13, 20, 21] Hence CSP methods have been evolving rapidly,[22] along with our understanding of why they may over-predict polymorphism.[23] They have been applied to about a dozen pharmaceutical molecules in collaboration with industrial pharmaceutical solid-state scientists,[1, 3] and shown to provide valuable insights into the crystallization behavior. However, a major barrier to the use of CSP

2

studies as a routine complement to industrial screening is the limitation in the number of flexible torsion angles in the molecule that can be tackled, even with considerable computer resources.

The importance of conformational flexibility can be seen from the approximation of the lattice energy (E_{latt}), defined as the energy for separating all the molecules in a static lattice at 0 K to an infinite distance in their lowest-energy conformation,[19] as:

$$E_{latt} = U_{inter} + \Delta E_{intra} \tag{1}$$

where U_{inter} is the intermolecular energy due to interactions between molecules and ΔE_{intra} is the intramolecular energy penalty for distorting the molecular conformation from its gas-phase global minimum.[19] This approximation of lattice energy, which is used in several CSP methods,[24] is effective for rigid molecules ($\Delta E_{intra} \sim 0$) but rapidly becomes more expensive as the molecule can adopt more conformations, particularly since the need for accurate lattice-energy differences (most known polymorphs differ in lattice energy by less than 6 kJ/mol)[5] means that electronic-structure *ab initio* calculations are needed. Since local optimization methods can only go to the nearest lattice-energy minima, the flexible torsion angles that can significantly change the shape and the packing capabilities of the molecule need to be included as search variables in the generation of a comprehensive set of plausible crystal structures, along with the cell parameters and molecular positions and orientations.

Hence, most successful CSP methods based on equation (1) can be broken down into three steps:[24]

1) The determination of the range of conformations that the molecule could adopt in the solid state.

2) The generation of a comprehensive set of plausible crystal structures, each corresponding to a separate lattice-energy minimum, usually generated using a

3

relatively cheap approximation for $E_{latt.}$. An example of a crystal-structure generation engine, which exploits the separation in equation (1), is CrystalPredictor,[17, 25, 26] which estimates ΔE_{intra} by interpolation over a multi-dimensional grid of molecular *ab initio* energies for those torsion angles that need to be treated as variables in the search.[25]

3) Optimization and ranking of the generated crystal structures using more accurate and expensive models for $E_{latt.}$.

Extending CSP to the more flexible pharmaceuticals requires speeding up steps (1) and (2). Although some CSP methods do not explicitly require step (1), since they perform global searches without the need of limiting the conformational space,[24, 27-29] those that have been successfully applied to larger flexible molecules implicitly include this through derivation of a molecule-specific force-field from *ab initio* calculations.[30] Successes in the blind tests of crystal structure prediction methods using cheap transferable force fields rather than using extensive *ab initio* calculations are rare.[24] In this paper, we seek to find tools that can assist with defining the conformational space for a specific molecule, which are faster than a set of *ab initio* calculations for determining how to combine the behavior of the individual torsion angles.

An informatics approach to studying crystalline conformations can be based on the Cambridge Structural Database (CSD),[31] which contains more than 850,000 experimentally determined crystal structures. Several informatics tools exist for retrieving and analyzing CSD conformational information, some of which have been successfully used to provide insightful information on pharmaceutical molecules.[32] The oldest CSD tool for investigating conformational preferences in the organic solid state is Mogul, which creates charts of the observed values of the bond-lengths, bond-angles and torsion-angles in the crystal structures of molecules containing a fragment of interest in the CSD.[33] Recently, the CCDC has

4

developed knowledge-based conformational libraries more suited to high-throughput analyses of geometric preferences of flexible molecules. There is a rotamer library for each rotatable bond, defined by the chemical environment, whose distribution covers the full 0-360° range.[34] Furthermore, the data extracted from these libraries can be analyzed by a kernel density estimation (KDE) to generate smooth probability density functions (PDFs), to allow a quantitative analysis of molecular conformational preferences; given the circular nature of torsion angles, the Von Mises kernel is the most suited for their analysis.[35] The CSD Conformer Generator (CG)[36] uses the rotamer libraries to generate conformers libraries, which were originally designed for use in life-science modeling such as docking ligands into active sites or pharmacophoric shape matching. The CG combines the frequencies of the torsion angle or flexible ring conformations within the CSD into a probability score, which is used to rank the final set of conformations.[36] In this work, we seek to determine whether these informatics tools can assist with defining the conformational space for specific flexible pharmaceutical-like molecules in CSP.

Hence, we propose a workflow for simplifying the definition of the conformational preferences[37] of flexible molecules, with the goal of reducing the computational cost of CSP studies without reducing their effectiveness in identifying potential conformational[38] polymorphs. The workflow determines how conformational information can be utilized to reduce the number of *ab initio* calculations that are required to generate plausible crystal structures. This approach is tested for the ability to generate all the low-energy crystal structures of five large molecules that were found in previous successful CSP studies where a large human and computational effort had been made to try to ensure comprehensive coverage of conformational space. The results show this approach can be a step towards a

routine use of CSP studies that can provide insights into conformational polymorphism of a wider range of pharmaceuticals.

2. **METHODS**

Molecule XXVI, 4947 CG conformations

GSK269984B, 9529 CG conformations

Molecule XX, 17374 CG conformations

Molecule XXIII, 14269 CG conformations

A-Mebendazole, 91 CG conformations
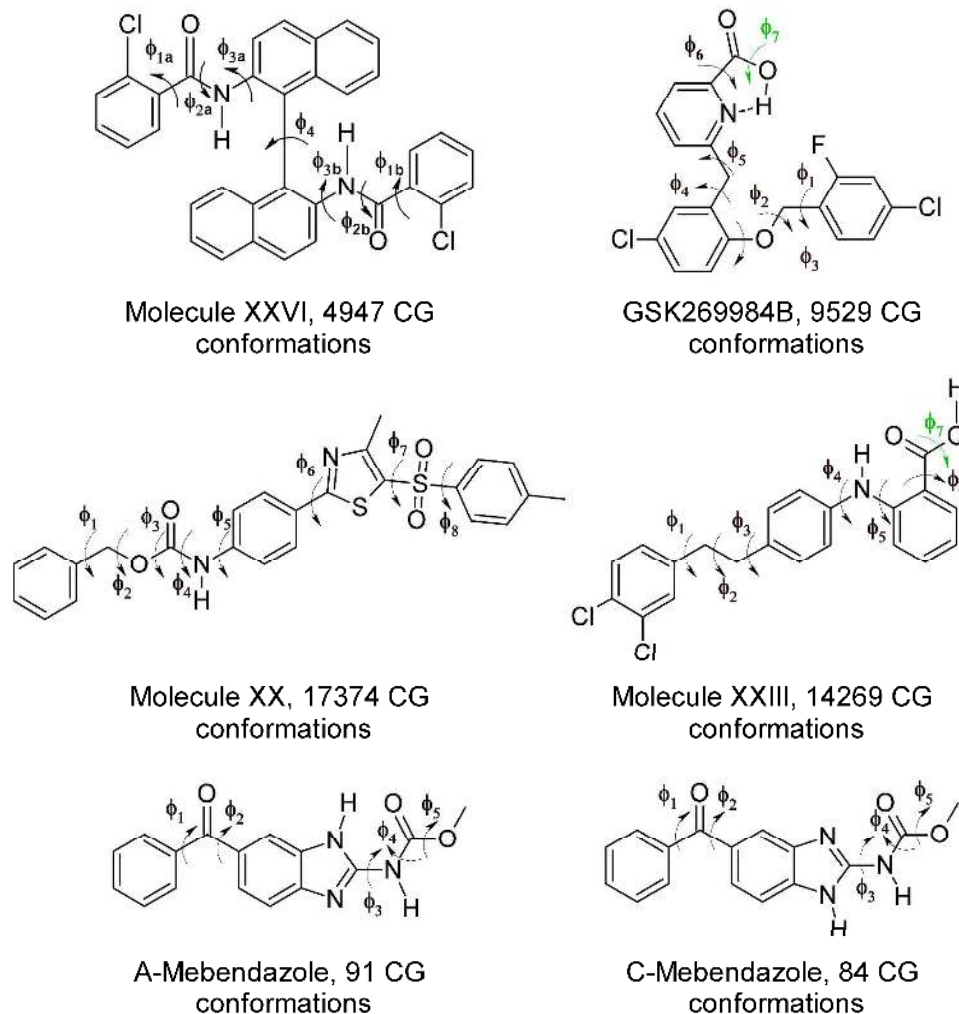
C-Mebendazole, 84 CG conformations

Figure 1. Chemical diagrams of the molecules used to test the applicability of CSD information to systems of pharmaceutical interest, showing the torsion angles that are identified as flexible by CG and the number of distinct conformations generated by the CG for each molecule. The additional angles not identified by CG are in green and define the position of polar hydrogen atoms.

6

Preliminary work was performed on a set of six small molecules to test the ability of the CG to generate all the likely conformations (see SI 1.1). It was verified that the CG-generated conformations included a close approximation to the observed solid-state conformers (SI Figure 3 and SI Table 1), and covered at least the range of conformational flexibility used in previous CSP studies. For these smaller molecules it could have been possible to do a CSP search with each of the CG conformations held rigid. However, this strategy cannot be applied for molecules with more flexible torsion angles, since the CG generated far too many conformations (Figure 1). The number cannot be reduced by just taking conformations with the best probability scores, as initial tests showed that the correlation between the *ab initio* energies and the CG probability score was too poor (SI 1.2): several conformations that would be included in a CSP search on energy grounds had poor probability scores, and vice versa. This reflects the fact that the lowest energy conformation of a flexible molecule does not correspond to the most probable, or lowest energy, value for each torsion angle. Furthermore, for the test molecules in Figure 1, the closest match between a CG generated conformation and the observed crystalline conformer is generally poorer than for the small molecules (see SI 1.3), with visual differences in the orientations of key peripheral bulky groups. This suggests that a CSP search with the closest CG conformation being treated as rigid may not find the experimental structure because none of the crystal structures generated by the search would locally minimize to the observed conformation.

The individual rotamer distributions fall into two classes, depending on how far observed torsion angles differ from the most probable values. Cases exemplified by 5-formyluracil and succinic acid (see SI Figures 2a and 2b) have all observations tightly clustered around a few values. These torsion angle values will be picked by the CG and can be kept as fixed in the search, with the different values defining separate conformational regions (CRs).

7

Alternatively, as exemplified by the fenamic acid (SI Fig 2d), there may be a wide range of

observed values, suggesting that these torsion angles need to be variables in the CSP search.

A related issue is that some torsion angles play a greater role in determining the overall shape

of the molecule than others, as exemplified in Figure 2. This implies that the effect of the

torsion angle on the shape of the entire molecule should be considered for an efficient CSP.
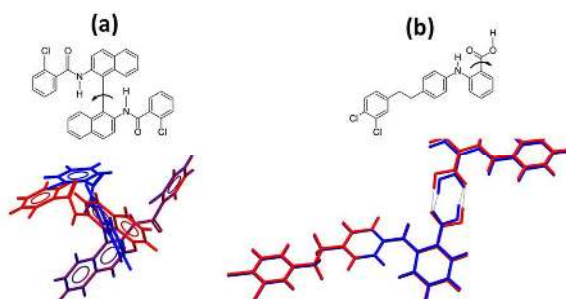
Figure 2. Comparison between the effects of 30° changes of (a) angle Φ_4 in molecule XXVI,

which has a large effect on the overall shape and (b) angle Φ_6 in XXIII, which has an

unnoticeable effect on the overall shape of the hydrogen-bonded dimer, except in the vicinity

of the carboxylic acid groups.

Hence we have developed a workflow to use information from the CSD Conformer

Generator and the underlying rotamer libraries in CSP studies to define the conformational

space of a specific molecule, based on the molecules in Figure 1.

2.1. The workflow.

1) *For each torsion angle, analyze the rotamer distributions and its effect on the shape

of the molecule.*   After extracting the rotamer distributions from the knowledge-based

libraries, the histograms of the rotatable torsion angles in Figure 1 were plotted and

superimposed with their PDFs calculated via Von Mises KDE using Python; SI sections 1

and 2 present a detailed description of the methodology. The effect of the torsion angle on

8

molecular shape was explored by ultra-fast shape recognition (USR), using the RDKit[39] and USRCAT[40] Python packages. The results of these analyses can be found in SI Figures 8-12 and Tables 4-9. Torsions defining the positions of hydrogen atoms that would be very important in determining hydrogen-bonding contributions to the lattice energy, or any other angle identified by CG as potentially flexible but with insufficient data in the CSD, have explicit *ab initio* potential-energy scans substituted. Hence, angles Φ_7 in GSK269984B and Φ_7 in XXIII were scanned with Gaussian 09[41] at the PBE0 6-31(d,p) level of theory (see SI Section 2.5).

2) *Assess the nature of each torsion angle.* The aim is to discriminate between torsion angles to be treated as explicitly flexible in the search and torsion angles to be constrained to CG values, hence defining separate conformational regions (CRs). The decision tree in Figure 3 was used for this purpose.
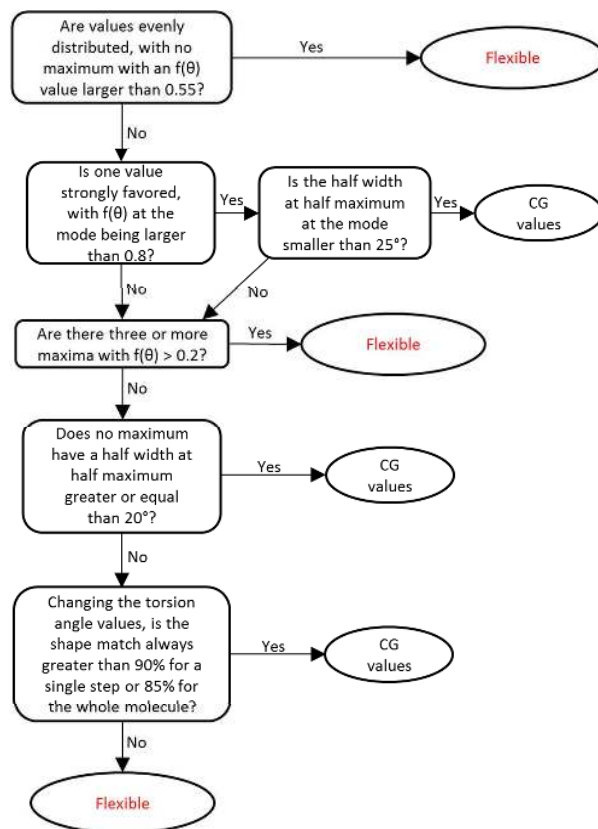
Figure 3. Decision tree used to discriminate between constrained torsion angles and explicitly flexible torsion angles on the basis of the PDF $f(\theta)$ and the changes in shape for varying the torsion angle (defined in SI section 2). An example its application is in SI section 4.

3) Determine the values or range of values for each torsion of angle. This involves:

a) Determine representative values for torsion angles that will be fixed in the search. This involved selecting a set of CG-generated conformations with different combinations of the constrained torsion angles. These selected conformations defined unique conformational regions (CRs), each of which requires a separate sub-search. Since the CG has a complex algorithm for picking the value of each torsion angle, aimed at maximizing the diversity of the conformations produced, many CG conformations do not have the angles that are at the peak of distributions. We wished to use this diversity while limiting the number of conformations (and associated regions) to be searched. Thus, the decision tree shown in Figure 4 was used to determine appropriate separation thresholds, which divide the full 360° range into intervals that define the most significant maxima of the distribution. These thresholds were then fed into an automated script to select those CG-generated conformations that were worth considering as separated CRs. The method is exemplified in SI Table 10.
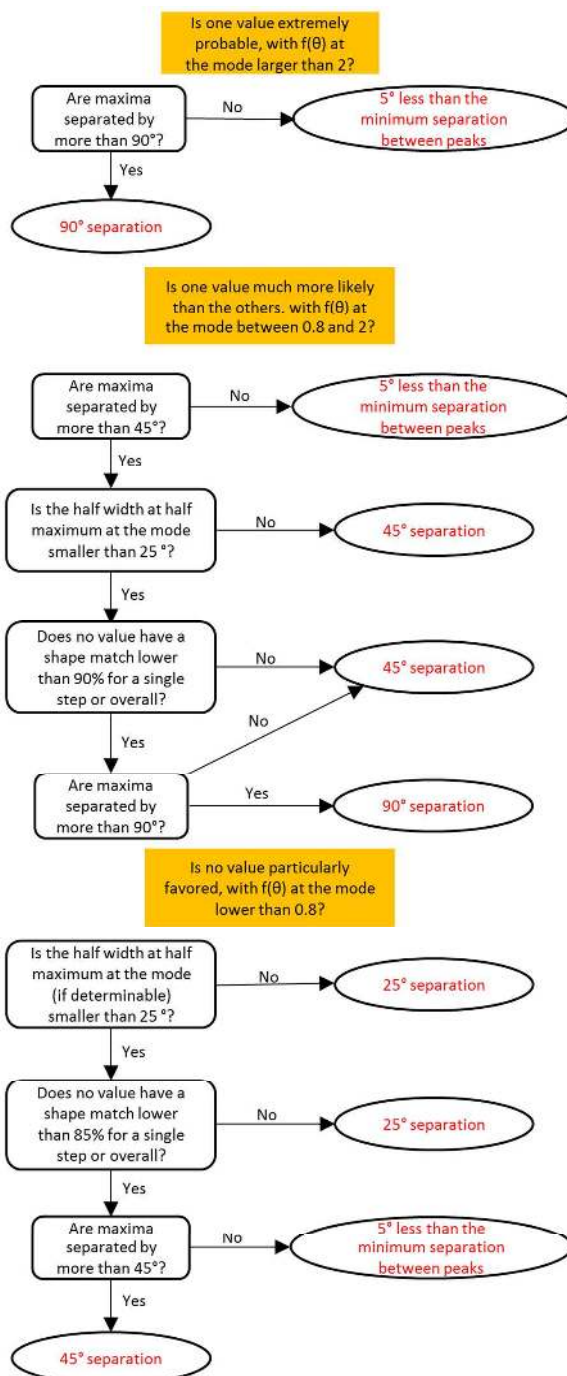
10

Figure 4. Decision tree used to choose the separation threshold of each torsion angle that is suitable for keeping fixed at CG values from the PDF $f(\theta)$ and shape-matching characteristics. An example of the application of this decision tree is in SI section 4.

11

b) Specification of the ranges for the flexible torsions. The range of the flexible degrees of freedom, and a suitable grid spacing for representing the energy variation, were selected using Mogul, as the distributions it produces are more chemistry-specific than the rotamer libraries.[34] Since the CG removes any gross steric clashes that can occur in large molecules, the torsion angle ranges produced by the CG can sometimes further reduce those observed in the Mogul distributions (see SI Figures 13-17). Step 4 refines the treatment of steric clashes and intramolecular stabilization.

c) Add polar hydrogen torsion angles. The same principles of analyzing the PDFs were applied to the *ab initio* torsion energy surfaces. Since the scans showed that both polar hydrogen angles had sharp energy minima, these torsion angles could be constrained to the values of the minima, which were added to the selected conformations. All the combinations defining selected CRs are tabulated for each molecule in SI Tables 11-16.

4) *Eliminate conformational regions that are energetically implausible.* The CG conformations representative of a conformational region (CR), with all the angles to be fixed during the search constrained at their initial, CG values, were optimized with Gaussian 09 at the PBE0 6-31(d,p) level of theory. This optimization of the flexible torsions and all the other bond-lengths, bond-angles and torsions not identified as flexible by CG, such as those within aromatic rings, was performed to calculate the energy of the nearest local minimum of the constrained potential-energy surface. Only CRs whose optimized energies, relative to the most stable fully optimized molecular structure, were plausible for crystalline conformers were kept after this step. We used a recently found threshold of $\Delta E_{Intra}^{CR} \leq 26$ kJ/mol.[6]

The workflow steps outlined thus far provide the low energy conformational space of the molecule, as a set of conformational regions and angles that can adopt a specified range of

12

possible values. Visualization of the conformational regions provides insight into how much the different parts of the molecule can interact with each other.

5) Perform CSP searches using the remaining conformational regions. Searches were performed with CrystalPredictor 1.8[25] for each conformational region. This requires a sufficiently realistic representation of the conformational energy for the molecule over the flexibility ranges for the torsions that are treated as search variables. Hence, grids of ΔE_{intra} values were calculated for the ranges and grid spacings determined in step 3b with Gaussian 09 at the PBE0 6-31(d,p) level of theory. CrystalPredictor interpolates these grids to estimate the conformational energy penalty relative to the minimum for each specific CR. The dimensionality of grids was reduced by identifying smaller surrogate molecules containing groups of torsion angles in the cases where the visualization showed it was reasonable to assume that their ΔE_{intra} values were not affected by the rest of the molecule (SI 2.4). This assumption that the overall conformational energy can sometimes be estimated by dividing a molecule into various surrogate molecules has been used in several successful CSP studies.[17, 24, 42] The chosen surrogate molecules (see SI Figures 18-22) had hydrogen atoms or methyl groups replacing certain non-torsional bonds to the rest of the molecule, and were large enough to represent the influence of the bonding environment on ΔE_{intra} for the group of torsion angles.

In the CrystalPredictor searches, the estimated ΔE_{intra} (from the ΔE_{intra} grids), was combined with the dominant intermolecular U_{inter} calculated as a sum of an electrostatic component derived from point charges (fitted using the CHELPG algorithm[43] from the optimized conformations) and a repulsion-dispersion component derived from the empirically fitted FIT potential.[44] Crystal structures were generated in the 59 most-common space groups in the CSD, which are listed in SI Table 18. The extent of the search was weighted according

13

to the lowest energy CR. 300,000 structures were generated in CRs with $\Delta E^{CR}_{Intra} \leq 4$ kJ/mol, 150,000 for those with $4 < \Delta E^{CR}_{Intra} \leq 17$ kJ/mol and 50,000 for those with $17 < \Delta E^{CR}_{Intra} \leq 26$ kJ/mol. By running the sub-searches in order of increasing ΔE^{CR}_{Intra}, it was possible to track the current global lattice-energy minimum for the overall search by adding ΔE^{CR}_{Intra} to the minimum lattice energy calculated by CrystalPredictor for that conformational region. Sub-searches that after generating at least 10,000 crystal structures had not produced any with a lattice energy within 25 kJ/mol from the global minimum could be terminated to save computational resources, with a high degree of confidence that the intermolecular lattice energy was unlikely to compensate for the ΔE_{intra} penalty and generate thermodynamically competitive structures. This contributed to the saving in computational resources.

2.2 Methodology assessment. This workflow for using crystal-structure informatics to define the conformational space needed in a CSP search can only be regarded as successful if it allows all the "significant" crystal structures that would be analyzed in the interpretation of the CSP search to be generated. These are the experimentally observed crystal structure(s) of each molecule and the low-energy unobserved putative polymorphs (PPMs), plus, for some molecules, a few higher-energy conformationally diverse computer-generated crystal structures that were explicitly examined in the original CSP study to exclude possible types of conformational polymorphs.

For each molecule, all of the generated crystal structures with E_{latt} (as calculated by CrystalPredictor) within 40 kJ/mol of the global minimum were checked to see whether the significant crystal structures were present among them. This range is typical of that used to determine which structures should be further refined with more accurate energy models. The generated crystal structures were compared with the fully optimized significant structures

14

found in the previous CSP study using the Crystal Packing Similarity Tool[45] in the CSD Python API. If the $RMSD_{15}$ for a 15 molecule cluster between the "significant" and the newly generated structure was smaller than 0.8 Å, which is the criterion used to test the success of the searches for the latest Blind Test,[24] the structure was considered to have been 'certainly found' (i.e. full optimization would lead to an almost exact match). If the $RMSD_{15}$ was larger than 0.8 Å, the structure was considered to have been 'probably found': it is highly likely, although not certain, that the generated structure would optimize to the corresponding significant one.
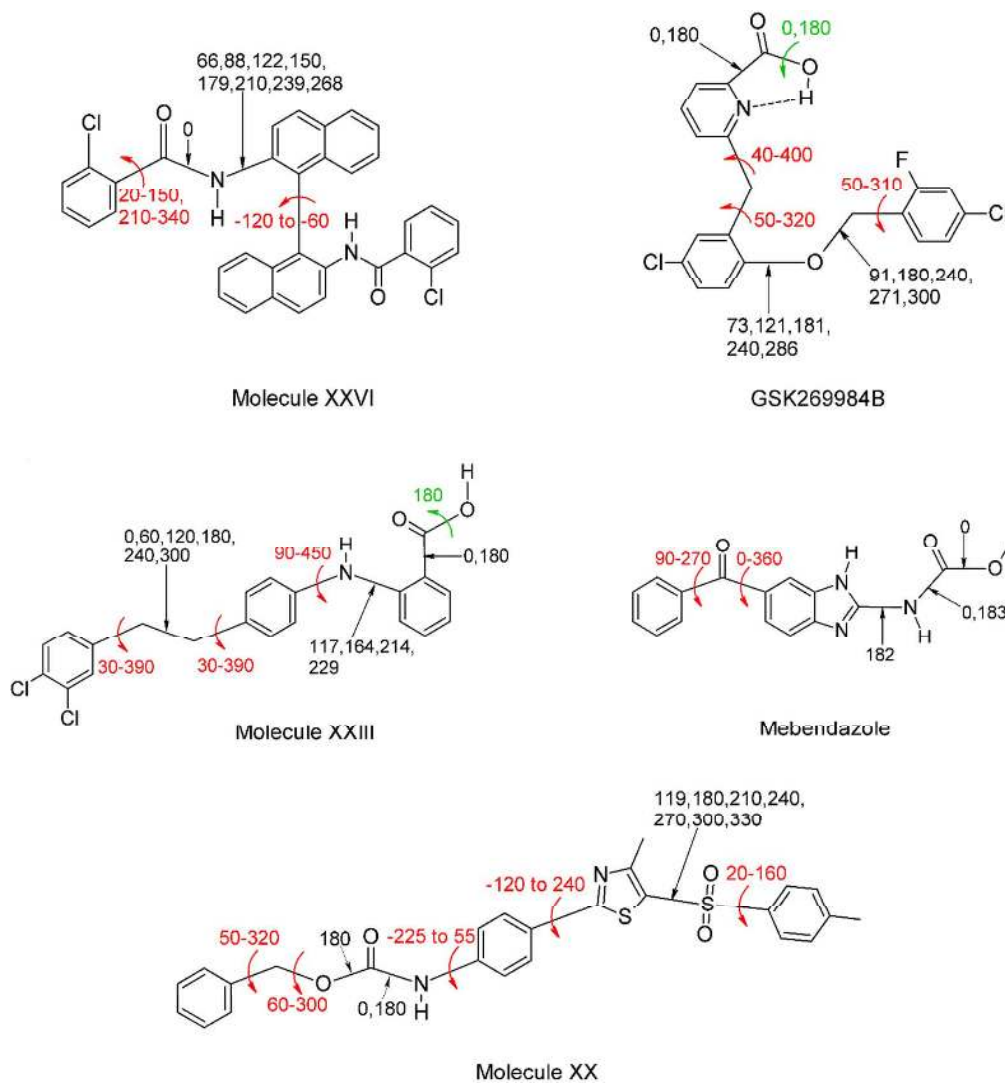
3. **RESULTS**

15

Figure 5: Summary of the application of the workflow on the molecules in Figure 1. The torsion angles in red were treated as flexible in the searches, covering the ranges given in degrees. Torsion angles in black were constrained to a set of CG values, with the values used in at least one search indicated; many combinations of these values were eliminated as energetically unfeasible in step 4. Torsion angles in green were constrained to the indicated values having been determined from an *ab initio* conformational-energy scan. The tautomers A and C of mebendazole were treated in the same way. Full details are shown in SI Table 17.

16

Figure 5 shows the result of the workflow for each molecule in Figure 1, indicating how each torsion angle in the search was treated, and Table 1 shows how the workflow reduced the conformational search space of each CSP study relative to that covered by the CG conformations.

Table 1: Summary of the CSP studies and their results.

| Molecule and original study	# CG conformations	# Selected CRs	# CRs ΔE_{intra} (CR) ≤ 26 kJ/mol			# ~ generated structures/ $\cdot 10^6$	Success in generating structures			Saving in CPU hours, relative to original CSP study
			0-4 kJ/mol	4-17 kJ/mol	17-26 kJ/mol		Found	Probably found	Not found	
XXVI[24]	4947	138	1	8	14	2.2	31	4	1	~14,000 (-50%)
GSK269984B[46]	9529	51	1	7	10	1.9	28	6	4	not recorded
XX[42, 47]	17347	21	2	8	3	2	19	2	0	~6,000 (-30%)
XXIII[24]	14269	127	3	6	7	2.1	37	12	4	~14,000 (-70%)
Mebendazole A[48]	91	4	1	1	0	0.45	28	4	0	~5,000 (-70%)
Mebendazole C[48]	84	4	1	0	1	0.35				

17
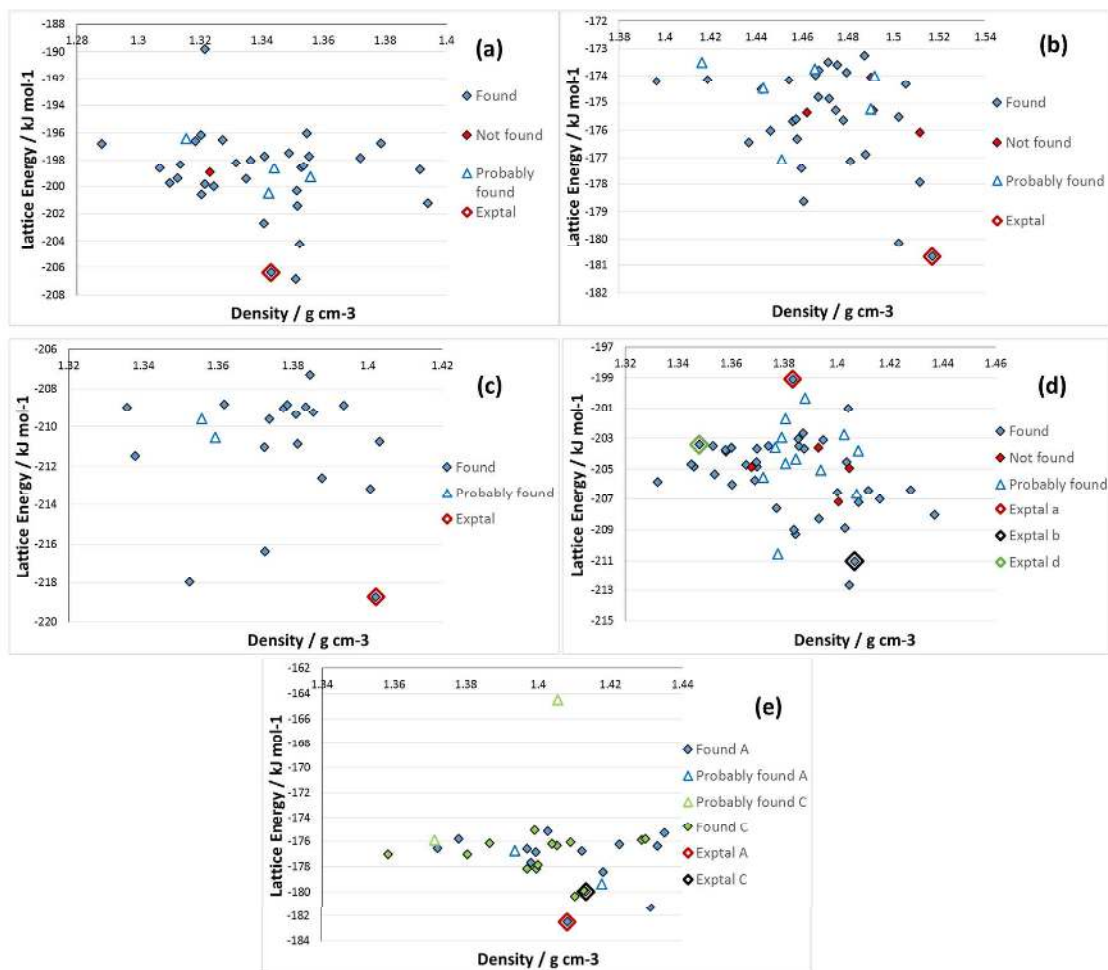
Figure 6. Plot of the significant CSP generated structures found in previous studies, classified as to whether they were found in the CSP workflow using the conformational space summarized in Figure 5 and SI Tables 19-23. All CSP crystal structures within a cutoff of 8-10 kJ/mol of the global minimum are presented in (a) for XXVI, (b) GSK269984B, (c) XX, (d) XXIII and (e) mebendazole. Higher energy structures were included for XXVI (a structure whose stability was very dependent on energy model), XX and mebendazole (a competitive crystal structure with a *cis* amide) and experimental Form A of XXIII. Open diamonds depict the experimental structures optimized with the same model for the lattice energy.

18

The CrystalPredictor searches using a combination of CSD informatics and *ab initio* calculations to limit the conformational space (Figure 5) generated millions of crystal structures (Table 1), though with a considerable saving in the computer resources for steps (1) and (2) of the CSP study. The evaluation of whether the search covered all plausible conformational and packing polymorphs is shown in Figure 6, by classifying the output of the original CSP searches as to whether the new workflow generated similar enough crystal structures. The vast majority of PPMs, including all the experimental and global energy minimum structures, were generated. For XX and mebendazole no significant crystal structures were missed, and only a few higher-energy PPMs were missed for XXVI, GSK269984B and XXIII. This was not due to any inadequacy in our method for defining the conformational space, as all the conformations in the missed structures were found in other structures generated in the searches. The structures were probably missed because they are located in particularly narrow wells on the lattice energy surface, which are rarely found[49] by the pseudo-random element of the CrystalPredictor searches. The accuracy of the conformational model was also evidenced by most of the CrystalPredictor generated structures being a good match with the corresponding fully optimized structures.

The workflow based on CSD informatics takes less than an hour to set up. The reduction in the computational cost (Table 1) came from replacing initial *ab initio* conformational analyses on individual torsion angles with the CSD informatics analysis, requiring fewer calculations for the grids of ΔE_{intra} values (see SI Tables 24-30), and from being able to terminate Crystal Predictor runs early in a systematic manner. However, the savings are dependent on the molecule.

19

4. **DISCUSSION**

The results of this work show that the CSD Conformer Generator is an effective method of describing the likely conformational space for pharmaceutical-like molecules, but the large number of conformations it generates makes a more-detailed analysis of both the individual torsional distributions and their effect on the overall shape of the molecule necessary. Some *ab initio* conformational energy estimates are also needed to define the plausible conformational space of the molecule. Overall, by setting up a CSP study as a set of searches with many fixed torsion angles fixed and some that are allowed to vary over defined ranges, the use of the conformational tools derived from the CSD vastly reduces the number of *ab initio* energy calculations required. The implementation of this workflow on a set of CrystalPredictor searches allowed the successful generation of most of the thermodynamically plausible polymorphs for a sample set of pharmaceutical-like molecules, with a 30 to 70% reduction in computational cost. Very few significant structures were not found (see Table 1), but this was not due to the definition of conformational space. Only a few of the generated structures poorly matched their optimized counterparts, having $RMSD_{15}$ values exceeding 1.5 Å, so it is probable, but not certain, that they would optimize to the same lattice-energy minima. Hence, since most of the significant crystal structures of these five molecules, including all the experimental and global energy minimum structures, were reproduced very well, the workflow has identified the range of potential conformational polymorphs.

This study has only addressed the question of using CSD informatics to reproduce the results of CSP studies based on the same assumptions, such as that sufficiently good *ab initio* calculations[50, 51] on the isolated molecule are effective in determining whether a conformation

20

is of low enough energy to plausibly occur in a crystal structure. As molecules become larger, and have more intramolecular hydrogen bonds that could become intermolecular within some crystal structures, the conformational energy cut-off values, such as the 26 kJ/mol used here to determine whether a conformational region is likely to be observed in a crystal structure, might need to be changed. Indeed, there is some evidence that higher energy but more extended, flatter conformations that can pack more densely are likely to be observed within crystal structures.[52] It may also be that the crystal structures in the CSD reflect more than relative thermodynamic stability [53, 54] However, the general approach outlined in this study will enable many more CSP studies of pharmaceutical-like molecules to be compared with the output of experimental polymorph screening.

The development of this workflow emphasizes that the conformational space of a flexible pharmaceutical is smaller than that obtained by just combining the behavior of the individual flexible torsion angles, with the CG providing a good starting point for defining this space. The use of shape matching criteria to determine which angles have a major effect on the overall shape, and the concept of dividing the angles into those which can take a wide range of values, or just a few probable ones defining separate conformational regions, are likely to be more widely applicable. The insights from this workflow can be tailored for use in other modeling applications such as the parameterization of force-fields or ligand docking. Finally, we believe that some of the procedures and tools discussed in this paper provide information that could be adapted for use in studies of larger, flexible organic molecules such as pharmaceuticals.

5. **CONCLUSION**

21

A methodology integrating CSD conformational information into CSP methods was found to generate the putative crystal structures of flexible pharmaceutical molecules, with the advantage of requiring less human expertise in conformational analysis and less computational resources. A method is proposed to use CSD torsion-angle distributions, molecular shape analysis and a limited number of *ab initio* calculations to reduce the conformations produced by the CSD Conformer Generator into a set of sufficiently low-energy conformational regions, defining which torsion angles should be used as search variables within specified ranges. The workflow effectively combines information on individual torsion angles and then eliminates the combinations that are too high in energy to be found in the solid state. Using this workflow in conjunction with CrystalPredictor generated all the experimentally known and almost all the putative polymorphs of five flexible pharmaceuticals produced by previous CSP studies, at a 30-70% lower computational cost. This approach provides some insights into how the molecular structure limits the range of low-energy conformational polymorphs, but more understanding of conformational polymorphism will come from the ability to contrast more CSP studies with solid-state screening results.

AUTHOR INFORMATION

Corresponding Author

*E-mail: s.l.price@ucl.ac.uk

Present Addresses

#School of Chemical Sciences, Dublin City University, Glasnevin, Dublin 9, Ireland

22

Funding Sources

SUPPORTING INFORMATION

Additional data illustrating the analysis of CSD conformational information and the results of the proposed CSP workflow, as well as a worked-though example of the generation of crystal structures of molecule XXIII. This information is available free of charge via the Internet at http://pubs.acs.org

ACKNOWLEDGMENT

REFERENCES

(1)   Price, S. L.; Reutzel-Edens, S. M., The potential of computed crystal energy landscapes to aid solid-form development. *Drug Discovery Today* **2016,** *21*, 912-923.

(2)   Lee, A. Y.; Erdemir, D.; Myerson, A. S., Crystal Polymorphism in Chemical Process Development. *Annu. Rev. Chem. Biomol. Eng., Vol 2* **2011,** *2*, 259-280.

23

(3)   Price, S. L.; Braun, D. E.; Reutzel-Edens, S. M., Can computed crystal energy landscapes help understand pharmaceutical solids? *Chem. Commun.* **2016,** *52*, 7065-7077.

(4)   Bernstein, J., *Polymorphism in Molecular Crystals*. ed.; Clarendon Press: Oxford, 2002.

(5)   Cruz-Cabeza, A. J.; Reutzel-Edens, S. M.; Bernstein, J., Facts and fictions about polymorphism. *Chem. Soc. Rev.* **2015,** *44*, 8619-8635.

(6)   Thompson, H.; Day, G., Which conformations make stable crystal structures? Mapping crystalline molecular geometries to the conformational energy landscape. *Chem. Sci.* **2014,** 5, 3173-3182.

(7)   Abramov, Y., Current Computational Approaches to Support Pharmaceutical Solid Form Selection. *Org. Process Res. Dev.* **2013,** *17*, 472-485.

(8)   Price, S. L., Crystal energy landscapes for aiding crystal form selection. In *Computational Pharmaceutics: Application of Molecular Modeling in Drug Delivery*, Ouyang, D.; Smith, S. C., Eds. Wiley: 2015.

(9)   Hilfiker, R., *Polymorphism in the Pharmaceutical Industry*. ed.; Wiley-VCH: Weinheim, 2006.

(10) Lee, E. H., A practical guide to pharmaceutical polymorph screening & selection. *Asian J. Pharm. Sci.* **2014,** *9*, 163-175.

(11) Nangia, A., Conformational Polymorphism in Organic Crystals. *Acc. Chem. Res.* **2008,** *41*, 595-604.

24

(12) McCrone, W. C., Polymorphism. In *Physics and Chemistry of the Organic Solid State*, Fox, D.; Labes, M. M.; Weissberger, A., Eds. Wiley Interscience: New York, 1965; Vol. II, pp 725-767.

(13) Neumann, M. A.; van de Streek, J.; Fabbiani, F. P. A.; Hidber, P.; Grassmann, O., Combined crystal structure prediction and high-pressure crystallization in rational pharmaceutical polymorph screening. *Nat. Commun.* **2015,** *6*, 7793.

(14) Pfund, L. Y.; Matzger, A. J., Towards Exhaustive and Automated High-Throughput Screening for Crystalline Polymorphs. *ACS Comb. Sci.* **2014,** *16*, 309-313.

(15) Threlfall, T., Crystallisation of polymorphs: Thermodynamic insight into the role of solvent. *Org. Process Res. Dev.* **2000,** *4*, 384-390.

(16) Morissette, S. L.; Almarsson, O.; Peterson, M. L.; Remenar, J. F.; Read, M. J.; Lemmo, A. V.; Ellis, S.; Cima, M. J.; Gardner, C. R., High-throughput crystallization: polymorphs, salts, co-crystals and solvates of pharmaceutical solids. *Adv. Drug Delivery Rev.* **2004,** *56*, 275-300.

(17) Vasileiadis, M.; Pantelides, C. C.; Adjiman, C. S., Prediction of the crystal structures of axitinib, a polymorphic pharmaceutical molecule. *Chem. Eng. Sci.* **2015,** *121*, 60-76.

(18) Price, S. L., The computational prediction of pharmaceutical crystal structures and polymorphism. *Adv. Drug Delivery Rev.* **2004,** *56,* 301-319.

(19) Price, S. L., Predicting crystal structures of organic compounds. *Chem. Soc. Rev.* **2014,** *43*, 2098-2111.

25

(20) Braun, D. E.; Oberacher, H.; Arnhard, K.; Orlova, M.; Griesser, U. J., 4-Aminoquinaldine monohydrate polymorphism: prediction and impurity aided discovery of a difficult to access stable form. *CrystEngComm* **2016,** *18*, 4053-4067.

(21) Srirambhatla, V. K.; Guo, R.; Price, S. L.; Florence, A. J., Isomorphous template induced crystallisation: a robust method for the targeted crystallisation of computationally predicted metastable polymorphs. *Chem. Commun.* **2016,** *52,* 7384-7386.

(22) Gavezzotti, A., Are Crystal-Structures Predictable? *Acc. Chem. Res.* **1994,** *27*, 309-314.

(23) Price, S. L., Why don't we find more polymorphs? *Acta Crystallogr. B* **2013,** *69*, 313-328.

(24) Reilly, A. M.; Cooper, R. I.; Adjiman, C. S.; Bhattacharya, S.; Boese, A. D.; Brandenburg, J. G.; Bygrave, P. J.; Bylsma, R.; Campbell, J. E.; Car, R.; Case, D. H.; Chadha, R.; Cole, J. C.; Cosburn, K.; Cuppen, H. M.; Curtis, F.; Day, G. M.; DiStasio Jr, R. A.; Dzyabchenko, A.; van Eijck, B. P.; Elking, D. M.; van den Ende, J. A.; Facelli, J. C.; Ferraro, M. B.; Fusti-Molnar, L.; Gatsiou, C.-A.; Gee, T. S.; de Gelder, R.; Ghiringhelli, L. M.; Goto, H.; Grimme, S.; Guo, R.; Hofmann, D. W. M.; Hoja, J.; Hylton, R. K.; Iuzzolino, L.; Jankiewicz, W.; de Jong, D. T.; Kendrick, J.; de Klerk, N. J. J.; Ko, H.-Y.; Kuleshova, L. N.; Li, X.; Lohani, S.; Leusen, F. J. J.; Lund, A. M.; Lv, J.; Ma, Y.; Marom, N.; Masunov, A. E.; McCabe, P.; McMahon, D. P.; Meekes, H.; Metz, M. P.; Misquitta, A. J.; Mohamed, S.; Monserrat, B.; Needs, R. J.; Neumann, M. A.; Nyman, J.; Obata, S.; Oberhofer, H.; Oganov, A. R.; Orendt, A. M.; Pagola, G. I.; Pantelides, C. C.; Pickard, C. J.; Podeszwa, R.; Price, L. S.; Price, S. L.; Pulido, A.; Read, M. G.; Reuter, K.; Schneider, E.; Schober, C.; Shields, G. P.; Singh, P.; Sugden, I. J.; Szalewicz, K.; Taylor, C. R.; Tkatchenko, A.; Tuckerman, M. E.;

26

Vacarro, F.; Vasileiadis, M.; Vazquez-Mayagoitia, A.; Vogt, L.; Wang, Y.; Watson, R. E.; de Wijs, G. A.; Yang, J.; Zhu, Q.; Groom, C. R., Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallogr. B*. **2016,** *72,* 439-459.

(25) Karamertzanis, P. G.; Pantelides, C. C., Ab initio crystal structure prediction. II. Flexible molecules. *Mol. Phys.* **2007,** *105,* 273-291.

(26) Habgood, M.; Sugdan, I. J.; Kazantsev, A. V.; Adjiman, C. S.; Pantelides, C., Efficient Handling of Molecular Flexibility in Ab Initio Generation of Crystal Structures. *J. Chem. Theory Comput.* **2015,** *11*, 1957-1969.

(27) Neumann, M. A. *GRACE (the Generation, Ranking and Characterisation Engine)*, 1.0; Avant-garde Materials Simulation Deutschland GmbH: 2007.

(28) van Eijck, B. P.; Kroon, J., Structure predictions allowing more than one molecule in the asymmetric unit. *Acta Crystallogr. B* **2000,** *56*, 535-542.

(29) Pickard, C. J.; Needs, R. J., Ab initio random structure searching. *J. Phys.: Condens. Matter* **2011,** 23, (5).

(30) Neumann, M. A., Tailor-made force fields for crystal-structure prediction. *J. Phys. Chem. B* **2008,** *112*, 9810-9829.

(31) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C., The Cambridge Structural Database. *Acta Crystallogr. B* **2016,** *72*, 171-179.

(32) Feeder, N.; Pidcock, E.; Reilly, A. M.; Sadiq, G.; Doherty, C. L.; Back, K. R.; Meenan, P.; Docherty, R., The integration of solid-form informatics into solid-form selection. *J. Pharm. Pharmacol.* **2015,** *67*, 857-868.

27

(33) Bruno, I. J.; Cole, J. C.; Kessler, M.; Luo, J.; Motherwell, W. D. S.; Purkis, L. H.; Smith, B. R.; Taylor, R.; Cooper, R. I.; Harris, S. E.; Orpen, A. G., Retrieval of Crystallographically-Derived Molecular Geometry Information. *J. Chem. Inf. Comput. Sci.* **2004,** *44,* 2133-2144.

(34) Taylor, R.; Cole, J.; Korb, O.; McCabe, P., Knowledge-Based Libraries for Predicting the Geometric Preferences of Druglike Molecules. *J. Chem. Inf. Model.* **2014,** *54,* 2500-2514.

(35) McCabe, P.; Korb, O.; Cole, J., Kernel Density Estimation Applied to Bond Length, Bond Angle, and Torsion Angle Distributions. *J. Chem. Inf. Model.* **2014,** *54,* 1284-1288.

(36) *CSD Conformer Generator, Version 1.0*, Cambridge Crystallographic Data Centre: Cambridge, UK, 2015.

(37) Schaerfer, C.; Schulz-Gasch, T.; Ehrlich, H. C.; Guba, W.; Rarey, M.; Stahl, M., Torsion Angle Preferences in Druglike Chemical Space: A Comprehensive Guide. *J. Med. Chem.* **2013,** *56,* 2016-2028.

(38) Cruz-Cabeza, A. J.; Bernstein, J., Conformational Polymorphism. *Chem. Rev.* **2014,** *114,* 2170-2191.

(39) RDKit: Cheminformatics and Machine Learning Software. http://www.rdkit.org

(40) Schreyer, A. M.; Blundell, T., USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. *J. Cheminf.* **2012,** *4,* 12.

(41) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J.

28

L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox *Gaussian 09, Revision E.01*, Gaussian, Inc.: Wallingford CT, 2009.

(42) Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C.; Price, S. L.; Galek, P. T. A.; Day, G. M.; Cruz-Cabeza, A. J., Successful prediction of a model pharmaceutical in the fifth blind test of crystal structure prediction. *Int. J. Pharm.* **2011,** *418*, 168-178.

(43) Breneman, C. M.; Wiberg, K. B., Determining Atom-Centered Monopoles From Molecular Electrostatic Potentials - The Need For High Sampling Density in Formamide Conformational-Analysis. *J. Comput. Chem.* **1990,** *11*, 361-373.

(44) Coombes, D. S.; Price, S. L.; Willock, D. J.; Leslie, M., Role of Electrostatic Interactions in Determining the Crystal Structures of Polar Organic Molecules. A Distributed Multipole Study. *J. Phys. Chem.* **1996,** *100*, 7352-7360.

(45) Chisholm, J. A.; Motherwell, S., COMPACK: a program for identifying crystal structure similarity using distances. *J. Appl. Crystallogr.* **2005,** *38*, 228-231.

(46) Ismail, S. Z.; Anderton, C. L.; Copley, R. C.; Price, L. S.; Price, S. L., Evaluating a Crystal Energy Landscape in the Context of Industrial Polymorph Screening. *Cryst. Growth Des.* **2013,** *13*, 2396-2406.

(47) Bardwell, D. A.; Adjiman, C. S.; Arnautova, Y. A.; Bartashevich, E.; Boerrigter, S. X. M.; Braun, D. E.; Cruz-Cabeza, A. J.; Day, G. M.; Della Valle, R. G.; Desiraju, G. R.; van Eijck, B. P.; Facelli, J. C.; Ferraro, M. B.; Grillo, D.; Habgood, M.; Hofmann, D. W. M.; Hofmann, F.; Jose, K. V. J.; Karamertzanis, P. G.; Kazantsev, A. V.; Kendrick, J.; Kuleshova, L. N.; Leusen, F. J. J.; Maleev, A. V.; Misquitta, A. J.; Mohamed, S.; Needs, R. J.; Neumann, M. A.; Nikylov, D.; Orendt, A. M.; Pal, R.; Pantelides, C. C.; Pickard, C. J.; Price, L. S.; Price, S. L.; Scheraga, H. A.; van de Streek, J.; Thakur, T. S.; Tiwari, S.; Venuti, E.; Zhitkov, I. K., Towards crystal structure prediction of complex organic compounds - a report on the fifth blind test. *Acta Crystallogr. B* **2011,** *67*, 535-551.

(48) Corpinot, M. K.; Iuzzolino, L.; Price, S. L.; Bučar, D.-K.; Screening work in progress, 2017.

(49) Case, D. H.; Campbell, J. E.; Bygrave, P. J.; Day, G. M., Convergence Properties of Crystal Structure Prediction by Quasi-Random Sampling. *J. Chem. Theory Comput.* **2016,** *12,* 910-924.

(50) van Mourik, T., Assessment of Density Functionals for Intramolecular Dispersion-Rich Interactions. *J. Chem. Theory Comput.* **2008,** *4*, 1610-1619.

(51) van Mourik, T.; Karamertzanis, P. G.; Price, S. L., Molecular conformations and relative stabilities can be as demanding of the electronic structure method as intermolecular calculations. *J. Phys. Chem. A* **2006,** *110*, 8-12.

(52) Thompson, H. P. G.; Day, G. M., Which conformations make stable crystal structures? Mapping crystalline molecular geometries to the conformational energy landscape. *Chem. Sci.* **2014,** *5*, 3173-3182.

(53) Dey, A.; Kirchner, M. T.; Vangala, V. R.; Desiraju, G. R.; Mondal, R.; Howard, J. A. K., Crystal structure prediction of aminols: Advantages of a supramolecular synthon approach with experimental structures. *J. Am. Chem. Soc.* **2005,** *127*, 10545-10559.

(54) Hylton, R. K.; Tizzard, G. J.; Threlfall, T. L.; Ellis, A. L.; Coles, S. J.; Seaton, C. C.; Schulze, E.; Lorenz, H.; Seidel-Morgenstern, A.; Stein, M.; Price, S. L., Are the Crystal Structures of Enantiopure and Racemic Mandelic Acids Determined by Kinetics or Thermodynamics? *J. Am. Chem. Soc.* **2015,** *137*, 11095-11104.

31

For Table of Contents Only

Use of crystal structure informatics for defining the conformational space needed for

predicting crystal structures of pharmaceutical molecules.

Luca Iuzzolino, Anthony M. Reilly, Patrick McCabe and Sarah L. Price

32