Expanding crystal structure prediction to larger and more flexible molecules of pharmaceutical interest.

Luca Iuzzolino

November 2018

Submitted in partial fulfilment of the requirements for the degree of

Doctor of Engineering (EngD)

to

University College London

Department of Chemistry, 20 Gordon Street, London, WC1H 0AJ, United Kingdom

Declaration

I, Luca Iuzzolino, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

_____

Luca Iuzzolino

November 2018

Abstract

The use of Crystal Structure Prediction (CSP) studies in the pharmaceutical industry is currently limited by computational cost, which scales badly with molecular size and flexibility. This thesis seeks to develop new methods that would allow to perform CSP studies on larger, more flexible pharmaceutical-like molecules.

First, a full CSP workflow was successfully used to predict the crystal structure of a large flexible molecule for the 6th Blind Test and in a joint computational-experimental study of the antihelminthic drug mebendazole. These CSP studies were integrated with three previously published computational analyses of flexible pharmaceuticals and used to benchmark the development of new methods.

Successively, knowledge-based conformational information retrieved from the Cambridge Structural Database (CSD) was used to facilitate the generation of candidate crystal structures of these five molecules. Millions of crystal structures were generated at a reduced computational cost, but with an equally effective coverage of the conformational search space, compared to the original CSP efforts.

The importance of treating conformational flexibility when optimising search-generated crystal structures was then assessed. This led to using dispersion-corrected density functional tight-binding (DFTB-D) as an intermediate step to minimise all intra- and intermolecular degrees of freedom of several thousands of search-generated crystal structures. DFTB-D reduced the cost of the final lattice energy evaluations by providing better starting points, and results of similar quality to the original CSP studies were obtained after optimising only the intermolecular interactions with a higher quality wave-function.

Finally, a CSD survey was performed to determine thresholds that can discriminate the great majority of polymorphs from duplicate determinations. These thresholds and comparison methods were implemented in a Python programme that can be used in CSP studies to perform clustering and to interpret the results more effectively. The prospects for expanding the use of CSP to pharmaceutical development are discussed.

Impact statement

The work shown in this thesis can bring a variety of benefits both inside and outside academia. First, the development of new methods to cut down the computational cost of predicting the crystal structure/s of large and flexible molecules could be used by researchers interested in surveying the possible solid-state forms of a molecule. These methods could help drug development, as they are targeted to large and flexible pharmaceutical-like molecules. Faster predictions of the crystal structure/s of drug molecules could be undertaken concurrently with experimental solid form screening and could cover the search space more completely, reducing the experimental work needed to minimise the risk of the late appearance of a new polymorph. This is important as unpredicted changes in the solid-state form of a pharmaceutical molecule occurring in late stages of drug development or after it has been marketed can be disastrous both to manufacturers and to patients.

This thesis shows an example of a joint computational and experimental effort aimed at finding new crystalline forms of the antihelminthic drug mebendazole. The experimental study, which has yet to be completed, may find new forms with better properties than the currently marketed one, overcoming some of the difficulties in delivering this important drug to patients. This can impact all those industrial and academic researchers who are interested in leveraging both theory and experiments in solid form screening.

Furthermore, criteria to discriminate the great majority of polymorphs from duplicates/redeterminations have been found. These criteria could be used both experimentally, to verify whether two crystal structures of the same molecule generated in different experiments are polymorphs, and computationally, to remove duplicates generated in crystal structures prediction studies.

The benefits of this work are being realised in a variety of ways. Writing computational scripts was vital to effectively perform the research illustrated in this thesis. For example, an efficient Python clustering algorithm was produced, as well a Bash script to monitor a crystal structure search and interrupt it once a sufficient coverage of the potential energy surface has been achieved. All those tools have been put at the disposal of my group and are also being shared with collaborators in industry and in academia.

Many of the methods and results developed in this thesis have been published in peer-reviewed scientific journals and have been disseminated through presentations at conferences or at specialised seminars, including the Gordon Research Conference on Crystal Engineering, the annual CCDC student days and the Bologna's Convention on Crystal Forms. They can be utilised by members of my research group, possibly in

collaboration with industry. Some of these methods have already been successfully tested on succinic acid, increasing their credibility; the next step will be the expansion to larger and more flexible molecules, which may require further adaptions.

Acknowledgments

The first person I would like to thank is my primary supervisor, Prof. Sarah L. (Sally) Price. Her continuous support and advice has been vital throughout the whole doctoral programme. Constant encouragement and feedback have been fundamental for me completing the many different components of this thesis.

My industrial sponsor, the Cambridge Crystallographic Data Centre (CCDC), has been indispensable. They have organised the 6th Blind Test, which was my first introduction to crystal structure prediction, and they curate the immense Cambridge Structural Database as well as numerous high quality computational tools for data retrieval and analysis that were used throughout my whole EngD. Their organisation of annual student days, where I was able to present my work and receive comments and suggestions, has been invaluable. I would like to thank in particular Dr. Jason Cole and Murray Read for providing several useful scripts and for the insightful talks. Finally my research would not have been possible without the financial support of both the CCDC and the Engineering and Physical Science Research Council (EPSRC) under the grant EP/G036675/1.

I would also like to thank my industrial (CCDC) supervisors, Dr. Anthony M. Reilly (September 2014-January 2017) and Dr. Patrick McCabe (from January 2017). They have been very supportive in guiding my research, proposing new ideas, helping with several computational tools and assisting with reports and publications.

My sincere gratitude also goes to Prof. Claire S. Adjiman, Prof. Costas Pantelides and Dr. Isaac Sugden at Imperial College London, who are responsible for developing and providing CrystalPredictor and CrystalOptimizer, which I have used constantly throughout my research.

Another special thanks goes to Dr. Louise S. Price, who first introduced me to the CSP algorithms I used in the thesis and whose deep experience has cleared many of my doubts. Many thanks to Dr. Jan G. Brandenburg, who introduced me to semi-empirical quantum mechanical methods and helped perform and analyse the DFTB calculations, as well as carrying out the DFTB phonon calculations. I would also like to thank Dr. Krešo Bučar and Merina Corpinot, who performed the experimental portion of the solid from screen on mebendazole and helped me comparing their results with those produced by my computational analysis.

Furthermore, the Centre for Doctoral Training in Molecular Modelling and Materials Science (CDTM3S) at UCL has provided extensive training in the first year of my EngD and has given me the chance of presenting my work at annual Industry Days in front of a diverse audience. Special thanks go to Dr. Zhimei Du.

I am very grateful to all the other current and former members of the Price research group based in G18: Dr. David H. Case, Dr. Rui Guo, Dr. Robert W. (Bob) Lancaster, Dr. Rona E. Watson, Dr.Rebecca K. Hylton and Alexander Aina. They have helped me in uncountable ways, from scientific advice to fun conversations.

Finally, I would like to thank my parents, family and friends for having been very patient and supportive throughout these challenging four years.

List of Publications

Main publications:

Iuzzolino, L.; Reilly, A. M.; McCabe, P.; Price, S. L., Use of Crystal Structure Informatics for Defining the Conformational Space Needed for Predicting Crystal Structures of Pharmaceutical Molecules. Journal of Chemical Theory and Computation 2017, 13 (10), 5163-5171.

Iuzzolino, L.; McCabe, P.; Price, S. L.; Brandenburg, J. G., Crystal structure prediction of flexible pharmaceutical-like molecules: density functional tight-binding as an intermediate optimisation method and for free energy estimation. *Faraday Discussions* 2018, *Advance Article.*

Further publications:

Lucaioli, P.; Nauha, E.; Gimondi, I.; Price, L. S.; Guo, R.; **Iuzzolino, L**.; Singh, I.; Salvalaglio, M.; Price, S. L.; Blagden, N., Serendipitous isolation of a disappearing conformational polymorph of succinic acid challenges computational polymorph prediction. CrystEngComm 2018, 20 (28), 3971-3977.

I performed a CSD survey on the conformational preferences of succinic acid in single-component crystal structures, cocrystals, solvates and salts.

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.; 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 Crystallographica Section B* 2016, *72* (4), 439-459

As a member of the Price group, I performed a CSP study that successfully predicted the crystal structure of molecule XXVI, the largest Blind Test target to date.

Table of contents

List of Figures

structure, labelled according to its space group. The structure that ended up matching the experimental form is indicated.

Figure 3.10: Lattice energy vs density plots showing (a) the 100 lowest energy crystal structures fully optimised with CrystalOptimizer submitted as a first list of predictions (b) the 100 lowest energy crystal structures after a rigid-body optimisation in a PCM with ε=3 and with the addition of the vibrational component to Helmholtz free energy at 298 K submitted as a second list of predictions. See Appendix Tables 3.3-3.4 for more details. Each point on the plots corresponds to a separate crystal structure, labelled according to its space group. The structures matching the target experimental form are indicated.

Figure 3.11: Overlay between the hydrogen bonded dimer in the experimental crystal structure of molecule XXVI (coloured by elements) and predicted structure 1600 in the first list (in green). The hydrogen bonds are coloured in purple. The $RMSD_{15}$ for the 15/15 overlay is 0.276 Å.

Figure 3.12: (above) Lattice energy vs density plot of the 100 lowest energy structures submitted as a first list of predictions (below) free energy vs density plot the 100 lowest energy structures submitted as a second list of predictions. They are labelled on whether they form intra- or intermolecular hydrogen bonds.

Figure 3.13: Overlay of the conformations of structures 1600 (coloured by elements), which matches the target experimental crystal structure, 675 (yellow), 421 (blue) and 2231 (red) (left) and the sheet common to all four structures (right).

Figure 3.14: (left) Overlay of the conformation of structure 3525 (coloured by elements), the global minimum in Elatt in the first list, and the isolated-molecule global minimum of molecule XXVI (in blue) (right) and its crystal structure.

Figure 3.15: Overlay of the conformations of structures 3104 (coloured by elements), 185 (blue), 1391 (red) and 7559 (yellow) (left) and the packing motif common to those 4 structures (right).

Figure 4.1: Chemical diagrams of (left) the A-tautomer of mebendazole (right) the C-tautomer of mebendazole

Figure 4.2: Intermolecular hydrogen bond motif of (above) mebendazole form A, where each molecule uses two donors and two acceptors, forming both a NH···O hydrogen bond with the $R_2^2(14)$ graph set motif and a NH···N hydrogen bond with the $R_2^2(8)$ graph set motif, and (below) mebendazole form C, where each molecule uses one donor and one acceptor, forming a NH···N hydrogen bond with the $R_2^2(8)$ graph set motif. The hydrogen bonds are coloured in purple.

Figure 4.3: Chemical diagram of the (above) A-tautomer (below) C-tautomer of mebendazole. The two tautomers are differentiated by where C14=O15 in the edge benzoyl substituent attaches to the central benzimidazole ring: in the A-tautomer it attaches three bonds away from the N1-H1a group, to atom C7, while in the C-tautomer it attaches four bonds away, to atom C8. The six torsion angles that wereconsidered in the initial analysis of conformational flexibility are indicated (see Appendix Table 4.1 for their definition). The scans were only performed on the C-tautomer.

Figure 4.4: Results of the angle scans of torsion angles a) Φ_1 (x-axis) and Φ_2 (y-axis) from 0° to 360° in 30° steps; this is a contour plot of the conformational energy penalty surface as a function of the values of these two torsion angles b) Φ_3 from 0° to 360° in 30° steps c) Φ_4 from 0° (*cis*) to 180° (*trans*) in 90° steps d) Φ_5 from 0° to 360° in 30° steps and e) Φ_6 from 0° to 360° in 30° steps. At each point on the plots, all the conformational degrees of freedom (CDFs) were relaxed with the exception of the scanned torsion angle/s. The optimisations were performed at the PBE0 6-31G(d,p) level of theory starting from the PBE0 6-31G(d,p) optimised global minimum gas-phase conformer. The red dots/lines indicate the values taken by the torsion angle/s in the conformation of form C of mebendazole, black dots/lines in the conformation of form A (see Appendix Table 4.2 for the actual values).

Figure 4.5: Surrogate molecules used to calculate the ΔE_{intra} grids of (left) Φ_1 and Φ_2 and (right) Φ_3 and Φ_5 of both tautomers of mebendazole.

Figure 4.6: Chemical diagram and atomic numbering of (above) the A-tautomer (below) the C-tautomer of mebendazole, showing the torsion angles (black arrows) and bond-angles (red arcs) treated as independent CDFs in the final refinement stage of this CSP study. Double arrows indicate that two definitions of torsion angles around the same central bond were treated as variables in the CrystalOptimizer optimisations. See Appendix Table 4.4 for the precise definition of the explicitly flexible torsion and bond-angles.

experimental conformations are coloured in red or in yellow. Polymorphs with very similar conformations are shown only once.

Figure 5.4: Chemical diagrams of the molecules used to test the applicability of CSD information to large and flexible targets, showing the torsion angles that are identified as flexible by the rotamer libraries and the number of distinct conformations generated by the CG. The additional angles not identified by the rotamers libraries are in green and define the position of polar hydrogen atoms. Atomic numbering can be found in Appendix Figure 5.1, and the definition of the torsion angles in Appendix Table 5.1.

Figure 5.5: Overlays of the experimental conformations of the molecules in Figure 5.4 (coloured by elements) with their best matches produced by the CG (in blue).

Figure 5.6: Visual comparison of the effects of changing by 30° (a) torsion angle Φ_4 in molecule XXVI, which strongly affects molecular shape and (b) torsion angle Φ_6 in molecule XXIII, which has a very small effect on the shape of the hydrogen bonded dimer, except in the proximity of the carboxylic acid functional groups.

Figure 5.7: Decision tree used to discriminate between constrained and explicitly flexible torsion angles on the basis of the PDF $f(\theta)$ values and the changes in shape associated with their variation.

Figure 5.8: Decision tree used to choose the separation threshold of each torsion angle defining a separate CR from the PDF $f(\theta)$ and shape-matching characteristics.

Figure 5.9: Chemical diagram and atomic numbering of succinic acid. The torsion angles identified as flexible by the CG are in black, while those in green define the position of polar hydrogen atoms. Their definition can be found in Appendix Table 5.8.

Figure 5.10: Summary of the application of the workflow on the molecules in Figure 5.4. 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; note that many combinations of these values were eliminated as energetically unfeasible (see Appendix Tables 5.10-5.15). 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 (the motivation for this assumption is illustrated in Chapter 4).

Figure 5.11: Plots of the significant CSP-generated crystal structures found in previous studies, classified as to whether they were found by the search workflow, of (a) molecule XXVI, (b) GSK269984B, (c) molecule XX, (d) molecule XXIII and (e) mebendazole. Higher energy structures were included for molecule XXVI (a structure whose stability was very dependent on the energy model), molecule XX and mebendazole (a competitive crystal structure with a *cis* amide for both molecules). Structures matching the experimentally known forms are indicated by open diamonds.

Figure 5.12: Summary of the application of the workflow to succinic acid. 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 (see Appendix Table 5.21). Torsion angles in green were constrained to the indicated values having been determined from an *ab initio* conformational-energy scan. Since the two halves of the molecule are symmetric, only one combination 180°-0° values for the green torsion angles Φ_1 and Φ_5 was considered, as they would represent identical molecules upon switching.

Figure 6.1: Chemical diagrams of the five molecules considered for this test. The black arrows indicate the $CDFs_{torsion}$.

Figure 6.2: Chemical diagram of the five molecules considered for this test, showing the torsion angles and bond-angles treated as independent in the CrystalOptimizer optimisation with the $CDFs_{AUTODOF}$. Black arrows indicate the torsion angles that are also present in the $CDFs_{torsion}$ and that are selected according to chemical intuition, while red arrows and red arcs indicate the torsion and bond-angles added by AUTODOF as they include polar hydrogen atoms.

Figure 6.3: Chemical diagrams of the five molecules considered for this test. The torsion angles indicated by black arrows and the bond-angles indicated by red arcs were present in the $CDFs_{AUTODOF}$ and varied significantly in at least one crystal structure optimised with CrystalOptimizer with the $CDFs_{all}$. Green arrows and arcs indicate torsion and bond-angles not included in the $CDFs_{AUTODOF}$ but that varied significantly in at least one crystal structure, while

those in turquoise are included in the $CDFs_{AUTODOF}$ but did not vary significantly in any crystal structure.

Figure 6.4: Chemical diagrams of the five molecules considered for this test, showing the torsion and bond-angles treated as independent in the CrystalOptimizer optimisation with the $CDFs_{AUTODOF+}$. The torsion angles indicated by red and black arrows and bond-angles indicate by red arcs were present in the $CDFs_{AUTODOF.}$ Blue arrows and arcs indicate the torsion and bond-angles that were manually added as they belong to one of the three categories of intuitively rigid CDFs that varied significantly in the CrystalOptimizer minimisations performed with the $CDFs_{all}$.

Figure 6.5: Comparison between the lattice energies obtained after full optimisations with CrystalOptimizer with the various sets of independent CDFs and those with the $CDFs_{all}$ for the 20 crystal structures of (a) molecule XXVI (b) molecule XXIII (c) the A-tautomer of mebendazole (d) the C-tautomer of mebendazole and (e) naproxen. For each molecule, the lattice energies are plotted as the difference with the E_{latt} value of the global minimum in the benchmark optimisations with the $CDFs_{all}$. The black lines indicate the lattice energies obtained in the optimisations with the $CDFs_{all}$, and the structures matching the experimentally-characterised forms are indicated. See Appendix Tables 6.6-6.10 for more detailed results. The presence of crystal structures with energies lower than those obtained with the $CDFs_{all}$ is probably due to the tolerances in the convergence criteria of the optimisations.

Figure 6.6: Comparison between the lattice energies obtained after full optimisations with CrystalOptimizer with the various sets of independent CDFs and those with the $CDFs_{all}$ for the 20 crystal structures of (a) molecule XXVI (b) molecule XXIII (c) the A-tautomer of mebendazole (d) the C-tautomer of mebendazole and (e) naproxen. For each crystal structure, the energy is shown as the difference with that obtained in the benchmark optimisations with the $CDFs_{all}$ for the same structure. The crystal structures that match experimentally-characterised forms are circled. See Appendix Tables 6.6-6.10 for more detailed results. The presence of crystal structures with energies lower than those obtained with the $CDFs_{all}$ is probably due to the tolerances in the convergence criteria of the optimisations.

Figure 6.7: For each molecule, average CPU cost of the optimisations as a function of the number of independent CDFs.

Figure 6.8: Fragments used to perform the CSD surveys with Conquest. The black arrow represents a torsion angle, red arcs bond-angles. Z indicates any non-H atom, while the 'a' subscript indicates that the atom is acyclic. Fragments (a), (b) and (c) represent categories 1-3 of intuitively rigid CDFs respectively (see section 6.2.5.2).

Figure 6.9: Fragments used to perform the CSD surveys with Conquest, representing torsion and bond-angles in rigid rings in the absence of heavy substituents in the central atoms. The black arrow represents a torsion angle, the red arc a bond-angle. Z indicates any non-H atom, X any atom, while the '1' subscript indicates that the atom forms only one bond.

Figure 6.10: Histogram showing the CSD distribution of the values of the torsion angle in the fragment in Figure6.8a.

Figure 6.11: Histogram showing the CSD distribution of the values of the torsion angle in the fragment in Figure6.9a.

Figure 6.12: Histogram showing the CSD distribution of the values of the bond-angle in the fragment in Figure 6.8b.

Figure 6.13: Histogram showing the CSD distribution of the values of the bond-angle in the fragment in Figure 6.9b.

Figure 6.14: Histogram showing the CSD distribution of the values of the bond-angle in the fragment in Figure6.8c.

Figure 7.1: Plots of the crystal energy landscapes obtained after the final optimisations with the $\Psi_{mol}^{PBE0+FIT}$ method of (a) molecule XXVI (b) GSK269984B (c) molecule XX (d) molecule XXIII and (e) mebendazole. Each point on the plots corresponds to a separate crystal structure, labelled according to its space group. The structures matching the known experimental forms are indicated. The relative energies of the Z'=2 polymorphs of XXIII (XAFPAY02 for form C and XAFPAY04 for form E), which were outside the scope of the searches performed in Chapter 5, were calculated independently and are only shown in Figure 7.3.

Figure 7.2: For some key CSP-generated crystal structures of each molecule, the relative vibrational contributions ΔF_{vib}, which can be added to E_{latt} to calculate the Helmholtz free energy

(A, see Equation 2.37). For each molecule, ΔF_{vib} is calculated relative to the structure which is the E_{latt} global minimum after the $\Psi_{mol}^{PBE0+FIT}$ optimisations, for which ΔF_{vib} = 0. The structures matching the experimentally known forms are indicated in green. The rigid-body modes are pure lattice modes, calculated with the $\Psi_{mol}^{PBE0+FIT}$ model using DMACRYS. Full details about the crystal structures and their energies can be found in Appendix Table 7.7.

Figure 7.3: Relative energies of the five polymorphs of molecule XXIII calculated in this study compared with those reported by the participants of the 6th Blind Test. Note that values linked by dashed lines denote changes from adding free energy estimates. The number in parentheses corresponds to the group identifier, with R denoting participants who only optimised crystal structures generated by others.

Figure 8.1: Comparison of the PXRD similarity and the number of molecules matched (out of 15) for the 3,925 pairs of crystal structures under consideration.

Figure 8.2: Decision tree to discriminate whether two organic crystal structures characterised at similar pressures are duplicates or polymorphs.

Figure 8.3: Schematics of the criteria used in the algorithm to decide whether two computer-generated crystal structures are duplicates or belong to different clusters. The blue boxes indicate steps where the user can define the thresholds for success or failure, while the green boxes represent hard-coded criteria.

Figure 8.4: Distribution of the number of molecules that could be overlaid for the 3,347 pairs of polymorphs. They correspond to the 3,371 pairs of crystal structures considered, excluding the 24 identified as duplicates.

Figure 8.5: 15-molecule overlay between the crystal structures of DIMETH01 (coloured by elements) and DIMETH06 (in green). The $RMSD_{15}$ is 0.226 Å. The publication clearly states that they are different polymorphs.

Figure 8.6: Comparison between the crystal structures of FORMAC (left) and FORMAC01 (right).

Figure 8.7: Summary of the results of the analysis of the polymorph-flagged CSD crystal structures. The pairs of false polymorphs are listed in Appendix Table 8.3 and the pairs of false duplicates in Appendix Table 8.4.

Figure 8.8: Summary of the results of the test on the polymorph-flagged CSD crystal structures for each step of the methodology in Figure 8.2.

Figure 8.9: 1-molecule overlay of the crystal structures of DBEZLM01 (coloured by elements) and DBEZLM05 (in green). See Appendix Table 8.3 for details.

Figure 8.10: Molecular conformations contained in (left) ANTCYB11 and (right) ANTCYB13. See Appendix Table 8.3 for details. The phenyl rings in ANTCYB11 are intuitively wrong.

Figure 8.11: Molecule and symmetry elements of (right) BENZID04 and (left) BENZID08. See Appendix Table 8.3 for details. Those crystal structures are both in the $P\bar{1}$ space group and a 15/15 molecule overlay is possible with an $RMSD_{15}$ of 0.181 Å. The yellow dots represent crystalline inversion centres, and in BENZID04 one of the inversion centre is located half the way through the central C-C bond.

Figure 8.12: 15-molecule overlay of DXYLEN14 (coloured by elements) and DXYLEN15 (in green). See Appendix Table 8.3 for details. Despite a clearly excellent agreement, the tool returns an unrealistic $RMSD_{15}$ of 2.335 Å. This has to be due to an error in the Crystal Packing Similarity tool.

Figure 8.13: 15-molecule overlays of (above) BIZWAI01 and BIZWAI02, with an $RMSD_{15}$ of just 0.007 Å, despite being explicitly indicated as polymorphs in the publication (below) SUWMIG and SUWMIG03, which require looser 40% distance and 40° angle tolerances to obtain a 15/15 overlay with an $RMSD_{15}$ of 1.16 Å, despite being explicitly indicated as redeterminations in the publication.

Figure 8.14: Histograms showing the simulated PXRD similarities of (above) the polymorphic pairs identified; the false duplicates are indicated by orange bars (below) the pairs of duplicates identified; false polymorphs are indicated by orange bars.

Figure 8.15: Lattice energy vs density plot obtained after the search with CrystalPredictor of molecule XXVI (above) before and (below) after clustering.

Figure 8.16: Lattice energy vs density plot obtained after the search with CrystalPredictor of the A-tautomer of mebendazole (above) before and (below) after clustering.

Figure 8.17: Lattice energy vs density plot obtained after the search with CrystalPredictor of the C-tautomer of mebendazole (above) before and (below) after clustering.

Figure 8.18: Overlay of structures 2494 and 2495 (right) after the search, where it is possible to overlay 15/15 molecules with an $RMSD_{15}$ of 0.006 (left) after the single-iteration of CrystalOptimizer, where it is only possible to overlay 5/15 molecules. Molecules of 2494 are coloured by elements, while for 2495 matched molecules are coloured in green, non-matched molecules in red.

List of Tables

List of Symbols and Abbreviations

Φ, ξ	Flexible torsion angle
Ψ	Wave-function
Ψ_{crys}	Modelling methods where the wave-function of the whole crystal structure is calculated
Ψ_{mol}	Modelling methods where the wave-function of each symmetry-independent molecular conformation/s in a crystal structure is calculated
A	Helmholtz free energy
ΔA	Relative difference in Helmholtz free energy between two crystal structures
API	Active pharmaceutical ingredient
B3LYP	Becke, three-parameter, Lee-Yang-Parr exchange-correlation functional
CCDC	Cambridge Crystallographic Data Centre
CDF	Conformational degree of freedom
CG	Conformer Generator
CPU	Central processing unit
CR	Conformational region
CSD	Cambridge Structural Database
CSP	Crystal structure prediction
D3	Grimmes's atom-pairwise dispersion correction
DFT	Density functional theory
DFT-D	Dispersion-corrected density functional theory
DFTB	Density functional tight binding
DFTB-D	Dispersion-corrected density functional tight binding
DMA	Distributed multipole analysis
DSC	Differential scanning calorimetry
ΔE_{intra}	Intramolecular (conformational) energy penalty
E_{latt}	Lattice energy
ΔE_{latt}	Relative difference in lattice energy between two crystal structures
F_{vib}	Vibrational component of the Helmholtz free energy
ΔF_{vib}	Relative differences in the vibrational component of the Helmholtz free energy between two crystal structures
KDE	Kernel density estimation
LAMs	Local approximate models
MBD	Many-body dispersion correction
NMR	Nuclear magnetic resonance

PBE	Perdew Burke and Ernzerhof exchange-correlation functional
PBE0	Perdew Burke and Ernzerhof exchange-correlation functional with 25% exact Hartree Fock exchange energy
PCM	Polarisable continuum model
PDF	Probability density function
PPM	Putative polymorph
PXRD	Powder X-ray diffraction
Python API	Python Application Programming Interface
QM	Quantum-mechanical
RMS	Root-mean square
$RMSD_x$	Root-mean square deviation for a cluster of *x* molecules
3OB	Third-order parametrisation for organic and biological systems
TMFF	Tailor-made force field
TS	Tkatchenko-Scheffler dispersion correction
U_{inter}	Intermolecular energy
USR	Ultrafast shape recognition
XDM	Exchange-hole dipole moment dispersion correction
Z	Number of molecules in the unit cell
Z’	Number of molecules in the asymmetric unit

Chapter 1: Introduction

1.1 Polymorphism

1.1.1 Definition and importance

Polymorphism occurs when molecules with the same chemical composition, covalent bonding and stereochemistry form different crystal structures.[1, 2] This definition excludes solvates and hydrates of a molecule, since they have different chemical compositions, as well as crystal structures containing different isomers and tautomers, since they have different covalent bonding and stereochemistry.[3] Although several definitions have been proposed to include these as special cases of polymorphism,[3, 4] in this thesis a strict demarcation is followed.

Polymorphism has been known for a long time. The first intuition of its existence dates back to 1822 when Mitscherlich noticed that different crystals of the same compounds displayed different chemical and physical properties.[3, 5] Although McCrone in 1965 famously stated that “every compound has different polymorphic forms and that, in general, the number of forms known is proportional to the time and money that has been spent in research on that compound”,[6] polymorphism was considered just a scientific curiosity until the last few decades, a rare phenomenon that affected a minority of compounds, while monomorphism was considered the norm.[7-9]

However, the interest in polymorphism has dramatically increased from the 1990s.[10] This can be explained by several concurring factors. First of all, the increasing understanding of how much a change in the arrangement of molecules can affect solid-state properties:[5] polymorphs can differ in melting point, colour, solubility, conductivity, physical and chemical stability, reflective index and mechanical properties.[2, 3, 11] ROY (from its red, orange, yellow crystals) provides a striking example of how a change in the crystal structure can affect the properties of solids formed of molecules with identical covalent bonding,[12, 13] as shown in Figure 1.1.[12]

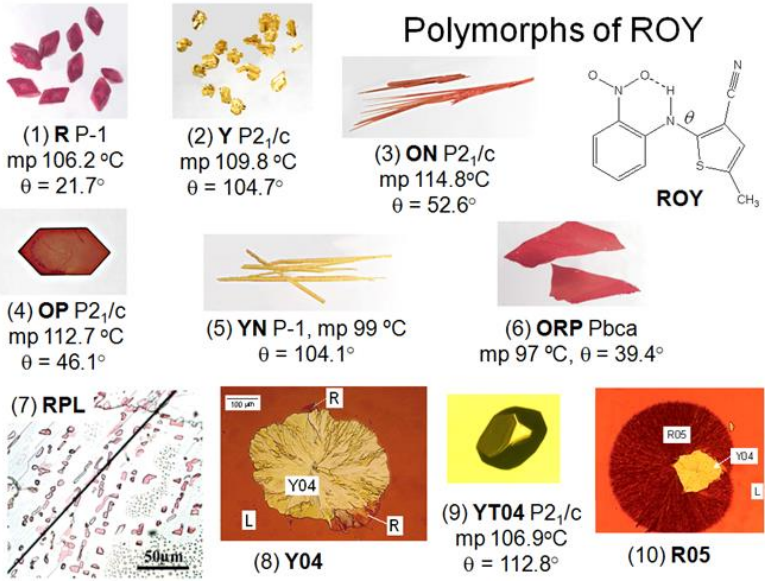


Figure 1.1: Molecular diagram of ROY and photos of its 10 known crystal structures. For those that have been solved the space group and the value taken by torsion angle θ are also indicated. The crystal structure of polymorph R05 has recently been solved:[14] it has two molecules in the asymmetric unit cell, with θ values of 44.9° and -34.0° respectively, and it crystallises in the $P2_1$ space group.

Secondly, the improvement in the experimental techniques to search for, characterise and compare polymorphs appears to confirm the validity of McCrone's statement;[5] polymorphism is now recognised as a prevalent phenomenon,[8, 9] and the majority of molecules that are thoroughly screened are found to form more than one single-component crystal structure.[10] Finally, there have been some highly publicised cases where polymorphism has led to manufacturing failures or to patent litigation,[5] which have upgraded this phenomenon from the subject of mere 'academic' scientific interest to an industrially relevant problem with important financial and legal implications.

1.1.2 Relevance to the pharmaceutical industry

Polymorphism is relevant for any industry that deals with solid-state products, such as agrochemicals, pigments, semiconductors and energy materials.[2, 15] The pharmaceutical industry is the one for which polymorphism can have the most severe consequences,[1, 3, 9, 16-18] and large sums of money and human time are invested to understand and analyse this phenomenon. The reason is that most drugs are formulated with the active pharmaceutical ingredients (APIs) in the crystalline state.[3, 18] This is because APIs in the form of solid crystal structures are easier to manufacture and more stable (*i.e.* less prone to degradation and phase transformations) than when they are in solution or in the amorphous state.[3, 18] Since the properties of a solid are dramatically affected by the ways molecules are arranged in space, when a candidate API moves from the discovery to the development stage the selection of an optimal crystalline form takes a fundamental importance.[2, 3, 17]

The most important polymorph-dependent properties for an API are bioavailability and chemical and physical stability.[3] Bioavailability is the proportion of a drug that has an effect on the patient, and it is related to solubility and dissolution rate.[2, 3, 19] More thermodynamically stable polymorphs are less soluble,[1, 3, 19] and the differences in solubility between different crystal structures can be large.[3, 19] Chemical stability is related to the reactivity of a molecule; unstable solid forms are undesirable as they lead to a loss of drug product and possibly to the formation of toxic degradants.[3] Physical stability is related to the tendency of a crystalline API to undergo transformation to another form, such as a different polymorph, a hydrate or an amorphous solid;[3] this is obviously undesirable as it alters the physical properties of the drug product. More thermodynamically stable solid forms generally have higher chemical and physical stability. Mechanical properties are also polymorph-dependent.[1, 3, 17] They can be relevant as they affect the mechanical strength and the tableting behaviour of a drug, but they are not as important as stability and bioavailability as they can be overcome by the choice of appropriate excipients and by the formulation process.[3]

The implications of polymorphism are so important that regulatory agencies require an accurate understanding of the solid-state structural behaviour and to have robust processes to produce the correct form of an API.[1, 3, 17] The risks of polymorphism for the pharmaceutical industry are well-known, and a review by Lee *et al.* written in 2011[3] lists five cases in which a change in the crystal structure of an API created problems to industrial manufacturers: Tegretol (carbamazepine) in 1988,[20] Norvir (ritonavir) in 1998,[21] Neupro (rotigotine) in 2008,[22] Avalide (irbesartan-hydrochlorothiazide)[23] and Coumadin (warfarin sodium 2-propanol solvate) in 2010.[3] It is likely that several other cases have occurred, which have not been publicly reported.[3] The most famous example is certainly the one of ritonavir, as it is emblematic in illustrating the financial, reputational and health-related risks that are associated with an incomplete understanding of the solid-state behaviour of an API. Only form I was initially found during drug development at Abbott. This form was introduced on the market in 1996, but in 1998 another polymorph, form II, appeared.[21] Form II contains a high-energy conformation, present in only 1% of the molecules in solution,[24] but it has a more complete hydrogen bonding network than form I that makes it more thermodynamically stable.[24] It became impossible to reliably produce the original form.[3] This new polymorph was significantly less soluble than the one that had originally been marketed,[24] compromising its oral bioavailability. Abbott had to withdraw ritonavir from the market,[3] and after huge research efforts it eventually came back as a soft-gelatin capsule.[25]

Finally, polymorphs are considered patentable inventions,[3] and the discovery of different solid forms can be exploited to extend the patent-life of a drug.[26]

1.1.3 Solid form screens

Given the importance of understanding the solid-state behaviour of a molecule, solid form (or polymorph) screens are a fundamental component of the drug development processes.[1-3, 17] The purpose of solid form screens is to identify all the possible solid-state forms of an API and to choose the one with the best balance of properties, which may eventually be marketed.[3]

Solid form screens rely on the principle that crystallising the target molecules under a comprehensive set of external conditions should produce all its polymorphs, as well as its solvates and hydrates and any amorphous form.[1, 3, 17] The set of external conditions that can be adjusted is very broad. Crystallisation can be performed from a variety of solvents and solvent mixtures, which can promote different intermolecular interactions in the solid-state.[2] Cooling rates, temperatures, pHs and degrees of supersaturation can also be varied to obtain different forms upon crystallisation from solution,[2] and different seeds and/or impurities can be used to promote heterogeneous nucleation.[2] Furthermore, crystallisation can be performed from the melt, from amorphous materials or from a slurry containing a solid-state form dissolved in a solvent.[2] Heat, pressure or mechanical deformation (such as liquid-assisted grinding) can then be applied to promote solid-solid transformations.[3] Finally, less conventional methods are often included in current solid form screening procedures, such as crystallisation in electrical or ultrasound fields, under laser beams or in nano-confined forms,[1, 17] as size can stabilise different polymorphs because of the increasing importance of surface energy for smaller crystals.[2, 27]

However, a trial and error approach must be utilised since the set of conditions that permit to find all the possible solid forms cannot be known prior to the screen.[1, 3, 17, 18] All molecules are different, and there is no pre-defined 'cookbook' approach that is guaranteed to find all important forms for any target.[1] Even high-throughput screening procedures that automatically test a multitude of different conditions can sometimes fail.[1, 3] Hence, polymorph screens have no clear end-point and their extent is arbitrary.[18] Crystallisation experiments are generally performed well after the last solid form has been identified to minimise the risk of missing anything important.[18] Because of the low success rate of drug development and the huge amount of money, time and human resources that are required to perform a comprehensive experimental solid form screen, this process is very costly to pharmaceutical manufacturers.[3, 18] The impossibility of determining the optimal extent of a screen causes the waste of precious resources.[18]

As a consequence, the ability to predict how a molecule crystallises, ideally not just as a single component, but also as a hydrate, or with a co-crystal former, is of tremendous value to the pharmaceutical industry.[1, 17, 18] Predictions could be used to confidently stop the polymorph screen once all important solid-state forms have been

found, as well as to direct the experimental effort towards those with the most desirable set of properties.[1, 17, 18] Fast and reliable predictions have the potential to revolutionise drug development.[18]

1.2 Crystal structure prediction (CSP)

1.2.1 Background

The prediction of the possible crystal structures a molecule could adopt, a very important goal for the pharmaceutical industry, requires programming in a computer code a model of what determines its crystallisation behaviour.[1] The art of using computer models to predict how a molecule crystallises takes the name of crystal structure prediction (CSP).[1, 7, 9, 16-18] The origins of CSP date back to the 1960s-1970s,[18] with the first proper CSP study performed on benzene in 1983;[28] however, the techniques were still immature, and in 1988 John Maddox famously stated that "One of the continuing scandals in the physical sciences is that it remains in general impossible to predict the structure of even the simplest crystalline solids from a knowledge of their chemical composition".[29] The progresses since that stimulating remark have been dramatic, thanks to the increasing power of computers and to improvements in CSP methodologies. However, scientific curiosity was still the main driving force for development, and CSP was considered a way to test the accuracy of theoretical models rather than as a technology to be applied to real-life problems.[18] It was not until the 5th Blind Test of CSP in 2010 (see Chapter 3.1.1 for details) that the pharmaceutical industry started to become interested in the commercial application of this technique;[18, 30, 31] in that occasion two groups successfully predicted the crystal structure of molecule XX (see Figure 1.2),[32, 33] whose size and level of flexibility are similar to those of small molecules in pharmaceutical development, giving CSP much needed credibility.

Since then, the commercial interest in CSP has drastically increased, to the point that most large pharmaceutical manufacturers are interested in including computational predictions in their drug development procedures.[18] As the title of a recent article points out, "Crystal structure prediction is changing from basic science to applied technology".[18]

1.2.2 Overview of the methodologies

The theoretical assumption behind all successful CSP methodologies is that molecules crystallise in thermodynamically stable crystal structures.[1, 7, 9, 18] In general, CSP methodologies follow the framework shown in Figure 1.2.

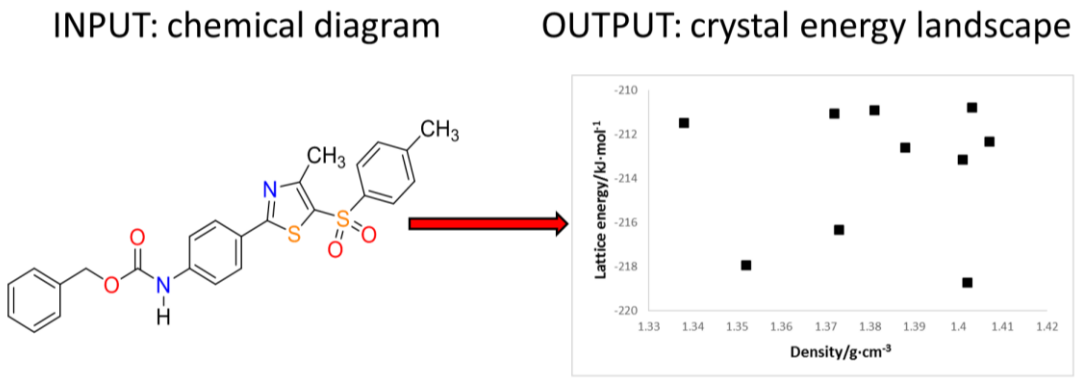


Figure 1.2: Framework followed by successful CSP methodologies based on finding the most thermodynamically stable crystal structures. The Blind Tests CSP study of molecule XX[32, 33] is taken as an example.

The input of a CSP methodology is only the chemical diagram of a molecule, while the output is the crystal energy landscape, which is the set of the computer-generated crystal structures that can be considered as thermodynamically plausible polymorphs and their energies.[1, 7, 17] A summary plot of a crystal energy landscape is shown in Figure 1.2, where each point represents a unique minimum on the potential energy surface (in this case modelled by lattice energy at 0 K, see Chapter 2 for details). Current CSP methodologies have become reliable at producing crystal energy landscapes, since known forms are in most cases found among the set of thermodynamically plausible crystal structures.[7, 18]

However, some problems remain. First of all, the potential energy surface is immense, and a severe limitation of the search space, in particular in the number of molecules in the asymmetric unit cell (*i.e.* Z'), is needed.[1, 18] Most CSP methodologies are restricted to Z'=1,[1, 7] with Z'=2 searches performed only rarely,[9, 19] and so a large portion of the potential energy surface with non-integer or large Z' values is not accessible by CSP. It is also common to introduce limitations in the range of molecular conformations, which may cause the missing of crystal structures containing strained conformations.[9, 18] Furthermore, CSP studies can only produce crystal structures with perfect infinite static lattices,[1, 16] excluding disordered (although disorder can be anticipated in some cases)[34, 35] and defective forms from its reach.[7, 18]

Nevertheless, despite these sources of under-prediction, CSP methodologies actually tend to over-predict polymorphism.[8, 18] One possible explanation is that temperature effects are not accurately modelled, and many of the crystal structures that are predicted as distinct minima at 0 K would not survive under real-life experimental conditions.[8, 18] However the main reason of this over-prediction is that it is not yet fully understood which of the thermodynamically plausible crystal structures are actually kinetically accessible via crystal nucleation and growth or via phase transition.[8, 18] Many of the predicted forms may simply not be accessible experimentally, or be so kinetically hard

to crystallise that they can only be produced with a very complicated set of experiments.[8, 18, 19, 36] Hence CSP can only produce a set of plausible but not necessarily accessible polymorphs.[18]

This information is extremely helpful in drug development, but a close collaboration between computational and experimental scientists is required, as well as a good level of understanding about what a crystal energy landscape can or cannot say about the crystallisation behaviour of a molecule.[1, 17, 18] Although there are some examples of CSP directing experimentalists towards the production of new polymorphs,[36-39] this remains a rarity.[18] Only once crystallisation kinetics is fully understood, and limitations to the search space are no longer required, CSP will become the definitive tool to design crystal structures with desired properties.[7, 18] Until then it will remain an important tool for the pharmaceutical industry, but subject to ongoing developments.

1.2.3 Molecular size and flexibility: effect on the computational cost of CSP

Another problem that hinders a more routine use of organic CSP studies as a complement to polymorph screening is computational cost.[1, 17] CSP studies are very computationally expensive, and the main reason is that the crystalline behaviour of organic molecules is dominated by weak intermolecular interactions,[40] whose accurate modelling requires costly electronic structure calculations.[1, 7, 9, 17]

The computational cost of CSP studies is strongly dependent on molecular size and flexibility.[1, 17, 41] As molecules get larger and more flexible the number of degrees of freedom that need to be modelled increases, causing an increment in the computational expense.[1, 17, 42, 43] Figure 1.3 shows how in the 6th Blind Test of CSP the successful predictions of the crystal structure of large and flexible molecule XXVI were on average more than one order of magnitude more computationally expensive those of small and rigid molecule XXII.[9]

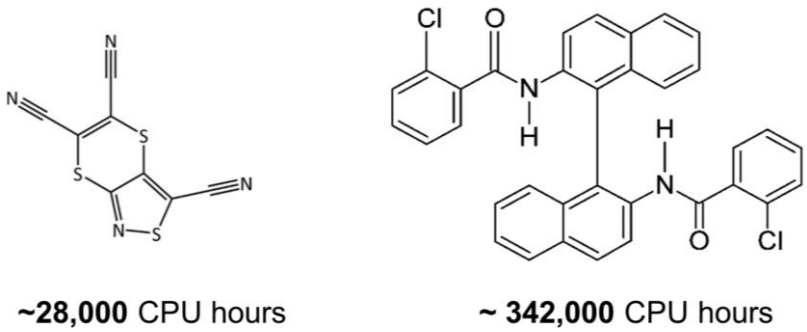


Figure 1.3: Chemical diagram and average CPU cost of the successful predictions of the crystal structures of (left) molecule XXII (right) molecule XXVI in the 6th Blind Test of organic CSP. The average is for the 13 successful predictions of molecule XXII and two of molecule XXVI for which the computational expense is reported in the Blind Test publication.[9]

Accurate CSP studies on molecules like XXVI require hundreds of thousands of CPU hours, which translates to weeks or months of use of computer clusters, depending on their size.[1, 9] This is problematic for the pharmaceutical industry: since drug molecules are chosen for their biological activity, they tend to be very conformationally flexible.[17, 44] High-quality CSP studies are simply unaffordable for many molecules in drug development, and using current methods with standard computational resources it would not be possible to perform a CSP study on a molecule like ritonavir.[17] Hence, a routine use of computational predictions in drug development requires more cost-effective CSP methodologies.

1.3 Signposting contents of the thesis

This thesis aims to develop ways to expand CSP studies towards larger and more flexible molecules of pharmaceutical interest, by reducing the computational cost of each portion of a CSP workflow and improving the analysis of the generated crystal structures to more effectively identify plausible polymorphs and separate them from duplicates.

Chapter 2 illustrates the theoretical and informatics methods used in this thesis to perform CSP and to compare and analyse crystal structures and molecular conformations. The functioning of the main computer algorithms is also described. In Chapter 3 the successful CSP study on large and flexible molecule XXVI that was undertaken in the context of the 6th Blind Test is discussed,[9] while Chapter 4 illustrates a computational study on the antihelminthic drug mebendazole that was carried out as part of an academic solid form screening effort. These studies were both performed with traditional CSP workflows, and some issues related to computational cost, limitation of search space and the effect of molecular size and flexibility are examined. Following from these studies, Chapter 5 discusses how a methodology to use conformational information retrieved from the Cambridge Structural Database (CSD)[45] to cut down the computational cost of CSP searches was developed and tested on molecule XXVI, mebendazole and three further flexible pharmaceutical molecules with previously published studies.[41] Chapter 6 analyses the importance of selecting which torsion and bond-angles are optimised as a function of solid-state interactions in the final minimisations of the crystal structures generated in CSP searches. This leads to Chapter 7, which examines the use of the semi-empirical quantum-mechanical method DFTB-D[46] to reduce the computational cost of the final refinement of the crystal structures that had been generated in the searches described in Chapter 5.[44] Chapter 8 then discusses how the CSD was mined to determine structural and crystallographic differences that can separate polymorphs from duplicates/redeterminations, and how a programme to cluster more efficiently CSP-generated crystal structures was developed and tested. Finally, in Chapter 9 the

main conclusions that can be drawn from this thesis are illustrated and future steps in this area of research are suggested.

1.4 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* (6), 912-923.
2. Lee, E. H., A practical guide to pharmaceutical polymorph screening & selection. *Asian Journal of Pharmaceutical Sciences* **2014,** *9* (4), 163-175.
3. Lee, A. Y.; Erdemir, D.; Myerson, A. S., Crystal Polymorphism in Chemical Process Development. *Annual Review of Chemical and Biomolecular Engineering, Vol 2* **2011,** *2*, 259-280.
4. Bernstein, J., Polymorphism - A Perspective. *Crystal Growth & Design* **2011,** *11* (3), 632-650.
5. Cruz-Cabeza, A. J.; Bernstein, J., Conformational Polymorphism. *Chemical Reviews* **2014,** *114* (4), 2170-2191.
6. 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.
7. Price, S. L., Is zeroth order crystal structure prediction (CSP_0) coming to maturity? What should we aim for in an ideal crystal structure prediction code? *Faraday Discussions* **2018,** *in press*.
8. Price, S. L., Why don't we find more polymorphs? *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **2013,** *69*, 313-328.
9. 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.; 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 Crystallographica Section B* **2016,** *72* (4), 439-459.
10. Cruz-Cabeza, A. J.; Reutzel-Edens, S. M.; Bernstein, J., Facts and fictions about polymorphism. *Chemical Society Reviews* **2015,** *44*, 8619-8635.
11. Abramov, Y., Current Computational Approaches to Support Pharmaceutical Solid Form Selection. *Organic Process Research & Development* **2013,** *17* (3), 472-485.
12. Yu, L., Polymorphism in Molecular Solids: An Extraordinary System of Red, Orange, and Yellow Crystals. *Accounts of Chemical Research* **2010,** *43* (9), 1257-1266.
13. Vasileiadis, M.; Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C., The polymorphs of ROY: application of a systematic crystal structure prediction technique. *Acta Crystallographica Section B-Structural Science* **2012,** *68*, 677-685.
14. Tan, M.; Shtukenberg, A. G.; Zhu, S.; Xu, W.; Dooryhee, E.; Nichols, Shane M.; Ward, M. D.; Kahr, B.; Zhu, Q., ROY revisited, again: the eighth solved structure. *Faraday Discussions* **2018**, *Advance article.*
15. Diao, Y.; Lenn, K. M.; Lee, W.-Y.; Blood-Forsythe, M. A.; Xu, J.; Mao, Y.; Kim, Y.; Reinspach, J. A.; Park, S.; Aspuru-Guzik, A.; Xue, G.; Clancy, P.; Bao, Z.; Mannsfeld, S. C. B., Understanding Polymorphism in Organic Semiconductor Thin Films through Nanoconfinement. *Journal of the American Chemical Society* **2014,** *136* (49), 17046-17057.
16. Price, S. L., Predicting crystal structures of organic compounds. *Chemical Society Reviews* **2014,** *43* (7), 2098-2111.
17. Price, S. L.; Braun, D. E.; Reutzel-Edens, S. M., Can computed crystal energy landscapes help understand pharmaceutical solids? *Chemical Communications* **2016,** *52*, 7065-7077.

18. Nyman, J.; Reutzel-Edens, S. M., Crystal structure prediction is changing from basic science to applied technology. *Faraday Discussions* **2018**, *Advance article.*
19. Neumann, M.; van de Streek, J., How many Ritonavir cases are there still out there? *Faraday Discussions* **2018**, *Advance article.*
20. Meyer, M. C.; Straughn, A. B.; Jarvi, E. J.; Wood, G. C.; Pelsor, F. R.; Shah, V. P., The bioinequivalence of carbamazepine tablets with a history of clinical failure. *Pharmaceutical Research* **1992,** *9* (12), 1612-1616.
21. Chemburkar, S. R.; Bauer, J.; Deming, K.; Spiwek, H.; Patel, K.; Morris, J.; Henry, R.; Spanton, S.; Dziki, W.; Porter, W.; Quick, J.; Bauer, P.; Donaubauer, J.; Narayanan, B. A.; Soldani, M.; Riley, D.; McFarland, K., Dealing with the impact of ritonavir polymorphs on the late stages of bulk drug process development. *Organic Process Research & Development* **2000,** *4* (5), 413-417.
22. Perez-Lloret, S.; Rey, M. V.; Ratti, P. L.; Rascol, O., Rotigotine transdermal patch for the treatment of Parkinson's Disease. *Fundamental & Clinical Pharmacology* **2013,** *27* (1), 81-95.
23. Pan, D. H.; Crull, G.; Yin, S.; Grosso, J., Low level drug product API form analysis - Avalide tablet NIR quantitative method development and robustness challenges. *Journal of Pharmaceutical and Biomedical Analysis* **2014,** *89*, 268-275.
24. Bauer, J.; Spanton, S.; Henry, R.; Quick, J.; Dziki, W.; Porter, W.; Morris, J., Ritonavir: An extraordinary example of conformational polymorphism. *Pharmaceutical Research* **2001,** *18* (6), 859-866.
25. Rossi, R. C.; Dias, C. L.; Donato, E. M.; Martins, L. A.; Bergold, A. M.; Fröehlich, P. E., Development and validation of dissolution test for ritonavir soft gelatin capsules based on in vivo data. *International Journal of Pharmaceutics* **2007,** *338* (1–2), 119-124.
26. Cabri, W.; Ghetti, P.; Pozzi, G.; Alpegiani, M., Polymorphisms and Patent, Market, and Legal Battles: Cefdinir Case Study. *Organic Process Research & Development* **2007,** *11* (1), 64-72.
27. Belenguer, A. M.; Lampronti, G. I.; Cruz-Cabeza, A. J.; Hunter, C. A.; Sanders, J. K. M., Solvation and surface effects on polymorph stabilities at the nanoscale. *Chemical Science* **2016,** *7* (11), 6617-6627.
28. Dzyabchenko, A. V., Theoretical Structures of Crystalline Benzene - the Search for A Global Minimum of the Lattice Energy in 4 Space-Groups. *Journal of Structural Chemistry* **1984,** *25* (3), 416-420.
29. Maddox, J., Crystals from 1St Principles. *Nature* **1988,** *335* (6187), 201-201.
30. 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. *Crystal Growth & Design* **2013,** *13* (6), 2396-2406.
31. Braun, D. E.; McMahon, J. A.; Koztecki, L. H.; Price, S. L.; Reutzel-Edens, S. M., Contrasting Polymorphism of Related Small Molecule Drugs Correlated and Guided by the Computed Crystal Energy Landscape. *Crystal Growth & Design* **2014,** *14* (4), 2056-2072.
32. 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 Crystallographica Section B-Structural Science* **2011,** *67*, 535-551.
33. Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C.; Price, S. L.; Galek, P. T.; Day, G. M.; Cruz-Cabeza, A. J., Successful prediction of a model pharmaceutical in the fifth blind test of crystal structure prediction. *International Journal of Pharmaceutics* **2011,** *418* (2), 168-178.
34. Habgood, M., Form II Caffeine: A Case Study for Confirming and Predicting Disorder in Organic Crystals. *Crystal Growth & Design* **2011,** *11* (8), 3600-3608.
35. Habgood, M.; Grau-Crespo, R.; Price, S. L., Substitutional and orientational disorder in organic crystals: a symmetry-adapted ensemble model. *Physical Chemistry Chemical Physics* **2011,** *13* (20), 9590-9600.
36. Braun, D. E.; Oberarcher, 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,** *DOI:10.1039/C5CE01758K.*
37. Arlin, J. B.; Price, L. S.; Price, S. L.; Florence, A. J., A strategy for producing predicted polymorphs: catemeric carbamazepine form V. *Chemical Communications* **2011,** *47* (25), 7074-7076.

38. 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. *Chemical Communications* **2016,** *52*, 7384-7386.
39. Neumann, M. A.; de Streek, J. V.; Fabbiani, F. P. A.; Hidber, P.; Grassmann, O., Combined crystal structure prediction and high-pressure crystallization in rational pharmaceutical polymorph screening. *Nature Communications* **2015,** *6*, 7793.
40. Stone, A. J., *The Theory of Intermolecular Forces*. Oxford University Press: Oxford, 2013; Vol. 2.
41. Iuzzolino, L.; Reilly, A. M.; McCabe, P.; Price, S. L., Use of Crystal Structure Informatics for Defining the Conformational Space Needed for Predicting Crystal Structures of Pharmaceutical Molecules. *Journal of Chemical Theory and Computation* **2017,** *13* (10), 5163-5171.
42. Vasileiadis, M.; Pantelides, C. C.; Adjiman, C. S., Prediction of the crystal structures of axitinib, a polymorphic pharmaceutical molecule. *Chemical Engineering Science* **2015,** *121*, 60-76.
43. Pantelides, C. C.; Adjiman, C. S.; Kazantsev, A. V., General Computational Algorithms for Ab Initio Crystal Structure Prediction for Organic Molecules. *Topics in Current Chemistry* **2014,** *345*, 25-58.
44. Iuzzolino, L.; McCabe, P.; Price, S. L.; Brandenburg, J. G., Crystal structure prediction of flexible pharmaceutical-like molecules: Density functional tight-binding as an intermediate optimization method and for free energy estimation. *Faraday Discussions* **2018**, *Advance article.*
45. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C., The Cambridge Structural Database. *Acta Crystallographica Section B-Structural Science Crystal Engineering and Materials* **2016,** *72*, 171-179.
46. Brandenburg, J. G.; Grimme, S., Accurate Modeling of Organic Molecular Crystals by Dispersion-Corrected Density Functional Tight Binding (DFTB). *Journal of Physical Chemistry Letters* **2014,** *5* (11), 1785-1789.

Chapter 2: Existing crystal structure prediction and informatics methods

2.1 Introduction

As mentioned in Chapter 1, successful crystal structure prediction (CSP) methodologies aim to determine which packing arrangements are the most thermodynamically stable. The thermodynamic stability of a crystal structure is determined by its free energy at given temperature and pressure conditions.[1-3] In most CSP applications the thermodynamic stability is however approximated by lattice energy (E_{latt}), *i.e.* the energy required separate all the molecules in a static lattice at 0 K to an infinite distance in their lowest energy conformation.[2, 4, 5] Although free energy estimates are becoming more established (see Chapter 2.3.3) they are generally performed only on some of the most promising crystal structures at the end of CSP workflows entirely based on E_{latt} minimisation.[2]

The accurate computation of E_{latt} requires a thorough understanding of the non-bonded interactions that occur between the molecules in the solid-state, as well as within the molecules themselves. Furthermore, an effective CSP study requires an accurate analysis and comparison of the generated crystal structures and their differences. This chapter outlines the methods that were used in this thesis to generate, rank, compare and analyse crystal structures of organic molecules.

2.2 Intermolecular forces

The presence of forces between the molecules that constitute matter is evident.[6] The existence of solid and liquid phases shows that there are attractive forces between the molecules they are constituted of, as otherwise they would not remain confined. On the other hand, the difficulty in compressing solid and liquid phases to smaller volumes demonstrates that at short-ranges molecules repel each other. These observations are summarised in Figure 2.1, which shows how the potential energy between two spherical rigid molecules (*U*) changes as a function of their distance (*R)*. The forces between molecules are related to the potential energy by:[6]

$$F = -\frac{dU}{dR} \qquad 2.1$$

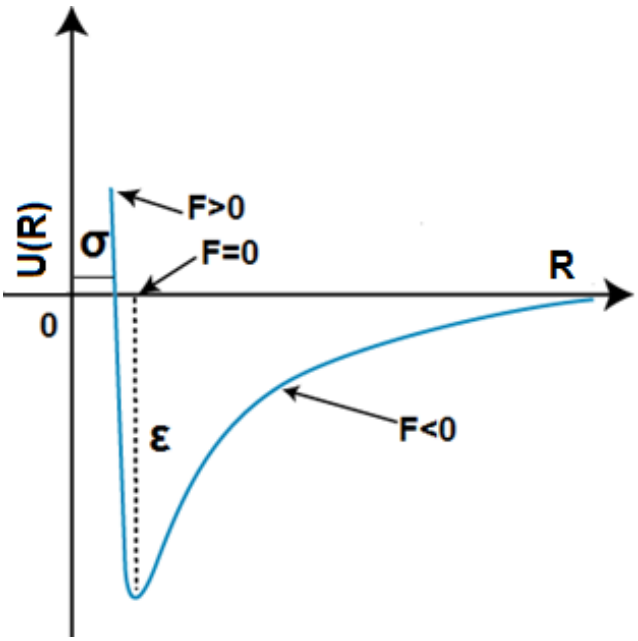


Figure 2.1: A typical energy vs distance plot for two rigid molecules; *σ* represents the distance at which *U=0*, and *ε* is the depth of the energy well.

Figure 2.1 shows that at long-range the forces between the molecules are attractive, as the derivative of *U* with respect to *R* is positive and *F* is negative, and at short-range they are repulsive. At equilibrium (the minimum in the curve), the forces acting on the molecules are null, and $U = -\varepsilon$.

Intermolecular energy and forces are essentially of quantum-mechanical (QM) origin.[6] Out of the four fundamental forces, strong and weak nuclear interactions are not relevant for objects of the size of molecules, while gravitational forces are present but so much weaker than electromagnetic forces that they can safely be ignored.[7] Furthermore, magnetic interactions are extremely weak between closed-shell molecules and can also usually be neglected.[6] Finally, resonance effects only arise for molecules in a degenerate electronic state, so they are not relevant for the ground state closed-shell molecules this thesis is focused on.[6]

Several distinct contributions to intermolecular forces exist, which are commonly separated into short- and long-range, depending on where they are the most relevant.

2.2.1 Long-range forces

Long-range intermolecular forces are easily described when molecules are at long distances from one another, *i.e.* their wave-functions do not overlap.[6] However, these forces act also at shorter distances, in a modified form that is more complex to describe (see Chapter 2.2.2.2). They decay as a function of an inverse power of distance, R^{-n}.[6]

Intermolecular forces can be understood in terms of perturbation theory, which finds the approximate solution to a problem from the exact solution to another problem.[8]

For two infinitely separated molecules A and B, their time independent Schrödinger's equations can be written as:

$$H_A^0 \Psi_A^0 = E_A^0 \Psi_A^0 \qquad 2.2$$

$$H_B^0 \Psi_B^0 = E_B^0 \Psi_B^0 \qquad 2.3$$

and for the combined system:

$$H_{AB}^0 \Psi_{AB}^0 = E_{AB}^0 \Psi_{AB}^0 \qquad 2.4$$

$$H_{AB}^0 = H_A^0 + H_B^0 \qquad 2.5$$

$$E_{AB}^0 = E_A^0 + E_B^0 \qquad 2.6$$

where H^0 indicates the Hamiltonian of the unperturbed ground state, E^0 the corresponding energy and Ψ^0 the unperturbed wave-function, which corresponds for the combined system corresponds to:[6]

$$\Psi_{AB}^0 = \Psi_A^0 \Psi_B^0 \qquad 2.7$$

Note that at long-range this wave-function does not require anti-symmetrisation, as there is no overlap of the wave-functions of molecules A and B (*i.e.* the electrons clearly belong to either A and B).[6] This complicates the theory at short-ranges, where anti-symmetrisation is instead required (see Chapter 2.2.2).

The perturbation from the ground state derives from the electrostatic interactions between the electrons and nuclei of molecules A and B. The perturbed Hamiltonian for the combined system can then be written as:[6]

$$H_{AB} = H_{AB}^0 + H' = H_A^0 + H_B^0 + H' \qquad 2.8$$

the perturbation term H' corresponds to:

$$H' = \sum_{a \in A} \sum_{b \in B} \frac{e_a e_b}{4\pi\varepsilon_0 R_{ab}} \qquad 2.9$$

where *a* and *b* are the electrons and nuclei in molecules A and B respectively, e_a and e_b are their charges, R_{ab} their distance and ε_0 is the permittivity of free space.

Using Rayleigh-Schrödinger perturbation theory,[9] the overall energy of the perturbed system can be expanded as:

$$E_{AB} = E_{AB}^0 + E'_{AB} + E''_{AB} + \ldots \qquad 2.10$$

the first order term (E'_{AB}) describes the classical electrostatic interactions between molecules A and B, and the second order term (E''_{AB}) describes the induction and dispersion interactions.[6] Higher order terms are neglected as the series converges rapidly and so they have a negligible contribution to E_{AB}.[6]

2.2.1.1 Electrostatics

Electrostatic interactions are the classical Coulombic interactions between the unperturbed molecular charge densities. Using the first order term in equation 2.10, the electrostatic energy can be defined as:[6]

$$U_{elec} = E'_{AB} = \langle \Psi_A^0 \Psi_B^0 | H' | \Psi_A^0 \Psi_B^0 \rangle = \int \frac{\rho_A(r)\rho_B(r')}{4\pi\epsilon_0 |r-r'|} d^3r\, d^3r' \qquad 2.11$$

where ρ is the charge density around molecules A and B and points in space r and r' respectively. U_{elec} can be either attractive (*i.e.* $U_{elec} < 0$) or repulsive (*i.e* $U_{elec} > 0$) depending on the orientation of ρ_A and ρ_B. As a classical Coulombic interaction, U_{elec} is strictly pairwise additive.[6] This means that for three molecules A, B and C:

$$U_{elec}^{ABC} = U_{elec}^{AB} + U_{elec}^{AC} + U_{elec}^{BC} \qquad 2.12$$

the same would be true for any number of molecules, and this property makes the computation of electrostatic interactions relatively simple. The charge density around molecules can be modelled in terms of isotropic atomic point charges (see Chapter 2.3.1.2.1). However, a more accurate approach that is used throughout most of this thesis is to model ρ in terms of a distributed multipole expansion around atomic sites; this is explained in detail in Chapter 2.3.1.2.2. In a multipole expansion of the charge density, the leading term describes the interaction between two ground-state charges and decays with R^{-1}.[6, 10]

2.2.1.2 Induction

Induction (also known as polarisation) interactions are a component of the second order term in Equation 2.10.[6] Continuing with the example of two molecules A and B, induction describes the interaction between the unperturbed charge density of A and the charge density of B modified by the presence of A, and vice-versa. Hence:

$$U_{ind}^{AB} = U_{ind}^{A} + U_{ind}^{B} = -\sum_{m\neq 0} \frac{\langle \Psi_A^0 \Psi_B^0 | H' | \Psi_A^m \Psi_B^0 \rangle \langle \Psi_A^m \Psi_B^0 | H' | \Psi_A^0 \Psi_B^0 \rangle}{E_A^m - E_A^0} - \sum_{n\neq 0} \frac{\langle \Psi_A^0 \Psi_B^0 | H' | \Psi_A^0 \Psi_B^n \rangle \langle \Psi_A^0 \Psi_B^n | H' | \Psi_A^0 \Psi_B^0 \rangle}{E_B^n - E_B^0} \qquad 2.13$$

where m and n are excited states of molecules A and B respectively. In the first term of Equation 2.13 the charge density of molecule B modifies the one of molecule A, which interacts with the electric field of molecule B, and vice-versa for the second term.

Induction interactions are always attractive and are not pairwise addictive.[6] The reason is that in the presence of multiple molecules, each has its charge density modified by the electric field caused by the presence of all the other molecules. The charge density around each molecule will then feel the changed electric field of its neighbours, and get modified as a consequence; the effect goes on until induction is completely converged.[11] In a multipole expansion, the leading term (charge-induced dipole) decays with R^{-4}.[6]

The strength of U_{ind} depends on the ease of distortion of the molecular charge density under the effect of an external electric field, or polarizability. For non-spherical molecules polarizabilities are very anisotropic, and they are defined by a tensor describing how much a molecular charge distribution can be modified by an electric field in the *x*, *y* and *z* directions.[6, 12] Polarizabilities can be calculated quantum-mechanically, making it possible for a many-body system like a crystal structure to calculate U_{ind} through an iterative process.[6] However, this is very computationally demanding, since high quality basis sets and wave-functions are required to compute accurate polarizabilities.[12] For this reason, in this thesis induction was not modelled explicitly. However, for some calculations its effect was estimated with a polarisable continuum model (PCM),[13] which is described in Chapter 2.3.1.4.

2.2.1.3 Dispersion

Dispersion is a universal attractive interaction of exclusive QM origin, and is a component of the second order term in Equation 2.10.[6] It derives from the constant fluctuation of the molecular electron densities, which creates instantaneous dipoles (as well as higher order multipoles, see Chapter 2.3.1.2.2) in a molecule. These dipoles induce dipoles in other molecules, creating an instantaneous correlation of their charge densities, which lowers the overall energy.[6] Without dispersion spherical neutral molecules could not form condensed states.[14] While induction is a static interaction, dispersion is frequency dependent. In terms of perturbation theory as, it is described as:

$$U_{disp} = -\sum_{m\neq 0}\sum_{n\neq 0}\frac{\left\langle \Psi_A^0\Psi_B^0\middle|H'\middle|\Psi_A^m\Psi_B^n\right\rangle\left\langle \Psi_A^m\Psi_B^n\middle|H'\middle|\Psi_A^0\Psi_B^0\right\rangle}{E_A^m+E_B^n-E_A^0-E_B^0} \qquad 2.14$$

dispersion interactions are approximately pairwise addictive, although many-body terms can have important effects.[6] The leading term in a multipole expansion (induced dipole-induced dipole) decays with R^{-6}.[6]

Similarly to induction, the molecule-specific dispersion coefficients can be calculated from the polarizabilities. However, in this thesis dispersion was modelled through empirically-fitted parameters;[3, 15] more details can be found in Chapter 2.3.1.2.3.

2.2.2 Short-range forces

Short-range interactions arise when molecules are at short distances from one another, so that their wave-functions overlap. They decay with the exponential $e^{-\alpha R}$.[6]

The perturbation theory for intermolecular forces at short distances is strongly complicated by the need to anti-symmetrise the wave-function in Equation 2.7,[6] and contrary to their long-range counterparts short-range interactions cannot be computed analytically.

2.2.2.1 Exchange-repulsion

Exchange-repulsion interactions derive from the overlap of the molecular electron clouds. The exchange term is attractive, and is due to electrons being free to occupy the orbitals over both molecules.[6] The repulsion term is, as the word suggests, repulsive, and derives from the Pauli exclusion principles: electrons try to occupy the same area of space, but they are forbidden to be in the same orbitals unless they have opposite spin.[6] Overall, the repulsion term dominates, which makes exchange-repulsion interactions increasingly repulsive at short-ranges (this explains the short-range wall in Figure 2.1):

$$U_{er} = U_{exch} + U_{rep} > 0 \qquad 2.15$$

U_{er} is approximately pairwise additive,[6] and is generally expressed in the form of a repulsive atom-atom potential parametrised from experimental data or *ab initio* calculations. Parametrisation to *ab initio* data is more accurate but more computationally demanding,[16, 17] and in this thesis short-range interactions were modelled with an empirically-fitted potential (see Section 2.3.1.2.3).

2.2.2.2 Other short-range terms

As mentioned in Chapter 2.2.1, the long-range interactions are also present at small distances, but they are modified by the overlap of the electron clouds around the molecules.

Charge-penetration is a short-range modification of the electrostatic energy. Where the electron densities overlap, the nuclei of one molecule are no longer completely shielded from the electrons of the other molecule.[18] Charge-penetration is generally attractive, although it becomes repulsive at very short distances, and is pairwise additive.[19]

Charge-transfer is a short-range modification of the induction energy, due to the interaction between donors and acceptors. Electrons are transferred from the high energy occupied orbitals of the donor to the low energy unoccupied orbitals of the acceptor. Charge-transfer interactions are attractive and are not pairwise addictive.[6]

Finally, damping is a correction to account for exchange-induction and exchange-dispersion, which are the modifications to induction and dispersion interactions due to the anti-symmetrisation of the wave-functions.[6]

Overall, the net effect of these interactions is small compared to U_{er} ,[6] and it is assumed they are absorbed in the parametrisation of the short-term repulsive potential.

2.3 Determining the thermodynamic stability of crystal structures

A realistic description of the interactions between molecules in organic crystalline solids, which are the subjects of this thesis, requires an accurate understanding of the electronic

structure. Since the exclusive use of standard transferable force-fields that consider only nuclear positions has proven unreliable in CSP,[15, 20, 21] at least some *ab initio* electronic structure calculations are required. In this section the two most common ways to model the energy of crystal structures in CSP are described, which are referred to as the Ψ_{mol} and the Ψ_{crys} methods.

2.3.1 The Ψ_{mol} method

In Chapter 2.2 intermolecular interactions were described through perturbation theory. It requires firstly to solve the Schrödinger's equation of the individual molecules, and then to calculate the intermolecular interactions as a perturbation of their ground-state molecular wave-functions. In most of this thesis, an approach based on this theory was used to determine the thermodynamic stability of crystal structures was utilised, and it can be referred to as the Ψ_{mol} method.[4, 22-24] After calculating the individual molecular wave-functions (*i.e.* the Ψ_{mol}), the intermolecular interactions are computed as the sum of the interactions between the atoms of different molecules. Crystalline symmetry simplifies the problem, since costly *ab initio* calculations are required only for molecules that are not related by symmetry operations[20] (symmetry related molecules have the same nuclear positions and so by definition the same ground-state wave-functions). Within the Ψ_{mol} formalism, E_{latt} can be defined as:[10, 15, 25]

$$E_{latt} = U_{inter} + \Delta E_{intra} \qquad 2.16$$

where the intermolecular energy term, U_{inter}, is the sum of all the interactions between molecules, and ΔE_{intra} is defined as the intramolecular energy penalty. Note that E_{latt} must be negative in a crystal structure, otherwise it would not form.

2.3.1.1 Intramolecular energy penalty (ΔE_{intra})

In the Ψ_{mol} method intermolecular interactions do not directly change the molecular conformation/s, as they are calculated from ground-state wave-functions.[6] Conformational distortions are negligible for very rigid molecules (*e.g.* benzene), for which $E_{latt} \approx U_{inter.}$[15, 26] However, flexible molecules can adopt different configurations that allow to minimise U_{inter} (and consequently E_{latt}) in Equation 2.16.[26, 27] Within the Ψ_{mol} framework, this is accounted for by the ΔE_{intra} term, which is the difference between the conformational energy of a molecule with a configuration corresponding to that in a crystal structure and that of the lowest energy conformer/s of the isolated molecule/s (also referred to as gas-phase energy).[15, 28, 29] If the conformational energy of the isolated-molecule global minimum is known, ΔE_{intra} can be calculated from the molecular wave-function; this calculation allows to also determine how the charge density varies for different conformations, which can be used to compute U_{inter}:

$$\Delta E_{intra} = E^{gas-phase}_{crystalline\ molecular\ conformation} - E^{gas-phase}_{global\ minimum} \qquad 2.17$$

ΔE_{intra} values are very dependent on the level of theory chosen for the calculation of the molecular wave-function.[30] In this thesis, density functional theory with the hybrid PBE0 6-31G(d,p) functional was utilised. PBE0 has proven to be successful in several high-quality CSP studies and provides a good balance between quality and computational cost.[2, 31-33] All the wave-function calculations with the Ψ_{mol} methods were performed with the Gaussian code.[34, 35]

As the name 'penalty' suggest, ΔE_{intra} is always a positive number. Despite being negligible for rigid molecules, no crystalline conformation is completely undistorted from an isolated-molecule conformer.[27] In most experimentally known crystal structures ΔE_{intra} is small, generally below 5-10 kJ·mol^{-1} and very rarely above 20-25 kJ·mol^{-1},[27, 36, 37] as improvements in intermolecular interactions cannot generally compensate for large conformational energy penalties. However, crystal structures with large ΔE_{intra} values exist, and this is often associated with a competition between intra- and intermolecular hydrogen bonds[27] or between folded and extended conformations.[36] In the presence of several torsion angles (or dihedral angles; in this thesis these two terms are used interchangeably) with low-energy barriers to rotation a wide range of low-energy conformations can exist, which adds to the complexity of the CSP problem.

In their review on conformational polymorphism, Cruz Cabeza and Bernstein[27] distinguished between conformational adjustment and conformational change. Conformational adjustment is defined as a slight change from a gas-phase conformer that a molecule undergoes in the crystal to minimise U_{inter}, but still within the same isolated-molecule conformational energy well (*i.e.* the crystalline geometry, when isolated, optimises to the same gas-phase minimum in conformational energy). On the other hand, conformational change requires that the molecule in the crystal structure belongs to a different conformational energy well (*i.e*. the crystalline geometry, when isolated, does not optimise to the same gas-phase minimum in conformational energy). Polymorphs related by conformational change can be classified as conformational polymorphs, while those related by adjustment simply as polymorphs.[27]

2.3.1.2 Intermolecular energy (U_{inter})

In the Ψ_{mol} method U_{inter} is modelled with a potential that calculates the intermolecular interactions of the molecules, treated as rigid-bodies,[10, 29] using some of the long- and short-range terms that have been listed in Chapter 2.2. There are several potentials that can be used to model U_{inter} within the Ψ_{mol} formalism. The most accurate ones are completely non-empirical and require *ab initio* calculations of polarizability to model induction and dispersion interactions analytically, as well as a parametrisation of the short-range interactions from electronic structure data.[16, 17] However, producing these accurate and theoretically compelling potentials is computationally expensive and by no

means routine, in particular for large and flexible molecules. Hence the potential utilised in this thesis to model U_{inter} contains a mixture of empirically-fitted and *ab initio*-derived terms:

$$U_{inter} = U_{elec} + U_{rep-disp} \qquad 2.18$$

where U_{elec} is the electrostatic energy calculated from atomic point charges or from distributed multipoles expanded around atom sites, which are derived from the *ab initio* wave-functions, and $U_{rep-disp}$ is a transferable empirically-fitted potential that describes repulsion and dispersion interactions; note that induction is not explicitly modelled in Equation 2.18.

2.3.1.2.1 U_{elec} from point charges

Electrostatic interactions are a fundamental component of intermolecular forces, and of very limited transferability.[38] The potential in Equation 2.18 is an atom-atom potential. Hence the calculation of U_{elec} requires to partition the molecular charge density into atom sites to allow the pairwise summation of the electrostatic interactions between each molecule under consideration.

The simplest and computationally cheapest way to partition the charge density is a division into isotropic (*i.e.* spherical) point-charges placed around each atom.[6] Within this model, the electrostatic energy between two atoms *i*, *j* in two molecules A and B can be written as:

$$U_{elec}^{ij} = \frac{1}{4\pi\varepsilon_0}\frac{q_i q_j}{R_{ij}} \qquad 2.19$$

Where q_i and q_j are the point charges at atoms *i* and *j* and R_{ij} is the distance between these atoms. This model is very simple and broadly independent of molecular conformation.[6] In this thesis, atomic point charges were used to compute the electrostatic component of intermolecular energy in the crystal structure searches, which were performed with CrystalPredictor (see Chapter 2.4.1.2 for details). However, this approach has the weakness of describing the charge density around each atom as a sphere, and as such it cannot describe directional features such as lone pairs and π electrons.[6, 15] This is particularly important for molecules that are close to one another, as they often are in crystals, while at longer-ranges the model becomes more adequate since higher order non-spherical terms decay more quickly with distance.[6] Additional expansion sites can be added to overcome some of the directionality problems, but this is an arbitrary correction that also increases the computational expense of the fitting.[6]

There are several methods to partition the charge density into atomic point charges. In this thesis is the CHELPG method (CHarges from Electrostatic Potential using a Grid based method)[39] was utilised. Atomic charges are fitted via a least-square

procedure from the *ab initio* calculated charge density evaluated on a grid of points around the molecule, guaranteeing that the total molecular charge is correct.[39]

2.3.1.2.2 U_{elec} from distributed multipoles

A more accurate way to represent the charge density is through multipole expansion.[6] The zeroth order (*i.e. l=0*) multiple moment is the total charge (*q*), which is independent of direction:

$$q = \sum_a e_a \qquad 2.20$$

where e_a is the charge on each electron and nucleus. No integration is required to calculate *q*. The first order moment (*l=1*) is the dipole (μ), which describes the separation of equal and opposite charges along a linear vector:[6]

$$\mu_\alpha = \int \rho(\boldsymbol{r}) r_\alpha d^3\boldsymbol{r} \qquad 2.21$$

where $\rho(\boldsymbol{r})$ is the charge density at point $\boldsymbol{r}$ and α stands for either *x*, *y* or *z.* A typical example of a dipole is the one formed by water, as shown in Figure 2.2.

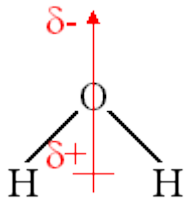


Figure 2.2: Schematics of the dipole moment in water.

The second order term (*l=2*) arises when two equal dipoles are located nearby so that their net dipole moment is zero, but as their positive charges overlap they have net non-zero moment called quadrupole (θ):[6]

$$\theta_{\alpha\beta} = \int \rho(\boldsymbol{r})(\tfrac{3}{2} r_\alpha r_\beta - \tfrac{1}{2} r_\alpha \delta_{\alpha\beta}) d^3\boldsymbol{r} \qquad 2.22$$

where α and β stand for either *x*, *y* or *z* and δ is the Kronecker delta. The $\theta_{\alpha\beta}$ matrix is traceless (*i.e.* $\theta_{xx}+\theta_{yy}+\theta_{zz}=0$), and symmetric with permutation of indices (*i.e.* $\theta_{xz}=\theta_{zx}$), meaning that only five independent quadrupole moments exist.[6] An example is the quadrupole moment around a benzene ring, shown in Figure 2.3.

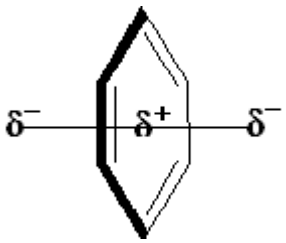


Figure 2.3: Schematics of the quadrupole moment around a benzene ring.

The third order term (*l=3*) is the octopole, which represents the overlap of the positive charges of four dipoles in a three-dimensional array, the fourth order term (*l=4*) is the hexadecapole and so on.[6]

In a multipole expansion, U_{elec} decays with the negative n^{th} power of distance between two multipoles *a* and *b*, where:[6]

$$n = l_a + l_b + 1 \qquad 2.23$$

hence U_{elec} between two charges decays with R_{AB}^{-1}, between a dipole and a charge with R_{AB}^{-2} and so on. Note that since multipole moments with $l > 0$ are orientation dependent, their interaction leads to torques and non-central forces.[10]

Although multipoles could be expanded around a single site for a whole molecule, better convergence is obtained when they are expanded around several specific sites, which are often atomic nuclei. This approach is called 'distributed multipole expansion'.[6] There are several ways of partitioning the molecular charge density to define the multipole expansion around each atom; the one that was utilised in this thesis is the Distributed Multipole Analysis (DMA) developed by Stone.[40] It requires that the molecular wave-function is calculated using a Gaussian basis set, *i.e.* that the basis set consists of a series of Gaussian orbitals around each atom. The DMA uses the properties of Gaussian orbitals to fit multipole moments at specific sites around each atom.[40] In this thesis, the programme Gaussian Distributed Multipole Analysis (GDMA)[41] was used to fit distributed multipoles from molecular wave-functions calculated with Gaussian up to fourth order terms (*i.e.* charges, dipoles, quadrupoles, octopoles and hexadecapoles).

This approach solves the lack of directionality of the point charge model, as multipoles around each atom are highly anisotropic; for example dipole moments can naturally represent lone pairs, and quadrupole moments can represent π electrons.[6, 15] Distributed multipoles have been proven to provide more accurate energy rankings in CSP studies than point charges.[42]

2.3.1.2.3 $U_{rep\text{-}disp}$

The remaining intermolecular interactions in Equation 2.18 were modelled in this thesis with an isotropic Buckingham atom-atom exp-6 potential.[3, 15, 29, 43] For two molecules A and B:

$$U_{rep-disp}^{AB} = U_{rep}^{AB} + U_{disp}^{AB} = \sum_{i\in A, k\in B} A_{\iota\kappa}\, exp(-B_{\iota\kappa}R_{\iota\kappa}) - \frac{C_{\iota\kappa}}{R_{\iota\kappa}^6} \qquad 2.24$$

where *ι* is the type of atom *i*, *κ* the type of atom *k*, and $R_{\iota\kappa}$ is the distance between these two atoms. The positive term in Equation 2.24 describes the short-range repulsive interactions, while the negative term the long-range attractive dispersion interactions. Other intermolecular interactions are assumed to be absorbed in the fitting.

A, B and C are the empirically fitted parameters, and they are specific to the atomic species and sometimes also to the atomic connectivity, *e.g.* an H atom bonded to N has different parameters from when it is bonded to C. Parameters for each pair of

atoms are not always fitted explicitly. Combining rules can be used to calculate the missing parameters,[44] as:

$$A_{\iota\kappa} = \sqrt{A_{\iota\iota} + A_{\kappa\kappa}};\ B_{\iota\kappa} = \frac{1}{2}(B_{\iota\iota} + B_{\kappa\kappa});\ C_{\iota\kappa} = \sqrt{C_{\iota\iota} + C_{\kappa\kappa}} \qquad 2.25$$

although combining rules can lead to some inaccuracies,[43] they are generally found to be broadly applicable.[6] In this thesis, the FIT parameters[45] were used to model $U_{rep\text{-}disp.}$ They were fitted to a variety of crystal structures: apolar hydrogen, carbon and nitrogen were fitted to azahydrocarbon crystal structures,[46] oxygen to oxohydrocarbon crystal structures,[47] and polar hydrogen to a set of azabenzenes, nitrobenzenes and other simple molecules that contain intermolecular hydrogen bonds.[45] Parameters were also fitted separately for chlorine[48] and fluorine.[49] In the FIT potential each atomic species is treated in the same way irrespective of the connectivity, with the exception of polar and apolar hydrogens.

The FIT potential has some shortcomings. First of all, induction interactions are not explicitly modelled, but their effect is absorbed in the empirical fitting. This increases the risk of double counting if induction were to be included in U_{inter}.[10, 12] The dispersion interactions are also not damped, which can lead to inaccurate short-term behaviours.[17] Furthermore, its parameters are fitted to crystal structures at room temperature, so they are not entirely consistent with a CSP procedure that calculates E_{latt} at 0 K, as they will have absorbed thermal expansion effects. The FIT parameters are also completely isotropic (*i.e.* each atom is treated as a sphere), and do not include many-body effects and higher order dispersion terms. Finally the FIT parameters were fitted in combination of U_{elec} values calculated from a wave-function determined at the HF/ 6-31G(d,p) level of theory,[45] and this complicates their transferability when charge densities are computed with different models. For these reasons new A, B and C parameters are being developed, either empirically-fitted with charge densities computed at specific levels of theory[43, 44] or in the form of non-empirical potentials.[16, 17]

However, fitting these more accurate potentials is not routine and very computationally demanding. Furthermore, the FIT potential has proven to be very effective in CSP, and has recently been found capable to reproduce absolute energies as accurately as some more complex and computationally expensive Ψ_{crys} (see Chapter 2.3.2) models.[50] Hence it was confidently used in this thesis to calculate U_{inter} within the Ψ_{mol} formalism.

2.3.1.2.4 The polarisable continuum model (PCM)

As already mentioned, the U_{inter} model in Equation 2.18 ignores induction interactions. However, induction can play a significant role in both absolute and relative energies of crystal structures, especially when they have different hydrogen bond patterns.[12, 51] Although induction can be modelled in the Ψ_{mol} method,[10] this was not done in this thesis.

An alternative way to estimate the effect of polarisation on E_{latt} is calculating the molecular wave-function/s with a polarisable continuum model (PCM). The polarisable continuum is generally used to study solvation of a solvent around a solute of interest (*i.e*. the electrostatic interactions that stabilise the solute losing a H^+ ion in a solvent).[13] PCM calculations can be performed with Gaussian, where a polarisable continuum surrounds a cavity defined by the solvent-excluded surface of the molecule (the solvent is assumed to be water). The cavity is calculated by atomic or group spheres plus extra spheres to smoothen the surface.[34] By calculating the wave-function in a PCM, the effect of induction on E_{latt} should be modelled without the need of introducing an explicit intermolecular term. However, the exact dielectric constant (ε) of the solid-state environment around the molecule is not exactly known, and the PCM simply introduces an average polarisation effect that is independent of the specific crystal structure.[52] Furthermore, there is a risk of double counting induction effects modelled by the PCM with those that have been accidentally included in the empirical fitting of $U_{rep\text{-}disp}$. Nevertheless, Cooper *et al.* showed how the inclusion of a PCM helps to improve the relative ranking of crystal structures matching experimental forms relative to other predicted but unfound alternatives.[52] Their suggested dielectric constant of ε=3 for neutral organic molecules,[52] which seems to be typical for this kind of crystals,[51] was used in Chapters 3 and 4 to recalculate E_{latt} values after the completion of the CSP workflow, with the aim of estimating the changes in relative stability that may arise from including induction. Note that the polarisable continuum was not used to fully-optimise crystal structures, but only to calculate polarised molecular wave-function/s and use them to recalculate ΔE_{intra} and to re-optimise U_{inter} with DMACRYS. Details of the functioning of DMACRYS can be found in Chapter 2.4.2.1.

2.3.2 The Ψ_{crys} method

The Ψ_{mol} method is based on separating intra- and intermolecular interactions. However, the crystal structure itself can be modelled as an assembly of nuclei and electrons, on which to perform *ab initio* calculations.[6] This can be referred to as the 'supermolecule' method, or $\Psi_{crys,}$ as crystalline and not molecular wave-functions are calculated.[6] Within this formalism, the lattice energy of can be calculated as:

$$E_{latt} = \frac{E_{crys}}{Z} - E_{global\ minimum}^{gas-phase} \qquad 2.26$$

where E_{crys} is the energy of the crystalline 'supermolecule' and Z is the number of molecules in the primitive unit cell.

There are several methods that can be used to perform calculations with the Ψ_{crys} method, ranging from periodic MP2 algorithms to Diffusion Monte Carlo and fragment based approaches.[53] However, successful CSP workflows have to perform Ψ_{crys}

calculations on hundreds, if not thousands, crystal structures and so they currently use cost-effective periodic density functional theory (DFT).[2]

2.3.2.1 Periodic density functional theory

Periodic DFT methods were not directly used in this thesis, mainly because of their relatively high computational expense. However, introducing them is fundamental as they are becoming increasingly prevalent in CSP: in the latest Sixth Blind Test of organic CSP (see Chapter 3.1.1), 12 groups used periodic DFT,[2] while only two groups had done the same in the Fifth Blind Test,[54] where Ψ_{mol} methods were still the norm.

With the Born-Oppenheimer approximation[55] (*i.e.* the position of nuclei is fixed), DFT calculates the energy as a functional of the electron density $\rho(r)$, using the method proposed by Kohn and Sham:[56]

$$E_{DFT}[\rho] = E_{NE}[\rho(r)] + Ts[\rho(r)] + J[\rho(r)] + E_{XC}[\rho(r)] \qquad 2.27$$

Hohenberg and Kohn proved that E_{DFT} is in principle exact.[57] The terms E_{NE} (the potential energy between nuclei and electrons), Ts (the non-interacting component of the kinetic energy of the electrons) and J (the classic Coulombic repulsion between electrons) can be calculated exactly. However, E_{XC} (the exchange and correlation functional), which describes non-classical interactions between electrons, is not known exactly and must be approximated.[53, 58] Several E_{XC} functionals have been developed. In CSP generalised gradient approximation (GGA) functionals are commonly used, which are dependent on $\rho(r)$ and its gradient.[58]

There are several reasons why periodic DFT is becoming increasingly prevalent in CSP. First of all, despite being more computationally demanding than Ψ_{mol} methods, periodic DFT is still relatively cheap compared to other Ψ_{crys} approaches such as periodic correlated wave-functions methods, while being more accurate, in principle, than periodic HF methods, because it includes some description of electron correlation.[53, 59] The increase in the power of computers and the presence of several high-quality, well-maintained and optimised codes[60-63] to perform periodic DFT is decreasing the cost-barrier for use in CSP. Also periodic DFT is generally more accurate than Ψ_{mol} methods.[50] There are several reasons: first of all, calculating the wave-function for a whole crystal structure does not require the somewhat artificial separation of inter- and intermolecular interactions, and all crystalline degrees of freedom can be optimised at the same time. On the other hands, Ψ_{mol} methods can only afford to optimise some selected molecular conformational degrees of freedom (CDFs) in response to solid-state interactions (the limitations of this approach are discussed in details in Chapter 6).[64] Furthermore, periodic DFT naturally calculates intra- and intermolecular interactions, including repulsion, electrostatics and induction, without needing any parametrisation to *ab initio* or experimental data, as well as approximations like the truncation of the

multipole series and pairwise additivity. This eliminates the need of selecting a potential that is adequate for the system of interest.

However, periodic DFT has some weaknesses. First of all, there is no systematic way to improve the quality of the wave-function,[65] and hence no guarantee that the quality of the calculations for one system would be the same if identical methods were applied to another. Energies of sub-kJ·mol^{-1} accuracy can only be obtained with extremely expensive periodic correlated wave-function methods such as coupled cluster.[66] Furthermore, most periodic DFT methods in CSP use plane-wave basis sets to compute the molecular wave-function,[2] which makes the use of functionals that include exact exchange, such as PBE0 or B3LYP, very computationally inefficient.[67, 68] In general, Ψ_{mol} methods can afford to compute wave-functions at higher levels of theory than Ψ_{crys} methods. Finally, the functionals used in periodic DFT to calculate exchange and correlation interactions depend on the local charge density. Hence they do not model dispersion interactions that derive from non-local electron correlation, and periodic DFT tends to be severely under-binding.[69] The most common solution is the use of dispersion-corrected DFT (DFT-D),[2, 70] where a dispersion correction is added to E_{DFT} in Equation 2.27:

$$E_{DFT-D} = E_{DFT} + E_{disp} \qquad 2.28$$

all common forms of E_{disp} are composed of an attractive asymptotic $-C_n/R^n$ term, where n is the order of the dispersion correction (often $n=6$ for dipole-dipole interactions, but sometimes higher order terms are included), and a damping parameter to limit the effect of the correction at short-ranges where the correlation functionals are sufficient.[70, 71] Older dispersion corrections are empirically fitted to experimental data,[72, 73] while in more state-of-the-art ones like XDM,[74] D3[71] and TS[75] the C_n parameters are geometry-dependent or calculated from *ab initio* polarizabilities. These corrections are all pairwise, and a more recent many-body dispersion (MBD) correction proposed by Tkatchenko *et al.* seems to improve the results even further.[76]

2.3.2.2 Semi-empirical quantum-mechanical methods

Although accurate, periodic DFT methods are computationally demanding, in particular for CSP studies of large and flexible molecules where the spectrum of possible packings is enormous and thousands crystal structures often need to be optimised.[77] As already mentioned, completely transferrable force-fields are not a suitable alternative. Semi-empirical QM methods could be a solution to affordably yet accurately optimise thousands of crystal structures in CSP workflows.[24, 77, 78] They start from standard *ab initio* methods and apply a minimal basis set and substantial systematic approximations.[79] The lowering in the quality of the wave-function is compensated by a

reduction of the computational expense of several orders of magnitude compared to standard periodic Ψ_{crys} methods.[79] Semi-empirical QM methods have been applied to large chemical and biochemical systems,[79] but they have never been used in CSP. There are several kinds of semi-empirical methods, but they can be broadly divided into two families: those derived from molecular orbital theory, and those derived from DFT, which take the name of density functional tight-binding (DFTB).[78, 79] DFTB methods were used in this thesis.

2.3.2.2.1 Density functional tight-binding

DFTB represents the DFT energy in Equation 2.27 as a Taylor expansion around a reference electron density ρ_0, which is calculated for neutral atoms.[79] Hence, the total energy is written in terms of a deviation δρ from this reference:

$$E[\rho] = E^0[\rho_0] + E^1[\rho_0, \delta\rho] + E^2[\rho_0, (\delta\rho)^2] + E^3[\rho_0, (\delta\rho)^3] + \cdots \qquad 2.29$$

DFTB1 only expands E to the first order term, DFTB2 to the second order and DFTB3, which was used in this thesis, to the third order.[79]

All terms in Equation 2.29 are subject to approximations. $E^0[\rho_0]$ represents the superimposition of the charge density of neutral atoms, and as such it is independent of the chemical environment and can be parametrised to DFT or experimental data.[79, 80] In DFTB, the zeroth order term takes the form of an atom-atom potential that describes short-range repulsive interactions between nearest-neighbours:[79]

$$E^0[\rho_0] \sim E_{rep} = \frac{1}{2}\sum_{AB} V_{AB}^{rep} \qquad 2.30$$

where V is the parametrised element and A and B are two atoms. The first order term $E^1[\rho_1]$ represents the energy of all the occupied orbitals:[79]

$$E^1[\rho_0, \delta\rho] = \sum_i \langle\psi_i|H_0|\psi_i\rangle = \sum_{iAB}\sum_{\mu\in A}\sum_{\nu\in B} n_i C_{\mu i} C_{\nu i} H^0_{\mu\nu} \qquad 2.31$$

where Ψ_i represents the molecular orbitals described by a minimal basis set constituted only of valence electrons, with occupancy n_i, μ and ν are atomic orbitals on atoms A and B respectively, C the basis set coefficients and H_0 the Hamiltonian of the zeroth order density.[80] The diagonal element of the Hamiltonian matrix (*i.e.* $H^0_{\mu\mu}$) describe the energy level of the free atoms; all the off-diagonal Hamiltonian matrix elements are 2-centred, as the 3 and 4-centred elements are neglected.[79] The Hamiltonian elements depend only on the atomic pairs, and so they are calculated in advance with DFT and tabulated for use in DFTB.[80]

In the second order term, deviations of charge density from the reference are approximated as interactions between atomic monopoles:[79, 80]

$$E^2[\rho_0, (\delta\rho)^2] = \frac{1}{2}\sum_{AB} \Delta q_A \Delta q_B \gamma_{AB} \qquad 2.32$$

where γ_{AB} is an analytical function that models the decay of coulombic interactions and replaces the charge-charge integrals. At long distances γ_{AB} approaches

$1/R_{AB}$, while at short distances it includes the electron-electron interactions and so it deviates from $1/R_{AB}$,[79] while the on-site values γ_{AA} and γ_{BB} represent Hubbard parameters U_A and U_B, which are related to the atomic hardness.[79] The γ_{AB} function works poorly for hydrogen, and a modified version γ_{AB}^{h} is used in DFTB3.[79, 80] The charges are iterated to self-consistency, which allows to estimate polarisation effects.[24]

Finally the third order term describes the change of the atomic hardness with respect to the charge state:[80]

$$E^3[\rho_0, (\delta\rho)^3] = \frac{1}{3}\sum_{AB}(\Delta q_A)^2 \Delta q_B \Gamma_{AB} \qquad 2.33$$

where Γ_{AB} is the derivative of γ_{AB} with respect to atomic charge.[80]

The main advantage of DFTB3 is that it can calculate all the properties that can be determined by DFT calculations, including energies, dipole moments, atomic forces and charge densities, at a computational cost up to three orders of magnitude lower.[79]

However, DFTB3 has some limitations that must be taken into account. First of all it inherits from DFT the lack of long-range dispersion, which in this work was adjusted by adding the D3 dispersion correction to the DFTB3 energies:[79]

$$E_{DFTB3-D3} = E_{DFTB3} + E_{disp}^{D3} \qquad 2.34$$

the damping of the D3 dispersion parameters has to be specifically fitted for use in DFTB3.[77] Secondly, the zeroth-order term in Equation 2.30 is approximated as a repulsive potential that is only a function of nearest neighbour distances: as a consequence, other long-range interactions are neglected, and so further corrections have to be included.[79] Furthermore, the neglect of three and four-centred integrals in the Hamiltonian matrix reduces the accuracy of the first order term in the expansion, and approximating the interactions of long-range charge densities by monopoles (Equation 2.32), excluding dipoles, quadrupoles *etc.,* worsens the description of intermolecular interactions.[79] Finally, the minimal basis set causes an underestimate of both polarizability and Pauli repulsion, which can lead to unrealistically short intermolecular distances.[79]

Despite these weaknesses, DFTB has been shown to perform reasonably well for molecular crystals. For example, DFTB3-D3 was capable to reproduce the absolute energies of the crystal structures in the X23 benchmark set to a good degree of accuracy.[77] Also, it has been shown that DFTB3-D3 can be an effective tool to estimate the vibrational energy component to free energy of carbamazepine.[81] Given these promising results, DFTB3-D3 was applied in Chapter 7 to test whether it could be integrated in CSP workflows. All DFTB3-D3 calculations were performed with the programme dftb+.[82]

2.3.3 Calculation of free energies

All methods that have been discussed so far calculate E_{latt} of static crystal structures. However, real crystals are not static, and their vibrations have fundamental effects on many crystalline properties: they expand crystal structures and can induce polymorphic phase transitions.[25, 83] Nyman and Day have calculated that for a sample of 475 polymorphic pairs, approximately 21% had their E_{latt} stability ranking reversed by vibrational effects at temperatures below their melting points. [83]

What determines the thermodynamic stability of crystal structures at given temperature and pressure conditions is the Gibbs free energy, *G*, defined as:

$$G(p,T) = U + pV - TS \qquad 2.35$$

where *U* is the internal energy of the system, *p* the pressure applied to the system, *V* its volume, *T* the temperature and S the entropy. Pressure effects can be easily included in CSP and in some cases calculations under pressure are performed to target specific polymorphic forms.[84] However, in most CSP studies the *pV* term is ignored as it only has an effect at high *p* values.[28] If the volume is kept constant, the thermodynamic stability can then be approximated by the Helmholtz free energy, *A*:

$$A(T) = U - TS \qquad 2.36$$

to calculate *A* in CSP, it is common practice to divide *A* into E_{latt} and F_{vib}, defined as the vibrational component to free energy:[25]

$$A(T) = E_{latt} + F_{vib}(T) \qquad 2.37$$

Where:

$$F_{vib}(T) = ZPE + U_{thermal}\,(T) + TS \qquad 2.38$$

ZPE represent the zero point energy (*i.e.* the lattice vibrations that occur at 0 K) and $U_{thermal}$ is the thermal component of internal energy.[83]

In this thesis, free energies were modelled using the rigid-body harmonic approximation, where F_{vib} is computed from the vibrational frequencies of harmonic phonons, which are collective non-interacting QM excitations of the lattice.[25, 85] Phonon frequencies (ω) can be calculated from the second derivatives of the energy with respect to the atomic displacements, and an accurate computation requires that the crystal structure is at an energy minimum (*i.e.* the forces are equal to zero, see Equation 2.1).[86, 87] Phonons repeat periodically within the reciprocal space, *k*, and their frequency varies at different points in the Brillouin zone, the smallest repeating unit in the reciprocal lattice.[85]

The quasi-harmonic approximation can be used to optimise *A* as a function of volume, and it allows to estimate the effect of thermal expansion.[1, 68, 83] This approximation was not utilised in this thesis, and its contribution to *A* is generally small.[83]

2.3.3.1 Calculation on free energies with the Ψ_{mol} model

In this thesis the calculation of F_{vib} values with the Ψ_{mol} method was performed with DMACRYS.[10] It determines the ω values by the analytical calculation of a matrix of second derivatives;[10] care must be taken that the lattice energy optimisation is complete, so that no imaginary frequency values are present.[87] Only low-frequency intermolecular phonon modes can be calculated with the Ψ_{mol} model, since the molecular conformation is held rigid in DMACRYS.[10] High-frequency intramolecular phonon modes are assumed to have a negligible effect on the relative energy ranking between crystal structures.[25]

DMACRYS only calculates the phonon modes at the Γ point, *i.e.* at the origin of the Brillouin zone (where k=0).[86] This decreases the accuracy of the results: only optical phonon modes can be found at the Γ-point, while acoustic modes, which can have a large effect on F_{vib}, are zero at the origin.[85] Accurate calculations of phonon frequencies requires the sampling of several k-points to guarantee that the phonon modes are converged.[25, 68] Nyman and Day have proposed a method to calculate more accurate F_{vib} values based on building several linear supercells of the same crystal structure.[25] This makes sure that a sufficient number of k-points is sampled. Relative free energies have been found to converge well if a k-point distance of at least 0.12 $Å^{-1}$ is guaranteed.[25] In Chapter 3, the standard DMACRYS approach to calculate F_{vib} from ω values at the Γ-point was utilised, while in Chapter 7 the linear supercell method was employed.

2.3.3.2 Calculation on free energies with the Ψ_{crys} model

Phonon vibrations calculated with Ψ_{crys} are more complete, as both molecular and lattice modes can be coupled.[24] Since the energy expression is much more complex than the intermolecular potential used with the Ψ_{mol} method (see Equation 2.18), the second derivatives are not computed analytically but from finite displacement: the position of the atoms are slightly perturbed and a reaction force is calculated.[87] A matrix of second derivatives is then computed form these reaction forces. Values of ω at different k-points can be calculated by integration of the Brillouin zone,[22] and supercells with minimum unit-cell lengths can be built to guarantee that the phonon modes are converged.[88] In this thesis, Ψ_{crys} phonon modes calculated with DFTB3-D3 were utilised in Chapter 7 to estimate the free energies of some crystal structures, which were compared to those calculated with DMACRYS using the linear supercell method.

2.4 Crystal structure prediction: methods and codes

CSP methodologies use various computational tools to explore the crystalline E_{latt} surface of molecules. This surface is multidimensional, and it is a function of the lattice lengths and angles of the unit cell (*i.e.* the smallest repeating unit in the crystalline lattice), the position of the centre of mass and the orientation of the molecules for a given space

group symmetry and number of molecules in the asymmetric unit, and the CDFs of all the symmetry independent molecules.[28, 29] The E_{latt} surface is immense, in particular for large molecules as each CDF increases its dimensionality, and cannot be fully explored with very accurate but computationally expensive methods. Hence, CSP algorithms require some approximations. Broadly speaking, all successful CSP methodologies use a hierarchical approach: an initial crystal structure search aimed at finding the most important local E_{latt} minima is followed by a refinement of the generated crystal structures performed with higher quality models.[2] Searches are generally carried out with potentials that can be empirically-fitted[89] or parametrised to reproduce *ab initio* data,[90] while the final refinement is performed either with periodic Ψ_{crys} methods or with Ψ_{mol} methods where a high-quality wave-function is calculated for each molecular conformation.[20]

In this section, the CSP workflow used in this thesis and the main algorithms are described.

2.4.1 Crystal structure search

2.4.1.1 Definition of conformational flexibility

Most (but not all)[91, 92] CSP methodologies require an initial assessment of conformational flexibility before performing a search,[2] to determine what conformations a molecule could realistically take in a crystal structure. Since the search space increases with the number of conformations that are considered,[29] limiting them to a set of the most relevant ones can keep the overall cost of CSP manageable. Furthermore, in a search the only CDFs that need to be considered are the most flexible torsion angles (*e.g.* those around acyclic bonds), as they can separate the important local minima in E_{latt}.[28]

A variety of methods to define the conformational flexibility of a molecule exists. They are generally based either on calculating the energies of different gas-phase configurations with *ab initio* methods or on using existing information on the conformational preferences of molecules in crystal structures, which are often retrieved from the Cambridge Structural Database (CSD).[93] Both approaches can be useful, and in Chapter 5 of this thesis an attempt to find an optimal balance between these two methods is discussed.

In general, molecules do not crystallise in conformations that have high-energies compared to the isolated-molecule global minimum in conformational energy (see Chapter 2.3.1.1). However, for flexible molecules the range of plausible conformations can be very broad. In this thesis *ab initio* calculations to define the search space were carried out in Chapters 3, 4 and 5 with Gaussian at the PBE0 6-31G(d,p) level of theory. The main shortcoming of this method is the high computational expense, in particular for very flexible molecules.

CSD surveys can also be helpful, as they provide information on the conformational preferences of molecules in actual crystal structures, not just in the gas-phase. However, care must be taken, as these preferences are deduced from molecules that contain similar fragments but could have a different balance of intra- and intermolecular interactions. Furthermore, a blind reliance on CSD data could bias CSP studies towards the 'already-known'[94] and prevent to access unusual but possibly relevant conformations and crystal structures.

2.4.1.2 Generating crystal structures: CrystalPredictor

In this thesis, CrystalPredictor[95, 96] was used to perform the crystal structure searches. This algorithm uses a deterministic low-discrepancy Sobol sequence[97] to perform an efficient search through the multidimensional E_{latt} surface, which is a function of the unit-cell lengths and angles, the position and orientation of the molecules (as allowed by the space group and number of molecules in the asymmetric unit) and a set of user-defined torsion angles that are treated as independent CDFs.[29, 95, 96] The Sobol sequence is used to generate the initial crystal structures and guarantees a better coverage of the search space compared to uniform or random (stochastic) samplings.[29]

The initial crystal structures are then automatically optimised with the Ψ_{mol} model (see Equation 2.16). U_{inter} is modelled as in Equation 2.18: U_{elec} is calculated from atomic point charges (Equation 2.19), as they are cheaper to compute and less dependent on conformation than distributed multipoles, while $U_{rep\text{-}disp}$ is calculated with the exp-6 potential in Equation 2.24. The atomic point charges can either be fixed at the values of the input conformation or calculated as a function of the values of the independent CDFs. In this thesis fixed point charges were used.

Two versions of CrystalPredictor exist, which differ only in the way they treat ΔE_{intra} and conformational flexibility (*i.e.* they work identically for rigid searches). In this thesis only CrystalPredictor 1 was used. It models ΔE_{intra} through a Hermite interpolation over a grid of intramolecular energies covering the flexibility ranges of the CDFs treated as search variables. The ΔE_{intra} values in the grid are calculated before the search through a set of *ab initio* optimisations of the isolated molecules, in which the independent CDFs are fixed at some user-defined values while the other CDFs are allowed to relax to the closest local minimum in conformational energy. The size of the ΔE_{intra} grid increases very rapidly with the number of independent CDFs. This problem can be limited by dividing a molecule into smaller more manageable surrogate molecules containing a subset of the independent CDFs. This convenient approximation allows a significant reduction in computational cost, and it was used in all the searches performed in Chapters 3, 4 and 5. The possibility of using surrogate models is indeed the main

reason why only CrystalPredictor 1 was used in this thesis. When surrogate molecules are used, ΔE_{intra} is then calculated as:[28]

$$\Delta E_{intra} = \sum_{a=1}^{N} \Delta E_{intra}^{a} \qquad 2.39$$

where *N* is the number of subsets. This approximation can be used only when it is reasonable to assume that the energy penalty for varying a subset of torsion angles is not affected by the rest of the molecule. In CrystalPredictor 1, only the independent CDFs are varied in the search, while the other dependant CDFs (bond-lengths, bond-angles and torsion angles treated as rigid) are kept at their input value (generally the lowest energy gas-phase conformer).

During the course of this thesis, a new version of this programme has been developed: CrystalPredictor 2, which models ΔE_{intra} using local approximate models (LAMs).[29, 64, 95] The intramolecular energy of each conformation is calculated by a second-order Taylor expansion from some pre-calculated points within the flexibility range of the independent CDFs; an example of LAMs is shown in Figure 2.4. LAMs are calculated in the same way as the grids in CrystalPredictor 1, but an expensive calculation of the second derivatives of conformational energy is also required. These LAMs can be stored and re-utilised in subsequent optimisations of E_{latt} with CrystalOptimizer (see Chapter 2.4.2.2), reducing the overall computational cost.[29]

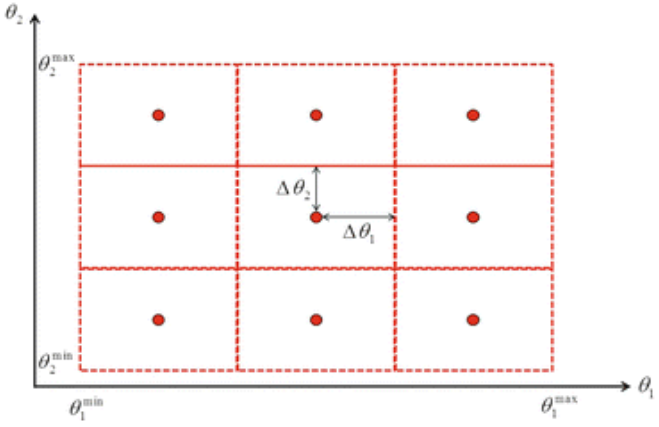


Figure 2.4: Schematics of LAMs of a molecule with two independent CDFs. *Ab initio* calculations are performed at the red LAM points, while the rectangles indicate the range of applicability of each LAM point. The energy at each point in the grid is calculated by a second order Taylor expansion from the closest LAM point.

A finer LAM spacing increases the accuracy of the model, but also the computational cost. Recently, improvements in the code to guide the user towards an optimal LAMs sampling have been added, which were partially inspired by the outcome of the Blind Test CSP study of molecule XXVI (see Chapter 3).[98] An advantage of CrystalPredictor 2 is that each LAM point corresponds to an optimised configuration, meaning that the CDFs that are not treated as search variables are not fixed at the values of the input conformation but can respond to changes of the independent CDFs, allowing a more accurate description of the molecular geometries. However, it is not currently

possible to utilise the approximation in Equation 2.39, meaning that for flexible molecules large multi-dimensional LAMs are required; also LAMs are specific to a certain input conformation, and have to be recalculated if a different input is used (*e.g.* two searches with *cis* and *trans* carboxyl groups), which is not required for the grids. These factors made CrystalPredictor 2 unaffordable for the large and flexible molecules that were studied in this thesis.

2.4.2 Crystal structure refinement

Since crystal structure searches are performed with approximate models, the generated crystal structures need to be refined to obtain more accurate geometries, as well as absolute and relative energies. The main crystal structure refinement programmes that were used in this thesis were DMACRYS, which optimises only the intermolecular interactions, and CrystalOptimizer, which minimises E_{latt} coupling DMACRYS with *ab initio* calculations on the molecular conformation/s.[64]

2.4.2.1 Optimisation of U_{inter}: DMACRYS

In this thesis, the optimisations of the intermolecular interactions were performed with DMACRYS. DMACRYS uses the Ψ_{mol} method: it treats the molecules in a crystal structure as rigid, and models U_{inter} as a sum of some of the terms described in Chapter 2.2. U_{inter} was modelled as in Equation 2.18, with U_{elec} calculated from distributed multipoles up to hexadecapoles (rank 4) fitted with GDMA (see Chapter 2.3.1.2.2) from the wave-function computed with Gaussian, and $U_{rep\text{-}disp}$ calculated with the exp-6 potential in Equation 2.24.

In DMACRYS, long-range electrostatic interactions (charge-charge, charge-dipole and dipole-dipole) are calculated using the Ewald summation method, and the remaining interactions are summed in real space up to a user-defined cut-off distance.[10] A Newton-Raphson scheme minimises the structure to the closest minimum in $U_{inter.}$[10] However, if the second-derivative matrix contains any negative eigenvalue DMACRYS warns the user that the crystal structure has converged to a saddle point. The crystalline symmetry can then be reduced to a sub-group of the initial space group, finding a lower-energy true E_{latt} minimum.[10]

DMACRYS can be sufficient as a final refinement tool in CSP only for very rigid molecules or if the molecular conformations have already been optimised with other methods (*e.g.* in Chapter 7 with DFTB3-D3). For flexible molecules DMACRYS is often not sufficient, and coupling with *ab initio* programmes is required to optimise both the intra- and intermolecular components of E_{latt} at the same time.[10]

2.4.2.2 Optimisation of $E_{latt:}$ CrystalOptimizer

In this thesis, CrystalOptimizer was the main programme that was used to perform full E_{latt} optimisations of the generated crystal structures. This algorithm uses the Ψ_{mol} method, and it optimises both components of E_{latt} (Equation 2.16) as a function of a set of user-defined independent CDFs. Since bond-lengths are not affected by the solid-state environment,[64] generally only the most flexible torsion and bond-angles are treated as independent CDFs. However, the choice of the independent CDFs is generally arbitrary, as there are no defined rules that allow an optimal selection of the torsion and bond-angles that need to be treated as variables in CrystalOptimizer minimisations. This issue is discussed in detail in Chapter 6.

CrystalOptimizer works by performing a two-level optimisation:[64]

$$min\, E_{latt}(CDFs_{independent}) = min[\Delta E_{intra} + min(U_{inter})] \qquad 2.40$$

in the outer-optimisation the independent CDFs are varied. For a given set of values of the independent CDFs Gaussian is called to relax the rest of the molecule/s to the closest local minimum in gas-phase energy, and ΔE_{intra} is calculated. Then, the inner-optimisation takes place: distributed multipoles are derived from the charge density of the optimised conformation with GDMA, and U_{inter} is minimised with DMACRYS.[64] The process is repeated until the outer-optimisation problem is complete; a quasi-Newton scheme is used to converge E_{latt} as a function of the independent CDFs.[64]

In principle, this scheme requires to perform expensive *ab initio* optimisations and charge density calculations at each point within the outer-optimisation scheme. However, the process is made more computationally efficient by using LAMs,[64] which have already been described in Chapter 2.4.1.2. *Ab initio* calculations are stored in LAMs databases, containing conformational energy values, their first and second derivatives with respect to the CDFs, and distributed multipoles.[64] If at any point in the outer-optimisation a similar set of independent CDF values in is present in the LAMs databases, within a user-defined tolerance, then the results of previous calculations are re-utilised. Otherwise, new *ab initio* calculations are performed on-the-fly and stored in the databases.[29, 64] This procedure is very efficient, as re-utilising *ab initio* calculations reduces the overall computational cost without decreasing accuracy.[64] This possibility of re-utilising previous calculations is indeed an advantage of the Ψ_{mol} method: the overall computational cost does not scale with the number of crystal structures that are optimised like for the Ψ_{crys} method, since as more optimisations are performed the LAMs databases cover a larger portion of the conformational space and less costly *ab initio* calculations are required.

2.5 Comparison of crystal structures and of molecular conformations

Since CSP studies tend to predict many more polymorphs than they are actually experimentally determined (as mentioned in Chapter 1.2.2),[99] it is necessary to determine whether structures matching known forms are present among those that have been computationally predicted and to find common motifs that can define structural trends[20] or hint to the presence of disorder.[100, 101] Furthermore, CSP methods tend to generate several duplicates of the same E_{latt} minima,[2, 15] and crystal structure comparison tools can be used to determine which forms are unique and which are just duplicates corresponding to a different definition of the same unit cell. Determining a degree of dissimilarity that experimental or computer-generated crystal structures must possess to be classified as polymorphs is a non-trivial task;[102, 103] this issue is discussed in Chapter 8. In this section, the tools that were used in this thesis for comparing crystal structures and molecular conformations are described.

2.5.1 The Crystal Packing Similarity tool

The Crystal Packing Similarity tool has developed from COMPACK, a programme written by the Cambridge Crystallographic Data Centre (CCDC) to quantify the similarity between crystal structures for use by both experimentalists and those involved in computational studies.[104] Crystal Packing Similarity is available through Mercury,[105] a programme developed by the CCDC that was used in this thesis mainly as a visualisation tool, and recently also through the CSD Python API (see Chapter 2.6.5). The idea behind the Crystal Packing Similarity tool (and COMPACK) is that relying on visual analyses to determine the similarity between crystal structures is a slow and arbitrary process, and quantitative parameters are key to achieve realistic comparisons.[104]

The Crystal Packing Similarity tool performs crystal structures comparisons based exclusively on the position of the atoms, and so it is independent of symmetry and unit cell information. It represents a crystal structure as a cluster of *N* molecules, where *N* is user-defined, built from a central reference molecule and the nearest *(N-1)* ones. This cluster is then used as a reference sub-structure search query, which is overlaid with another crystal structure to find the three-dimensional coordinates that lead to highest possible number of molecules matching. Two molecules are considered to match if the distances and the angles of the triangles in the reference cluster and in the comparison structure differ by no more than user-defined tolerances; the default values are 20% distance and 20° angle tolerances.[104] The level of structural similarity is then quantified in terms of the root-mean-square-deviation ($RMSD_x$), which is a measure of the average distance between the atoms of the *x* molecules that can be matched. This value is calculated as:

$$RMSD_x = \sqrt{\frac{\sum_{i=1}^{M} \delta_i}{M}} \qquad 6672.41$$

where *M* is the total number of atoms of the *x* molecules that can be matched and δ is the distance between these atoms. The user can decide whether or not to account for the position of hydrogen atoms, which are generally neglected when comparing experimental crystal structures since they are very weak X-ray scatterers. When *x* = *N*, each molecule can be overlaid with the reference cluster. If the $RMSD_x$ value is very low, within numerical noise, the two clusters can be considered to be identical, while large values can indicate more significant differences, *e.g.* thermal expansion for crystal structures determined at different temperatures.

The choice of *N* depends on the purpose of the comparison. If *N=1*, only one molecule between two crystal structures is compared; for crystal structures with only one molecule in the asymmetric unit cell (*i.e.* Z'=1) this is equivalent to performing a comparison of the molecular conformations. On the other hand, larger values of *N* mean that wider portions of the crystal structures are compared, which could better capture differences and similarities. The default value of *N=15* is commonly used, as in the original testing of COMPACK it was found sufficient to discriminate between polymorphs and duplicates of crystal structures containing only one molecular species, which are the subject of this thesis.[104] The analyses in Chapter 8 confirm these results.

2.5.2 Simulated powder X-ray diffraction (PXRD) pattern similarity

Another way to obtain a single quantitative measure of structural similarity is to compare simulated powder X-ray diffraction (PXRD) patterns. PXRD patterns depend only on the distances between atoms.[106] They consist of a 1-dimensional function that simulates the diffraction patterns of a crystal structure for a specific X-ray wavelength and for a given range of 2θ values. There are several ways to quantify the similarity between powder patterns. Point-by-point measures that calculate the differences in peak intensity at identical 2θ values are extremely sensitive to peak positions.[106] Since the position of peaks in a simulated PXRD patterns depends on the unit cell dimensions, a small difference, *e.g.* due to variations in pressure or temperature, can give poor similarities for two determinations of identical crystal structures.[106] Hence, the CCDC tool used in this thesis calculates PXRD similarity with a weighted cross-correlation function developed by de Gelder,[107] which is less sensitive to peak shifts. PXRD similarities are quantified in terms of a number that goes from 0 to 1, where 1 indicates identical simulated patterns. This measure used to be common in CSP, and it is still utilised to compile lists of unique polymorphs within the Cambridge Structural Database (CSD).[108] However, further studies have shown that PXRD similarities are sometimes insensitive to certain crystalline modifications, as they can be close to 1 for different crystal

structures;[109] their inadequacy was also shown in the analyses in Chapter 8. Hence, in this thesis PXRD similarities were never used on their own but only as a preliminary method to filter out very dissimilar crystal structures. The Crystal Packing Similarity tool (or its earlier version COMPACK) was always used to perform realistic comparisons.

2.5.3 Ultrafast shape recognition (USR)

Recognising the shape similarity between molecules is a vital part of drug discovery, as shape differences can hinder the interaction between a drug molecule and its target.[110] Ultrafast shape recognition (USR) was developed to perform a fast quantitative shape comparison of large databases of molecules, using simple descriptors that to do not rely on slow molecular overlays (*e.g.* calculating the $RMSD_1$ with the Crystal Packing Similarity tool, see Chapter 2.5.1), which can also be the source of several errors.[110] USR characterises the shape of a molecule as a set of 1-dimensional parameters that carry 3-dimensional information. Twelve moments are computed: they are the average, the variance and the skewness of the distribution of the distances of the atoms to four locations, namely "the molecular centroid (ctd), the closest atom to ctd (cst), the farthest atom to ctd (fct) and the farthest atom to fct (ftf)".[110] This method provides a good balance between accuracy and computational cost. A unique percentage similarity score ($S^{\%}$) between two molecules A and B can then be calculated as:[110]

$$S_{AB}^{\%} = \left(\frac{1}{1+\frac{1}{12}\sum_{i=1}^{12} M_i^A - M_i^B} \right) \cdot 100 \qquad 2.42$$

where M_i represents the i^{th} moment. In Chapter 5, USR was used to quantify the effect of varying a torsion angles on the overall shape of the molecule. The programme USRCAT[111] was employed for this purpose.

2.6 Crystal structure informatics

CSP methods use calculations to predict the packing possibilities of a molecule, without any prior knowledge. However, we possess an enormous amount of information on real crystal structures, which can be used to validate and possibly integrate CSP workflows.

Any experimentally-determined crystal structure is a source of information on the intra- and intermolecular interactions that occur in the solid-state.[112] Analysing samples of crystal structures can provide knowledge about the packing preferences of molecules with certain characteristics, such as hydrogen bond donors and acceptors, size, functional groups, *etc*. Fast and efficient search and analysis tools allow to retrieve and analyse information from large datasets. In the context of this thesis, the most useful database is the CSD, which currently contains more than 900,000 experimentally-determined organic crystal structures. The predictive power of CSD information is proven by how it has been recently used to perform 'healthchecks' on

pharmaceutical crystal structures and assess the likelihood of more stable polymorphs existing,[112] and even to perform full CSP studies.[113] In this section, the main tools that can be used to extract information from the CSD are described, and the ways this information was utilised in this thesis are indicated.

2.6.1 ConQuest

ConQuest is the primary programme for mining the CSD.[112, 114] It allows to search crystal structures based on several factors, including compound name, molecular formula, crystalline properties and/or literature reference. It can also perform searches based on chemical substructures, which can easily be drawn. Data of interest on crystal structures that meet the search criteria can be extracted, *e.g.* specific torsion angle distributions, distributions of bond-lengths and of hydrogen bond or other non-bonded contacts. Furthermore ConQuest allows the restriction of searches using parameters like R-factor or disorder and to consider only structures in specific CSD-subsets, such as lists of unique polymorphs.

ConQuest is a very flexible tool, and it was used in Chapters 3, 6 and 8 to find crystal structures with specific molecular and packing characteristics as well as to extract the distributions of some key torsion and bond-angles.

2.6.2 Mogul

Mogul is a programme for extracting conformational information from the CSD and analysing geometrical preferences.[112, 115] It takes a fragment of interest as input and mines the CSD for molecules containing the same fragment in a comparable chemical environment. It finally outputs distributions of observed values of the bond-lengths, bond-angles and/or torsion angles in the form of histograms. Mogul classifies the value of a CDF as usual or unusual depending on its CSD frequency, and this has been used within the 'healthchecks' to evaluate the likelihood of polymorphism: an unusual molecular geometry indicates that alternate crystal structures with more common values for the conformation may exist.[112]

In this thesis, Mogul was used in Chapter 5 to assess the ranges of conformational flexibility for generating the ΔE_{intra} grids (see Chapter 2.4.1.2) of some torsion angles to be treated as flexible in CrystalPredictor searches.

2.6.3 CSD knowledge-based conformational libraries

Recently, an approach has been developed by the CCDC for analysing the geometrical preferences of organic molecules using knowledge-based conformational libraries, which contain information about the CSD distributions of bond-lengths, bond-angles and dihedrals.[116] These libraries are derived from Mogul, but have some significant

differences from their parent programme that make them more suited for the rapid and fully automatized analysis of conformational preferences that is required in applications such as generating key molecular conformations (see Chapter 2.6.4) and geometrical optimisations. First of all, key conformational features are automatically identified, and then a "cascade" search approach is used to extract the distribution of their values from the CSD; this approach is faster than the one used by Mogul.[116] A first search is performed for libraries that match precisely the chemical environment of the fragment of interest, and if enough entries are not found the search is repeated for libraries that describe it less accurately and so on. If no library remains, the search fails, otherwise the most chemically precise library containing a sufficient number of entries is utilised. Furthermore the CSD conformational libraries do not provide information about each definition of flexible torsion angles, but they consider rotamers, selecting one specific 'reference torsion angle' around each rotatable bond to avoid redundancies.[116] Finally, the rotamer distributions cover the full 0-360° range, accounting for chirality.[116] There are also other changes compared to Mogul in the way flexible rings are handled, but none of the molecules explored in this thesis contained this kind of feature.

The CSD knowledge-based conformational libraries were used in Chapter 5 to assess the ranges of conformational flexibility of the main flexible torsion angles to choose their treatment in the crystal structures searches.

2.6.3.1 Kernel density estimation (KDE)

Conformational information is generally analysed through histograms. However, histograms have weaknesses, as they are not smooth, non-differentiable and depend strongly on the bin size. Hence, despite the ease of visualisation, histograms are not well-suited to perform quantitative analysis of conformational distributions.[117] A statistical density estimator on the other hand can generate probability density functions (PDFs), which are smooth functions that are related to the frequency of a given value in a distribution. In particular, kernel density estimation (KDE) is a very general approach for the generation of PDFs, which does not require parameters like mean or standard deviation of the distribution and depends only on the data points.[117] It is important that the kernel is suited to the conformational data at hand. In Chapter 5 KDE was used to generate PDFs of torsion angle distributions, which are circular in nature, and the Von Mises kernel was chosen, following the method used by McCabe *et al*.[117] With the Von Mises kernel, the probability density function *f* is calculated as:

$$f(\theta) = \frac{1}{2\pi I_0(\nu)} \sum_{i=1}^{n} exp[\nu \, cos(\theta - \theta_i)] \qquad 2.43$$

where θ is a torsion angle value ranging from 0 to 2π, n is the size of the sample, I_0 is the zeroth order modified Bessel function of the first kind, θ_i represents the i^{th}

observed value, and *v* is a shape parameter that is related to the level of smoothing of the PDF. The *v* parameter was calculated via a set of equations proposed by Fisher,[118] which optimise the level of smoothing depending on the shape of the distribution.

2.6.4 The CSD Conformer Generator

The CSD Conformer Generator (CG) produces plausible conformations for an input molecule, using the information in the CSD knowledge-based conformational libraries to guide this selection.[119] It was originally developed for use in life-science modelling applications such as protein-ligand-docking and pharmacophoric shape matching.

As a first step the CG optimises bond-lengths and bond-angles of the molecular input conformation using a modified version of the Tripos force field. Successively, the distributions of each rotamer and flexible ring (if present) in the molecule are retrieved from the CSD conformational libraries. Through sampling these distributions, a diverse set of conformations is generated, each having a different combination of rotamer and flexible ring values, if present. Similar conformations are clustered using a combination of a torsion dissimilarity coefficient and atomic root-mean-squared deviations. In addition, structures that feature unphysical close contacts or more than a specified number of unusual rotamer values are removed. For each generated conformation, the CG assigns a probability score, on which the conformations are ranked:[119]

$$score = \frac{ln(p_{max}) - ln(p)}{ln(p_{max}) - ln(p_{min})} \qquad 2.44$$

where p is the probability of the conformation, and p_{max} and p_{min} are the probabilities of the most and least probable conformations respectively. These p values are calculated as:

$$p = \prod_{i=1}^{n} p_i \qquad 2.45$$

where p_i is the probability assigned to each of the n parameters (rotamers and flexible rings) that the CG uses to produce the conformations. Despite some weaknesses, such as the use of a heuristic clash term and the lack of any electrostatic interaction like internal hydrogen bonds, the CG has been found to perform well when it was compared to competitors on identical sets of molecules.[119] In this thesis, the CG was used in Chapter 5 to aid the selection of the conformational regions that were searched with CrystalPredictor.

2.6.5 The CSD Python API

Although the CSD contains an enormous amount of information, which can be systematically retrieved and analysed, it was not until recently that it could be used in programming. The release of the CSD Python API has brought an enormous advance, as it allows to access data more quickly and effectively, and to interface the retrieved

information with other tools and to build workflows.[93] Almost all CCDC informatics tools can be accessed through the Python API. The CSD Python API was used throughout this thesis to perform a variety of analyses and to interface information retrieved from the CSD with external programmes.

2.7 References

1. Cervinka, C.; Beran, G. J. O., Ab initio prediction of the polymorph phase diagram for crystalline methanol. *Chemical Science* **2018,** *9*, 4622-4629.
2. 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.; 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 Crystallographica Section B* **2016,** *72* (4), 439-459.
3. Price, S. L.; Braun, D. E.; Reutzel-Edens, S. M., Can computed crystal energy landscapes help understand pharmaceutical solids? *Chemical Communications* **2016,** *52*, 7065-7077.
4. Price, S. L., Is zeroth order crystal structure prediction (CSP_0) coming to maturity? What should we aim for in an ideal crystal structure prediction code? *Faraday Discussions* **2018,** *in press*.
5. Neumann, M.; van de Streek, J., How many Ritonavir cases are there still out there? *Faraday Discussions* **2018**, *Advance article.*
6. Stone, A. J., *The Theory of Intermolecular Forces*. Oxford University Press: Oxford, 2013; Vol. 2.
7. Maitland, G. C.; Rigby, M.; Smith, E. B.; Wakeham, W. A., *Intermolecular Forces. Their Origin and Determination*. Clarendon Press: Oxford, 1981.
8. Kumar, M.; Parul, Methods for solving singular perturbation problems arising in science and engineering. *Mathematical and Computer Modelling* **2011,** *54* (1), 556-575.
9. Dalgarno, A.; Drake, G. W. F., Rayleigh-Schrödinger multiple perturbation theory. *Chemical Physics Letters* **1969,** *3* (6), 349-350.
10. Price, S. L.; Leslie, M.; Welch, G. W. A.; Habgood, M.; Price, L. S.; Karamertzanis, P. G.; Day, G. M., Modelling Organic Crystal Structures using Distributed Multipole and Polarizability-Based Model Intermolecular Potentials. *Physical Chemistry Chemical Physics* **2010,** *12* (30), 8478-8490.
11. Karamertzanis, P. G.; Day, G. M.; Welch, G. W. A.; Kendrick, J.; Leusen, F. J. J.; Neumann, M. A.; Price, S. L., Modeling the interplay of inter- and intramolecular hydrogen bonding in conformational polymorphs. *Journal of Chemical Physics* **2008,** *128* (24), 244708-244717.
12. Welch, G. W. A.; Karamertzanis, P. G.; Misquitta, A. J.; Stone, A. J.; Price, S. L., Is the induction energy important for modeling organic crystals? *Journal of Chemical Theory and Computation* **2008,** *4* (3), 522-532.
13. Tomasi, J.; Mennucci, B.; Cancès, E., The IEF version of the PCM solvation method: an overview of a new method addressed to study molecular solutes at the QM ab initio level. *Journal of Molecular Structure: THEOCHEM* **1999,** *464* (1–3), 211-226.
14. Dunitz, J. D.; Gavezzotti, A., How molecules stick together in organic crystals: weak intermolecular interactions. *Chemical Society Reviews* **2009,** *38* (9), 2622-2633.
15. Price, S. L., Predicting crystal structures of organic compounds. *Chemical Society Reviews* **2014,** *43* (7), 2098-2111.
16. Aina, A. A.; Misquitta, A. J.; Price, S. L., From dimers to the solid-state: Distributed intermolecular force-fields for pyridine. *The Journal of Chemical Physics* **2017,** *147* (16), 161722.

17. Misquitta, A. J.; Stone, A. J., Ab Initio Atom-Atom Potentials Using CAMCASP: Theory and Application to Many-Body Models for the Pyridine Dimer. *Journal of Chemical Theory and Computation* **2016,** *12* (9), 4184-4208.
18. Ryno, S. M.; Risko, C.; Brédas, J.-L., Noncovalent Interactions and Impact of Charge Penetration Effects in Linear Oligoacene Dimers and Single Crystals. *Chemistry of Materials* **2016,** *28* (11), 3990-4000.
19. Wang, Q.; Rackers, J. A.; He, C.; Qi, R.; Narth, C.; Lagardere, L.; Gresh, N.; Ponder, J. W.; Piquemal, J.-P.; Ren, P., General Model for Treating Short-Range Electrostatic Penetration in a Molecular Mechanics Force Field. *Journal of Chemical Theory and Computation* **2015,** *11* (6), 2609-2618.
20. Price, S. L.; Reutzel-Edens, S. M., The potential of computed crystal energy landscapes to aid solid form development. *Drug Discovery Today* **2016,** *21* (6), 912-923.
21. Day, G. M.; Motherwell, W. D. S.; Jones, W., A strategy for predicting the crystal structures of flexible molecules: the polymorphism of phenobarbital. *Physical Chemistry Chemical Physics* **2007,** *9* (14), 1693-1704.
22. Buchholz, H. K.; Hylton, R. K.; Brandenburg, J. G.; Seidel-Morgenstern, A.; Lorenz, H.; Stein, M.; Price, S. L., Thermochemistry of Racemic and Enantiopure Organic Crystals for Predicting Enantiomer Separation. *Crystal Growth & Design* **2017,** *17* (9), 4676-4686.
23. Price, S. L.; Brandenburg, J. G., Chapter 11 - Molecular Crystal Structure Prediction. In *Non-Covalent Interactions in Quantum Chemistry and Physics*, Elsevier: 2017; pp 333-363.
24. Iuzzolino, L.; McCabe, P.; Price, S. L.; Brandenburg, J. G., Crystal structure prediction of flexible pharmaceutical-like molecules: Density functional tight-binding as an intermediate optimization method and for free energy estimation. *Faraday Discussions* **2018**, *Advance article.*
25. Nyman, J.; Day, G. M., Static and lattice vibrational energy differences between polymorphs. *Crystengcomm* **2015,** *17* (28), 5154-5165.
26. Thompson, H.; Day, G., Which conformations make stable crystal structures? Mapping crystalline molecular geometries to the conformational energy landscape. *Chemical Science* **2014,** *5* (8), 3173-3182.
27. Cruz-Cabeza, A. J.; Bernstein, J., Conformational Polymorphism. *Chemical Reviews* **2014,** *114* (4), 2170-2191.
28. Vasileiadis, M.; Pantelides, C. C.; Adjiman, C. S., Prediction of the crystal structures of axitinib, a polymorphic pharmaceutical molecule. *Chemical Engineering Science* **2015,** *121*, 60-76.
29. Pantelides, C. C.; Adjiman, C. S.; Kazantsev, A. V., General Computational Algorithms for Ab Initio Crystal Structure Prediction for Organic Molecules. *Topics in Current Chemistry* **2014,** *345*, 25-58.
30. Uzoh, O. G.; Galek, P. T. A.; Price, S. L., Analysis of the conformational profiles of fenamates shows route towards novel, higher accuracy, force-fields for pharmaceuticals. *Physical Chemistry Chemical Physics* **2015,** *17* (12), 7936-7948.
31. Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C.; Price, S. L.; Galek, P. T.; Day, G. M.; Cruz-Cabeza, A. J., Successful prediction of a model pharmaceutical in the fifth blind test of crystal structure prediction. *International Journal of Pharmaceutics* **2011,** *418* (2), 168-178.
32. 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. *Crystal Growth & Design* **2013,** *13* (6), 2396-2406.
33. Spencer, J.; Patel, H.; Deadman, J. J.; Palmer, R. A.; Male, L.; Coles, S. J.; Uzoh, O. G.; Price, S. L., The unexpected but predictable tetrazole packing in flexible 1-benzyl-1H-tetrazole. *CrystEngComm* **2012,** *14* (20), 6441-6446.
34. 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. 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, D. J. *Gaussian 09, Revision D.01*, 2009.
35. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.;

Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; 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.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Gaussian Inc.: Wallingford CT, 2004.

36. Thompson, H. P. G.; Day, G. M., Which conformations make stable crystal structures? Mapping crystalline molecular geometries to the conformational energy landscape. *Chemical Science* **2014,** *5* (8), 3173-3182.

37. Nangia, A., Conformational Polymorphism in Organic Crystals. *Accounts of Chemical Research* **2008,** *41* (5), 595-604.

38. Price, S. L., Applications of realistic electrostatic modelling to molecules in complexes, solids and proteins. *Journal of the Chemical Society-Faraday Transactions* **1996,** *92* (17), 2997-3008.

39. Breneman, C. M.; Wiberg, K. B., Determining Atom-Centered Monopoles From Molecular Electrostatic Potentials - The Need For High Sampling Density in Formamide Conformational-Analysis. *Journal of Computational Chemistry* **1990,** *11* (3), 361-373.

40. Stone, A. J., Distributed Multipole Analysis, or How to Describe a Molecular Charge Distribution. *Chemical Physics Letters* **1981,** *83* (2), 233-239.

41. Stone, A. J., Distributed multipole analysis: Stability for large basis sets. *Journal of Chemical Theory and Computation* **2005,** *1* (6), 1128-1132.

42. Day, G. M.; Motherwell, W. D. S.; Jones, W., Beyond the isotropic atom model in crystal structure prediction of rigid molecules: Atomic multipoles versus point charges. *Crystal Growth & Design* **2005,** *5* (3), 1023-1033.

43. Gatsiou, C. A.; Adjiman, C.; Pantelides, C. C., Repulsion-dispersion parameters for the modelling of organic molecular crystals containing N, O, S and Cl. *Faraday Discussions* **2018**, *Advance article.*

44. Pyzer-Knapp, E. O.; Thompson, H. P. G.; Day, G. M., An optimized intermolecular force field for hydrogen bonded organic molecular crystals using atomic multipole electrostatics. *Acta Crystallographica Section B-Structural Science Crystal Engineering and Materials* **2016,** *72*, 477-487.

45. 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. *Journal of Physical Chemistry* **1996,** *100* (18), 7352-7360.

46. Williams, D. E.; Cox, S. R., Nonbonded Potentials For Azahydrocarbons: the Importance of the Coulombic Interaction. *Acta Crystallographica Section B - Structural Science* **1984,** *40* (8), 404-417.

47. Cox, S. R.; Hsu, L. Y.; Williams, D. E., Nonbonded Potential Function Models for Crystalline Oxohydrocarbons. *Acta Crystallographica Section A - Crystal Physics, Diffraction, Theoretical and General Crystallography* **1981,** *37* (MAY), 293-301.

48. Hsu, L. Y.; Williams, D. E., Intermolecular Potential-Function Models for Crystalline Perchlorohydrocarbons. *Acta Crystallographica Section A - Crystal Physics, Diffraction, Theoretical and General Crystallography* **1980,** *36* (MAR), 277-281.

49. Williams, D. E.; Houpt, D. J., Fluorine Nonbonded Potential Parameters Derived From Crystalline Perfluorocarbons. *Acta Crystallographica Section B - Structural Science* **1986,** *42* (JUN), 286-295.

50. Nyman, J.; Pundyke, O. S.; Day, G. M., Accurate force fields and methods for modelling organic molecular crystals at finite temperatures. *Physical Chemistry Chemical Physics* **2016,** *18* (23), 15828-15837.

51. Habgood, M.; Price, S. L.; Portalone, G.; Irrera, S., Testing a Variety of Electronic-Structure-Based Methods for the Relative Energies of 5-Formyluracil Crystals. *Journal of Chemical Theory and Computation* **2011,** *7* (9), 2685-2688.

52. Cooper, T. G.; Hejczyk, K. E.; Jones, W.; Day, G. M., Molecular Polarization Effects on the Relative Energies of the Real and Putative Crystal Structures of Valine. *Journal of Chemical Theory and Computation* **2008,** *4* (10), 1795-1805.

53. Beran, G. J. O., Modeling Polymorphic Molecular Crystals with Electronic Structure Theory. *Chemical Reviews* **2016,** *116* (9), 5567-5613.

54. 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 Crystallographica Section B-Structural Science* **2011,** *67*, 535-551.
55. Woolley, R. G.; Sutcliffe, B. T., Molecular structure and the born—Oppenheimer approximation. *Chemical Physics Letters* **1977,** *45* (2), 393-398.
56. Kohn, W.; Sham, L. J., Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Review* **1965,** *140* (4A), A1133-A1138.
57. Hohenberg, P.; Kohn, W., Inhomogeneous Electron Gas. *Physical Review* **1964,** *136* (3B), B864-B871.
58. Perdew, J. P.; Burke, K.; Ernzerhof, M., Generalized gradient approximation made simple. *Physical Review Letters* **1996,** *77* (18), 3865-3868.
59. McKinley, Jessica L.; Beran, G. J. O., Identifying pragmatic quasi-harmonic electronic structure approaches for modeling molecular crystal thermal expansion. *Faraday Discussions* **2018**, *Advance article.*
60. Kresse, G.; Furthmuller, J., Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B* **1996,** *54* (16), 11169-11186.
61. Dovesi, R.; Orlando, R.; Erba, A.; Zicovich-Wilson, C. M.; Civalleri, B.; Casassa, S.; Maschio, L.; Ferrabone, M.; De La Pierre, M.; D'Arco, P.; Noël, Y.; Causà, M.; Rérat, M.; Kirtman, B., CRYSTAL14: A program for the ab initio investigation of crystalline solids. *International Journal of Quantum Chemistry* **2014,** *114* (19), 1287-1317.
62. Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L.; Cococcioni, M.; Dabo, I.; Dal Corso, A.; de Gironcoli, S.; Fabris, S.; Fratesi, G.; Gebauer, R.; Gerstmann, U.; Gougoussis, C.; Kokalj, A.; Lazzeri, M.; Martin-Samos, L.; Marzari, N.; Mauri, F.; Mazzarello, R.; Paolini, S.; Pasquarello, A.; Paulatto, L.; Sbraccia, C.; Scandolo, S.; Sclauzero, G.; Seitsonen, A. P.; Smogunov, A.; Umari, P.; Wentzcovitch, R. M., QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *Journal of Physics-Condensed Matter* **2009,** *21* (39), 395502.
63. Clark, S. J.; Segall, M. D.; Pickard, C. J.; Hasnip, P. J.; Probert, M. J.; Refson, K.; Payne, M. C., First principles methods using CASTEP. *Zeitschrift fur Kristallographie* **2005,** *220* (5-6), 567-570.
64. Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C., Efficient Handling of Molecular Flexibility in Lattice Energy Minimization of Organic Crystals. *Journal of Chemical Theory and Computation* **2011,** *7* (6), 1998-2016.
65. Orio, M.; Pantazis, D. A.; Neese, F., Density functional theory. *Photosynthesis Research* **2009,** *102* (2), 443-453.
66. Yang, J.; Hu, W.; Usvyat, D.; Matthews, D.; Schutz, M.; Chan, H., Ab initio determination of the lattice energy in crystalline benzene to sub-kilojoule per mole accuracy. *Science* **2014,** *345* (6197), 640-643.
67. Broqvist, P.; Alkauskas, A.; Pasquarello, A., Hybrid-functional calculations with plane-wave basis sets: Effect of singularity correction on total energies, energy eigenvalues, and defect energy levels. *Physical Review B* **2009,** *80* (8), 085114.
68. Hoja, J.; Tkatchenko, A., First-principles stability ranking of molecular crystal polymorphs with the DFT+MBD approach. *Faraday Discussions* **2018,** *Advance article*.
69. Kristyán, S.; Pulay, P., Can (semi)local density functional theory account for the London dispersion forces? *Chemical Physics Letters* **1994,** *229* (3), 175-180.
70. Grimme, S., Density functional theory with London dispersion corrections. *Wiley Interdisciplinary Reviews-Computational Molecular Science* **2011,** *1* (2), 211-228.
71. Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H., A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J Chem Phys* **2010,** *132* (15), 154104.
72. Grimme, S., Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *Journal of Computational Chemistry* **2006,** *27* (15), 1787-1799.
73. Neumann, M. A.; Perrin, M. A., Energy ranking of molecular crystals using density functional theory calculations and an empirical van der Waals correction. *Journal of Physical Chemistry B* **2005,** *109* (32), 15531-15541.
74. Becke, A. D.; Johnson, E. R., Exchange-hole dipole moment and the dispersion interaction revisited. *The Journal of Chemical Physics* **2007,** *127* (15), 154108.

75. Tkatchenko, A.; Scheffler, M., Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Physical Review Letters* **2009,** *102* (7), 073005.
76. Tkatchenko, A.; DiStasio, R. A. J.; Car, R.; Scheffler, M., Accurate and efficient method for many-body van der Waals interactions. *Physical Review Letters* **2012,** *108* (23), 236402-236402.
77. Brandenburg, J. G.; Grimme, S., Accurate Modeling of Organic Molecular Crystals by Dispersion-Corrected Density Functional Tight Binding (DFTB). *Journal of Physical Chemistry Letters* **2014,** *5* (11), 1785-1789.
78. Brandenburg, J. G.; Hochheim, M.; Bredow, T.; Grimme, S., Low-Cost Quantum Chemical Methods for Noncovalent Interactions. *Journal of Physical Chemistry Letters* **2014,** *5* (24), 4275-4284.
79. Christensen, A. S.; Kubař, T.; Cui, Q.; Elstner, M., Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications. *Chemical Reviews* **2016,** *116* (9), 5301-5337.
80. Elstner, M.; Seifert, G., Density functional tight binding. *Philosophical Transactions of the Royal Society A* **2014,** *372* (2011).
81. Brandenburg, J. G.; Potticary, J.; Sparkes, H. A.; Price, S. L.; Hall, S. R., Thermal Expansion of Carbamazepine: Systematic Crystallographic Measurements Challenge Quantum Chemical Calculations. *Journal of Physical Chemistry Letters* **2017,** *8* (17), 4319-4324.
82. Aradi, B.; Hourahine, B.; Frauenheim, T., DFTB+, a Sparse Matrix-Based Implementation of the DFTB Method. *The Journal of Physical Chemistry A* **2007,** *111* (26), 5678-5684.
83. Nyman, J.; Day, G. M., Modelling temperature-dependent properties of polymorphic organic molecular crystals. *Physical Chemistry Chemical Physics* **2016,** *18* (45), 31132-31143.
84. Neumann, M. A.; de Streek, J. V.; Fabbiani, F. P. A.; Hidber, P.; Grassmann, O., Combined crystal structure prediction and high-pressure crystallization in rational pharmaceutical polymorph screening. *Nature Communications* **2015,** *6*, 7793.
85. Dove, M. T., *Introduction to Lattice Dynamics*. Cambridge University Press: Cambridge, 1993.
86. Day, G. M.; Price, S. L.; Leslie, M., Atomistic Calculations of Phonon Frequencies and Thermodynamic Quantities for Crystals of Rigid Organic Molecules. *The Journal of Physical Chemistry B* **2003,** *107* (39), 10919-10933.
87. Togo, A.; Tanaka, I., First principles phonon calculations in materials science. *Scripta Materialia* **2015,** *108*, 1-5.
88. Reilly, A. M.; Tkatchenko, A., Understanding the role of vibrations, exact exchange, and many-body van der Waals interactions in the cohesive properties of molecular crystals. *The Journal of Chemical Physics* **2013,** *139* (2), 024705-024705.
89. Kim, S.; Orendt, A. M.; Ferraro, M. B.; Facelli, J. C., Crystal Structure Prediction of Flexible Molecules Using Parallel Genetic Algorithms with a Standard Force Field. *Journal of Computational Chemistry* **2009,** *30* (13), 1973-1985.
90. Neumann, M. A., Tailor-made force fields for crystal-structure prediction. *Journal of Physical Chemistry B* **2008,** *112* (32), 9810-9829.
91. Pickard, C. J.; Needs, R. J., Ab initio random structure searching. *Journal of Physics: Condensed Matter* **2011,** *23* (5).
92. van Eijck, B. P.; Kroon, J., Structure predictions allowing more than one molecule in the asymmetric unit. *Acta Crystallographica Section B - Structural Science* **2000,** *56*, 535-542.
93. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C., The Cambridge Structural Database. *Acta Crystallographica Section B-Structural Science Crystal Engineering and Materials* **2016,** *72*, 171-179.
94. Kendrick, J.; Stephenson, G. A.; Neumann, M. A.; Leusen, F. J. J., Crystal Structure Prediction of a Flexible Molecule of Pharmaceutical Interest with Unusual Polymorphic Behavior. *Crystal Growth & Design* **2013,** *13* (2), 581-589.
95. Habgood, M.; Sugden, I. J.; Kazantsev, A. V.; Adjiman, C. S.; Pantelides, C., Efficient Handling of Molecular Flexibility in Ab Initio Generation of Crystal Structures. *Journal of Chemical Theory and Computation* **2015,** *11* (4), 1957-1969.
96. Karamertzanis, P. G.; Pantelides, C. C., Ab initio crystal structure prediction. II. Flexible molecules. *Molecular Physics* **2007,** *105* (2-3), 273-291.
97. Sobol, I. M., *USSR Computational Mathematics and Mathematical Physics* **1967,** *7*, 86-112.
98. Sugden, I.; Adjiman, C. S.; Pantelides, C. C., Accurate and efficient representation of intramolecular energy in ab initio generation of crystal structures. I. Adaptive local approximate

models. *Acta Crystallographica Section B-Structural Science Crystal Engineering and Materials* **2016,** *72*, 864-874.
99. Price, S. L., Why don't we find more polymorphs? *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **2013,** *69*, 313-328.
100. Habgood, M.; Grau-Crespo, R.; Price, S. L., Substitutional and orientational disorder in organic crystals: a symmetry-adapted ensemble model. *Physical Chemistry Chemical Physics* **2011,** *13* (20), 9590-9600.
101. Habgood, M., Form II Caffeine: A Case Study for Confirming and Predicting Disorder in Organic Crystals. *Crystal Growth & Design* **2011,** *11* (8), 3600-3608.
102. Gavezzotti, A., A solid-state chemist's view of the crystal polymorphism of organic compounds. *Journal of Pharmaceutical Sciences* **2007,** *96* (9), 2232-2241.
103. Coles, S. J.; Threlfall, T. L.; Tizzard, G. J., The Same but Different: Isostructural Polymorphs and the Case of 3-Chloromandelic Acid. *Crystal Growth & Design* **2014,** *14* (4), 1623-1628.
104. Chisholm, J. A.; Motherwell, S., COMPACK: a program for identifying crystal structure similarity using distances. *Journal of Applied Crystallography* **2005,** *38*, 228-231.
105. Macrae, C. F.; Bruno, I. J.; Chisholm, J. A.; Edgington, P. R.; McCabe, P.; Pidcock, E.; Rodriguez-Monge, L.; Taylor, R.; van de Streek, J.; Wood, P. A., Mercury CSD 2.0 - new features for the visualization and investigation of crystal structures. *Journal of Applied Crystallography* **2008,** *41*, 466-470.
106. van de Streek, J.; Motherwell, S., Searching the Cambridge Structural Database for polymorphs. *Acta Crystallographica Section B - Structural Science* **2005,** *61*, 504-510.
107. de Gelder, R.; Wehrens, R.; Hageman, J. A., A generalized expression for the similarity of spectra: Application to powder diffraction pattern classification. *Journal of Computational Chemistry* **2001,** *22* (3), 273-289.
108. van de Streek, J., Searching the Cambridge Structural Database for the 'best' representative of each unique polymorph. *Acta Crystallographica Section B - Structural Science* **2006,** *62*, 567-579.
109. Copley, R. C. B.; Barnett, S. A.; Karamertzanis, P. G.; Harris, K. D. M.; Kariuki, B. M.; Xu, M. C.; Nickels, E. A.; Lancaster, R. W.; Price, S. L., Predictable disorder versus polymorphism in the rationalization of structural diversity: A multidisciplinary study of eniluracil. *Crystal Growth & Design* **2008,** *8* (9), 3474-3481.
110. Ballester, P. J.; Richards, W. G., Ultrafast shape recognition to search compound databases for similar molecular shapes. *J Comput Chem* **2007,** *28*.
111. Schreyer, A. M.; Blundell, T., USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. *Journal of Cheminformatics* **2012,** *4*, 12.
112. 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. *Journal of Pharmacy and Pharmacology* **2015,** *67* (6), 857-868.
113. Cole, J. C.; Groom, C. R.; Read, M. G.; Giangreco, I.; McCabe, P.; Reilly, A. M.; Shields, G. P., Generation of crystal structures using known crystal structures as analogues. *Acta Crystallographica Section B-Structural Science Crystal Engineering and Materials* **2016,** *72*, 530-541.
114. Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kessler, M.; Macrae, C. F.; McCabe, P.; Pearson, J.; Taylor, R., New software for searching the Cambridge Structural Database and visualizing crystal structures. *Acta Crystallographica Section B - Structural Science* **2002,** *58*, 389-397.
115. 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. *Journal of Chemical Information and Computer Sciences* **2004,** *44* (6), 2133-2144.
116. Taylor, R.; Cole, J.; Korb, O.; McCabe, P., Knowledge-Based Libraries for Predicting the Geometric Preferences of Druglike Molecules. *Journal of Chemical Information and Modeling* **2014,** *54* (9), 2500-2514.
117. McCabe, P.; Korb, O.; Cole, J., Kernel Density Estimation Applied to Bond Length, Bond-angle, and Torsion Angle Distributions. *Journal of Chemical Information and Modeling* **2014,** *54* (5), 1284-1288.
118. Fisher, N. I., Smoothing a sample of circular data. *Journal of Structural Geology* **1989,** *11* (6), 775-778.
119. Cole, J. C.; Korb, O.; McCabe, P.; Read, M. G.; Taylor, R., Knowledge-Based Conformer Generation Using the Cambridge Structural Database. *Journal of Chemical Information and Modeling* **2018,** *58* (3), 615-629.

Chapter 3: Successful prediction of molecule XXVI for the 6th Blind Test of Crystal Structure Prediction

3.1 Introduction

3.1.1 The CCDC Blind Tests of crystal structure prediction methods

CSP studies are rarely carried out without any previous experimental knowledge. This makes the real quality of an apparently effective CSP methodology dubious, since a user may be inclined to tweak the study to reproduce the known form as a low energy one. The Blind Tests periodically organised by the Cambridge Crystallographic Data Centre (CCDC)[1, 2] are an invaluable tool to test the credibility of CSP techniques, as well as to stimulate the development of new methodologies.[2] Their main purpose is to verify the reliability of CSP methods in predicting unknown crystal structures.

Crystallographers are asked in advance to provide some unpublished, high-quality crystal structures without disorder of molecules that meet certain criteria in terms of size, flexibility, atom type and number of molecules in the asymmetric unit (Z'). After selecting a set of molecules, the CCDC keeps the experimental crystal structures confidential. The molecular diagrams and the crystallisation conditions are released to the participating computational groups at the beginning of the challenge, together with a deadline to submit the predictions. At the end of the challenge, the predictions are evaluated and compared with the experimental forms, and a workshop is held, which results in a joint publication summarising the state-of-the-art of CSP.[2]

Six Blind Tests have been held so far. The first three, held in 1999,[3] 2001[4] and 2004,[5] were open only to invited participants, while the following three, held in 2007,[6] 2010[7] and 2014-15[2] were open to anyone who was interested in the challenge. The results of the first three challenges were of mixed quality, but in general they showed the immaturity of CSP methods, with no approach being consistently reliable; the results for the more flexible systems were particularly discouraging. The 4th Blind Test[6] saw the first substantial success, although limited to small molecules: GRACE,[8] developed by Neumann *et al*., was the first method to correctly predict all the target crystal structures. The 5th Blind Test[7] was another important step forward: two groups successfully predicted the crystal structure of molecule XX, whose size (55 atoms) and flexibility are comparable to those of small molecules in drug development.[9] These successes drastically increased industrial interest in CSP, leading to several collaborations between computational groups and pharmaceutical manufacturers.[10, 11]

Almost all successful CSP methodologies used in the Blind Tests aim to determine which crystal structures are the most thermodynamically stable.[12] A problem inherent with the Blind Tests is that there is no guarantee that the molecules are monomorphic, let alone that the target crystal structures are the most stable forms. Indeed, there have been instances of polymorphs being found during or after the completion of the Blind Tests, in some cases more stable than the target form,[3, 13] in others metastable.[14] If the target crystal structure is metastable, there is a high chance that it could be missed or ranked poorly by the participating groups. Hence, a proper assessment of the possible crystalline forms of the target molecules prior to the start of the Blind Test considerably improves the scientific conclusiveness of the challenge.

This chapter discusses how the crystal structure of one of the target molecules of the 6th Blind Test was successfully predicted, as a validation of the CSP methodology utilised throughout most of this thesis. It also illustrates some of the challenges that can be encountered when performing CSP on large and flexible molecules, which need to be addressed to make computational studies a routine component of industrial polymorph screens. The information on the experimental forms that is outlined in this chapter was not known when the CSP study was carried out; all comparisons with experimental data were performed after the release of the target crystal structures at the end on the Blind Test.

3.1.2 The 6th Blind Test

The 6th Blind Test took place between September 2014 and August 2015. The five target molecules are shown in Figure 3.1.

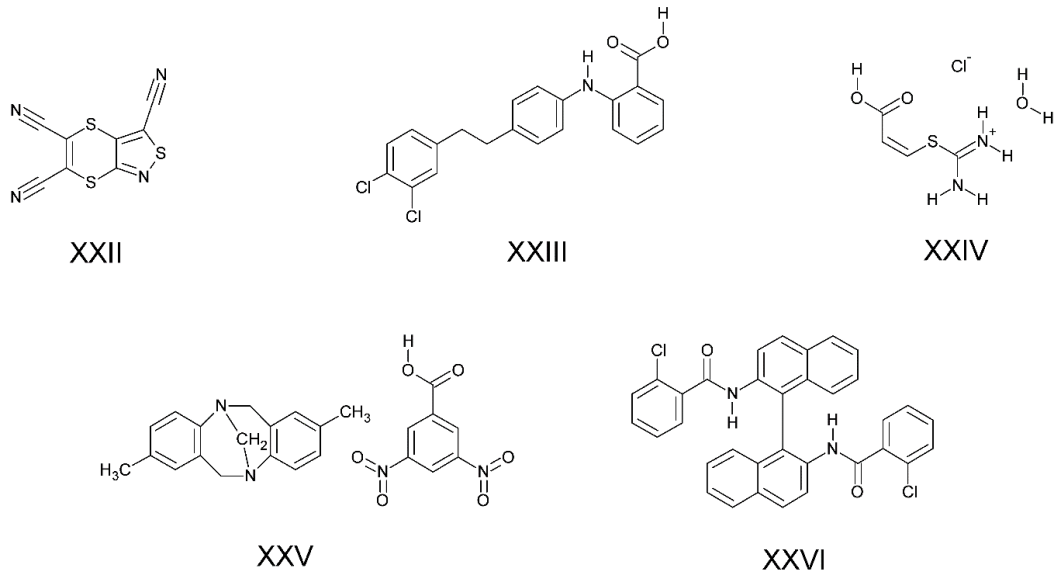


Figure 3.1: Chemical diagrams five target molecules for the 6th Blind Test of CSP.

This set of molecules reflects the progress undertaken by CSP methods, with the addition of targets with unprecedented complexities for Blind Tests, such as the first three component salt (XXIV) and the largest Blind Test molecule to date (XXVI). Particularly

interesting is moderately flexible molecule XXIII, for which participants were challenged to predict five polymorphs, two of which were Z'=2.

Acknowledging that polymorphism is a very prevalent phenomenon in organic solids, the submission criteria for this 6th Blind Test were changed compared to the previous challenges: participants were no longer invited to submit three unique predictions, but two lists of up to 100 structures, ranked in terms of a scoring function (energy, in most cases).[2] This gave participants greater flexibility, allowing the submission of predictions obtained with more than one model, for example with or without the inclusion of thermal effects. The Price group sent predictions for all five molecules, with each group member dedicated to a different target. A set of 1,000 putative crystal structures for each molecule was also sent for optimisation and ranking by other groups developing accurate periodic DFT-D methods, but who did not have the tools for generating candidate crystalline forms. This chapter is focused on the molecule I worked on: large and flexible molecule XXVI. Details about the CSP studies of the other molecules can be found in the Blind Test publication.[2]

3.1.3 Molecule XXVI

3.1.3.1 Molecule and crystallisation conditions

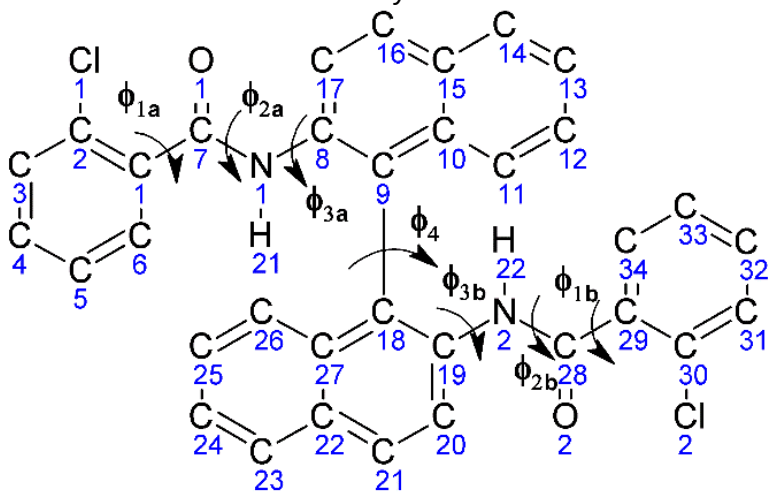


Figure 3.2: Chemical diagram of molecule XXVI. The arrows define the torsion angles considered as the flexible in the CSP study. Φ_{1a} and Φ_{1b} (C2-C1-C7-O1 and C34-C29-28-O2) are 0° in the diagram above, Φ_{2a} and Φ_{2b} (O1-C7-N1-H21 and O2-C28-N2-H22) are 180°, Φ_{3a} and Φ_{3b} (H21-N1-C8-C17 and H22-N2-C19-C20) are 180° and Φ_4 (C10-C9-C18-C19) is 180°.

N,N'-([1,1'-Binaphthalene]-2,2'-diyl)bis(2-chlorobenzamide) ($C_{34}H_{22}C_{12}N_2O_2$) has two halves (C1-C17 and C18-C34) that are related by symmetry. Although 1,1'-binaphthyl compounds can be axially chiral, this molecule was crystallised as a racemic mixture.[2] Hence, it was not possible to assume it could only crystallise in the chiral space groups. It was crystallised by slow evaporation from a 1:1 mixture of

hexanes and dichloromethane; this information was not used as part of the CSP process by any participant.[2]

The main challenges for performing CSP on molecule XXVI were the large size, the numerous flexible rotatable bonds,[15] which drastically increase the search space, and the bulky congested shape that can hinder the formation of directional intermolecular interactions.

3.1.3.2 The experimental forms

The only fully-solved single-component crystal structure of molecule XXVI was the target of this Blind Test study. This crystal structure is stable at room temperature, has one molecule in the asymmetric unit cell and it is in the triclinic $P\bar{1}$ space group; it was later deposited in the CSD[1] with refcode XAFQIH.[2] A polymorph screen performed by Johnson Matthey (Pharmorphix) revealed that the only solved crystal structure (form 1) undergoes a phase transition to another crystalline polymorph, referred to as form 11, taking place at 428 K; form 11 was characterised through high-resolution powder diffraction, but the crystal structure has not been solved yet.[2] Nine solvates of molecule XXVI, referred to as forms 2-10, were also found in the polymorph screen.[2] These solvates are currently not deposited in the CSD.

3.2 Methods

3.2.1 Analysis of conformational flexibility

3.2.1.1 Torsion angle scans

The first step of this CSP study was the determination of a plausible range of solid-state conformations. The seven torsion angles shown in Figure 3.2 were identified as flexible through chemical intuition and a set of isolated molecule optimisations. To determine the range of values that each torsion angle could realistically take in the solid-state, the conformational energy penalty for varying torsion angles Φ_{1a}, Φ_{2a}, Φ_{3a} and Φ_4 was calculated via constrained 1-dimensional angle scans performed with Gaussian 03[16] at the PBE0 6-31G(d,p) level of theory. In each angle scan, one of these torsion angles was fixed at one value, while the other conformational degrees of freedom (CDFs) were relaxed to minimise conformational energy. Once the conformation was fully optimised, the same torsion angle was fixed at another value and the rest of the molecule was relaxed; the process was repeated until the entire range of the scan was covered.

The ranges and size steps of the scans were chosen from chemical intuition and an analysis of possible CSD values (see Section 3.2.1.2). Angles Φ_{1b}, Φ_{2b} and Φ_{3b} were not explicitly scanned since this would have given the same results as their symmetry-related counterparts; the energy profile of angle Φ_4 was assumed to be symmetric about

0° and so it was only scanned in the -180 to 0° range. Each scan was initiated from the lowest energy isolated-molecule conformation at the PBE0 6-31G(d,p) level of theory, since this has been shown to be the most suitable starting point for performing angle scans on flexible molecules.[17]

3.2.1.2 CSD surveys

The relaxed angle scans were complemented by a CSD survey of all molecules containing the fragments shown in Figure 3.3.

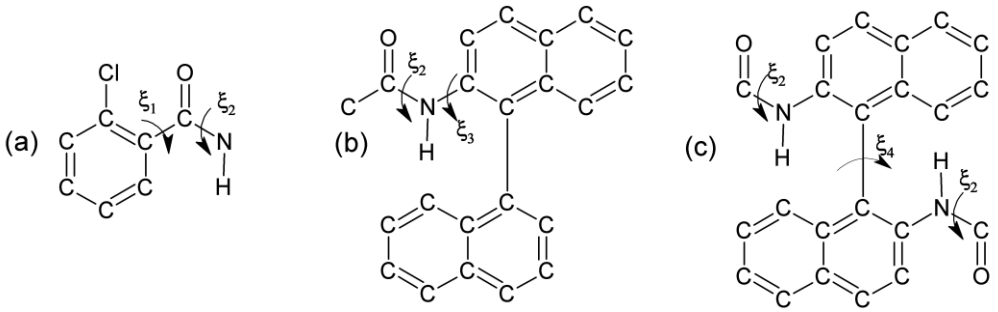


Figure 3.3: Search fragments used in Conquest to perform the CSD surveys. Angles ξ_1, ξ_3 and ξ_4 were considered analogues of Φ_{1a} and Φ_{1b}, Φ_{3a} and Φ_{3b} and Φ_4 respectively. Also all these fragments contain angle ξ_2 that is an analogue of torsion angles Φ_{2a} and Φ_{2b}.

As already mentioned in Chapter 2, CSD surveys can integrate the analysis of conformational flexibility, by highlighting any weakness in the chosen level of theory[17] and determining the solid-state geometrical preference.[18] The experimentally-determined crystal structures of molecules containing the fragments shown in Figure 3.3, which mimic the functional groups of molecule XXVI, were retrieved with Conquest.[19] When more than one entry belonged to the same refcode family, the structures were checked manually. If they were redeterminations, only the one with the lowest R-factor was considered, while polymorphs were all kept.

The fragment in Figure 3.3a was found in 99 unique crystal structures, the one in Figure 3.3b in 67 and the one in Figure 3.3c in 33. The values taken by ξ_1, ξ_2, ξ_3 and ξ_4 in the CSD-retrieved crystal structures were extracted with Mercury.[20]

The results of the angle scans and of the CSD surveys are summarised in Figure 3.4, where the values taken by each dihedral in the conformation of target experimental form of molecule XXVI are also indicated (see Appendix Table 3.1 for the specific values).

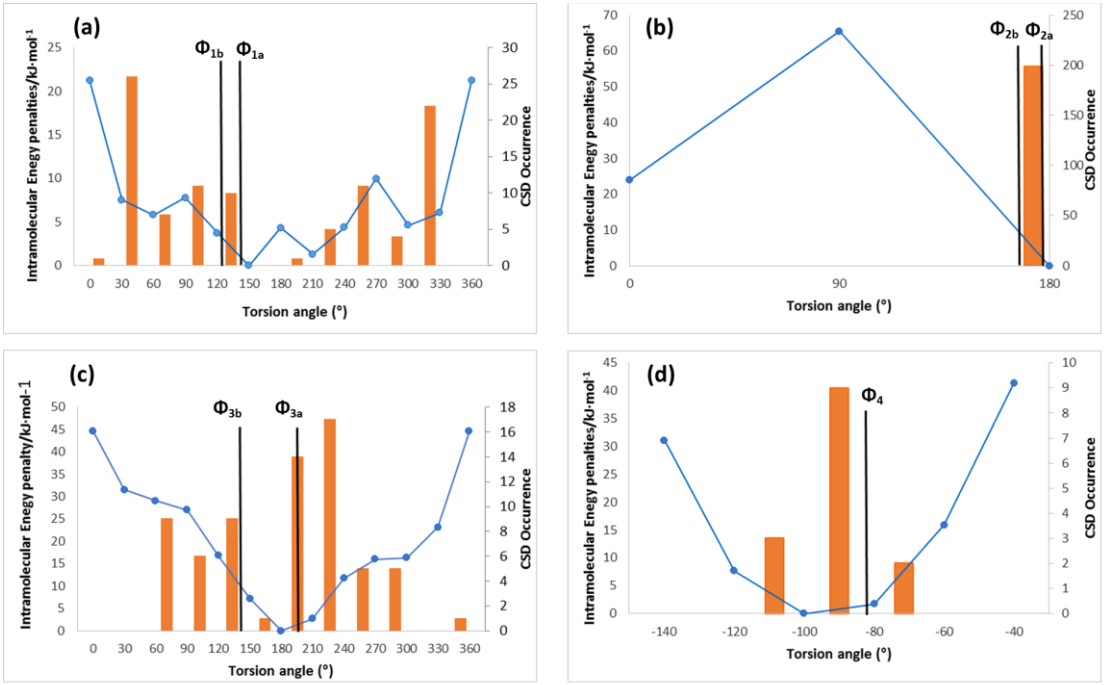


Figure 3.4: Results of the isolated-molecule scans of torsion angle a) Φ_{1a} from 0° to 360° in 30° steps; this is also valid for Φ_{1b} b) Φ_{2a} from 0° to 180° in 90° steps, this is also valid for Φ_{2b} c) Φ_{3a} from 0° to 360° in 30° steps; this is also valid for Φ_{3b} d) Φ_4 from -40° to -140° in 20° steps. The blue points indicate the relative conformational energy when the torsion angle took a certain value; at each point, all the CDFs were relaxed with the exception of the scanned torsion angle. All the calculations were performed at the PBE0 6-31G(d,p) level of theory starting from the PBE0 6-31G(d,p) optimised global minimum gas-phase conformer. The orange bars indicate the frequency of each value in the CSD. The black lines indicate the values taken by the torsion angles in the conformation of the target experimental crystal structure.

The constrained angle scans and the CSD surveys provided consistent results. With the exception of Figure 3.4a, it is clear that low-energy regions are the most populated in the CSD, and only a few torsion angle values with a calculated conformational energy penalty exceeding 20 kJ·mol^{-1} can be found in solid-state conformations. The main inconsistency is for $\Phi_{1a\text{-}1b}$, where few solid-state conformations can be found in the apparently low-energy region between 150 and 210°; this may be due to the specific characteristics and interactions of molecule XXVI. It is interesting to note that the scans for angles Φ_{1a} and Φ_{3a} are not exactly symmetric around 180°: this is due to the interactions between the scanned torsion angles and the rest of the molecule, including the other symmetric half.

In summary, Figure 3.4 shows that the isolated-molecule scans could be confidently used to limit the conformational search space in this CSP study. The effectiveness of this initial analysis is confirmed by all torsion angles of the target crystalline conformation taking values with calculated energy penalties smaller than 20 kJ·mol^{-1} and found in at least some crystal structures of similar molecules.

3.2.2 Crystal structure generation

After analysing the torsion angle scans and the CSD distributions shown in Figure 3.4, a flexible crystal structure search was performed, with only five torsion angles ($\Phi_{1a,}$ Φ_{1b}, $\Phi_{3a,}$ Φ_{3b} and Φ_4) treated as explicitly flexible CDFs. This was done because the scan of amide torsion angles Φ_{2a} and Φ_{2b} showed the presence of a large energy barrier for moving away from the *trans* configuration (*i.e.* 180°), with the competitive *cis* configuration (*i.e.* 0°) being approximately 24 $kJ \cdot mol^{-1}$ higher in energy. Furthermore, all the retrieved solid-state conformations had the amide group (torsion angle ξ_2 in Figure 3.3a-c) in the *trans* configuration. Hence, it was decided to constrain both amide torsion angles Φ_{2a} and Φ_{2b} to their values in the isolated-molecule global minimum in conformational energy (181.21° and 180.59° respectively) during the search, and no crystal structure was generated with *cis* amides. However, both amide torsion angles were treated as independent CDFs in the final refinement stage (see Chapter 3.2.3) to allow them to take the most suitable configuration in the crystalline environment.

The search was performed with CrystalPredictor 1.6, which estimates ΔE_{intra} with a crude model based on the interpolation of a grid of *ab initio* calculated intramolecular energies for each CDF treated as an explicit search variable (for details of the methodology, see Chapter 2.4.1.2).[21] The other CDFs, including Φ_{2a} and Φ_{2b}, were constrained at the values in the isolated-molecule global minimum in conformational energy. This version of CrystalPredictor was used because an immense multidimensional grid would have been required to produce the local approximate models (LAMs) utilised by CrystalPredictor 2[22] for estimating ΔE_{intra}, drastically increasing the computational expense, as explained in Chapter 2.4.1.2. On the other hand, CrystalPredictor 1.6 allows the grids to be constructed more cheaply by dividing a molecule into several appropriate surrogate molecule each containing a subset of the torsion angles. Note that this assumption is realistic only for torsion angles that define the positions of groups that do not interact strongly with one another.[24]

Molecule XXVI was broken down into three surrogate molecules, shown in Figure 3.5, and two grids of ΔE_{intra} values were produced: one for the central portion of the molecule, including angles Φ_{3a}, Φ_{3b} and Φ_4, and one for the two identical edge portions containing angles Φ_{1a} and Φ_{1b}. Methyl groups were added at the extremities of each surrogate molecule to avoid the presence of unphysical free bonds.

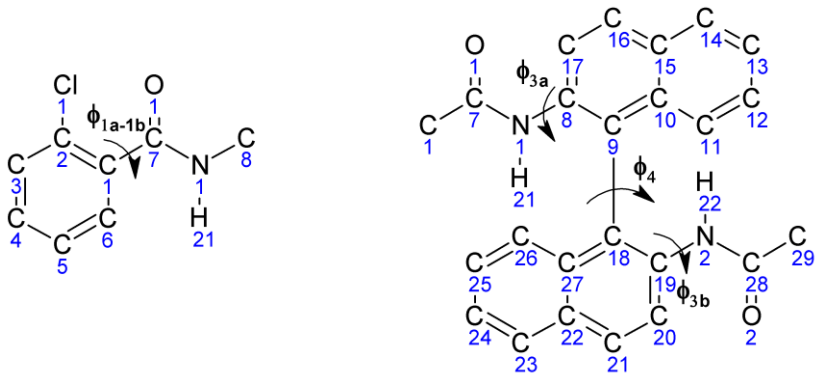


Figure 3.5: Surrogate molecules used to calculate the ΔE_{intra} grids of (left) Φ_{1a} and Φ_{1b} and (right) Φ_{3a}, Φ_{3b} and Φ_4 of molecule XXVI.

The good agreement between the energy penalty for varying the torsion angles in the surrogate molecules and in the whole molecule (see Appendix Figure 3.1) indicates that this was a sensible approximation.

The two grids for each surrogate molecule were calculated with Gaussian 03 at the PBE0 6-31G(d,p) level of theory. The grid ranges shown in Table 3.1 were selected from the analysis of the angle scans and the CSD surveys (see Figure 3.4). The values of the torsion angles in the conformation of the target experimental crystal structures (see Figure 3.4 and Appendix Table 3.2) are all within the ranges covered by the grids.

Table 3.1: Dimensionality of the ΔE_{intra} grids used to perform the crystal structure search with CrystalPredictor. The grids were calculated from the surrogate molecules in Figure 3.5.

Central surrogate molecule					
Torsion angle label	**Torsion Angle Definition**	**Minimum of search range/°**	**Maximum of search range/°**	**Step Size/°**	**Number of grid points**
Φ_{3a}	H21-N1-C8-C17	120	300	20	10
Φ_{3b}	H22-N2-C19-C20	120	300	20	10
Φ_4	C19-C18-C9-C8	-130	-70	20	4
				Number of Grid Points	400
Edge surrogate molecule					
Torsion angle label	**Torsion Angle Definition**	**Minimum of search range/°**	**Maximum of search range/°**	**Step Size/°**	**Number of grid points**
Φ_{1a} and Φ_{1b}	C2-C1-C7-O1 and C30-C29-C28-O2	20	340	20	17

In the search, ΔE_{intra} was calculated from the grids, while U_{inter} was modelled as the sum of an electrostatic component from fixed point charges calculated at the PBE0 6-31G(d,p) level of theory on the isolated-molecule global minimum conformer and a repulsion-dispersion component calculated with the empirically-fitted FIT potential.[25] A total of 1,000,000 crystal structures were generated in the 59 most common space groups, listed in Appendix Table 3.1. Although the challenge stated that the molecule could have crystallised in Z'=2, it was decided to limit the study to structures with only one molecule in the asymmetric unit to limit the overall computational cost. This seemed to be a sensible assumption since molecule XXVI does not possess the characteristics

that correlate with a higher likelihood of crystallisation with Z'>1, which were determined in a recent survey of the crystal structures in the CSD with more than one molecule in the asymmetric unit (~10% of the total), as it is large, flexible and not crystallised as homo-chiral.[26]

3.2.3 Refinement of the generated crystal structures

The final stage of this CSP study was the optimisation and re-ranking of the most promising generated crystal structures. The 9,400 generated structures with E_{latt} up to 40 kJ·mol^{-1} of the global minimum in CrystalPredictor energy were taken to this final stage. A large energy window was necessary because of the simple and relatively inaccurate model used in the search. The flexible nature of molecule XXVI meant that both ΔE_{intra} and U_{inter} had to be optimised for each crystal structure, and CrystalOptimizer[27] (see Chapter 2.4.2.2) was used to carry out the all the E_{latt} minimisations. However, in order to limit the computational cost, not all the 9,400 generated crystal structures were fully optimised, but a hierarchal approach was used. Firstly, an intermediate optimisation and re-ranking step was performed with a single iteration of CrystalOptimizer, and only the most promising candidates were then fully optimised. This same approach has been used in other successful CSP studies.[11, 28]

3.2.3.1 Intermediate optimisation of the generated crystal structures

A single-iteration of CrystalOptimizer does not produce fully converged crystal structures, but improves their geometries and the energy ranking because of the greater accuracy of the CrystalOptimizer Ψ_{mol} model compared to the CrystalPredictor one.[11]

All the 9,400 generated crystal structures within 40 kJ·mol^{-1} of the global minimum in CrystalPredictor energy underwent this intermediate optimisation step. ΔE_{intra} was improved optimising the molecular conformation with Gaussian03 at the PBE0 6-31G(d,p) level of theory as a function of the seven torsion angles in Figure 3.2, which were treated as independent CDFs. U_{inter} was also improved by optimising the intermolecular interactions with a more accurate distributed multipoles electrostatic model with GDMA 2.2[29] and DMACRYS (see Chapter 2.4.2.1).[30] All the electronic-structure calculations performed in this intermediate step to calculate ΔE_{intra} and the multipoles were stored into LAMs databases to be re-utilised for further CrystalOptimizer optimisations, reducing in this way the computational cost of the successive CSP stage.[23, 27] When all the single iterations of CrystalOptimizer were completed, duplicates were removed with COMPACK.[31] Crystal structures were considered as duplicates if they had an energy difference smaller than 2.5 kJ·mol^{-1}, a density difference smaller than 0.05 g·cm^{-3}, and if it was possible to overlay 20/20

molecules, with 20% distance and 20° angle tolerances, with a root mean square deviation ($RMSD_{20}$) smaller than 0.65 Å.

3.2.3.2 Final optimisation of the most promising crystal structures

The 1,322 unique crystal structures within 30 $kJ \cdot mol^{-1}$ of the global minimum after the single iteration of CrystalOptimizer were taken forward. All these structures were fully optimised with CrystalOptimizer at the PBE0 6-31G(d,p) level of theory; the seven torsion angles in Figure 3.2 were once again treated as independent CDFs, and were optimised under the influence of packing forces. A few structures were found not to be true E_{latt} minima, and their symmetry was reduced to lower energy Z'=2 structures. The fully optimised crystal structures were finally clustered with COMPACK to remove any duplicate. This time, crystal structures were considered as duplicates if they had an energy difference smaller than 2.85 $kJ \cdot mol^{-1}$, a density difference smaller than 0.05 $g \cdot cm^{-3}$, and if it was possible to overlay 30/30 molecules, with 20% distance and 20° angle tolerances, with a root mean square deviation ($RMSD_{30}$) smaller than 0.65 Å.

3.2.4 Estimate of the effect of polarisation on lattice energies and calculation of free energies at 298 K

The final step of the CSP study was the estimate of the effect of polarisation and temperature on the energies of the predicted crystal structures. The effect of polarisation on E_{latt} was estimated by re-calculating ΔE_{intra} and the charge density of the fully optimised crystal structures in a PCM with ε=3 (see Chapter 2.3.1.2.4), a value typical of organic solids, and then performing a rigid-body optimisation of the intermolecular interactions with DMACRYS[32]. The same level of theory, PBE0 6-31G(d,p), and repulsion-dispersion potential, FIT, as for the CrystalOptimizer optimisations were used. Temperature effects were then estimated with DMACRYS by calculating the vibrational component (F_{vib}) of the Helmholtz free energy (A) at 298 K from the rigid-body k=0 phonon modes (see Chapter 2.3.3.1).[30, 33]

3.2.5 The two submitted lists

Each group taking part in this Blind Test challenge was allowed to submit two lists containing up to 100 crystal structures. For molecule XXVI, the first list (see Appendix Table 3.3) contained the 100 lowest energy crystal structures after the full optimisations with CrystalOptimizer. This list covered an energy range of approximately 16.1 $kJ \cdot mol^{-1}$, larger than the typical polymorphic energy range of 10 $kJ \cdot mol^{-1}$.[34, 35]

The second list (see Appendix Table 3.4) was constituted by the 100 structures that were considered as the most likely to crystallise at room temperature. They were ranked on the Helmholtz free energy at 298 K, whose E_{latt} component (see Equation

2.37) was calculated with the PCM. Starting from the bottom of the energy ranking, structures that upon human judgment looked too similar to nucleate and grow independently without transforming into lower-energy forms were removed. This process continued until 100 individual crystal structures were found, which covered an energy range of approximately 22.2 kJ·mol^{-1}.

3.3 Results and Discussion

Both submitted lists contained a close match to the target experimental crystal structure of molecule XXVI.[2]

3.3.1 Crystal structure search

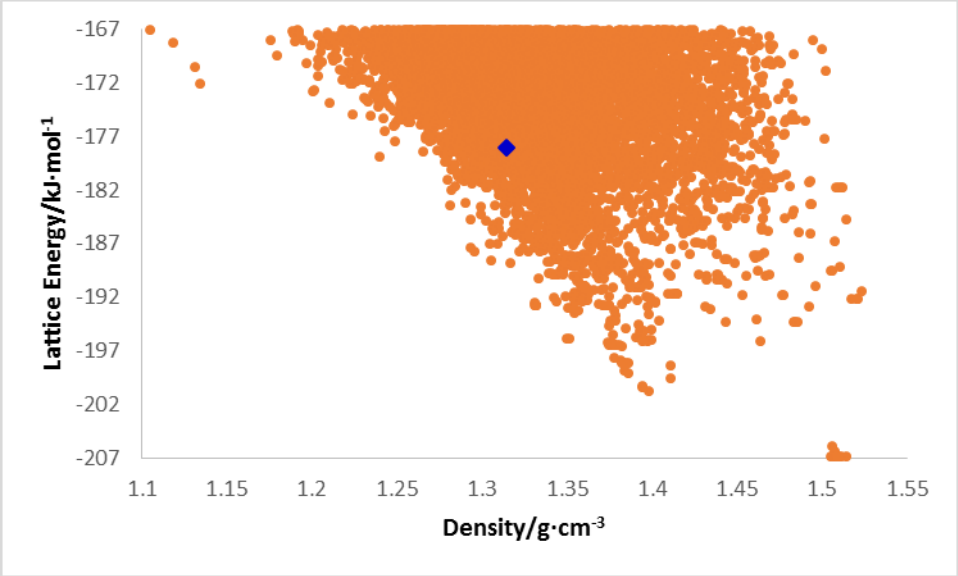


Figure 3.6: Lattice energy vs density plot obtained after the search with CrystalPredictor 1.6. ΔE_{intra} was calculated from the *ab initio* grids, while U_{inter} was modelled with the atomic point charges and the FIT potential. The structure that ended up matching the experimental form is indicated in blue. Each point on the plot corresponds to a separate crystal structure.

The computer-generated structures were named according to their ranking after the CrystalPredictor search. The structure that ended up matching the experimental form is structure 1600, *i.e.* the structure that was ranked 1600th at the search stage.

Treating the most flexible torsion angles as flexible in the search was key to the success of this CSP study. Rigid searches performed only on the isolated-molecule global minimum or on a set of local minima in conformational energy would have failed to generate crystal structures that could have optimised to the target experimental form. Figure 3.7 shows an overlay between the experimental conformation and the closest isolated-molecule local minimum conformer optimised at a PBE0 6-31G(d,p) level of theory. The overlay is poor, with an $RMSD_1$ of 0.624 Å and a noticeable difference in the orientation of the edge groups.

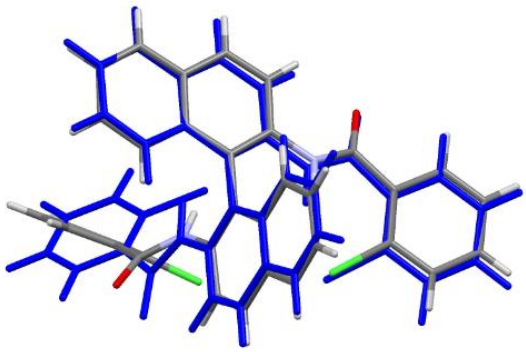


Figure 3.7: Overlay of the conformation in the experimental crystal structure of molecule XXVI (coloured by elements) and the closest gas-phase optimised conformer (in blue). The $RMSD_1$ calculated with the Crystal Packing Similarity tool is 0.624 Å.

The experimental conformation has one amide group distorted from planarity with the naphthalene group, which Figure 3.4c shows being a relatively high-energy configuration. This illustrates how improved intermolecular interactions can distort a crystalline conformation from optimised gas-phase geometries.[35, 36] The selection of the search space for the flexible search (see Table 3.1) was effective: both amide groups are in their *trans* configuration, and the experimental values of the five torsion angles that were treated as independent variables in the search (see Figure 3.4 and Appendix Table 3.2) are all within the grid ranges.

The CrystalPredictor energies were poor, as shown by the ranking of structure 1600, ~29 $kJ \cdot mol^{-1}$ above the global minimum. Furthermore, an overlay of search-generated structure 1600 and the target experimental form performed with the Crystal Packing Similarity tool[31] (see Chapter 2.5.1) with its standard settings only matches 9/15 molecules. Nonetheless, the conformations are much more similar, as shown in Figure 3.8, proving how vital the correct treatment of molecular flexibility in the search was for the success of this CSP study.

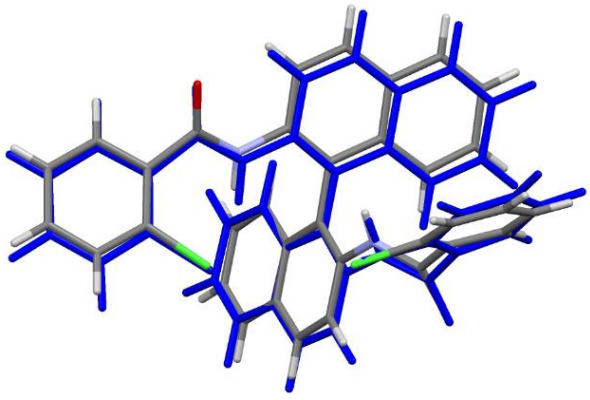


Figure 3.8: Overlay of the conformation of the experimental crystal structure of molecule XXVI (coloured by elements) and the conformation of search-generated structure 1600 (in blue). The $RMSD_1$ calculated with the Crystal Packing Similarity tool is 0.282 Å.

3.3.2 Refinement of the generated crystal structures

3.3.2.1 Intermediate optimisation of the generated crystal structures

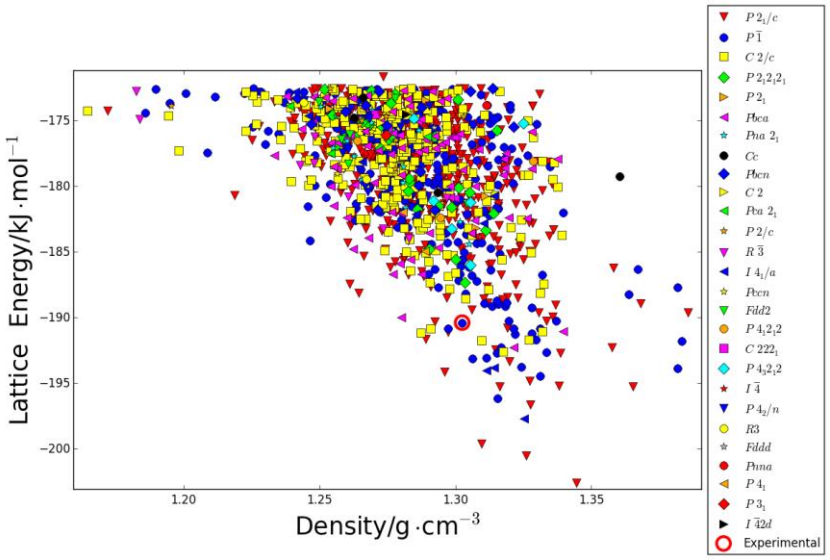


Figure 3.9: Lattice energy vs density plot obtained after the intermediate optimisations with a single-iteration of CrystalOptimizer. Each point on the plot corresponds to a separate crystal structure, labelled according to its space group. The structure that ended up matching the experimental form is indicated.

The intermediate optimisation with a single-iteration of CrystalOptimizer improved the energies and geometries substantially, as expected given the more accurate and expensive energy model. The energy ranking was drastically modified, and the global minimum in Figure 3.9 was ranked 675th by CrystalPredictor. Structure 1600 came down to the 45th place, ~12 kJ·mol^{-1} less stable than the global minimum. Its geometry was also improved: after the intermediate optimisation structure 1600 had a 15/15 molecule match with the target experimental crystal structure, with an $RMSD_{15}$ of just 0.422 Å. The conformation was also improved by the single iteration of CrystalOptimizer, with an $RMSD_1$ with its experimental counterpart of 0.206 Å.

This intermediate step was key to the success of the CSP study, since it allowed a drastic reduction in the number of structures requiring full optimisations and provided a more accurate starting point for further calculations. Although not exactly cheap (see Table 3.3), it made the CSP study feasible within its time and resource constraints. This underlines how important an intermediate step bridging the gap between the search and the final optimisation can be for screening several thousands of computer-generated crystal structures, a topic explored more in depth in Chapter 7.

3.3.3.2 The two submitted lists

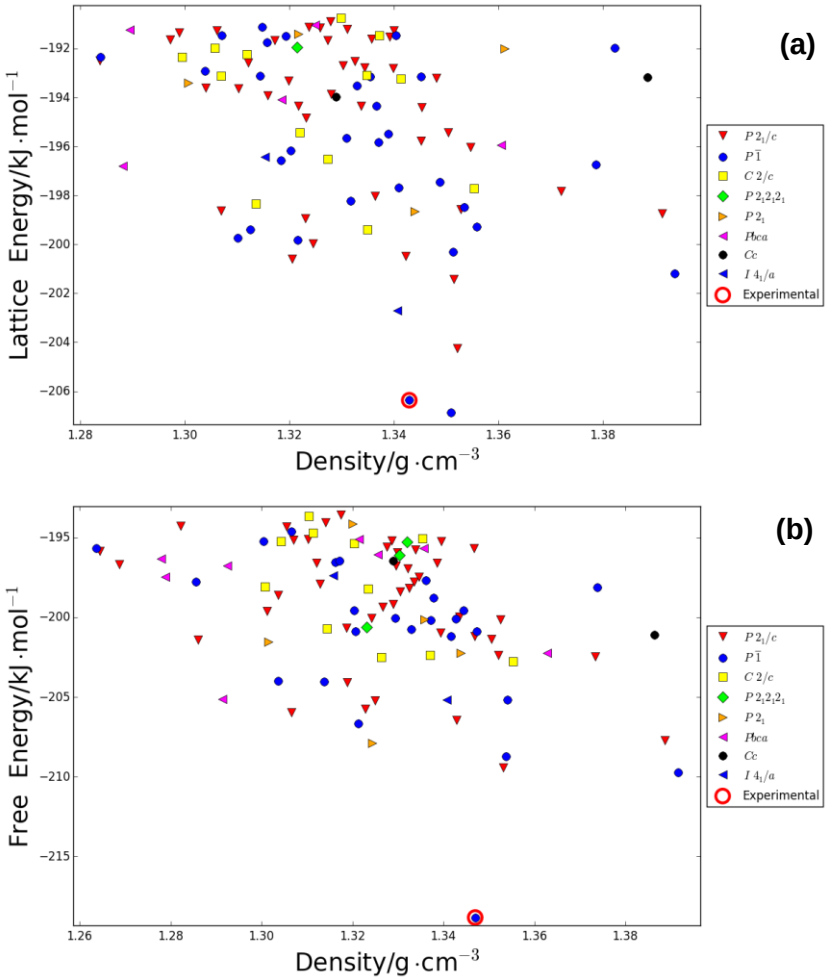


Figure 3.10: Lattice energy vs density plots showing (a) the 100 lowest energy crystal structures fully optimised with CrystalOptimizer submitted as a first list of predictions (b) the 100 lowest energy crystal structures after a rigid-body optimisation in a PCM with ε=3 and with the addition of the vibrational component to Helmholtz free energy at 298 K submitted as a second list of predictions. See Appendix Tables 3.3-3.4 for more details. Each point on the plots corresponds to a separate crystal structure, labelled according to its space group. The structures matching the target experimental form are indicated.

A total of 1,322 unique crystal structures were fully optimised with CrystalOptimizer, and the ones in Figure 3.10a were submitted as the first list of predictions. The re-ranking was once again drastic. Structure 1600 ranked 2nd, only ~0.5 kJ·mol^{-1} above the global minimum, structure 3525. A high-quality 15/15 molecule overlay with the target experimental crystal structure was achieved, with an $RMSD_{15}$ of 0.276 Å and an $RMSD_1$ of 0.126 Å. The importance of allowing the amide torsion angles to respond to packing forces in the final optimisation is outlined by Φ_{2b} taking a value of 167.4°, as opposed to 180.6° in the input conformation for the search. This illustrates the importance of carefully analysing the conformational space before a search is performed, and that CSD analyses can be an effective tool to guide the process of determining the most relevant conformations.

The crystal energy landscape obtained after estimating the effect of polarisation and temperature (Figure 3.10b) is remarkably different. Structure 1600 was now clearly the most favourable form in terms of free energy, with a ~8.3 kJ·mol^{-1} gap with the second most competitive one. Most of the re-ranking was due the inclusion of polarisation, which was the main stabilising factor for structure 1600. This stabilisation was mainly due to the presence of an intermolecular hydrogen bond in structure 1600: when calculating the molecular wave-function in a PCM, the charge distribution of the molecule polarises the dielectric continuum, which in turn induces a change in the molecular electron density compared to the gas-phase in a way that stabilises strong intermolecular electrostatic interactions (e.g. strengthening the dipole moments, which reinforces hydrogen bonds);[37] the same phenomenon occurred for other low-energy predicted crystal structures (see Chapter 3.3.4.2). The rigid-body vibrational contribution to free energy was much smaller, and the low-energy crystal structures had small variations in terms of F_{vib}, usually in the order of 2-4 kJ·mol^{-1}, indicating that E_{latt} differences dominated the free energy ranking.[33, 38] Structure 1600 in the second list is nearly identical to its counterpart in the first list, with an $RMSD_{15}$ of just 0.019 Å.

Although both crystal energy landscapes contain the target experimental form as a low-energy crystal structure, they provide very different insights in terms of the possible polymorphism of molecule XXVI and show how sensitive energy differences are on theoretical models. While Figure 3.10a suggests that the known form is one of the most promising structures within several competitive putative polymorphs (PPMs), Figure 3.10b shows a typical monomorphic crystal energy landscape,[39] which indicates that the known crystal structure is much more stable than any competitor. The monomorphic landscape in Figure 3.10b seems to be refuted by molecule XXVI undergoing a high-temperature phase transition to another polymorph (form 11) at 428 K. The crystal structure of form 11 is currently unknown. If form 11 were a different crystal structure from form 1 and found as a separate minimum in this study, then the energy gap in Figure 3.10b would be incorrect, since the vibrational component of free energy is unlikely to cause a sufficient re-ranking even at high temperatures[33] for any other predicted crystal structures to become more stable. However, these calculations of the free energy did not account for thermal expansion[40, 41] and/or the coupling of lattice and molecular vibrations[42] (see Chapter 2.3.3 for details). On the other hand, if form 11 were a distinct polymorph missed in this study or outside of its boundaries (for example with Z'>1), then it would mean the CSP search was not complete enough. The energy of form 11 would have to be calculated with both models to understand which one provides a more accurate picture of the solid-state behaviour of molecule XXVI. It is also possible that the phase transition from form 1 to form 11 is associated with no significant change in crystal structure, meaning that form 11 is an isostructural polymorph of the known form[43] (a topic

discussed in Chapter 8); in this case, CSP methods would not be capable to predict form 11, and the existence of this polymorph would not disprove the monomorphic energy landscape in Figure 3.10b. Finally, it is unlikely that form 11 is higher symmetry phase averaging over lower symmetry E_{latt} minima, since both crystal energy landscapes do not reveal the presence of groups of crystal structures with similar energies and common motifs.

3.3.3 Analysis of the generated crystal structures

3.3.3.1 Reproduction of the experimental form

The structural and crystallographic parameters of the experimental form and structure 1600 in both lists are compared in Table 3.2. The overlay between the experimental form and predicted structure 1600 optimised with the model used to compile the first list is shown in Figure 3.11.

Table 3.2: Comparison between the crystallographic and structural parameters in the experimental crystal structure of XXVI and in predicted structure 1600.

Structure	Density /g·cm^3	Packing coefficient/ %	Space group	Z'	a/Å	b/Å	c/Å	α/°	β/°	γ/°
XAFQIH	1.346	65.5	$P\bar{1}$	1	10.40	11.03	14.18	76.83	73.33	63.47
1600 (1st list)	1.343	67.4	$P\bar{1}$	1	10.26	11.17	14.23	78.54	73.53	62.86
1600 (2nd list)	1.347	67.7	$P\bar{1}$	1	10.26	11.17	14.20	78.59	73.40	62.87

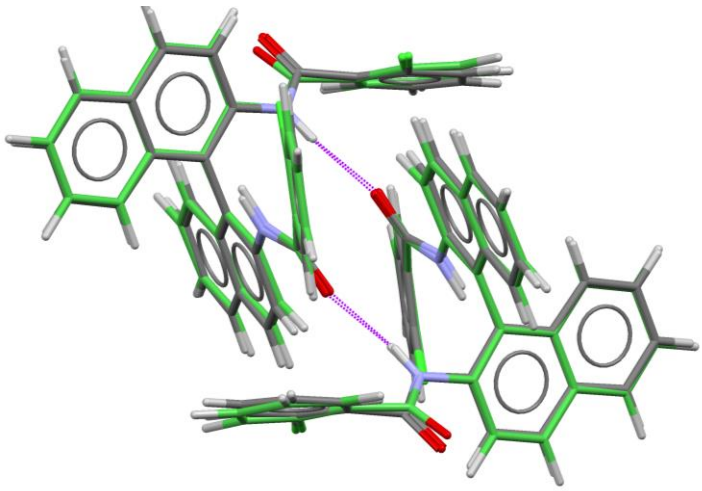


Figure 3.11: Overlay between the hydrogen bonded dimer in the experimental crystal structure of molecule XXVI (coloured by elements) and predicted structure 1600 in the first list (in green). The hydrogen bonds are coloured in purple. The $RMSD_{15}$ for the 15/15 overlay is 0.276 Å.

Table 3.2 shows an excellent agreement between the structural and crystallographic parameters of the experimental crystal structure and the forms predicted with both models, well within the margin of error of the CSP methodology. This is confirmed by the good match shown in Figure 3.11. Although the experimental form is denser than structure 1600 in the first list, it has a smaller calculated packing coefficient.

This shows the limitation of the grid-based method used to estimate the packing coefficient of crystal structures.

This analysis confirms that CrystalOptimizer was accurate at reproducing the target experimental crystal structure and at ranking it relative to its predicted competitors. The success of this CSP study was due to the quality of the CrystalOptimizer Ψ_{mol} model, which accurately balanced intra- and intermolecular interactions, even if only seven CDFs were explicitly optimised under the influence of packing forces. However, the generation of a similar enough crystal structure in the search was equally as fundamental.

3.3.3.2 Other competitive low-energy crystal structures

Although only one experimental single-component form is known, an analysis of the other predicted low-energy crystal structures of molecule XXVI reveals interesting information about the crystallisation possibilities of this molecule. Beside structure 1600, 27 other predicted crystal structures could be found within the typical polymorphic energy range of 10 $kJ \cdot mol^{-1}$ in the first list and 3 in the second list. Only 5 of the structures in the first list and 6 in the second list had their symmetry reduced to Z'=2, the lowest energy one being structure 4201, ranked 18th in the first list and 19th in the second list.

XXVI seems to struggle to pack well with itself: most low-energy predicted crystal structures have packing coefficients of 67-68% and even the densest PPMs do not exceed 70.6%. This is probably due to the awkward shape of the molecule, which hinders the formation of dense packings. The polymorph screen that found nine solvates (forms 2-10) seems to confirm this hypothesis: the presence of many solvated forms is often associated with the inability of a molecule to pack densely.[2, 44]

Conformational flexibility is also a clear feature of the low-energy forms. There is no dominating conformation: using the criteria proposed by Cruz Cabeza and Bernstein for determining whether two crystal structures are related by conformational change (see Chapter 2.3.1.1),[36] the 100 crystal structures in the first and second list can be grouped into 40 and 44 clusters of conformational polymorphs respectively.

Hydrogen bonds are a fundamental feature of organic crystals, when hydrogen bond donor and acceptors are present. Figure 3.12 shows which structures in the two lists form intra or intermolecular hydrogen bonds.

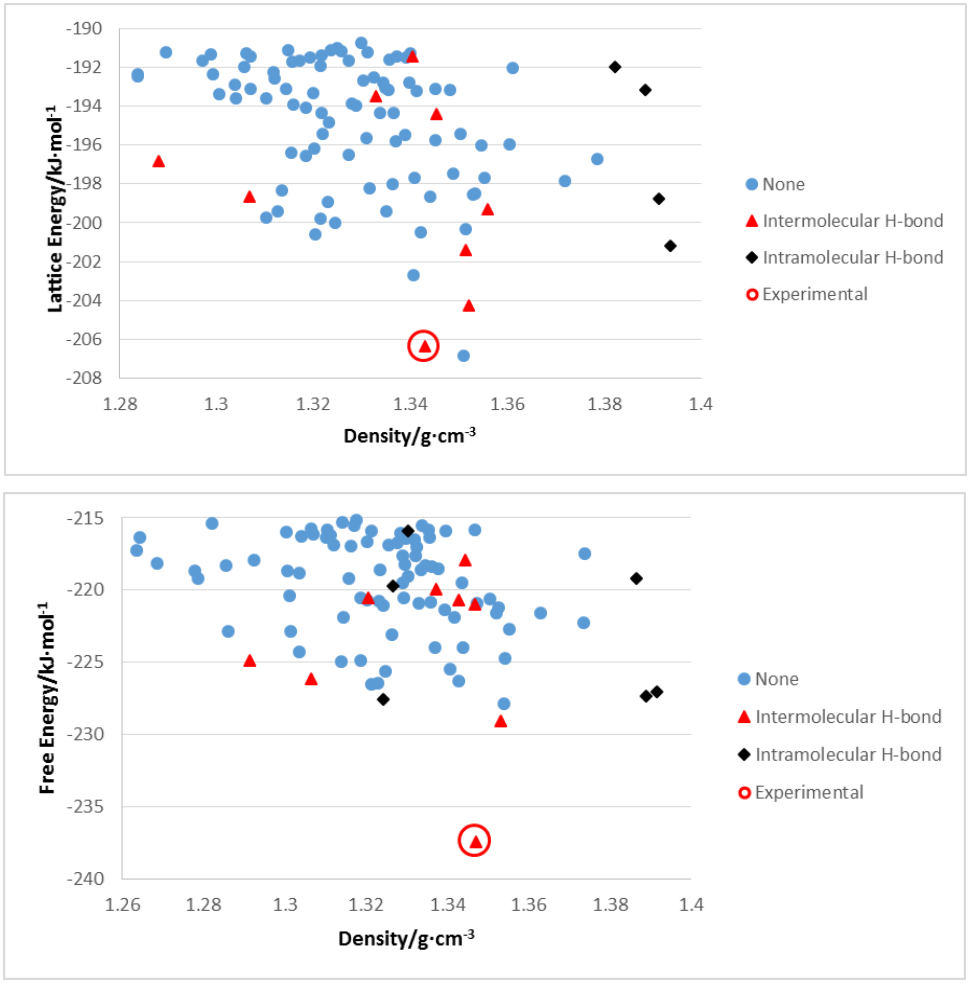


Figure 3.12: (above) Lattice energy vs density plot of the 100 lowest energy structures submitted as a first list of predictions (below) free energy vs density plot the 100 lowest energy structures submitted as a second list of predictions. They are labelled on whether they form intra- or intermolecular hydrogen bonds.

A comparison between these two plots reveals that the most important effect of the PCM with ε=3 was to stabilise structure 1600, but it also stabilised other crystal structures that formed hydrogen bonds, either intra- or intermolecular, relative to the ones that did not form any.

Although the target experimental form forms an intermolecular NH···O hydrogen bond, few predicted crystal structures share this feature. The reason is the bulky shape of molecule XXVI, with steric effects preventing the formation of an extensive hydrogen bond network.[2] The formation of hydrogen bonds requires that the amide groups are not co-planar with the naphthalene groups, which leads to a high conformational energy penalty, as shown in Figure 3.4c. Structure 1600 has a high relative ΔE_{intra} of ~17.5 kJ·mol^{-1} for the PBE0 6-31G(d,p) isolated molecule, while the inclusion of the PCM with ε=3 decreases the energy penalty to ~12.7 kJ·mol^{-1}.

Three interesting PPMs forming intermolecular hydrogen bonds are structures 675 (ranked 3rd in the first list and 2nd in the second), 421 (ranked 5th in the first list, and removed from the second list because considered similar to structure 675) and 2231

(ranked 14th in the first list, and removed from the second list because considered similar to structure 675). They all contain similar conformations to structure 1600 and form the same hydrogen bonded and π-π stacked sheet, as shown in Figure 3.13. As expected, these structures are all characterised by high ΔE_{intra} values of 13 to 20 kJ·mol^{-1} in the first list, compensated by favourable intermolecular interactions.

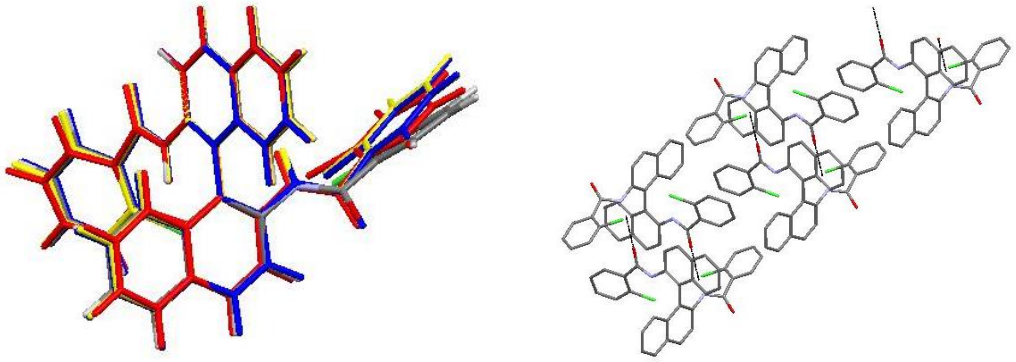


Figure 3.13: Overlay of the conformations of structures 1600 (coloured by elements), which matches the target experimental crystal structure, 675 (yellow), 421 (blue) and 2231 (red) (left) and the sheet common to all four structures (right).

On the other hand the global minimum in E_{latt} in the first list (ranked 3rd in the second list), structure 3525, contains a conformation similar to the isolated-molecule global minimum, with a low ΔE_{intra} value of ~1.2 kJ·mol^{-1} in the first list and ~2.2 kJ·mol^{-1} in the second list. Its crystal structure is characterised by sheets of π-π stacked chlorobenzene rings with Cl atoms oriented in opposite directions, as shown in Figure 3.14. This π-π stacking motif is present in 46 of the 100 structures in the first list of predictions.

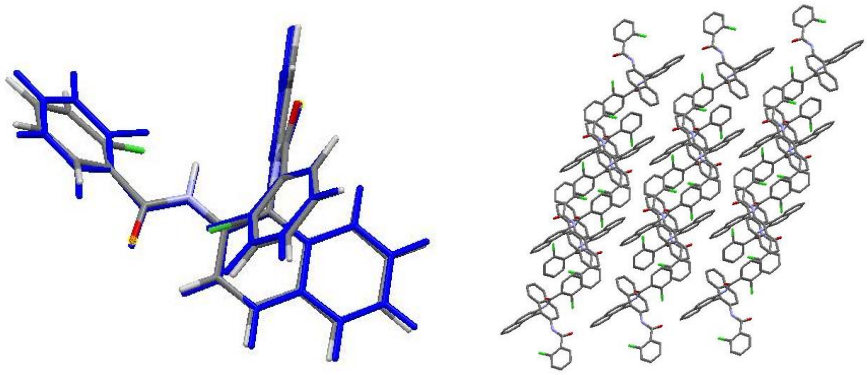


Figure 3.14: (left) Overlay of the conformation of structure 3525 (coloured by elements), the global minimum in E_{latt} in the first list, and the isolated-molecule global minimum of molecule XXVI (in blue) (right) and its crystal structure.

Other structures form intramolecular hydrogen bonds: structure 3104 (ranked 6th in both lists), 185 (ranked 17th in the first list and 5th in the second list), 1391 (ranked 59th in the first list and 48th in the second list) and 7559 (ranked 77th in the first list, and removed from the second list because considered similar to structure 3104). They all

have high ΔE_{intra} values of ~17-19 kJ·mol^{-1} in the first list, while in the second list they are slightly stabilised by the PCM with ΔE_{intra} values of ~13-14 kJ·mol^{-1}. The reason for these high conformational energies is that the stabilisation provided by the internal hydrogen bond does not fully compensate for the penalty of distorting both amide groups from co-planarity with naphthalene. However, their densities are the highest among all the low-energy crystal structures, and so they are stabilised by dispersion.[39] These four structure have similar conformations and share a similar stacking, shown in Figure 3.15.

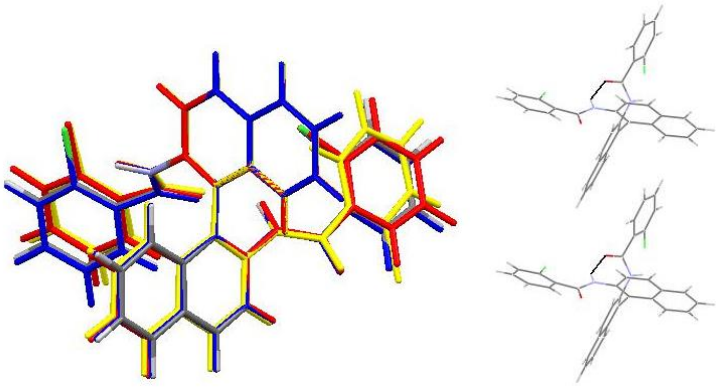


Figure 3.15: Overlay of the conformations of structures 3104 (coloured by elements), 185 (blue), 1391 (red) and 7559 (yellow) (left) and the packing motif common to those 4 structures (right).

In summary, this CSP study has shown that this molecule could form several alternative crystal structures, although the two lists give a very different picture of their relative competitiveness. The energetic and structural characteristics of these crystal structures provide several insights about this molecule and its possible solid-state behaviour.

3.3.4 Overall computational cost

The overall computational cost of this CSP study is broken down in Table 3.3.

Table 3.3: Breakdown of the computational cost of the Blind Test prediction of the crystal structure of molecule XXVI.

	CPU time (hours)	% of total time
Flexibility analysis	11,000	4.2
Grid Generation	11,000	4.2
Crystal structure search with CrystalPredictor 1.6	5,300	2.0
Single iterations with CrystalOptimizer	138,000	52.6
Full optimisations with CrystalOptimizer	96,000	36.6
Estimate of polarisation and temperature effects	1,100	0.4
Total time	262,400	100.0

The most expensive steps were the ones involved in the refinement of the search-generated crystal structures (see Chapter 3.2.3), which combined accounted for almost 90% of the overall computational cost. Although the flexibility analysis and the search accounted for slightly more than 10% of the overall CPU cost, those steps required the largest amount of human effort. The CPU cost shown in Table 3.3 is lower than the one given in the Blind Test publication, since in this more accurate estimate only the

calculations that led to the final predictions are accounted for.[2] However, the overall expense for predicting the crystal structure of this molecule remains rather large, highlighting an important problem related to performing CSP on large and flexible molecules. Some solutions to tackle the high computational cost of CSP studies of large and flexible molecules are proposed in Chapters 5 and 7.

3.3.5 DFT-D optimisation of the structures generated in this study

The 1,000 lowest energy crystal structures after the single-iteration of CrystalOptimizer were sent to two groups that did not have the codes to generate candidate structures, in order to give them the opportunity to use their high-quality Ψ_{crys} DFT-D methods (see Chapter 2.3.2.1) to optimise and re-rank them. As mentioned in Chapter 3.3.3.1, this list contained structure 1600 as the 45th structures in the E_{latt} ranking, which was already a close match to the target experimental form, as it was possible to match 15/15 molecules with an $RMSD_{15}$ of 0.422 Å.

Tkatchenko *et al.* did not send predictions for molecule XXVI due to time and resource constraints. In their post-analysis, they performed periodic DFT-D calculation with the PBE functional, first with a single-point E_{latt} calculation on each structure using the pairwise Tkatchenko-Scheffler (TS)[45] dispersion correction, then fully optimising 20 the lowest energy ones at the PBE+TS level of theory, and finally re-calculating the energies using the many-body dispersion (MBD)[46] correction. Structure 1600 was found to be the global minimum in E_{latt}, confirming the results of my CSP study, although with a small gap of just 0.7 kJ·mol^{-1} from the second lowest energy predicted form.

Brandenburg and Grimme used a multi-stage hierarchical approach, from cheaper semi-empirical methods (see Chapter 2.3.2.2) to more accurate electronic structure calculations. The first list was purely based on semi-empirical methods, while in the second one the structures were further optimised at the TPSS+D3[47, 48] level of theory. Structure 1600 was not present in either submitted list. Post-analysis revealed it was lost because it was slightly above the ranking cut-off that was adopted after one of the intermediate stages with semi-empirical methods; if it had been kept, it would have been the ranked 18th in the first list and 1st in the second (and more accurate) list, consistently with the post-analysis by Tkatchenko *et al.*

These high-quality calculations confirm that structure 1600 is one of the most stable computer-generated crystal structures in terms of E_{latt}. It is promising how the less sophisticated and cheaper Ψ_{mol} method employed in this study gave similar answers to more expensive Ψ_{crys} methods. The relative ranking of the other predicted structures is different, but an objective assessment of the quality of different crystal energy landscapes is impossible because of the absence of any other solved experimental polymorph of molecule XXVI. The fact that these very accurate methodologies failed to

provide a successful prediction for the Blind Test because of cost and resource constraints reinforces the idea that any effective CSP methodology must combine accuracy with cost feasibility.

3.4 Comparison with submissions by other groups

3.4.1 Molecule XXVI

Because of its complexity, only 12 out of the 25 participants submitted predictions for molecule XXVI. Only three of these 12 groups, including ours, successfully predicted the target crystal structure in their lists. Many of the failures were due to the inclusion of assumptions and/or limitations that were needed to keep the computational cost manageable.[2]

Both other groups that included the experimental crystal structure of XXVI in their predicted lists had it ranked very competitively. Neumann *et al*. used the well-known and successful GRACE method.[8] A tailor-made force field (TMFF)[49] was parametrised from DFT-D calculations, which was used to evaluate the energies of both bonded and non-bonded interactions. A Monte-Carlo parallel tempering algorithm was utilised to generate a set of crystal structure from the TMFF E_{latt} surface, in order to keep the computational cost manageable. The most promising candidates were then optimised and re-ranked with DFT-D, using the PBE functional and the empirically-fitted Neumann-Perrin dispersion correction. The DFT-D minimisations were performed firstly with more loose convergence criteria, forming the first list of predictions, and then with more stringent ones, forming the second list. The structure matching the experimental form was ranked first in both lists. The computational cost of their successful CSP study was larger than the one shown in Table 3.3 (~356,000 vs ~262,000 CPU hours); however, the cost of the methodologies is hardly comparable since GRACE produced some structures with Z'=2, although within a less complete search.

Elking and Fust-Molnar used a multi-step procedure in their CSP study. Their method consisted in the random generation of crystal structures in 32 space groups with 1 or 2 molecules in the asymmetric unit, the most promising of which were then optimised and re-ranked with DFT-D methods. The random generation was based on an estimate of E_{latt} in which ΔE_{intra} was calculated with the Merck Molecular Force Field,[50] while U_{inter} was estimated as the sum of an electrostatic component modelled with distributed multipoles calculated with GAMESS[51] and an repulsion-dispersion component calculated from an empirically fitted potential.[52] Several rigid searches were performed fom isolated-molecule conformers. After the crystal structure generation, the distributed multipoles of the top 2,000 lowest energy generated structures were re-calculated, and used to perform a flexible optimisation and energy re-ranking using the same E_{latt} model as in the search. The 50 lowest energy structures after this intermediate optimisation were

then fully minimised and re-ranked with DFT-D at the PBE+XDM[53] level of theory with Quantum Espresso.[54] The computational cost of their study and the criteria to produce the two lists are unknown. A structure matching the experimental form was ranked eighth in their first list and first in the second.

The other methods were not successful at predicting the target experimental form. Pantelides, Adjiman *et al.* used a similar method to the one utilised in this CSP study. CrystalPredictor (although version 2 rather than 1.6) was used to perform the search and CrystalOptimizer was utilised for the final refining the generated crystal structures. Post-analysis revealed that the failure to find a match to the experimental form was caused by the value of torsion angle Φ_{3b} (see Appendix Table 3.2) being outside the flexibility range in the search. After extending the search space in their post-analysis, a crystal structure matching the experimental form was found.[55] This shows that an accurate and comprehensive assessment of the conformational search space is vital to successful CSP studies.

Another interesting example is that of Day *et al*., who performed a set of rigid searches with the crystal structure generation engine G_{LEE}.[56] The searches were performed from conformers generated via a low-mode conformational search[57] using the OPLS2005 force field[58] and then optimised in the gas-phase with Gaussian09[59] at the B3LYP 6-311G(d,p)+D3 level of theory. The rigid searches were limited to the distinct isolated-molecule energy minima with a relative conformational energy lower than 30 $kJ \cdot mol^{-1}$. No match to the target experimental form was found, because it is highly distorted from the closest isolated-molecule conformational energy minimum (as shown in Figure 3.7). This hints to the need of not limiting the search space to isolated-molecule minima, since intermolecular forces can cause significant conformational adjustment.

3.4.2 Other molecules

The overall results of the Blind Test were quite promising. All molecules had at least one correct prediction, with the exception of the Z'=2 form E of molecule XXIII. The methodology used in our group was successful for most molecules, and the only Z'=1 crystal structure that was not present in either list was form A of molecule XXIII. Also, the two Z'=2 polymorphs of molecule XXIII (forms C and E) and the crystal structure of the three component salt XXIV were not found. The GRACE commercial methodology utilised by Neumann *et al*. was also very successful, since it only missed form E of molecule XXIII, and it was the only method that performed better than our simpler Ψ_{mol} alternative. This Blind Test has been considered a big advance in CSP,[60] which is becoming a procedure for understanding the solid form landscapes of organic molecules. However, the computational cost currently hinders a larger-scale applicability of CSP,

particularly on large and flexible molecules of pharmaceutical interest. Hence, it is worth to develop CSP methods to make them more widely applicable.

3.5 Conclusion

The CSP methodology discussed in this chapter successfully predicted the crystal structure of molecule XXVI in both lists submitted for the Blind Test. In the first list, ranked purely on E_{latt}, the structure matching the experimental form was ranked second, only 0.5 $kJ \cdot mol^{-1}$ above the global minimum, and was one among several competitive PPMs. In the second list, in which estimates of polarisation and temperature effects were included in the energy ranking, the structure matching the experimental form was ranked first, with a large energy gap from its closest competitor. Although the existence of a high-temperature polymorph of molecule XXVI may suggest that this large gap is incorrect, this cannot really be confirmed until this form is fully characterised, since it may be a structure not present in the crystal energy landscape or very similar to the known form.

The results from this study illustrate some points that are very important in the overall context of this thesis. First of all, the experimental form contains a strained high-energy conformation, which is stabilised by intermolecular hydrogen bonds. This shows how important the initial assessment of the conformational space is: trying to limit the computational cost of a flexible search by excluding conformational ranges that are relatively high in energy can often lead to incomplete searches, as shown by the CSP study by Pantelides, Adjiman *et al.* On the other hand, limiting the cost just by performing rigid searches on a set of isolated-molecule local minima is not adequate, as shown by the study by Day *et al*., since molecules can crystallise in distorted conformations that must be explicitly included in the searches. Hence a balance between computational cost and coverage of the conformational space must be found.[24]

Another important point is that for large and flexible molecules the final refinement procedure requires a cost-effective methodology. Since the search is bound to use approximate models, the energy ranking it provides is often poor. Hence hundreds or even thousands of structures often need to be considered, and it is currently impossible to minimise each of them with the most accurate models. An accurate method that cannot optimise a sufficient number of structures can be ineffective for use in CSP for large and flexible molecules, as illustrated by the failure of both groups that re-ranked crystal structures with DFT-D to submit a correct prediction for molecule XXVI despite being provided with a set of structures containing a close match to the target experimental form. Hence, accurate but cost effective optimisation and re-ranking methods are desirable.[42]

Finally, this study proves how sensitive energy differences are to the theoretical models. However, the quality of any energy ranking is difficult to judge in the absence of an experimental benchmark.[61] Overall, the 6th Blind Test showed that several CSP methods can predict and rationalise the crystallisation behaviour of organic molecules that are similar in terms of size and flexibility to the smaller ones in pharmaceutical development. The trade-off between computational cost and accuracy remains an important problem, and the expansion of CSP to even larger and more flexible molecules requires a limitation of the expense without a decrease in the quality of the predictions.

3.6 References

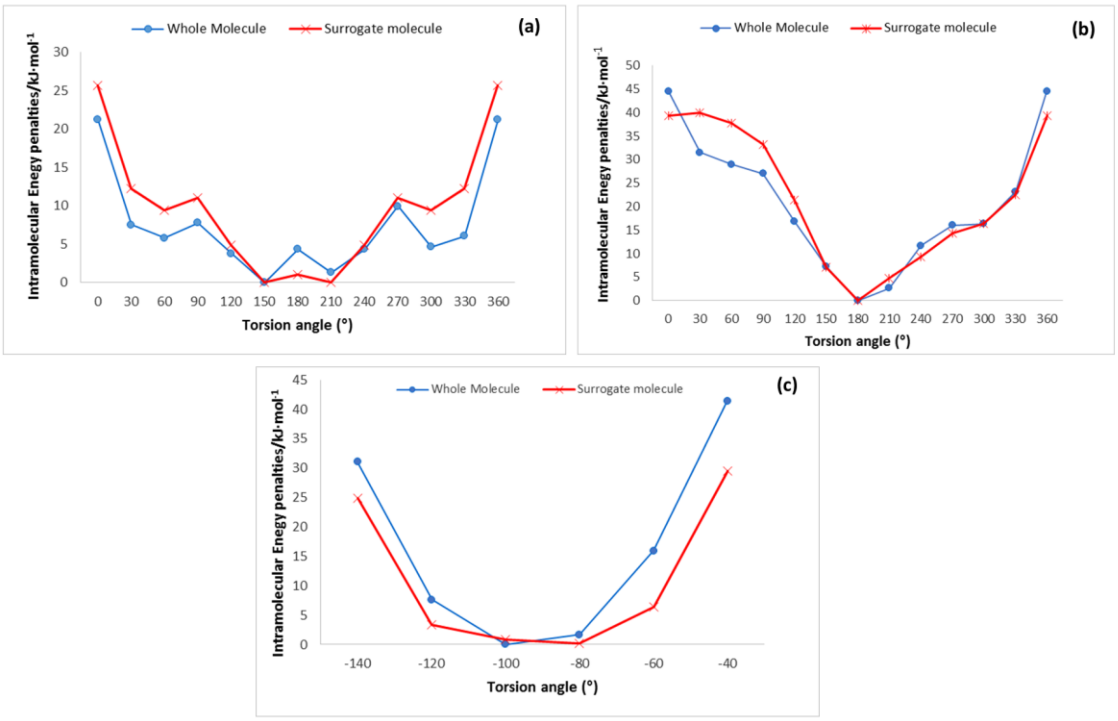
1. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C., The Cambridge Structural Database. *Acta Crystallographica Section B-Structural Science Crystal Engineering and Materials* **2016,** *72*, 171-179.
2. Reilly, A. M.; Cooper, R. I.; Adjiman, C. S.; Bhattacharya, S.; Boese, A. D.; Brandenburg, 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.; Vacarro, F.; Vasileiadis, M.; Vazquez-Mayagoitia, A.; Vogt, L.; Wang, Y.; Watson, R. E.; de Wijs, G. A.; Yang, J. Z.; Zhu, Q.; Groom, C. R., Report on the sixth blind test of organic crystal-structure prediction methods. *Acta Crystallographica Section B - Structural Science* **2016**.
3. Lommerse, J. P. M.; Motherwell, W. D. S.; Ammon, H. L.; Dunitz, J. D.; Gavezzotti, A.; Hofmann, D. W. M.; Leusen, F. J. J.; Mooij, W. T. M.; Price, S. L.; Schweizer, B.; Schmidt, M. U.; van Eijck, B. P.; Verwer, P.; Williams, D. E., A test of crystal structure prediction of small organic molecules. *Acta Crystallographica Section B - Structural Science* **2000,** *56*, 697-714.
4. Motherwell, W. D. S.; Ammon, H. L.; Dunitz, J. D.; Dzyabchenko, A.; Erk, P.; Gavezzotti, A.; Hofmann, D. W. M.; Leusen, F. J. J.; Lommerse, J. P. M.; Mooij, W. T. M.; Price, S. L.; Scheraga, H.; Schweizer, B.; Schmidt, M. U.; van Eijck, B. P.; Verwer, P.; Williams, D. E., Crystal structure prediction of small organic molecules: a second blind test. *Acta Crystallographica Section B - Structural Science* **2002,** *58*, 647-661.
5. Day, G. M.; Motherwell, W. D. S.; Ammon, H. L.; Boerrigter, S. X. M.; Della Valle, R. G.; Venuti, E.; Dzyabchenko, A.; Dunitz, J. D.; Schweizer, B.; van Eijck, B. P.; Erk, P.; Facelli, J. C.; Bazterra, V. E.; Ferraro, M. B.; Hofmann, D. W. M.; Leusen, F. J. J.; Liang, C.; Pantelides, C. C.; Karamertzanis, P. G.; Price, S. L.; Lewis, T. C.; Nowell, H.; Torrisi, A.; Scheraga, H. A.; Arnautova, Y. A.; Schmidt, M. U.; Verwer, P., A third blind test of crystal structure prediction. *Acta Crystallographica Section B - Structural Science* **2005,** *61* (5), 511-527.
6. Day, G. M.; Cooper, T. G.; Cruz-Cabeza, A. J.; Hejczyk, K. E.; Ammon, H. L.; Boerrigter, S. X. M.; Tan, J.; Della Valle, R. G.; Venuti, E.; Jose, J.; Gadre, S. R.; Desiraju, G. R.; Thakur, T. S.; van Eijck, B. P.; Facelli, J. C.; Bazterra, V. E.; Ferraro, M. B.; Hofmann, D. W. M.; Neumann, M.; Leusen, F. J. J.; Kendrick, J.; Price, S. L.; Misquitta, A. J.; Karamertzanis, P. G.; Welch, G. W. A.; Scheraga, H. A.; Arnautova, Y. A.; Schmidt, M. U.; van de Streek, J.; Wolf, A.; Schweizer, B., Significant progress in predicting the crystal structures of small organic molecules - a report on the fourth blind test. *Acta Crystallographica Section B - Structural Science* **2009,** *65* (2), 107-125.
7. 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 Crystallographica Section B-Structural Science* **2011,** *67*, 535-551.
8. Neumann, M. A. *GRACE (the Generation, Ranking and Characterisation Engine)*, 1.0; Avant-garde Materials Simulation Deutschland GmbH: 2007.
9. 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. *International Journal of Pharmaceutics* **2011,** *418* (2), 168-178.
10. Price, S. L.; Braun, D. E.; Reutzel-Edens, S. M., Can computed crystal energy landscapes help understand pharmaceutical solids? *Chemical Communications* **2016,** *52*, 7065-7077.
11. Vasileiadis, M.; Pantelides, C. C.; Adjiman, C. S., Prediction of the crystal structures of axitinib, a polymorphic pharmaceutical molecule. *Chemical Engineering Science* **2015,** *121*, 60-76.
12. Price, S. L.; Reutzel-Edens, S. M., The potential of computed crystal energy landscapes to aid solid form development. *Drug Discovery Today* **2016,** *21* (6), 912-923.
13. Jetti, R. K. R.; Boese, R.; Sarma, J.; Reddy, R. S.; Vishweshwar, P.; Desiraju, G. R., Searching for a polymorph: Second crystal form of 6-amino-2-phenylsulfonylimino-1,2-dihydropyridine. *Angewandte Chemie-International Edition* **2003,** *42*, 1963-1967.
14. Hulme, A. T.; Johnston, A.; Florence, A. J.; Fernandes, P.; Shankland, K.; Bedford, C. T.; Welch, G. W. A.; Sadiq, G.; Haynes, D. A.; Motherwell, W. D. S.; Tocher, D. A.; Price, S. L., Search for a predicted hydrogen bonding motif - A multidisciplinary investigation into the polymorphism of 3-azabicyclo[3.3.1]nonane-2,4-dione. *Journal of the American Chemical Society* **2007,** *129* (12), 3649-3657.
15. 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. *Journal of Chemical Information and Computer Sciences* **2004,** *44* (6), 2133-2144.
16. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; 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.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Gaussian Inc.: Wallingford CT, 2004.
17. Uzoh, O. G.; Galek, P. T. A.; Price, S. L., Analysis of the conformational profiles of fenamates shows route towards novel, higher accuracy, force-fields for pharmaceuticals. *Physical Chemistry Chemical Physics* **2015,** *17* (12), 7936-7948.
18. Taylor, R.; Cole, J.; Korb, O.; McCabe, P., Knowledge-Based Libraries for Predicting the Geometric Preferences of Druglike Molecules. *Journal of Chemical Information and Modeling* **2014,** *54* (9), 2500-2514.
19. Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kessler, M.; Macrae, C. F.; McCabe, P.; Pearson, J.; Taylor, R., New software for searching the Cambridge Structural Database and visualizing crystal structures. *Acta Crystallographica Section B - Structural Science* **2002,** *58*, 389-397.
20. Macrae, C. F.; Bruno, I. J.; Chisholm, J. A.; Edgington, P. R.; McCabe, P.; Pidcock, E.; Rodriguez-Monge, L.; Taylor, R.; van de Streek, J.; Wood, P. A., Mercury CSD 2.0 - new features for the visualization and investigation of crystal structures. *Journal of Applied Crystallography* **2008,** *41*, 466-470.
21. Karamertzanis, P. G.; Pantelides, C. C., Ab initio crystal structure prediction. II. Flexible molecules. *Molecular Physics* **2007,** *105* (2-3), 273-291.
22. 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. *Journal of Chemical Theory and Computation* **2015,** *11* (4), 1957-1969.

23. Pantelides, C.; Adjiman, C.; Kazantsev, A., General Computational Algorithms for Ab Initio Crystal Structure Prediction for Organic Molecules. In *Prediction and Calculation of Crystal Structures*, Springer International Publishing: 2014; Vol. 345, pp 25-58.
24. Iuzzolino, L.; Reilly, A. M.; McCabe, P.; Price, S. L., Use of Crystal Structure Informatics for Defining the Conformational Space Needed for Predicting Crystal Structures of Pharmaceutical Molecules. *Journal of Chemical Theory and Computation* **2017,** *13* (10), 5163-5171.
25. 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. *Journal of Physical Chemistry* **1996,** *100* (18), 7352-7360.
26. Steed, K.; Steed, J., Packing problems: High Z' crystal structures and theri relationship to co-crystals, inclusion compounds and polymorphism. *Chemical Reviews* **2015**.
27. Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C., CrystalOptimizer. An efficient Algorithm for Lattice Energy Minimisation of Organic Crystal using Isolated-Molecule Quantum Mechanical Calculations. In *Molecular System Engineering*, Adjiman, C. S.; Galindo, A., Eds. WILEY-VCH Verlag GmbH & Co.: Weinheim, 2010; Vol. 6, pp 1-42.
28. Braun, D.; Lingireddy, S.; Beidelschies, M.; Guo, R.; Muller, P.; Price, S.; Reutzel-Edens, S., Unraveling Complexity in the Solid Form Screening of a Pharmaceutical Salt: Why so Many Forms? Why so Few? *Crystal Growth & Design* **2017,** *17* (10), 5349-5365.
29. Stone, A. J., Distributed multipole analysis: Stability for large basis sets. *Journal of Chemical Theory and Computation* **2005,** *1* (6), 1128-1132.
30. Price, S. L.; Leslie, M.; Welch, G. W. A.; Habgood, M.; Price, L. S.; Karamertzanis, P. G.; Day, G. M., Modelling Organic Crystal Structures using Distributed Multipole and Polarizability-Based Model Intermolecular Potentials. *Physical Chemistry Chemical Physics* **2010,** *12* (30), 8478-8490.
31. Chisholm, J. A.; Motherwell, S., COMPACK: a program for identifying crystal structure similarity using distances. *Journal of Applied Crystallography* **2005,** *38*, 228-231.
32. Habgood, M.; Price, S. L.; Portalone, G.; Irrera, S., Testing a Variety of Electronic-Structure-Based Methods for the Relative Energies of 5-Formyluracil Crystals. *Journal of Chemical Theory and Computation* **2011,** *7* (9), 2685-2688.
33. Nyman, J.; Day, G. M., Static and lattice vibrational energy differences between polymorphs. *Crystengcomm* **2015,** *17* (28), 5154-5165.
34. Cruz-Cabeza, A. J.; Reutzel-Edens, S. M.; Bernstein, J., Facts and fictions about polymorphism. *Chemical Society Reviews* **2015,** *44*, 8619-8635.
35. Thompson, H. P. G.; Day, G. M., Which conformations make stable crystal structures? Mapping crystalline molecular geometries to the conformational energy landscape. *Chemical Science* **2014,** *5* (8), 3173-3182.
36. Cruz-Cabeza, A. J.; Bernstein, J., Conformational Polymorphism. *Chemical Reviews* **2014,** *114* (4), 2170-2191.
37. Cooper, T. G.; Hejczyk, K. E.; Jones, W.; Day, G. M., Molecular Polarization Effects on the Relative Energies of the Real and Putative Crystal Structures of Valine. *Journal of Chemical Theory and Computation* **2008,** *4* (10), 1795-1805.
38. Nyman, J.; Day, G., Modelling temperature-dependent properties of polymorphic organic molecular crystals. *Physical Chemistry Chemical Physics* **2016,** *18* (45), 31132-31143.
39. Price, S. L., Predicting crystal structures of organic compounds. *Chemical Society Reviews* **2014,** *43* (7), 2098-2111.
40. Nyman, J.; Day, G. M., Modelling temperature-dependent properties of polymorphic organic molecular crystals. *Physical Chemistry Chemical Physics* **2016,** *18* (45), 31132-31143.
41. Cervinka, C.; Beran, G. J. O., Ab initio prediction of the polymorph phase diagram for crystalline methanol. *Chemical Science* **2018**.
42. Iuzzolino, L.; McCabe, P.; Price, S. L.; Brandenburg, J. G., Crystal structure prediction of flexible pharmaceutical-like molecules: density functional tight-binding as an intermediate optimisation method and for free energy estimation. *Faraday Discussions* **2018***, Advance article*.
43. Coles, S. J.; Threlfall, T. L.; Tizzard, G. J., The Same but Different: Isostructural Polymorphs and the Case of 3-Chloromandelic Acid. *Crystal Growth & Design* **2014,** *14* (4), 1623-1628.
44. Bhardwaj, R. M.; Price, L. S.; Price, S. L.; Reutzel-Edens, S. M.; Miller, G. J.; Oswald, I. D. H.; Johnston, B.; Florence, A. J., Exploring the Experimental and Computed Crystal Energy Landscape of Olanzapine. *Crystal Growth & Design* **2013,** *13* (4), 1602-1617.
45. Tkatchenko, A.; Scheffler, M., Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Physical Review Letters* **2009,** *102* (7), 073005.

46. Tkatchenko, A.; DiStasio, R. A. J.; Car, R.; Scheffler, M., Accurate and efficient method for many-body van der Waals interactions. *Physical Review Letters* **2012,** *108* (23), 236402-236402.
47. Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H., A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *Journal of Chemical Physics* **2010,** *132* (15).
48. Grimme, S.; Ehrlich, S.; Goerigk, L., Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *Journal of Computational Chemistry* **2011,** *32* (7), 1456-1465.
49. Neumann, M. A., Tailor-made force fields for crystal-structure prediction. *Journal of Physical Chemistry B* **2008,** *112* (32), 9810-9829.
50. Halgren, T. A., Merck molecular force field .1. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry* **1996,** *17* (5-6), 490-519.
51. Gordon, M. S.; Schmidt, M. W., Advances in electronic structure theory: GAMESS a decade later. In *Theory and Applications of Computational Chemistry: the first forty years*, Dykstra, C. E.; Frenking, G.; Kim, K. S.; Scuseria, G. E., Eds. Amsterdam, 2005; pp 1167-1189.
52. Elking, D. M.; Fusti-Molnar, L.; Nichols, A., Crystal structure prediction of rigid molecules. *Acta Crystallographica Section B-Structural Science Crystal Engineering and Materials* **2016,** *72*, 488-501.
53. Becke, A. D.; Johnson, E. R., Exchange-hole dipole moment and the dispersion interaction revisited. *The Journal of Chemical Physics* **2007,** *127* (15), 154108.
54. Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L.; Cococcioni, M.; Dabo, I.; Dal Corso, A.; de Gironcoli, S.; Fabris, S.; Fratesi, G.; Gebauer, R.; Gerstmann, U.; Gougoussis, C.; Kokalj, A.; Lazzeri, M.; Martin-Samos, L.; Marzari, N.; Mauri, F.; Mazzarello, R.; Paolini, S.; Pasquarello, A.; Paulatto, L.; Sbraccia, C.; Scandolo, S.; Sclauzero, G.; Seitsonen, A. P.; Smogunov, A.; Umari, P.; Wentzcovitch, R. M., QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *Journal of Physics-Condensed Matter* **2009,** *21* (39), 395502.
55. Sugden, I.; Adjiman, C.; Pantelides, C., Accurate and efficient representation of intramolecular energy in ab initio generation of crystal structures. I. Adaptive local approximate models. *Acta Crystallographica Section B-Structural Science Crystal Engineering and Materials* **2016,** *72*, 864-874.
56. Case, D. H.; Campbell, J. E.; Bygrave, P. J.; Day, G. M., Convergence Properties of Crystal Structure Prediction by Quasi-Random Sampling. *Journal of Chemical Theory and Computation* **2016,** *12* (2), 910-924.
57. Kolossváry, I.; Guida Wayne, C., Low-mode conformational search elucidated: Application to C39H80 and flexible docking of 9-deazaguanine inhibitors into PNP. *Journal of Computational Chemistry* **1999,** *20* (15), 1671-1684.
58. Banks Jay, L.; Beard Hege, S.; Cao, Y.; Cho Art, E.; Damm, W.; Farid, R.; Felts Anthony, K.; Halgren Thomas, A.; Mainz Daniel, T.; Maple Jon, R.; Murphy, R.; Philipp Dean, M.; Repasky Matthew, P.; Zhang Linda, Y.; Berne Bruce, J.; Friesner Richard, A.; Gallicchio, E.; Levy Ronald, M., Integrated Modeling Program, Applied Chemical Theory (IMPACT). *Journal of Computational Chemistry* **2005,** *26* (16), 1752-1780.
59. 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. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; 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, D. J. *Gaussian 09*, Gaussian, Inc.: Wallingford, CT, USA, 2009.
60. Gibney, E., Software predicts slew of fiendish crystal structures. *Nature* **2015,** *527* (7576), 20-21.
61. Brandenburg, J. G.; Grimme, S., Organic crystal polymorphism: a benchmark for dispersion-corrected mean-field electronic structure methods. *Acta Crystallographica Section B-Structural Science Crystal Engineering and Materials* **2016,** *72*, 502-513.

3.7 Appendix



Appendix Figure 3.1: Comparison of the energy penalties for varying torsion angles (a) Φ_{1a} and Φ_{1b} (b) Φ_{3a} and Φ_{3b} (c) Φ_4 in the whole molecule and in the surrogate molecules used to calculate the ΔE_{intra} grids. All the calculations were performed at the PBE0 6-31G(d,p) level of theory starting from the PBE0 6-31G(d,p) optimised gas-phase conformations of both the whole molecule and the surrogate molecules. Φ_{1b} and Φ_{3b} were not scanned explicitly, as their energy profile would be identical to their symmetry related counterparts.

Appendix Table 3.1: List of the 59 space groups considered in the crystal structure search performed with CrystalPredictor 1.6.

P1	$P\bar{1}$	$P2_1$	$P2_1/c$	$P2_12_12$	$P2_12_12_1$	$Pna2_1$	$Pca2_1$	Pbca	Pbcn
C2/c	Cc	C2	Pc	Cm	$P2_1/m$	C2/m	P2/c	$C222_1$	$Pmn2_1$
$Cmc2_1$	Aba2	Fdd2	Iba2	Pnna	Pccn	Pbcm	Pnnm	Pmmn	Pnma
Cmcm	Cmca	Fddd	Ibam	$P4_1$	$P4_3$	$I\bar{4}$	P4/n	$P4_2/n$	I4/m
$I4_1/a$	$P4_12_12$	$P4_32_12$	$P\bar{4}2_1c$	$I\bar{4}2d$	$P3_1$	$P3_2$	R3	$P\bar{3}$	$R\bar{3}$
$P3_121$	$P3_221$	R3c	$R\bar{3}C$	$P6_1$	$P6_3$	$P6_3/m$	$P2_13$	$PA\bar{3}$	

Appendix Table 3.2: Values taken by the seven torsion angles in Figure 3.2 in the conformation of the target experimental crystal structure of molecule XXVI.

Torsion angle label	Torsion Angle Definition	Experimental conformer value
Φ_{1a}	C2-C1-C7-O1	145.6
Φ_{1b}	C30-C29-C28-O2	124.1
Φ_{2a}	O1-C7-N1-H21	179.9
Φ_{2b}	O2-C28-N2-H22	167.4
Φ_{3a}	H21-N1-C8-C17	194.3
Φ_{3b}	H22-N2-C19-C20	140.9
Φ_4	C19-C18-C9-C8	-78.4

Appendix Table 3.3: Lattice energies, structural and crystallographic parameters of the 100 crystal structures of molecule XXVI that were submitted as a first list of prediction for the 6th Blind Test. The structure matching the experimental form is highlighted in orange.

Structure	E_{latt} /kJ·mol^{-1}	Density /g·cm^3	Packing coefficient/ %	Space group	Z'	a/Å	b/Å	c/Å	α/°	β/°	γ/°
3525	-206.86	1.351	68.3	$P\bar{1}$	1	10.55	10.84	14.97	71.52	69.52	61.07
1600	-206.37	1.343	67.4	$P\bar{1}$	1	10.26	11.17	14.23	78.54	73.53	62.86
675	-204.25	1.352	68.7	$P2_1/c$	1	11.67	14.21	16.91	90.00	100.2	90.00
38	-202.70	1.341	67.5	$I4_1/a$	1	10.90	23.24	23.24	86.85	76.43	76.43
421	-201.43	1.352	68.6	$P2_1/c$	1	11.56	14.34	16.91	90.00	100.2	90.00
3104	-201.20	1.394	70.5	$P\bar{1}$	1	10.50	10.95	13.72	105.0	100.5	112.3
615	-200.58	1.321	66.8	$P2_1/c$	1	10.42	14.26	20.19	90.00	90.00	109.7
239	-200.48	1.342	68.0	$P2_1/c$	1	10.78	13.75	18.95	98.39	90.00	90.00
2930	-200.30	1.351	68.7	$P\bar{1}$	1	10.74	10.83	14.64	88.66	70.78	60.44
354	-199.98	1.325	67.0	$P2_1/c$	1	10.21	15.69	19.12	113.2	90.00	90.00
851	-199.81	1.322	66.7	$P\bar{1}$	1	11.16	11.18	11.94	86.52	73.76	80.74
6460	-199.74	1.310	65.8	$P\bar{1}$	1	10.89	11.46	13.66	72.83	69.52	64.82
6335	-199.41	1.335	67.6	$C2/c$	1	10.69	10.90	28.19	78.98	79.07	60.65
221	-199.40	1.313	66.4	$P\bar{1}$	1	11.09	11.22	13.90	69.59	89.40	63.01
2231	-199.29	1.356	68.6	$P\bar{1}$	1	10.61	10.87	14.39	74.95	73.13	61.12
2496	-198.93	1.323	67.4	$P2_1/c$	1	10.53	13.85	19.65	100.2	90.00	90.00
185	-198.75	1.391	70.7	$P2_1/c$	1	11.95	13.74	17.80	90.00	90.00	113.5
4201	-198.65	1.344	67.9	$P2_1$	2	9.95	10.20	29.41	90.00	90.00	111.6
314	-198.63	1.307	65.8	$P2_1/c$	1	10.84	14.31	19.45	90.00	90.00	108.9
508	-198.56	1.353	68.5	$P2_1/c$	1	11.12	13.20	18.84	90.00	94.36	90.00
4946	-198.48	1.354	68.8	$P\bar{1}$	1	10.56	11.10	14.39	103.8	98.13	118.1
6879	-198.35	1.314	66.3	$C2/c$	1	10.69	10.69	28.58	94.89	96.61	117.7
506	-198.23	1.332	67.8	$P\bar{1}$	1	10.85	11.39	14.01	85.79	67.59	61.94
4842	-198.03	1.336	67.8	$P2_1/c$	1	10.85	14.66	17.79	90.00	99.49	90.00
43	-197.84	1.372	69.5	$P2_1/c$	1	8.84	16.17	19.01	90.00	90.60	90.00
1236	-197.71	1.355	68.3	$C2/c$	1	14.93	14.93	15.87	103.4	107.0	116.2
1537	-197.69	1.341	67.8	$P\bar{1}$	1	10.38	10.43	14.40	94.31	91.70	116.3
188	-197.46	1.349	68.3	$P\bar{1}$	1	9.00	10.75	16.26	104.9	94.63	111.8
5126	-196.81	1.288	65.1	$Pbca$	1	13.54	19.66	21.75	90.00	90.00	90.00
444	-196.75	1.379	69.7	$P\bar{1}$	1	10.38	10.43	14.27	101.3	101.3	111.0
544	-196.57	1.318	66.8	$P\bar{1}$	1	11.15	11.22	11.98	73.18	86.57	80.28
686	-196.52	1.327	67.4	$C2/c$	1	13.14	15.37	15.41	92.81	90.00	115.3
89	-196.42	1.315	66.3	$I4_1/a$	1	19.93	19.93	19.93	91.50	119.1	119.1
20	-196.16	1.320	67.0	$P\bar{1}$	1	10.12	11.07	13.39	88.56	80.15	72.88
83	-196.04	1.355	68.4	$P2_1/c$	1	10.85	15.53	16.34	90.00	90.09	90.00
3075	-195.96	1.361	69.0	$Pbca$	1	14.44	17.91	21.20	90.00	90.00	90.00
4199	-195.83	1.337	67.6	$P\bar{1}$	1	9.97	10.80	14.41	95.32	94.86	114.2
1249	-195.77	1.345	68.2	$P2_1/c$	1	12.26	15.21	15.39	90.00	105.0	90.00
177	-195.67	1.331	67.1	$P\bar{1}$	1	10.84	11.20	12.19	107.2	92.98	95.99
2012	-195.49	1.339	67.3	$P\bar{1}$	1	9.60	11.00	14.73	94.10	96.01	114.8
3860	-195.43	1.322	66.8	$C2/c$	1	10.93	11.66	25.21	95.63	90.00	117.9
1237	-195.43	1.351	68.3	$P2_1/c$	1	11.25	13.71	17.91	90.00	91.44	90.00
579	-194.83	1.323	66.6	$P2_1/c$	1	10.64	13.84	19.74	90.00	90.00	104.0
361	-194.40	1.345	67.9	$P2_1/c$	1	11.81	14.27	16.70	90.00	99.87	90.00
985	-194.35	1.337	67.4	$P\bar{1}$	1	10.08	11.15	14.01	82.16	83.35	63.63
371	-194.35	1.334	67.4	$P2_1/c$	2	16.71	17.04	21.10	90.00	111.4	90.00
694	-194.34	1.322	66.9	$P2_1/c$	1	10.81	14.41	18.24	90.00	96.90	90.00
102	-194.08	1.319	66.8	$Pbca$	1	16.24	18.32	19.02	90.00	90.00	90.00
5147	-193.98	1.329	67.2	Cc	2	10.62	11.08	27.22	92.41	90.00	118.6
8917	-193.91	1.316	66.5	$P2_1/c$	1	10.65	14.36	19.50	90.00	90.00	108.0
1098	-193.86	1.328	67.1	$P2_1/c$	1	11.06	14.00	18.14	90.00	90.00	90.38
1378	-193.62	1.310	66.6	$P2_1/c$	1	10.34	13.72	21.02	90.00	90.00	107.2
1662	-193.60	1.304	66.3	$P2_1/c$	1	10.97	13.22	20.06	100.6	90.00	90.00
1369	-193.52	1.333	66.9	$P\bar{1}$	1	10.50	11.01	13.15	72.81	74.73	88.80
1019	-193.40	1.301	65.8	$P2_1$	2	10.50	10.84	28.40	90.00	90.00	117.4
6282	-193.32	1.320	66.6	$P2_1/c$	1	10.70	14.40	19.30	90.00	90.00	108.1

88	-193.22	1.341	67.9	*C2/c*	1	13.50	13.50	15.62	96.37	96.37	98.03
125	-193.20	1.348	68.2	$P2_1/c$	1	11.47	13.42	18.62	90.00	90.00	105.2
1391	-193.19	1.389	70.4	*Cc*	1	10.76	10.76	13.89	103.9	103.9	112.0
6915	-193.15	1.336	67.2	$P\bar{1}$	1	10.55	10.76	14.02	83.48	69.50	69.46
203	-193.14	1.345	68.1	$P\bar{1}$	1	10.83	11.36	12.96	78.29	83.03	62.67
1749	-193.13	1.307	65.6	*C2/c*	1	10.75	10.75	29.06	90.95	99.79	119.6
2339	-193.12	1.314	66.0	$P\bar{1}$	1	10.59	10.91	14.59	106.1	97.43	114.1
277	-193.08	1.335	67.4	*C2/c*	1	13.51	13.51	17.30	71.19	71.19	75.97
8620	-192.92	1.304	66.2	$P\bar{1}$	1	10.89	11.46	11.74	92.43	98.87	98.31
485	-192.82	1.340	67.8	$P2_1/c$	1	13.67	14.24	14.30	90.00	91.30	90.00
576	-192.77	1.334	67.5	$P2_1/c$	1	11.12	15.67	16.04	90.00	90.98	90.00
709	-192.69	1.330	67.3	$P2_1/c$	1	11.85	13.03	18.49	90.00	100.8	90.00
1275	-192.59	1.312	66.1	$P2_1/c$	1	10.59	14.21	19.74	90.00	90.00	106.9
3196	-192.51	1.333	67.7	$P2_1/c$	1	11.55	13.83	17.96	90.00	102.6	90.00
1881	-192.49	1.284	65.1	$P2_1/c$	1	10.56	13.36	21.66	90.00	90.00	108.0
595	-192.35	1.299	65.7	*C2/c*	1	14.42	14.42	14.60	103.0	103.0	91.58
4822	-192.34	1.284	64.7	$P\bar{1}$	1	9.80	11.21	14.07	75.37	81.15	77.82
833	-192.25	1.312	66.2	*C2/c*	1	11.10	11.10	24.42	87.75	87.75	71.17
692	-192.02	1.361	69.0	$P2_1$	1	9.98	11.23	12.23	92.28	90.00	90.00
2129	-191.99	1.306	66.2	*C2/c*	1	10.73	10.73	28.49	80.19	80.19	62.78
7559	-191.98	1.382	69.9	$P\bar{1}$	1	10.68	11.13	13.64	103.5	103.6	113.0
1244	-191.94	1.322	67.0	$P2_12_12_1$	1	10.39	11.56	23.50	90.00	90.00	90.00
7458	-191.74	1.316	66.5	$P\bar{1}$	1	11.36	11.53	13.52	84.72	65.88	62.13
2763	-191.68	1.327	67.1	$P2_1/c$	1	11.44	13.98	17.62	90.00	94.12	90.00
3441	-191.68	1.317	66.1	$P2_1/c$	1	11.17	14.77	17.17	90.00	90.00	91.86
500	-191.64	1.297	65.6	$P2_1/c$	1	10.42	14.70	18.78	90.00	90.27	90.00
545	-191.60	1.336	67.5	$P2_1/c$	1	10.92	14.77	17.62	100.7	90.00	90.00
174	-191.52	1.339	67.7	$P2_1/c$	1	10.95	14.97	18.05	90.00	90.00	109.8
4800	-191.50	1.319	66.8	$P\bar{1}$	1	10.47	10.95	14.96	69.21	77.70	61.98
4000	-191.47	1.307	65.8	$P\bar{1}$	1	11.21	11.55	12.37	104.4	93.42	111.1
5656	-191.47	1.340	67.5	$P\bar{1}$	1	11.35	11.78	12.08	95.99	117.6	97.74
119	-191.46	1.337	67.8	*C2/c*	1	10.51	10.51	25.56	96.10	96.10	92.56
1825	-191.41	1.322	66.9	$P2_1$	2	10.87	11.08	26.56	90.00	90.00	118.1
455	-191.35	1.299	65.8	$P2_1/c$	1	12.84	14.90	16.14	111.5	90.00	90.00
3136	-191.28	1.306	66.3	$P2_1/c$	1	10.71	14.98	17.82	90.00	90.00	92.33
4915	-191.28	1.340	67.9	$P2_1/c$	1	11.18	12.93	19.31	90.00	94.41	90.00
779	-191.24	1.290	65.3	*Pbca*	1	14.66	19.82	19.91	90.00	90.00	90.00
1377	-191.21	1.331	67.5	$P2_1/c$	1	10.81	14.98	18.05	106.6	90.00	90.00
201	-191.16	1.326	67.0	$P2_1/c$	1	10.70	16.31	16.72	90.00	105.5	90.00
4638	-191.14	1.315	66.1	$P\bar{1}$	1	10.01	10.27	15.57	97.28	100.5	112.5
1920	-191.13	1.324	66.9	$P2_1/c$	1	10.58	13.60	19.96	90.00	90.00	101.3
1684	-191.04	1.325	67.1	*Pbca*	1	14.01	18.77	21.41	90.00	90.00	90.00
31	-190.89	1.328	67.4	$P2_1/c$	1	8.71	16.36	20.61	106.9	90.00	90.00
4958	-190.75	1.330	67.3	*C2/c*	1	14.44	14.44	16.50	107.0	107.4	108.7

Appendix Table 3.4: Free energies, structural and crystallographic parameters of the 100 structures of molecule XXVI that were submitted as a second list of prediction for the 6th Blind Test. The structure matching the experimental form is highlighted in orange.

Structure	A /kJ·mol^{-1}	Density /g·cm^3	Packing coefficient/ %	Space group	Z'	a/Å	b/Å	c/Å	α/°	β/°	γ/°
1600	-237.40	1.347	67.7	$P\bar{1}$	1	10.2	16.24	11.17	59.0	117.13	93.6
675	-229.09	1.353	68.8	$P2_1/c$	1	16.9	14.17	11.69	90.0	79.90	90.0
3525	-227.85	1.354	68.5	$P\bar{1}$	1	10.8	10.86	29.89	59.6	69.36	121.
2591	-227.59	1.324	67.4	$P\bar{1}$	1	11.8	10.79	11.54	90.0	106.74	90.0
185	-227.32	1.389	70.6	$P2_1/c$	1	11.9	17.81	13.73	90.0	113.55	90.0
3104	-227.04	1.392	70.4	$P\bar{1}$	1	10.5	22.34	10.93	117.	67.73	62.5
851	-226.49	1.321	66.7	$P\bar{1}$	1	11.1	11.16	11.95	73.5	86.55	80.8
2496	-226.41	1.323	67.3	$P2_1/c$	1	13.8	10.54	19.64	90.0	79.80	90.0
239	-226.30	1.343	68.0	$P2_1/c$	1	18.9	10.80	13.73	90.0	81.69	90.0
314	-226.15	1.307	65.8	$P2_1/c$	1	10.8	19.49	14.86	90.0	65.49	90.0

354	-225.61	1.325	67.0	$P2_1/c$	1	15.7	10.20	19.35	90.0	65.08	90.0
38	-225.46	1.341	67.5	$I4_1/a$	1	31.9	31.96	10.89	90.0	90.00	90.0
221	-224.93	1.314	66.5	$P\bar{1}$	1	11.2	13.91	11.08	89.2	117.09	110.
5126	-224.86	1.291	65.3	*Pbca*	1	21.7	13.53	19.66	90.0	90.00	90.0
615	-224.84	1.319	66.7	$P2_1/c$	1	10.4	20.21	14.53	90.0	67.25	90.0
4946	-224.71	1.354	68.9	$P\bar{1}$	1	10.5	14.54	11.11	100.	61.92	76.8
8620	-224.26	1.304	66.1	$P\bar{1}$	1	10.8	11.75	11.45	87.3	81.92	98.8
6335	-224.01	1.337	67.7	*C2/c*	1	28.2	10.68	18.99	90.0	77.15	90.0
4201	-223.99	1.344	67.9	$P\bar{1}$	2	11.3	29.46	9.95	90.0	123.25	90.0
686	-223.08	1.326	67.4	*C2/c*	1	27.7	13.14	15.42	90.0	87.25	90.0
1019	-222.86	1.302	65.9	*P21*	2	11.0	28.43	10.50	90.0	60.08	90.0
1881	-222.84	1.286	65.2	$P2_1/c$	1	13.3	21.66	10.56	90.0	71.82	90.0
1236	-222.73	1.355	68.3	*C2/c*	1	25.3	15.77	19.09	90.0	133.86	90.0
43	-222.28	1.373	69.6	$P2_1/c$	1	18.9	16.21	8.84	90.0	89.26	90.0
6879	-221.90	1.314	66.3	*C2/c*	1	18.3	11.04	29.66	90.0	71.16	90.0
1537	-221.88	1.342	67.9	$P\bar{1}$	1	10.3	14.43	10.41	85.6	63.78	91.8
508	-221.58	1.352	68.5	$P2_1/c$	1	18.8	13.21	11.11	90.0	85.69	90.0
3075	-221.55	1.363	69.1	*Pbca*	1	17.8	21.22	14.45	90.0	90.00	90.0
4842	-221.33	1.339	68.0	$P2_1/c$	1	10.8	14.62	19.29	90.0	114.50	90.0
83	-221.22	1.353	68.2	$P2_1/c$	1	10.8	15.49	19.64	90.0	56.76	90.0
694	-221.05	1.324	67.0	$P2_1/c$	1	10.8	14.41	18.22	90.0	83.14	90.0
361	-220.97	1.347	68.0	$P2_1/c$	1	11.8	14.25	18.77	90.0	118.65	90.0
506	-220.92	1.333	67.9	$P\bar{1}$	1	15.2	11.39	11.47	123.	118.89	61.5
188	-220.88	1.347	68.2	$P\bar{1}$	1	9.00	11.17	16.65	68.2	76.71	63.3
803	-220.85	1.336	67.8	$P2_1$	2	10.6	28.86	9.82	90.0	67.66	90.0
1244	-220.78	1.323	67.1	$P2_12_12_1$	1	11.5	10.40	23.46	90.0	90.00	90.0
203	-220.68	1.343	67.9	$P\bar{1}$	1	12.9	11.58	10.78	119.	96.89	94.6
20	-220.65	1.320	67.0	$P\bar{1}$	1	13.4	10.12	11.06	72.8	88.52	80.1
1237	-220.58	1.351	68.3	$P2_1/c$	1	17.9	13.70	11.25	90.0	88.60	90.0
177	-220.56	1.329	67.1	$P\bar{1}$	1	11.2	10.85	12.20	93.0	72.60	84.1
7902	-220.55	1.321	67.0	$P\bar{1}$	1	11.1	10.98	13.23	104.	83.73	67.0
8917	-220.54	1.319	66.6	$P2_1/c$	1	10.6	19.49	14.98	90.0	65.44	90.0
455	-220.41	1.301	65.9	$P2_1/c$	1	14.8	12.83	17.47	90.0	120.69	90.0
1369	-219.94	1.337	67.1	$P\bar{1}$	1	13.0	10.54	11.01	91.4	72.71	105.
4273	-219.73	1.327	67.3	$P2_1/c$	1	13.1	11.35	20.74	90.0	65.60	90.0
1098	-219.51	1.329	67.1	$P2_1/c$	1	11.0	18.17	13.99	90.0	89.50	90.0
1249	-219.45	1.344	68.0	$P2_1/c$	1	12.2	15.23	15.41	90.0	74.93	90.0
1391	-219.18	1.387	70.3	*Cc*	1	12.0	17.84	13.89	90.0	115.58	90.0
89	-219.17	1.316	66.3	$I4_1/a$	1	20.2	20.21	27.76	90.0	90.00	90.0
6909	-219.16	1.279	65.0	*Pbca*	1	18.9	21.26	14.50	90.0	90.00	90.0
709	-219.01	1.330	67.3	$P2_1/c$	1	11.8	13.02	18.51	90.0	79.03	90.0
1662	-218.84	1.304	66.3	$P2_1/c$	1	20.0	10.98	13.22	90.0	79.68	90.0
1378	-218.76	1.313	66.7	$P2_1/c$	1	10.3	21.05	14.49	90.0	64.43	90.0
595	-218.69	1.301	65.7	*C2/c*	1	20.1	20.67	14.57	90.0	71.15	90.0
1786	-218.69	1.278	64.9	*Pbca*	1	14.5	18.91	21.17	90.0	90.00	90.0
3860	-218.62	1.323	66.9	*C2/c*	1	25.2	10.93	20.59	90.0	83.51	90.0
576	-218.56	1.334	67.5	$P2_1/c$	1	16.0	15.64	11.12	90.0	88.79	90.0
4199	-218.54	1.338	67.6	$P\bar{1}$	1	16.8	10.80	9.94	65.8	121.28	108.
6915	-218.35	1.336	67.3	$P\bar{1}$	1	10.7	14.01	10.56	110.	69.33	96.6
371	-218.32	1.335	67.4	$P2_1/c$	2	16.6	17.04	21.09	90.0	68.67	90.0
4822	-218.27	1.286	64.8	$P\bar{1}$	1	15.8	9.80	11.21	102.	84.44	61.2
2763	-218.18	1.330	67.2	$P2_1/c$	1	11.4	13.94	17.60	90.0	86.11	90.0
5110	-218.11	1.269	63.9	$P2_1/c$	1	10.8	19.66	15.56	90.0	117.87	90.0
779	-217.92	1.292	65.5	*Pbca*	1	19.8	19.90	14.65	90.0	90.00	90.0
5656	-217.89	1.344	67.7	$P\bar{1}$	1	11.3	11.78	12.10	84.4	62.03	97.8
1377	-217.65	1.332	67.6	$P2_1/c$	1	14.9	10.81	19.85	90.0	60.47	90.0
5147	-217.59	1.329	67.2	*Cc*	2	19.4	10.62	27.22	90.0	87.27	90.0
1850	-217.59	1.339	67.9	$P2_1/c$	1	13.4	18.60	11.63	90.0	74.10	90.0
444	-217.46	1.374	69.5	$P\bar{1}$	1	10.4	10.41	14.29	101.	78.72	69.0
9156	-217.27	1.264	64.3	$P\bar{1}$	1	11.3	10.49	14.61	89.9	83.63	121.
3196	-217.00	1.332	67.6	$P2_1/c$	1	11.5	13.84	17.94	90.0	77.35	90.0
2002	-216.91	1.316	66.7	$P\bar{1}$	1	10.9	10.14	14.78	85.4	98.04	119.
844	-216.87	1.326	67.0	*Pbca*	1	28.5	11.98	16.46	90.0	90.00	90.0

7450	-216.85	1.312	66.4	$P2_1/c$	1	13.1	11.05	19.60	90.0	84.18	90.0
1666	-216.73	1.328	67.3	$P2_1/c$	1	8.87	31.28	11.02	90.0	66.71	90.0
2656	-216.62	1.320	66.8	$I2/c$	1	17.5	18.26	18.09	90.0	76.95	90.0
3124	-216.49	1.332	67.6	$P2_12_12_1$	1	24.0	11.42	10.21	90.0	90.00	90.0
3937	-216.40	1.330	67.6	$P2_1/c$	1	18.8	10.59	15.23	90.0	67.62	90.0
3136	-216.38	1.310	66.5	$P2_1/c$	1	10.7	17.82	14.95	90.0	87.64	90.0
3892	-216.35	1.264	64.5	$P2_1/c$	1	20.7	13.35	10.74	90.0	83.14	90.0
4171	-216.34	1.336	67.6	$Pbca$	1	18.3	27.77	10.97	90.0	90.00	90.0
1833	-216.24	1.304	66.1	$C2/c$	1	30.3	9.92	19.53	90.0	76.56	90.0
833	-216.23	1.311	66.2	$C2/c$	1	18.0	12.90	24.46	90.0	92.90	90.0
2384	-216.16	1.307	66.1	$P2_1/c$	1	10.8	11.93	22.23	90.0	82.31	90.0
31	-216.03	1.329	67.4	$P2_1/c$	1	20.6	8.72	16.32	90.0	73.11	90.0
4819	-216.01	1.300	66.1	$P\bar{1}$	1	10.1	10.46	15.45	95.8	78.13	66.0
8710	-215.94	1.330	67.2	$P2_12_12_1$	1	11.4	10.89	22.57	90.0	90.00	90.0
773	-215.91	1.340	67.5	$P2_1/c$	1	12.1	14.14	16.47	90.0	79.74	90.0
102	-215.91	1.321	67.0	$Pbca$	1	18.9	18.32	16.23	90.0	90.00	90.0
882	-215.85	1.310	66.1	$C2/c$	1	16.4	11.62	33.17	90.0	63.68	90.0
119	-215.85	1.335	67.7	$C2/c$	1	14.5	15.20	25.60	90.0	81.25	90.0
316	-215.80	1.347	68.1	$P2_1/c$	12	23.2	16.37	9.06	90.0	53.52	90.0
1825	-215.76	1.320	66.8	$P2_1$	2	10.8	26.55	11.09	90.0	118.27	90.0
9140	-215.73	1.307	66.2	$P\bar{1}$	1	10.0	10.86	15.46	74.4	78.41	61.5
7458	-215.52	1.317	66.6	$P\bar{1}$	1	15.0	11.51	11.36	62.1	88.19	119.
12	-215.50	1.334	67.0	$P2_1/c$	1	13.7	17.09	12.80	90.0	68.24	90.0
911	-215.38	1.282	64.8	$P2_1/c$	1	10.7	12.28	22.01	90.0	89.53	90.0
836	-215.32	1.314	66.0	$P2_1/c$	1	10.4	15.05	20.99	90.0	120.60	90.0
1798	-215.28	1.305	66.1	$P2_1/c$	1	17.2	11.64	14.30	90.0	86.65	90.0
1246	-215.19	1.318	66.5	$P2_1/c$	1	10.8	21.88	11.98	90.0	94.28	90.0

Chapter 4: Crystal Structure Prediction of mebendazole

4.1 Introduction

4.1.1 CSP as a complement to pharmaceutical solid form screening

In the previous chapter, the ability of CSP methodologies to predict the crystal structure of organic molecules under Blind Test[1] conditions was assessed. However, CSP studies on pharmaceuticals do not generally take place under these conditions, as their main purpose is to aid experimental solid form screens that are performed in parallel with computational analyses.[2-4] Solid form screens are by their nature limited, as they cannot cover any possible crystallisation condition;[2, 4-6] a detailed discussion on this subject can be found in Chapter 1.1.3. Computational modelling can effectively complement experimental studies in several ways. The most intuitive contribution of CSP is to reassure manufacturers that the most thermodynamically stable form is known,[5, 7-9] minimising the risk of ritonavir-like catastrophes.[10] In some cases other competitive polymorphs can be proposed, ideally with desirable properties (solubility, morphology, porosity, *etc*.) either stable or metastable with the respect to the known form/s. Strategies to crystallise further polymorphs can sometimes be identified, *e.g.* seeding,[11] templating,[12, 13] crystallisation under pressure[4] or from solvents.[14] Furthermore, in the absence of single crystals, CSP data can help solve crystal structures from powder X-ray diffraction (PXRD) data,[15-17] nuclear magnetic resonance (NMR)[18-20] or electron diffraction.[21] Finally, a joint analysis of the findings of a CSP study and of an experimental polymorph screen can enhance the understanding of the crystallisation behaviour of a molecule: it can rationalise the presence of disorder, establish a link between its single-component behaviour and the formation of solvates and salts, or explain the role of kinetics in the formation of metastable crystal structures.[2, 5] All these applications are of extreme value to pharmaceutical manufacturers,[2, 3] and this is leading to an increasing industrial interest in CSP (see also Chapter 1.2.1) and to the development of several in-house CSP methodologies, *e.g.* the one in ref.3.[3]

This chapter describes a CSP study undertaken with the purpose of aiding an academic experimental solid form screen by two UCL collaborators (Merina Corpinot and Dr. Krešo Bučar) on the drug molecule mebendazole.

4.1.2 Properties of mebendazole and of its known experimental forms

Mebendazole ($C_{16}H_{13}N_3O_3$),[22] trade name Vermox,[23] is an antihelminthic (*i.e.* against worms) drug molecule used to treat infestations by a broad range of parasites (ascaris,

threadworms, hookworms, whipworms, pinworms, roundworms, tapeworms, etc.).[24-26] Recently, it has also shown some interesting anti-cancer properties against various forms of tumours.[22] Infestations by parasites are a major problem in poor countries with lower hygiene standards. First synthesised in 1971,[27] it is included in the WHO list of essential medicines[28] and produced for the international market by generic manufacturers. It is taken in the form of oral capsules, and it has a low bioavailability (only 10%)[25] due to the poor solubility of its single-component forms. There have been some successful attempts to improve its solubility via co-crystallisation,[25] formation of a nitrate salt,[29] promoieties[30] and dispersion in oils.[31]

Three single-component crystal structures were known at the beginning of this joint experimental-computational study: forms A, B and C.[24, 25, 29, 30, 32] Only forms A and C have solved crystal structures deposited in the CSD,[33] with refcodes TUXPEJ[24] and YULGIW[34] respectively. These two forms are not polymorphs since they are constituted of two different tautomers of mebendazole[32] (from now on the A-tautomer and C-tautomer respectively, see Figure 4.1), as the protons in the imidazole rings are bonded to different nitrogen atoms. Defining crystal structures containing different isomers as polymorphs is incorrect, although some ambiguity does exist for tautomers as they can rapidly interconvert in solution or in the melt.[35] Indeed, form C of mebendazole is known to convert to more stable form A in the solid-state, implying that the tautomers are closely related and can easily interconvert.[36] However, as mentioned in Chapter 1.1.1, in this thesis a strict definition of polymorphism, which is also used in the CSD,[37] is utilised. This strict discrimination is necessary in this chapter as CSP methods do not allow tautomeric interconversion.

Figure 4.1: Chemical diagrams of (left) the A-tautomer of mebendazole (right) the C-tautomer of mebendazole

The crystal structure of form A (TUXPEJ) was determined via synchrotron PXRD on a commercial sample at room temperature (~300K).[24] Form A has one molecule in the asymmetric unit cell (*i.e*. Z'=1), it is in the $P\bar{1}$ space group, it has a density of 1.402 g/cm^3 and a 70.7% packing coefficient. Form C (YULGIW) was characterised via single-crystal X-ray diffraction at 150 K, and it was crystallised from a methanolic solution.[34] It is a Z'=1 crystal structure in the $P\bar{1}$ space group, with a density of 1.444 g/cm^3 and a 71.9% packing coefficient.

The two fully-characterised crystal structures of mebendazole have different intermolecular hydrogen bond motifs. In form C each molecule uses one hydrogen bond

donor and one acceptor (N acceptor in the imidazole group and NH in the amide group, with the $R_2^2(8)$ graph set motif), while in form A two donors and two acceptors per molecule are utilised, as there is an additional hydrogen bond between the O acceptor in the benzophenone group and NH in the imidazole ring with the $R_2^2(14)$ graph set motif. The two intermolecular hydrogen bond motifs of forms A and C are shown in Figure 4.2.

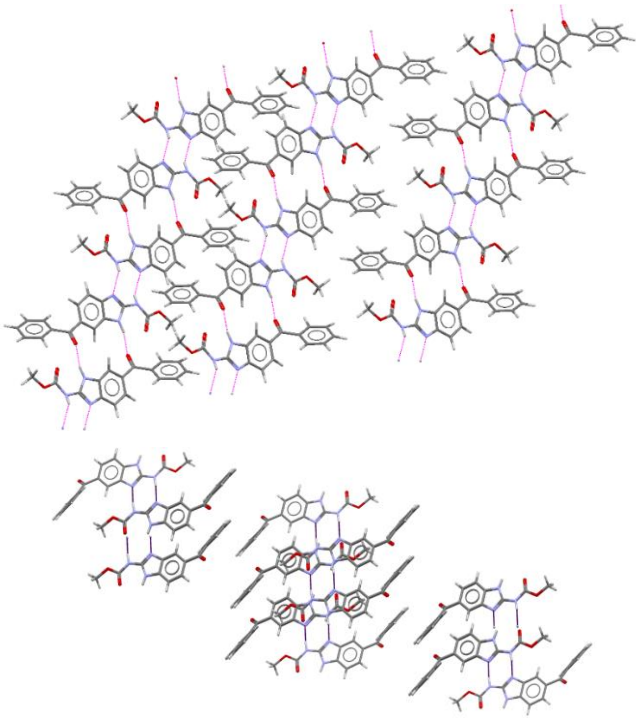


Figure 4.2: Intermolecular hydrogen bond motif of (above) mebendazole form A, where each molecule uses two donors and two acceptors, forming both a NH···O hydrogen bond with the $R_2^2(14)$ graph set motif and a NH···N hydrogen bond with the $R_2^2(8)$ graph set motif, and (below) mebendazole form C, where each molecule uses one donor and one acceptor, forming a NH···N hydrogen bond with the $R_2^2(8)$ graph set motif. The hydrogen bonds are coloured in purple.

Form B has only been characterised via PXRD and no solved crystal structure of this form exists.[38] The most soluble form is B, followed by C and A.[24, 36] This, combined with the well-known tendency of form C to transform into form A[36] and with DSC data showing that B has the lowest melting temperature and A the highest,[39] indicates that A is the most thermodynamically stable single-component form of mebendazole (probably because of its more extensive hydrogen bonding network, which compensates the lower density), followed by C and B. Metastable form C is the commercial form because it has the best trade-off between solubility and stability;[39, 40] form B is known to be toxic, while the solubility of A is so low that it has no antihelminthic effects when it is alone or when it is present in concentrations above 30% in mixtures with different forms.[36, 40] Hence, the transformation of tablets of form C into thermodynamically more stable form A at the environmental conditions typical of poor tropical countries is an important problem as it decreases the shelf-life and the efficacy of this drug.[36]

4.1.3 Reasons for this joint computational-experimental solid form screen

This CSP study was undertaken in parallel to the experimental effort, with the aim of determining whether the most stable form of mebendazole was known and if additional forms were plausible. This could prove important as the known forms possess sub-optimal properties. The hoped outcome was to find a more soluble polymorph of mebendazole that is stable under environmental conditions. Furthermore, this study also aimed at finding and solving the crystal structure of form B of mebendazole, to determine whether it is a different unique form or mixture of various phases. Finally, the polymorph screen also aimed at improving the understanding of the overall solid-state behaviour of mebendazole.

The most thermodynamically stable computer-generated crystal structures were provided to the experimental collaborators to complement their effort in some of the ways described in Chapter 4.1.1. Furthermore, the results of this additional CSP study on a flexible pharmaceutical molecule were used to benchmark alternative methods in Chapters 5-7.

4.2 Methods

4.2.1 Analysis of conformational flexibility

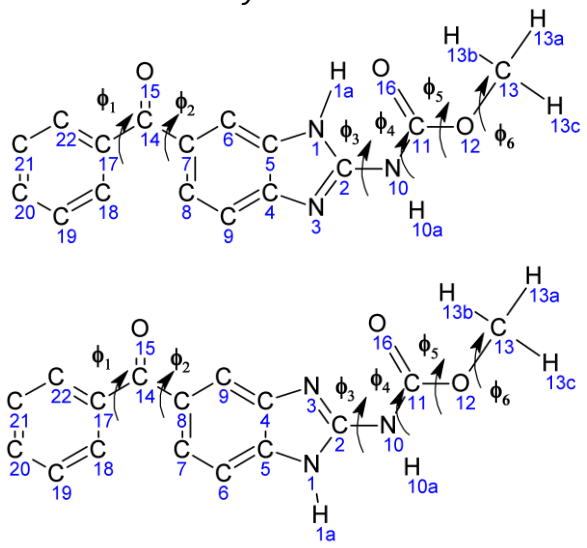


Figure 4.3: Chemical diagram of the (above) A-tautomer (below) C-tautomer of mebendazole. The two tautomers are differentiated by where C14=O15 in the edge benzoyl substituent attaches to the central benzimidazole ring: in the A-tautomer it attaches three bonds away from the N1-H1a group, to atom C7, while in the C-tautomer it attaches four bonds away, to atom C8. The six torsion angles that were considered in the initial analysis of conformational flexibility are indicated (see Appendix Table 4.1 for their definition). The scans were only performed on the C-tautomer.

The first step of this CSP study was to determine the range of conformations that the two tautomers of mebendazole are likely to take in the solid-state. A set of isolated molecule

optimisations combined with chemical intuition led to the identification of the six torsion angles in Figure 4.3 as those that are the most flexible and so could deviate the most in the crystal structure/s from the isolated molecule conformational energy minima.

For the purpose of analysing conformational flexibility only the C-tautomer was considered. This is because the two tautomers are distinguished only by the position of the edge benzoyl substituent (C14-C21) relative to proton H1a, as explained in the caption to Figure 4.3. Since the flexible torsion angles in the benzoyl substituent are well-separated from the remaining ones, the fragments they define can be assumed not interact strongly with one another. Hence in the methodology used in this CSP study, where most torsion angles were analysed independently in the initial assessment of the conformational flexibility and surrogate molecules were used to estimate ΔE_{intra} values in the searches (see Chapter 4.2.2), the main torsion angles could be assumed to have indistinguishable energy profiles in the two tautomers. This assumption is confirmed by the gas-phase minimum conformer of the A-tautomer being only ~0.5 $kJ \cdot mol^{-1}$ higher in energy than that of the C-tautomer. Note that the definitions of the torsion angles in Figure 4.3 (see Appendix Table 4.1) differ only for Φ_1 and Φ_2, and this is due to the different position of the benzoyl substituent.

The conformational energy penalties for varying torsion angles Φ_1-Φ_6 relative to the isolated-molecule global minimum were calculated on the C-tautomer via constrained 1-dimensional angle scans at the PBE0 6-31G(d,p) level of theory performed with Gaussian09.[41] The values that were obtained were deemed to be valid also for the A-tautomer. The procedure for performing angle scans is described in Chapter 3.2.1.1. All torsion angles were optimised from 0° to 360° in 30° steps, with the exception of the amide torsion angle Φ_4 that was scanned in 90° steps from 0° (*i.e.* its *cis* configuration) to 180° (*i.e.* its *trans* configuration). Each scan started from the global minimum in isolated-molecule conformational energy determined at the PBE0 6-31G(d,p) level of theory, since this is the most suitable starting point for angle scans on flexible molecules.[42] Angles Φ_1 and Φ_2 were scanned together since they are adjacent and strongly interacting. The results of the torsion angle scans are shown in Figure 4.4.

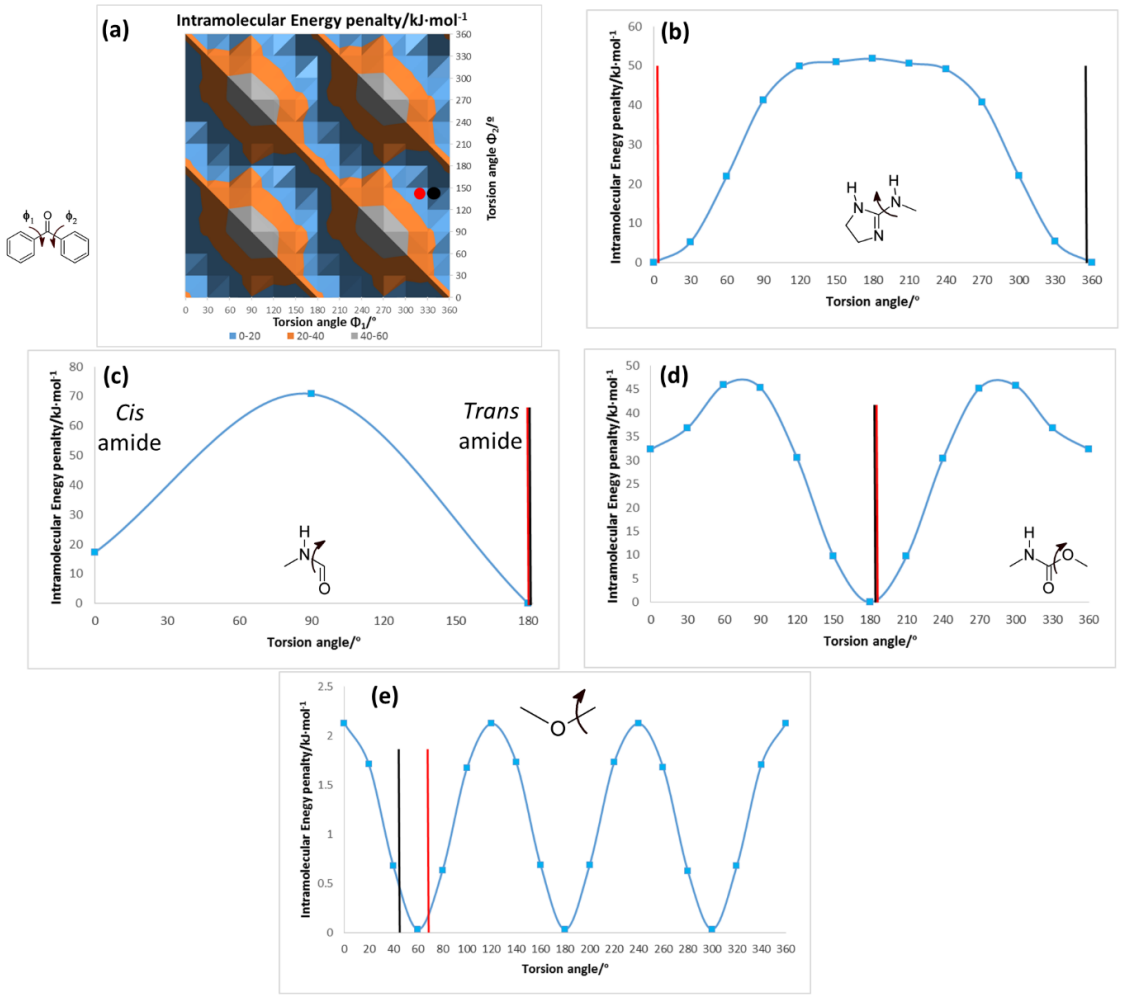


Figure 4.4: Results of the angle scans of torsion angles a) Φ_1 (x-axis) and Φ_2 (y-axis) from 0° to 360° in 30° steps; this is a contour plot of the conformational energy penalty surface as a function of the values of these two torsion angles b) Φ_3 from 0° to 360° in 30° steps c) Φ_4 from 0° (*cis*) to 180° (*trans*) in 90° steps d) Φ_5 from 0° to 360° in 30° steps and e) Φ_6 from 0° to 360° in 30° steps. At each point on the plots, all the conformational degrees of freedom (CDFs) were relaxed with the exception of the scanned torsion angle/s. The optimisations were performed at the PBE0 6-31G(d,p) level of theory starting from the PBE0 6-31G(d,p) optimised global minimum gas-phase conformer. The red dots/lines indicate the values taken by the torsion angle/s in the conformation of form C of mebendazole, black dots/lines in the conformation of form A (see Appendix Table 4.2 for the actual values).

The results of the constrained angle scans indicate the geometric preferences of mebendazole. No CSD survey was performed, as the plots in Figure 4.4 were deemed sufficient to confidently limit the conformational search space. However, the analysis of the CSD conformational preferences of dihedrals Φ_1-Φ_5 was performed in Chapter 5.

The confidence in these results was enhanced by both experimentally known forms taking values in low-energy regions, with conformational energy penalties lower than 10 kJ·mol^{-1}. The scan of the torsion angle describing the rotation of the methyl group (i.e. angle Φ_6, Figure 4.4e) shows that its variation has a negligible effect on

conformational energy, which confirms the findings of previous analyses on this functional group.[43]

4.2.2 Crystal structure generation

An analysis of the isolated molecule angle scans suggested that the most effective way to generate crystal structures for mebendazole was to perform flexible searches on both tautomers. In particular, it was decided to treat only torsion angles Φ_1, Φ_2, Φ_3 and Φ_5 as explicitly flexible in the searches. On the other hand, the amide torsion angle Φ_4 was constrained in both tautomers at two values corresponding to its *cis* (*i.e.* 0°) and *trans* (*i.e.* 180°) configurations, while the methyl torsion angle Φ_6 was constrained at the values of the respective isolated-molecule minima in conformational energy because of its negligible effect on energy and intermolecular interactions (this is a common practice in CSP).[43-45]

Hence four searches were carried out: for each tautomer two searches were performed with the amide torsion angle Φ_4 in the *cis* and *trans* configurations respectively. The searches were performed with CrystalPredictor 1.8, which estimates ΔE_{intra} interpolating grids of *ab initio* calculated energies for the four explicitly flexible CDFs (see Chapter 2.4.1.2).[46] The reason why this older version of CrystalPredictor was utilised was the same as for the Blind Test (see Chapter 3.2.2), *i.e.* the immense local approximate models (LAMs) that would need to be built for use with CrystalPredictor 2.[47] Also, for mebendazole this problem was exacerbated by the fact that the LAMs would have had to be calculated separately for each tautomer, further increasing the computational cost, while the ΔE_{intra} grids required in CrystalPredictor 1.8 could be assumed to be valid for both tautomers since the energy differences for varying the main torsion angles were deemed to be indistinguishable, as explained in Chapter 4.2.1.

In order to produce the grids at a reasonable computational cost mebendazole was broken down into surrogate molecules. This assumption is generally accurate as long as the torsion angles that define the position of strongly interacting groups are scanned together.[43, 48, 49] Hence, the two surrogate molecules shown in Figure 4.5, which were deemed to be valid for both tautomers, were considered. One contains torsion angles Φ_1 and Φ_2, and the other one contains torsion angles Φ_3 and Φ_5. Hydrogen atoms were added at the edges of each group to avoid the presence of unphysical free bonds.

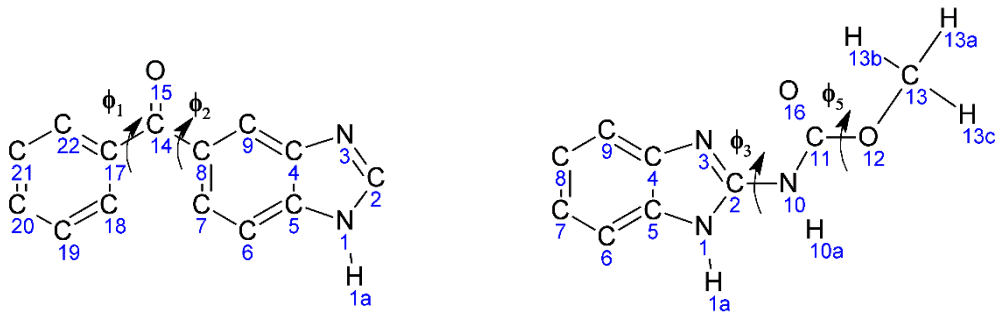


Figure 4.5: Surrogate molecules used to calculate the ΔE_{intra} grids of (left) Φ_1 and Φ_2 and (right) Φ_3 and Φ_5 of both tautomers of mebendazole.

For each surrogate molecule, the ΔE_{intra} grids were calculated with Gaussian09 at the PBE0 6-31G(d,p) level of theory; for the surrogate molecule containing torsion angles Φ_3 and Φ_5 two separate grids were calculated, one with the amide group in the *cis* configuration and one in the *trans*. The grid ranges are shown in Table 4.1, and they include all values of the four torsion angles with a conformational energy penalty smaller than 25 kJ·mol^{-1} (see Figure 4.4).

Table 4.1: Dimensionality of the ΔE_{intra} grids used to perform the crystal structure searches with CrystalPredictor. The grids were calculated from the surrogate molecules in Figure 4.5.

First surrogate molecule					
Torsion angle label	**Torsion Angle Definition**	**Minimum search range/°**	**Maximum search range/°**	**Step Size/°**	**Steps Number**
Φ_1	C18-C17-C14-C7	90	270	20	10
Φ_2	O15-C14-C7-C8	0	340	20	18
				Number of Grid Points	180
Second surrogate molecule					
Torsion angle label	**Torsion Angle Definition**	**Minimum search range/°**	**Maximum search range/°**	**Step Size/°**	**Steps Number**
Φ_3	N1-C2-N10-C11	285	85	20	9
Φ_5	N10-C11-O12-C13	125	245	20	7
				Number of Grid Points	63 x 2

Each of the four searches generated 1,000,000 structures with Z`=1, in the 59 most common space groups in the CSD (listed in the Appendix Table 4.3), for a total of 4,000,000 generated structures. In the searches, U_{inter} was estimated as the sum of an electrostatic component calculated from fixed point charges computed at the PBE0 6-31G(d,p) level of theory on the isolated-molecule conformers of both tautomers and an exp-6 repulsion-dispersion component from the empirically fitted FIT potential.[50] No search was performed with Z'>1 to keep the overall cost manageable. This seemed a sensible approximation as both known forms are Z'=1 and mebendazole does not have characteristics that correlate with a tendency of forming crystal structures with more than one molecule in the asymmetric unit cell, as it is a relatively large and flexible achiral molecule.[51]

4.2.3 Refinement of the generated crystal structures

The final refinement was performed on all the crystal structures within 15 kJ·mol^{-1} of the global minimum in CrystalPredictor E_{latt} for each tautomer. This relatively small energy window was utilised as the CrystalPredictor energy model, albeit simple, successfully located both known forms as global minima in E_{latt} for each tautomer, suggesting that the atomic point charges+FIT model produced a reliable energy ranking. A total of 1,819 crystal structures were present within this energy window: 855 for the A-tautomer with the *trans* amide group, 932 for the C-tautomer with *trans* amide and 32 for the C-tautomer with *cis* amide; no generated crystal structure with the A-tautomer and the *cis* amide was low enough in energy to be taken forward. As mebendazole is a flexible molecule, both U_{inter} and ΔE_{intra} had to be optimised in the final refinement stage of CSP, and CrystalOptimizer[52] (see Chapter 2.4.2.2) was used for this purpose. Following the success of the Blind Test CSP study, the same hierarchal approach was utilised to limit the overall computational expense: first structures underwent an intermediate optimisation and re-ranking with a single iteration of CrystalOptimizer, and the most stable ones were then fully optimised.

4.2.3.1 Determination of the independent CDFs

In order to guarantee an accurate final refinement of the generated crystal structures, it is important to treat as explicitly flexible not only those torsion angles that are treated as such in the searches (either as explicitly flexible, *i.e*. Φ_1-Φ_3 and Φ_5 for mebendazole, or as defining separate conformational regions, *i.e.* Φ_4), but also other torsion and bond-angles that can impact the balance between intra- and intermolecular interactions.[43, 49, 53] On the other hand, bond-lengths can generally be safely ignored, as they are not significantly affected by the crystalline environment.[54] Ideally, all conformational CDFs should be treated as independent in a fully atomistic optimisation, but this is often too computationally expensive as the cost of CrystalOptimizer minimisations increases drastically with the number of explicitly flexible CDFs.[54] Hence, a limitation is required (all these issues are discussed in detail in Chapter 6).

In the Blind Test (see Chapter 3.2.2), the independent CDFs were limited to the seven torsion angles of molecule XXVI that are intuitively flexible. Although this approach was effective, it is not ideal, because it can lead to neglect some CDFs that can improve the balance between ΔE_{intra} and U_{inter}. Hence a more comprehensive approach was utilised for determining the independent CDFs of mebendazole, which was affordable because of its smaller size compared to XXVI: for each tautomer, the 20 most competitive crystal structures in CrystalPredictor E_{latt} were fully optimised with CrystalOptimizer considering each torsion and bond-angle as an independent CDF.

Those torsion and bond-angles whose values changed by more than 5º and 1º respectively in one or more optimisation were considered to be sensitive to packing forces and so were selected for treatment as independent CDFs in the final refinement stage of this CSP study. It is interesting to note that although the selected CDFs are very similar for the two tautomers, there are some differences, as shown in Figure 4.6. These differences are probably due to the different point at which C14=O15 in the edge benzoyl group attaches to the benzimidazole rings, causing variations in intermolecular interactions that change the degree to which each CDF is affected by the crystalline environment.

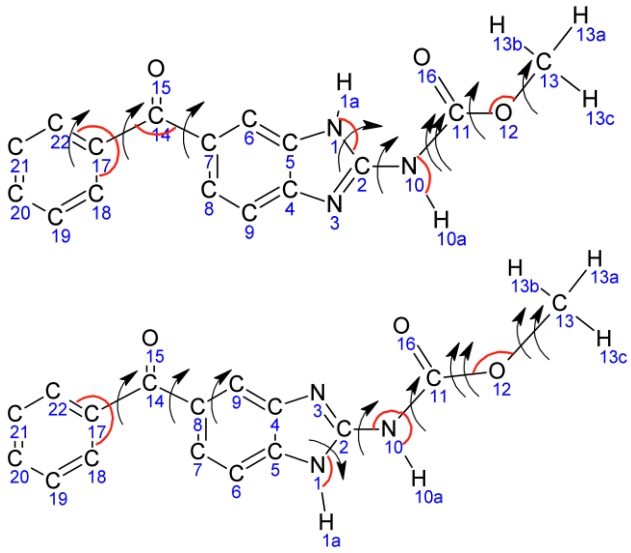


Figure 4.6: Chemical diagram and atomic numbering of (above) the A-tautomer (below) the C-tautomer of mebendazole, showing the torsion angles (black arrows) and bond-angles (red arcs) treated as independent CDFs in the final refinement stage of this CSP study. Double arrows indicate that two definitions of torsion angles around the same central bond were treated as variables in the CrystalOptimizer optimisations. See Appendix Table 4.4 for the precise definition of the explicitly flexible torsion and bond-angles.

4.2.3.2 Intermediate optimisation of the generated crystal structures

The 855 crystal structures within 15 kJ·mol^{-1} of the CrystalPredictor lattice energy global minimum of the A-tautomer of mebendazole, and the 964 of the C-tautomer, underwent the intermediate optimisation with the single-iteration of CrystalOptimizer. ΔE_{intra} was calculated optimising the molecular conformations at the PBE0 6-31G(d,p) level of theory with Gaussian09 as a function of the torsion and bond-angles shown in Figure 4.6, while U_{inter} was determined with DMACRYS,[55] modelling it as a sum of electrostatic component from distributed multipoles calculated with GDMA 2.2[56] and a repulsion-dispersion component calculated with the empirically fitted FIT potential. All the *ab initio* calculations

were stored in local approximate models (LAMs) databases of both tautomers, and re-utilised in later stages of the CSP procedure to limit the computational expense.

Once the single-iterations were complete, COMPACK[57] was used to remove duplicates. Structures were considered as duplicates if they had an energy difference smaller than 2.5 $kJ \cdot mol^{-1}$, a density difference smaller than 0.05 $g \cdot cm^{-3}$, and if it was possible to overlay 30/30 molecules, with a 20% distance tolerance and a 20° angles tolerance, with a root mean square deviation ($RMSD_{30}$) below 0.65 Å.

4.2.3.3 Final optimisation of the most promising crystal structures

All unique structures within 20 $kJ \cdot mol^{-1}$ of the E_{latt} global minimum for each tautomer after the intermediate minimisations were then fully optimised. A larger energy window was applied than the 15 $kJ \cdot mol^{-1}$ one after the searches. This was done because the single- iterations of CrystalOptimizer and the successive clustering step had spread-out the energy values: a smaller number of low-energy crystal structures were present after the intermediate optimisations than after the searches, making the consideration of a larger energy window feasible. Furthermore, after all the minimisations were completed the variation in E_{latt} between the intermediate and the full optimisations was analysed in order to verify whether the 20 $kJ \cdot mol^{-1}$ window was sufficient. Since the maximum change in energy was ~15 $kJ \cdot mol^{-1}$ for structures with the C-tautomer and ~17.5 $kJ \cdot mol^{-1}$ for those with the A-tautomer, the 20 $kJ \cdot mol^{-1}$ threshold was deemed to be appropriate.

A total of 198 crystal structures with the A-tautomer and 155 with the C-tautomer were fully optimised with CrystalOptimizer at the PBE0 6-31G(d,p) level of theory for both ΔE_{intra} and charge density calculations, treating all torsion and bond-angles in Figure 4.6 as independent CDFs. A few structures were not true E_{latt} minima, and their symmetry was reduced to more stable Z'=2 structures. Only one crystal structure with the amide group in its *cis* configuration was taken to this final optimisation stage: this indicates that these structures were not competitive, mainly because *cis* amides are associated with large conformational energy penalties (~17 $kJ \cdot mol^{-1}$, see Figure 4.4c).

Once again, COMPACK was used to cluster the fully optimised crystal structures of both tautomers and remove duplicates. The clustering parameters were identical to those used after the intermediate optimisations (see Chapter 4.2.3.2), with the exception that a maximum energy difference of 2.85 rather than 2.5 $kJ \cdot mol^{-1}$ was allowed for two structures to be considered as duplicates. After the single-iterations of CrystalOptimizer a smaller energy difference had been permitted to avoid the removal of structures that could have converged to different minima at later stages of the CSP process.

4.2.4 Estimate of the effect of polarisation

After all the optimisations were completed, the effect of polarisation on the energy ranking of the crystal structures was estimated. In order to do so, the E_{latt} values of the fully optimised unique crystal structures were re-determined calculating both ΔE_{intra} and the charge density in a PCM with ε=3,[58] and then optimising U_{inter} with DMACRYS. The same level of theory, PBE0 6-31G(d,p), and repulsion-dispersion potential, FIT, as for the CrystalOptimizer optimisations were utilised.

In this CSP study, no estimate of free energy was performed, as this generally has a small effect on relative rankings. Indeed in Chapter 7 free energies were calculated for a few low energy crystal structures of mebendazole, and they were shown to have a very limited effect on the relative thermodynamic stabilities.

4.3 Results and Discussion

4.3.1 Crystal structure search and intermediate optimisation of the generated structures

The CSP-generated crystal structures were named depending on the tautomer they contained and on the ranking after the CrystalPredictor search. For example, structure A70 was the 70th crystal structure in CrystalPredictor E_{latt} ranking for the A-tautomer. The few crystal structures containing *cis* amide groups for the C-tautomer that were taken to the refinement stage are named CisC followed by their ranking in the respective search.

The E_{latt} global minima of both tautomers after the searches (*i.e.* structures A1 and C1) were close matches to the two known experimental forms. Using the Crystal Packing Similarity tool with its standard settings, for A1 it is possible to overlay 15/15 molecules with experimental form A (refcode TUXPEJ), with an $RMSD_{15}$ of 0.458 Å, while for C1 and form C (refcode YULGIW) the same overlay produces an $RMSD_{15}$ of 0.367 Å. This result was much better than that obtained at the search stage for molecule XXVI (see Chapter 3.3.1), and this was probably due to both known forms having conformations with low-energy values for all the torsion angles that were considered as independent CDFs (see Figure 4.4). Indeed, both experimental conformations are very similar to the closest optimised isolated-molecule minima, as shown in Appendix Figure 4.1. Although limiting the search to isolated-molecule minima would have probably produced both known forms, this would have also limited the range of predicted polymorphs. Several duplicates of A1 and C1, matching the known experimental forms, were also generated at higher energies. This suggests that these wells on the lattice energy surface are very broad, with several minima corresponding to almost identical crystal structures.[59] Indeed, Chapter 8.3.3 shows how clustering search-generated structures would have removed several duplicates.

The intermediate optimisation produced similar results, although some of the duplicates mentioned above optimised to energies lower than those of A1 and C1. Since the clustering procedure only keeps the lowest energy crystal structure among those that are deemed to belong to the same cluster, both A1 and C1 were removed as duplicates. However, a structure matching experimental form A, A14, was ranked 1st for the crystal structures containing the A-tautomer, and it had a 15/15 molecule overlay with TUXPEJ with an $RMSD_{15}$ of 0.285 Å. A structure matching experimental form C, C5, was ranked 2nd for the for the crystal structures containing C-tautomer (only 0.4 kJ·mol^{-1} above the global minimum at this stage, C248), having a 15/15 molecule overlay with YULGIW with and $RMSD_{15}$ of 0.282 Å. The improvements in the overlays were due to the more accurate energy model of CrystalOptimizer compared to CrystalPredictor. The intermediate optimisations spread-out the relative E_{latt} values, which led to the exclusion of many non-competitive crystal structures, limiting the overall computational expense.

4.3.2 Combined crystal energy landscape of both tautomers

4.3.2.1 Crystal energy landscape obtained after the full optimisations

A total of 353 crystal structures were fully optimised with CrystalOptimizer, which converged to 220 distinct minima. The 211 crystal structures within 20 kJ·mol^{-1} of the global minimum are plotted in Figure 4.7, which summarises the combined crystal energy landscapes of both tautomers of mebendazole; more details can be found in Appendix Table 4.5. All the ΔE_{intra} values were calculated relative to the gas-phase minimum energy of the C-tautomer, which is lower than that of the A-tautomer by ~0.5 kJ·mol^{-1}.

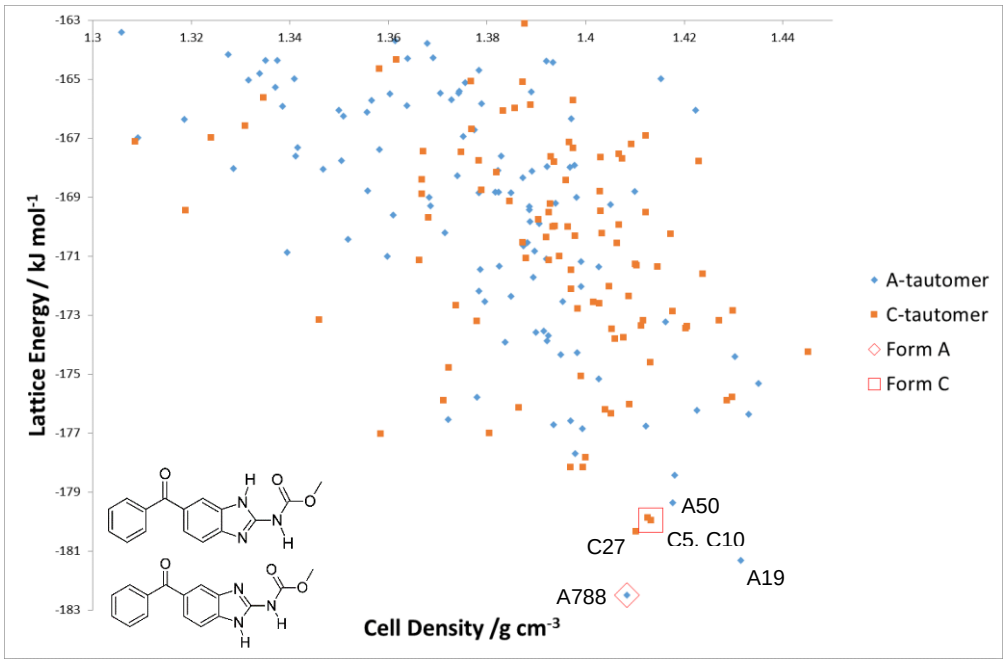


Figure 4.7: Plot summarising the combined crystal energy landscape of both tautomers of mebendazole. See Appendix Table 4.5 for more details. Some of the low-energy plausible structures are labelled, and those matching the known forms are indicated.

The combined crystal energy landscape shows that structure A788, a match to form A, was the global minimum in E_{latt}, ~1 kJ·mol^{-1} more stable than the closest competitor (A19). On the other hand structure C5, a match to form C, was ranked 4th, ~2.5 kJ·mol^{-1} above A788 and ~0.4 kJ·mol^{-1} above C27, the lowest energy predicted crystal structure among those containing the C-tautomer. The overlay between A788 and form A is accurate, with an $RMSD_{15}$ of 0.300 Å and an $RMSD_1$ of 0.113 Å. The overlay between C5 and form C is also very good, with an $RMSD_{15}$ of 0.276 Å and an $RMSD_1$ of 0.052 Å. Both overlays are shown in Figure 4.8.

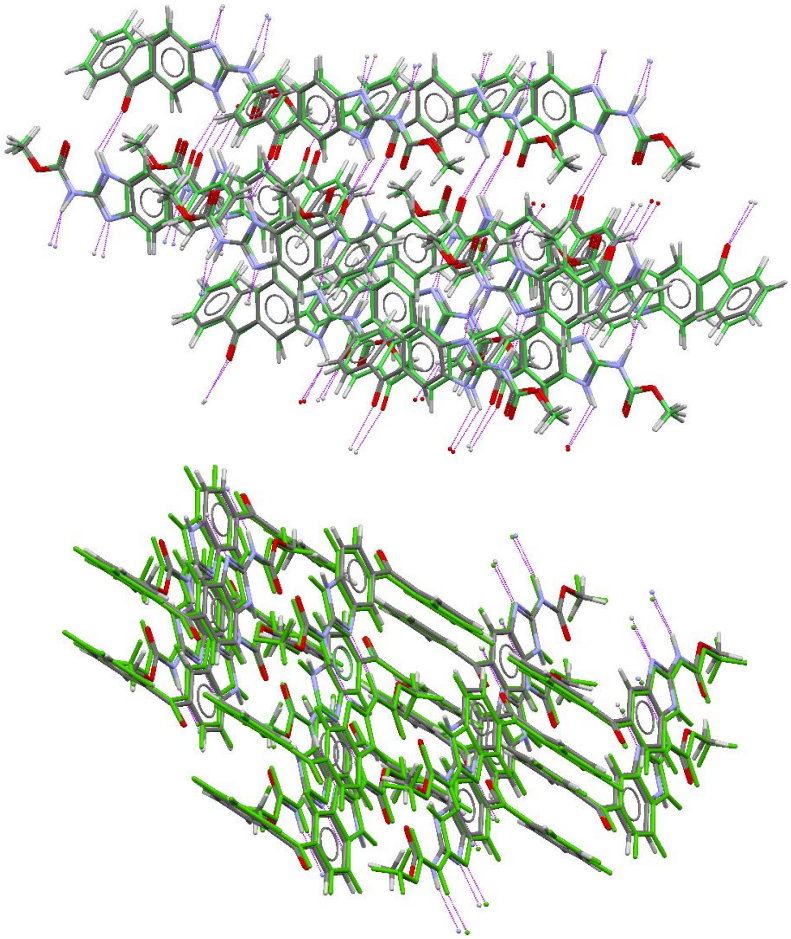


Figure 4.8: 15-molecule overlay between (above) the experimental crystal structure of mebendazole form A, coloured by elements, and structure A788, in green, with an $RMSD_{15}$ of 0.300 Å and (below) the experimental crystal structure of mebendazole form C, coloured by elements, and structure C5, in green, with an $RMSD_{15}$ of 0.276 Å. The hydrogen bonds are coloured in purple.

These data confirm that form A of mebendazole is more thermodynamically stable than form C, and suggest that it is unlikely that a lower energy single-component crystal structure than form A exists. Although the existence of a lower energy as yet unfound form cannot be completely ruled out, given the presence of several competitive crystal structures within the margin of error of the model, it is unlikely than it would be significantly more stable than the known ones, showing that there is little risk of a ritonavir-like phenomenon.

4.3.2.2 Estimate of the effect of polarisation on lattice energy

The results of recalculating the lattice energy of the fully minimised unique crystal structures of both tautomers (see Figure 4.7) with a PCM with ε=3 are shown in Figure 4.9. More details can once again be found in Appendix Table 4.5.

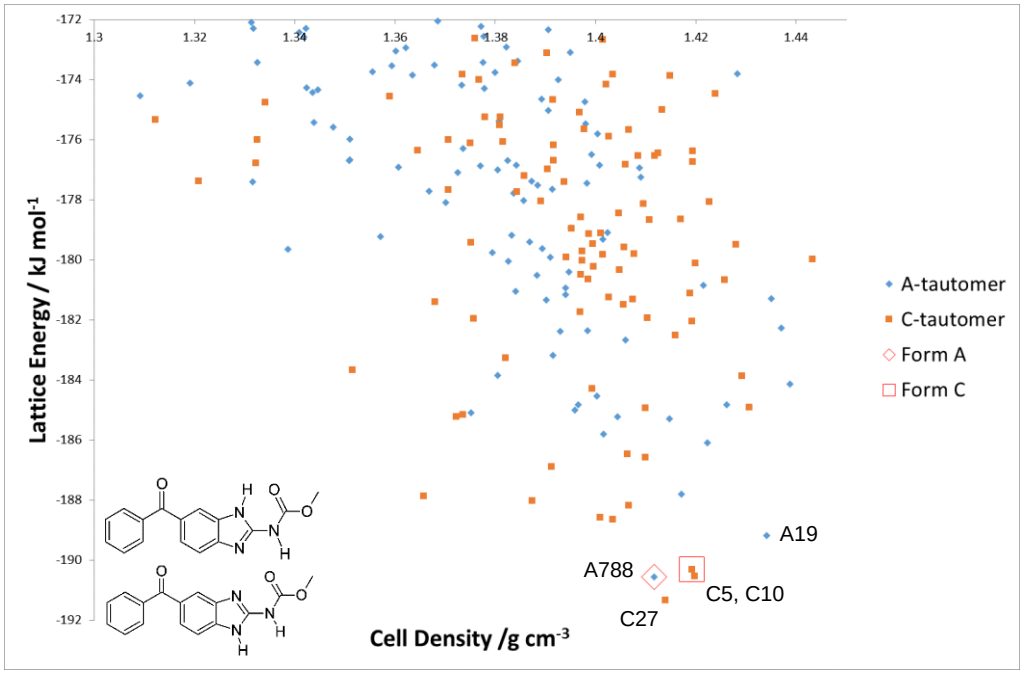


Figure 4.9: Plot summarising the combined crystal energy landscape of both tautomers of mebendazole obtained after recalculating the lattice energy of structures in Figure 4.7 with PCM. See Appendix Table 4.5 for more details. Some of the low-energy plausible structures are labelled.

The global minimum was now structure C27, ~0.8 kJ·mol^{-1} lower in energy than A788, which ranked second. A788 remained more energetically favourable than C5, although the application of the PCM decreased the energy difference to ~0.3 kJ·mol^{-1}. Structure C5 ranked 3rd among the structures containing from the C-tautomer, ~0.2 kJ·mol^{-1} less stable than structure C10 and ~1.0 kJ·mol^{-1} than structure C27.

Including an estimate of polarisation slightly stabilised the predicted crystal structures containing the C-tautomer with respect to those with the A-tautomer, while the relative rankings within the same tautomer remained very similar (this can also be observed by comparing Figure 4.10 with Appendix Figure 4.2a and Figure 4.13 with Appendix Figure 4.2b). Overall, the re-ranking caused by the application of the PCM was much smaller than in the Blind Test CSP study on molecule XXVI (see Chapter 3.3.3.2). Although these results were more at odds with experimental evidence than those shown in Figure 4.7, as the E_{latt} global minimum was an unfound crystal structure and not a match to form A, the energy differences remained well within the margins of error of the models, confirming that the existence of a significantly more stable polymorph is unlikely.

The correct experimental energy ranking of the two known forms was also maintained. Since the relative stability of the CSP-generated crystal structures was not drastically affected by the inclusion of the PCM, except a slight stabilisation of those with the C-tautomer, these calculations increased the confidence in the results.

4.3.3 Analysis of the low energy predicted crystal structures after the full optimisations

Although the results shown in Chapter 4.3.2 show that the two experimentally known forms are thermodynamically stable relative to competitors, both Figure 4.7 and Figure 4.9 indicate that there is a large number of alternative crystal structures that are plausible candidates for crystallisation if the right experimental conditions were to be obtained. Hence, an analysis of the low energy forms is very important to understand the packing behaviour of mebendazole as well as the possible range of crystalline conformations.

4.3.3.1 Crystal structures containing the A-tautomer

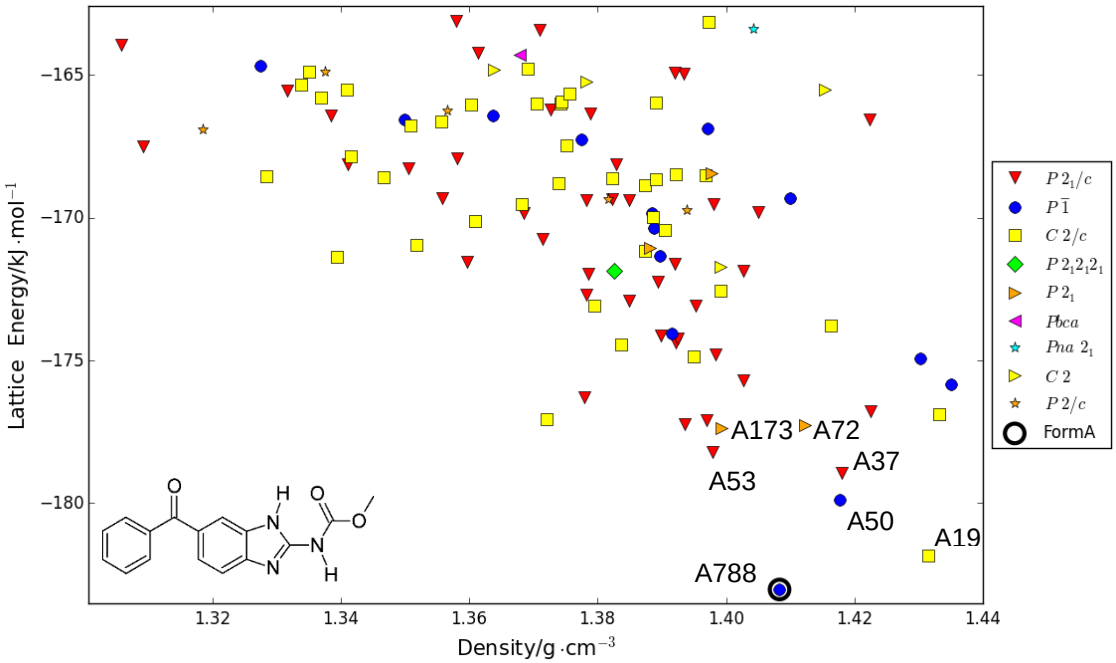


Figure 4.10: Plot summarising the crystal energy landscape of the A-tautomer of mebendazole after the full optimisations with CrystalOptimizer, showing all the crystal structures within 20 kJ·mol^{-1} of the global minimum in E_{latt}. Each point on the landscape corresponds to a separate crystal structure, labelled according to its space group. The structure matching experimental form A is circled, and some of the most competitive structures are also labelled. The crystal energy landscape calculated with the PCM is summarised in Appendix Figure 4.2a, and more details can be found in Appendix Table 4.5.

The crystal structures that contain the A-tautomer and are within 20 kJ·mol^{-1} of the global minimum in E_{latt} do not differ much in terms of molecular conformations. In particular, the only really flexible conformational degrees of freedom appear to be Φ_1, Φ_2

and Φ_6, but Φ_6 is a methyl rotation that has a minor effect on ΔE_{intra} and on the overall molecular shape. Using the criteria developed by Cruz-Cabeza and Bernstein for identifying conformational polymorphs (see Chapter 2.3.1.1),[60] the 112 crystal structures containing the A-tautomer in Figure 4.10 can be grouped into just three clusters of conformational polymorphs. The low-energy crystal structures tend to have low ΔE_{intra} values, with none exceeding 8.5 kJ·mol^{-1}. In all the low-energy crystal structures, the conformations have the benzophenone carbonyl group (C14-O15 in Figure 4.3) on the same side as the NH imidazole group (N1-H1a in Figure 4.3).

The A-tautomer of mebendazole seems to be able to pack well: all the structures within 20 kJ·mol^{-1} of the global minimum have packing coefficients between ~67% and ~73% (see Appendix Table 4.5). This means that the A-tautomer of mebendazole should not have a problem to form stable single-component crystal structures.[44]

The global minimum for the A-tautomer (A788) matches the experimental crystal structure of mebendazole form A, as shown in Figure 4.8. In experimental form A two hydrogen bond acceptors and two hydrogen bond donors are utilised (see Figure 4.2). This intermolecular hydrogen bonding motif is shared by A788 other 3 low energy structures: A90, ranked 10th among the crystal structures containing the A-tautomer in Figure 4.10, A291, ranked 11th, and A143, ranked 14th.

A total of 105/112 crystal structures in Figure 4.10 form an intermolecular hydrogen bond between the N donor in the imidazole group and the NH in the amide group, including all 26 structures in the lowest 10 kJ·mol^{-1}. In particular, 82 of these structures form this hydrogen bond with the $\mathrm{R}_2^2(8)$ graph set motif. The remaining 23 structures, including A72, ranked 7th, form this hydrogen bond with the $\mathrm{C}_1^1(4)$ graph set motif, with each donor and acceptor bonding to two different molecules. These motifs are shown in Figure 4.11.

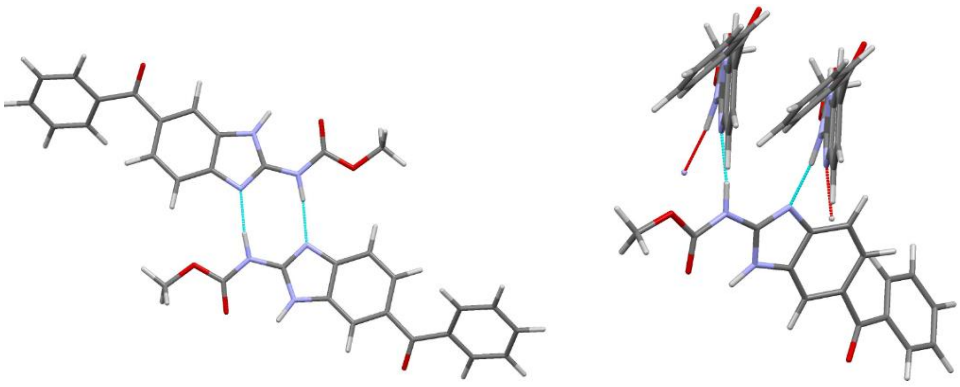


Figure 4.11: **NH···N intermolecular hydrogen bond with (left) the $\mathrm{R}_2^2(8)$ graph set motif common to 82 low-energy structures containing the A-tautomer of mebendazole (right) the $\mathrm{C}_1^1(4)$ graph set motif common to 23 low-energy structures .**

Although all three oxygen atoms in mebendazole are potential hydrogen bond acceptors, they are only used only in 47/112 low-energy crystal structures. The oxygen

atoms that are most commonly used as hydrogen bond acceptors are those in the amide and benzophenone C=O groups (*i.e.* atoms O15 and 16, see Figure 4.3); atom O12 is only utilised as an acceptor in three low-energy CSP-generated crystal structures.

Two structures were found after symmetry reduction to be Z`=2, in the $P2_1$ space group. A173, ranked 6th, is very interesting, because of its competitiveness and the unique intermolecular hydrogen bond pattern: on top of the usual NH···N hydrogen bond with the $R_2^2(8)$ graph set motif (see Figure 4.11), the imidazole NH group in one molecule in the asymmetric unit forms an intermolecular hydrogen bond with the benzophenone O acceptor of the other. This motif is shown in Figure 4.12.

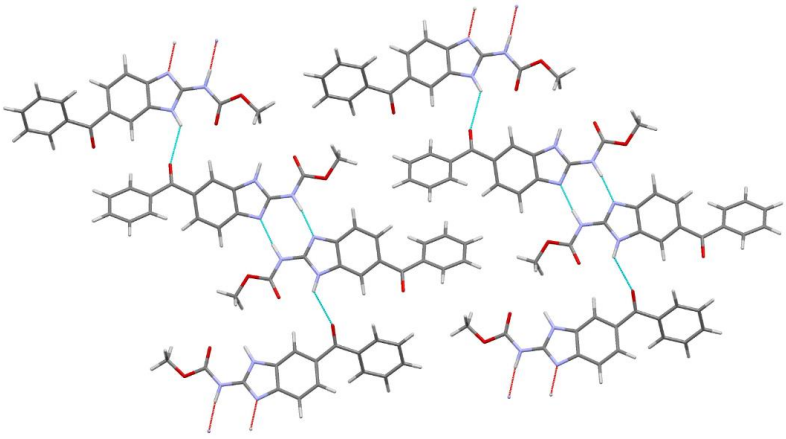


Figure 4.12: **Unique intermolecular hydrogen bond motif in A173.**

In summary, it is unlikely that the crystal structures containing the A-tautomer of mebendazole could form conformational polymorphs. The conformations in the low-energy crystal structures are all broadly similar, and have low ΔE_{intra} values.

In terms of intermolecular interactions, the most common feature of the crystal structures containing the A-tautomer of mebendazole is the NH···N intermolecular hydrogen bond, as it is present in almost all the low-energy CSP-generated structures (including A788). However, the spectrum of low-energy structures suggests that packing polymorphs of this tautomer are possible. The NH···N hydrogen bond is in fact formed with two possible graph set motifs, $R_2^2(8)$ and $C_1^1(4)$, and other hydrogen bonds using any of the three O atoms as acceptors are also present among the thermodynamically competitive crystal structures.

4.3.3.2 Crystal structures containing the C-tautomer

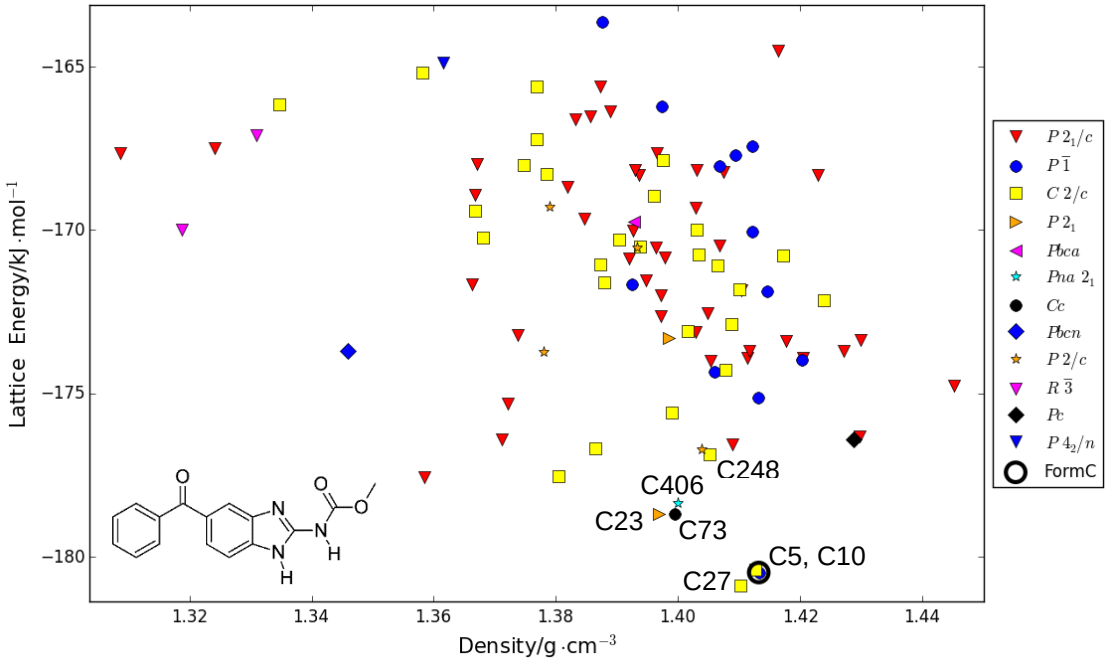


Figure 4.13: Plot summarising the crystal energy landscape of the C-tautomer of mebendazole after the full optimisations with CrystalOptimizer, showing all the crystal structures within 20 kJ·mol^{-1} of the global minimum in E_{latt}. Each point on the landscape corresponds to a separate crystal structure, labelled according to its space group. The structure matching experimental form C is circled, and some of the lowest energy structures are also labelled. The crystal energy landscape calculated with the PCM is summarised in Appendix Figure 4.2b, and more details can be found in Appendix Table 4.5.

All the 99 CSP-generated crystal structures containing the C-tautomer that were not removed as duplicates are within the bottom 20 kJ·mol^{-1} in E_{latt}. The lowest 10 kJ·mol^{-1} is more populated than for the A-tautomer, as 61 structures are present rather than 26. Five of these low-energy structures were found to be Z'=2 after symmetry reduction, though the two independent molecules did not differ much in their conformations and intermolecular interactions. They include C23, ranked 4th among the crystal structures containing the C-tautomer in Figure 4.13, C73, ranked 5th and C406, ranked 6th.

The only crystal structure with *cis* amide taken to this stage, CisC32, ranks 98th, ~16 kJ·mol^{-1} above the global minimum. Hence, putative crystal structures with *cis* amide seem not to be viable candidates for polymorphs, although CisC32 has favourable intermolecular interactions that compensate for a high ΔE_{intra} value of ~18 kJ·mol^{-1}. If the *cis* conformation were to be produced in the synthesis, CisC32 would become a serious candidate.

The predicted low energy crystal structures containing the C-tautomer of mebendazole can also pack well, with packing coefficients ranging from ~66.5 to ~74.3%

(see Appendix Table 4.5). For this tautomer the main torsional degrees of freedom are once again Φ_1, Φ_2 and Φ_6 (*i.e.* methyl rotation). In terms of conformations, the criteria developed by Cruz-Cabeza and Bernstein produce three possible clusters of conformational polymorphs. One cluster contains only CisC32, with the *cis* amide configuration. In 37/98 of the low-energy crystal structures with *trans* amide the benzophenone C=O group (C14-O15 in Figure 4.3) is on the same side as the imidazole NH (N1-H1a in Figure 4.3), while in 61/98 it is in the opposite side. From now on they will be called C-conformers 1 and 2 respectively. Although C-conformer 2 is more frequent, the lowest energy structures contain C-conformer 1. However, these geometries do not differ drastically in terms of conformational energy: no low-energy crystal structure with *trans* amide has ΔE_{intra} values above 10 kJ·mol^{-1}. Figure 4.14 shows one example for each of these two conformers.

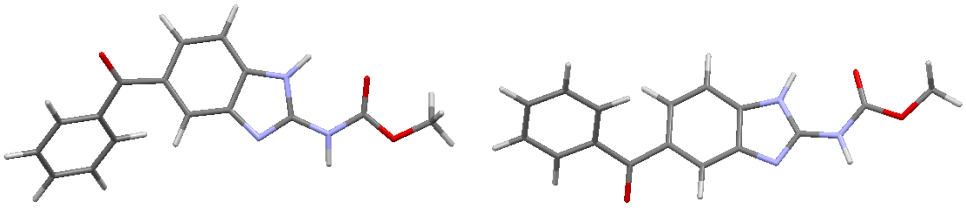


Figure 4.14: **Example of conformations belonging to the conformational energy wells of C-conformer 1 (left) and C-conformer 2 (right).**

Structure C5, which matches experimental form C of mebendazole (see Figure 4.8) is ranked 2nd in energy, ~0.4 kJ/mol above the global minimum C27. Both contain C-conformer 1.

In terms of intermolecular hydrogen bonding motifs, all the 99 structures within 20 kJ·mol^{-1} of the global minimum form an intermolecular hydrogen bond between the N acceptor in imidazole group and the NH in the amide group with the $\mathrm{R}_2^2(8)$ graph set. This motif is shown in Figure 4.15.

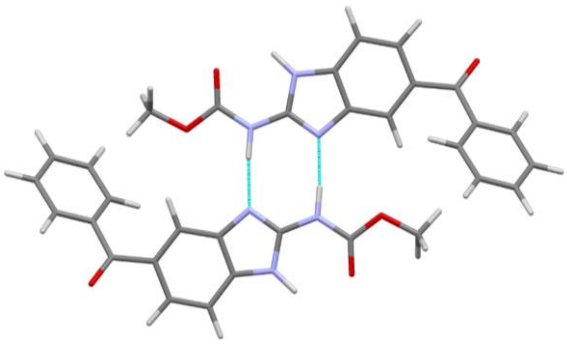


Figure 4.15: **NH···H hydrogen bond with the $\mathrm{R}_2^2(8)$ graph set motif common to all low-energy crystal structures containing the C-tautomer of mebendazole.**

Note that this motif is identical to the one that dominates the most competitive CSP-generated structures containing the A-tautomer (Figure 4.11), with the edge

benzoyl substituent attached four rather than three bonds away from the imidazole NH group. Structure C5 only forms this hydrogen bond (see also Figure 4.2).

Although three oxygen atoms could be used as intermolecular hydrogen bond acceptors, this only happens in 17/99 low-energy structures; the global minimum C27 is one of them, as it combines the hydrogen bond shown in Figure 4.15 with another one between the amide O acceptor and imidazole NH with the $\mathrm{R}_2^2(12)$ graph set motif, as shown in Figure 4.16.

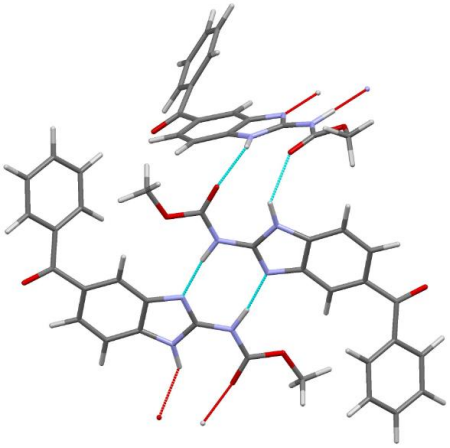


Figure 4.16: Hydrogen bond motif of the global minimum structure C27.

This hydrogen bonding pattern is shared by other five low-energy structures, including C406, ranked 6th, C46, ranked 10th and C24, ranked 11th. All these structures contain C-conformer 1. Five other low-energy structures, in this case containing C-conformer 2, utilise the same donors and acceptors, but the amide O acceptor and the imidazole NH are hydrogen bonded to two different molecules, forming the $\mathrm{C}_1^1(6)$ graph set motif: C248, ranked 9th, C115, ranked 12th, C509, ranked 13th, C908, ranked 16th and C244, ranked 17th. This motif is shown in Figure 4.17.

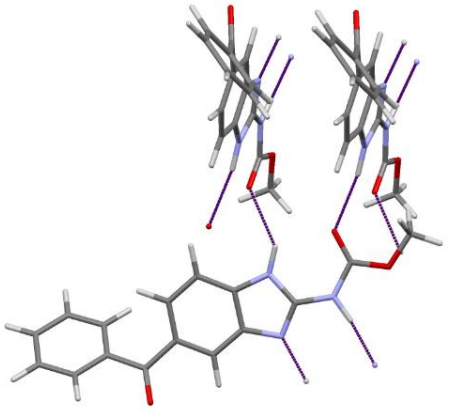


Figure 4.17: Hydrogen bond motif common to C248, C115, C509, C908 and C244.

Many low-energy structures contain a common sheet formed by several molecules with C-conformer 1 connected by the NH···N hydrogen bond motif shown in Figure 4.15. Nine out of the ten lowest energy crystal structures (including structure C5

and the global minimum) contain this sheet, which is shown in Figure 4.18. This could hint to the possibility of static disorder.[53, 61]

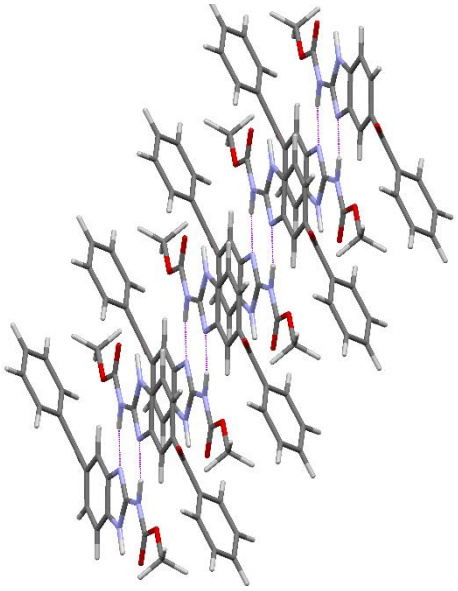


Figure 4.18: **Sheet common to 9/10 lowest energy structures, including the global minimum (C27) and the match to experimental form C (C5).**

In summary, the crystal structures containing the C-tautomer of mebendazole could form conformational polymorphs, as two distinct groups of conformations are present among low energy structures. Furthermore, packing polymorphs are also a possibility: although all thermodynamically competitive crystal structures form an intermolecular NH···H hydrogen bond with the $R_2^2(8)$ graph set motif, other additional motifs using any of the three O atoms as acceptors are present. Finally, the presence of a common sheet in several of the lowest computer-generated energy crystal structures hints to the possibility of disorder.

4.3.4 Computational cost and importance of molecular flexibility

A breakdown of the computational expense for undertaking this CSP study is shown in Table 4.2.

Table 4.2: Breakdown of the computational cost of the CSP study of mebendazole.

	CPU time (hours)	% of total time
Flexibility analysis	700	5.2
Grid Generation	200	1.5
Crystal structure search with CrystalPredictor	5,800	43.3
Single iterations with CrystalOptimizer	1,800	13.5
Full optimisations with CrystalOptimizer	4,800	35.9
Estimate of polarisation	80	0.6
Total time	13,380	100.0

The computational cost of this study was much smaller than the one of molecule XXVI (see Chapter 3.3.5), despite the need of performing separate searches for the two

tautomers of mebendazole and the consideration of both *cis* and *trans* amide configurations (for molecule XXVI only *trans* amide configurations were considered, see Chapter 3.2.2). Compared to the Blind Test CSP study, the search contributed more to the overall computational cost (43.3 vs 2%), while the refinement had a smaller role (49.4 vs 89.2%). Both the smaller overall computational cost and the lower relative contribution of final refinement stage were due to the limited size of mebendazole compared to molecule XXVI (35 vs 62 atoms, and 22 vs 40 non-hydrogenic atoms). Optimising smaller molecules is inherently cheaper,[47] and fewer competitive crystal structures that required refinement with expensive methods were generated, both because the conformational search space was smaller and the CrystalPredictor energies were deemed to be more realistic.

Although mebendazole has only one less intuitively flexible torsion angle than molecule XXVI, only two dihedrals (Φ_1 and Φ_2 in Figure 4.3) show a significant degree of flexibility and affect the overall shape of the molecule; torsion angle Φ_6 is very flexible but describes a methyl rotation with a negligible effect on molecular shape and packing capabilities. The high energy penalty for distorting torsion angles Φ_3-Φ_5 away 0° and/or 180° (see Figure 4.4b-d) makes mebendazole a very planar molecule both when isolated and in the energetically plausible computer-generated crystal structures: there is no trade-off between intramolecular dispersion favouring folded conformers in the gas-phase and intermolecular dispersion favouring planar conformations in the solid-state,[62] and no need of distorting the molecule to obtain an extensive hydrogen bond network (differently from molecule XXVI,[1] see Chapter 3). This drastically reduced the conformational search space of mebendazole. Only three clusters of conformational polymorphs were present among the low-energy predicted crystal structures of each tautomer, while more than 40 could be found for molecule XXVI. Furthermore most low-energy CSP-generated crystal structures of both tautomers have low ΔE_{intra} values, indicating that the degree of conformational adjustment that occurs in the thermodynamically competitive structures is limited.

This is very important in the overall context of this thesis, as it shows that level of molecular flexibility cannot be easily understood from the molecular diagram but it has a drastic effect on the computational expense. Hence it is important to develop new approaches that speed up the analysis of conformational flexibility, which will be discussed in Chapter 5.

4.4 Experimental polymorph screen

The experimental polymorph screen was carried out by Corpinot and Bučar. Mebendazole was purchased 98% pure from Cambridge Bioscience.

Unfortunately no single crystals of mebendazole could be produced, and characterisations could only be performed with PXRD. Qualitative PXRD data were collected using a Stoe StadiP diffractometer in transmission geometry using monochromated CuKα1 radiation (λ = 1.54056 Å) generated at 40 kV and 30 mA. Data were collected at room temperature using a 2-60° 2θ range. Rietveld refinements were performed with the TOPAS academic programme.

A brief outline of the experimental work and its results is reported.

4.4.1 Solvent-mediated phase transformation

Mebendazole (*m* ≈ 200 mg) was suspended in a variety of solvents (*V* ≈ 2 mL) and slurried for 7-14 days. The solutions were then filtered and the obtained solids were dried on filter paper for several minutes. Solvent-mediated phase transformation experiments yielded predominantly phase pure powders consisting of form A. Two exceptions were noted where physical mixtures were obtained. In particular, the experiment involving nitromethane yielded a mixture of form A and an unidentified phase, while the experiment involving hexane yielded a mixture of forms A and C. The results of the slurry experiments suggest that form A is the thermodynamically most stable, consistently with the outcome of the CSP study.

4.4.2 Crystallisation by slow solvent evaporation

Mebendazole (*m* ≈ 200 mg,) was dissolved under reflux conditions in a variety of solvents (*V* ≈ 5-23 mL) for 10-30 min in glass vials. The mixtures were filtered through a *Millex*® syringe filter and let evaporate at ambient conditions for 1-14 days. The results obtained from crystallisation by slow solvent evaporation were more varied, and yielded a complex solid form landscape.

The outcome of these experiments is summarised in Table 4.3, and the powder patterns are shown in Figure 4.19 and Figure 4.20. Note that two studies by Swanepoel *et al.*[39] and Kumar *et al.*[63] published distinct powder patterns for what they defined as mebendazole form B; the diffractograms were in both cases of poor quality and no 2θ values were reported. Some crystallisation experiments yielded a solid that resembles the form B powder pattern reported by Kumar *et al.*, while others resulted in the formation of a solid that exhibits a diffractogram similar to that reported by Swanepoel *et al.* The comparison between the two diffractograms is shown in Appendix Figure 4.3: they correspond to two different forms, and so they are hereafter (tentatively) referred to as forms B (grey) and D (blue). Furthermore, there are two variations of form D that have similar powder patterns, meaning they could be structurally related: they are (tentatively) referred to as forms D1 and D2.

Table 4.3: Overview of all mebendazole crystal forms and the solvents that were used in their preparation.

Single-component forms	Solvent
A	ethanol, xylene
B	dichloroethane, chloroform, acetonitrile
C	2-propanol, 2-butanol
D1	ethyl acetate, chloroethanol, 2-propanol
D2	methanol, dichloromethane
E	acetone
F	dioxane, pyridine
G	dichloroethane (phase transition from form B within 6 month)
Solvates	**Solvent**
THF	THF
DMF	DMF

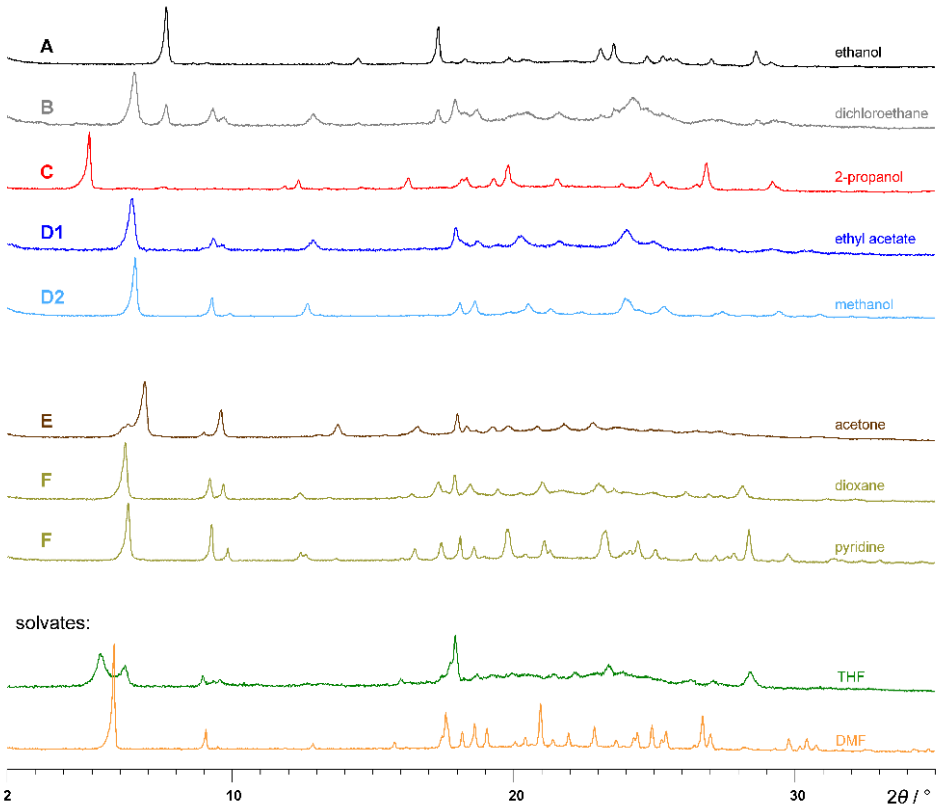


Figure 4.19: Diffractograms of forms A (black), B (grey), C (red), D1 (dark blue), D2 (light blue), E (brown) and F (olive). These patterns are compared to the diffractograms of the THF and DMF solvates (shown in green and orange).

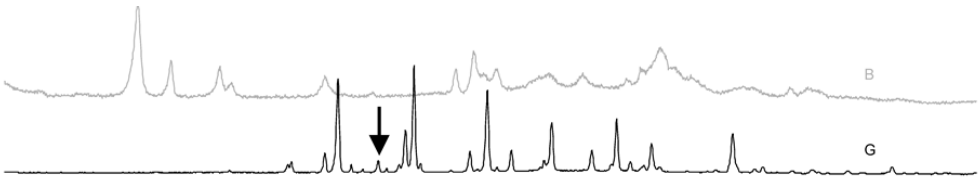


Figure 4.20: Diffractogram of form G (black) obtained through a phase transition of form B after six months at ambient conditions.

Mebendazole does not appear to possess a strong tendency to form solvates, which is consistent with the results of the CSP study showing that both the A and the C-tautomers can easily form dense packings (see Chapters 4.3.3.1 and 4.3.3.2).

It has not yet been possible to solve the crystal structures of form B, the five new single component forms and the two solvates. Only the powder pattern of form G could be indexed and several different unit cells were proposed; the best fit is a mixture of two orthorhombic phases.

The 211 CSP-generated crystal structures in the crystal energy landscape (see Figure 4.7, Figure 4.9, and Appendix Table 4.5 for more details) were used to try to refine the PXRD patterns of form B and the five new single-component forms. However, no match could be found.

4.4.3 Comparison with the computational results

From a computational viewpoint, it is worrying that none of the PXRD patterns of the five new single-components forms of mebendazole and previously known but unsolved form B could be matched with the crystal structures in the crystal energy landscape. However, until the crystal structures of these experimental forms are solved, allowing the determination of the atomic positions, the space groups and Z' values, it will not be possible to know why they were missed.

Nonetheless, a number of reasons can be anticipated. These forms might be outside the scope of the searches: they may have Z'>1, contain mixtures of tautomers, or contain other tautomers of mebendazole (Martins *et al.*[34] listed three other possible tautomers on top of the A- and C- ones, see Appendix Figure 4.4) that were not included in the searches. This would be the least concerning scenario, as it would indicate that the CSP procedure did not miss any experimental form that was within its reach. However, this would be a reminder that the CSP algorithms must be improved to allow a broader coverage of the E_{latt} surface, such has high Z' values or tautomeric mixtures, without making the computational cost unmanageable. A more worrying possibility would be that structures matching form B and/or the five new single-component forms were lost during the CSP process because they were outside the thresholds used to decide which structures to carry forward, they were not found because they contain strained conformations not included in the conformational search space or the searches

themselves were interrupted before they could completely explore the E_{latt} surface. This would mean that the CSP study was incomplete, and that its parameters should be corrected in the future. Form B and the five new forms appear to be metastable, and this increases the chance that they were missed because of an incomplete coverage of the search space or limitations in the number of crystal structures optimised with the most accurate methods. Chapters 5 and 7 propose new workflows that could solve some of these problems.

Finally, these forms might be disordered or not phase pure (this appears to be the case for form G): in such cases, they would be outside the reach of the adopted CSP methodology, although it would be interesting to know whether the disordered or mixed components are present among the computer-generated crystal structures.[64] Further experimental work is ongoing: if successful, it will allow to determine why CSP failed to predict these forms and whether they could actually be predicted.

4.5 Conclusion

The CSP study on two tautomers of the antihelminthic drug mebendazole successfully reproduced the only two solved crystal structures, forms A and C. Form A was correctly predicted to be more thermodynamically stable than C, and it was found as the global minimum in E_{latt} without an estimate of polarisation and as the second most stable form, 0.8 $kJ \cdot mol^{-1}$ above the global minimum, when the PCM was included. These results were confirmed by the experimental solid form screen, which found form A as the most thermodynamically stable and form C as a competitive crystal structure that can be produced under the right experimental conditions.

The solid form screen found the previously known but unsolved form B of mebendazole, as well as five new single-component forms. Unfortunately it was not possible to solve any of these crystal structures. Suitable single crystals could not be obtained, and only PXRD patterns could be produced. Quite worryingly, none of these PXRD patterns could be matched with those simulated from the CSP-generated crystal structures.

This study provides some insights that are very useful in the context of this thesis. First of all, it illustrates that CSP studies can provide useful information to experimentalists: the crystal energy landscape confirms that it is unlikely that a single-component crystal structure of mebendazole significantly more stable form than form A exists, and a ritonavir-like occurrence can be excluded. Furthermore, it correctly anticipates that mebendazole is not a promiscuous solvate former since it can pack well with itself. However, it also shows some of the limitations of current CSP methodologies: although the reasons for missing form B and the five new single-component forms are not known, this may have been caused by the need of limiting the search in terms of

conformational ranges and Z' values. Furthermore, this combined CSP-experimental effort shows that every molecule is a new challenge: mebendazole revealed to be very difficult to crystallise making the experimental work inconclusive.

Finally this chapter illustrates how analysing the quality of a computer-generated crystal energy landscape using partial experimental information can be misleading: if the extra forms had not been discovered, this would have been considered a completely successful CSP study, since matches both the known solved forms of mebendazole were found among the most competitive computer-generated crystal structures. However, a thorough solid form screen revealed that some experimental forms were indeed missed, making this CSP study only partially successful. This is a well-known problem, as the possibility of unreported polymorphs has often cast a shadow of doubt on Blind Test results.[64] This partial failure may indicate that computational methodologies must be made more cost-effective to allow a broader coverage of the multidimensional E_{latt} surface.

4.6 References

1. 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.; Vacarro, F.; Vasileiadis, M.; Vazquez-Mayagoitia, A.; Vogt, L.; Wang, Y.; Watson, R. E.; de Wijs, G. A.; Yang, J. Z.; Zhu, Q.; Groom, C. R., Report on the sixth blind test of organic crystal-structure prediction methods. *Acta Crystallographica Section B - Structural Science* **2016**.
2. Price, S. L.; Reutzel-Edens, S. M., The potential of computed crystal energy landscapes to aid solid form development. *Drug Discovery Today* **2016,** *21* (6), 912-923.
3. Nyman, J.; Reutzel-Edens, S., Crystal structure prediction is changing from basic science to applied technology. *Faraday Discussions* **2018,** *Advance article*.
4. Neumann, M. A.; de Streek, J. V.; Fabbiani, F. P. A.; Hidber, P.; Grassmann, O., Combined crystal structure prediction and high-pressure crystallization in rational pharmaceutical polymorph screening. *Nature Communications* **2015,** *6*, 7793.
5. Price, S. L.; Braun, D. E.; Reutzel-Edens, S. M., Can computed crystal energy landscapes help understand pharmaceutical solids? *Chemical Communications* **2016,** *52*, 7065-7077.
6. Lee, E. H., A practical guide to pharmaceutical polymorph screening & selection. *Asian Journal of Pharmaceutical Sciences* **2014,** *9* (4), 163-175.
7. Neumann, M.; van de Streek, J., How many Ritonavir cases are there still out there? *Faraday Discussions* **2018,** *Advance article*.
8. Price, S. L., The computational prediction of pharmaceutical crystal structures and polymorphism. *Advanced Drug Delivery Reviews* **2004,** *56* (3), 301-319.
9. Abramov, Y., Current Computational Approaches to Support Pharmaceutical Solid Form Selection. *Organic Process Research & Development* **2013,** *17* (3), 472-485.
10. Chemburkar, S. R.; Bauer, J.; Deming, K.; Spiwek, H.; Patel, K.; Morris, J.; Henry, R.; Spanton, S.; Dziki, W.; Porter, W.; Quick, J.; Bauer, P.; Donaubauer, J.; Narayanan, B. A.;

Soldani, M.; Riley, D.; McFarland, K., Dealing with the impact of ritonavir polymorphs on the late stages of bulk drug process development. *Organic Process Research & Development* **2000,** *4* (5), 413-417.
11. Bucar, D. K.; Day, G. M.; Halasz, I.; Zhang, G. G. Z.; Sander, J. R. G.; Reid, D. G.; MacGillivray, L. R.; Duer, M. J.; Jones, W., The curious case of (caffeine).(benzoic acid): how heteronuclear seeding allowed the formation of an elusive cocrystal. *Chemical Science* **2013,** *4* (12), 4417-4425.
12. Arlin, J. B.; Price, L. S.; Price, S. L.; Florence, A. J., A strategy for producing predicted polymorphs: catemeric carbamazepine form V. *Chemical Communications* **2011,** *47* (25), 7074-7076.
13. 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. *Chemical Communications* **2016,** *52*, 7384-7386.
14. Pulido, A.; Chen, L. J.; Kaczorowski, T.; Holden, D.; Little, M. A.; Chong, S. Y.; Slater, B. J.; McMahon, D. P.; Bonillo, B.; Stackhouse, C. J.; Stephenson, A.; Kane, C. M.; Clowes, R.; Hasell, T.; Cooper, A. I.; Day, G. M., Functional materials discovery using energy-structure-function maps. *Nature* **2017,** *543* (7647), 657-664.
15. Wu, H.; Habgood, M.; Parker, J. E.; Reeves-McLaren, N.; Cockcroft, J. K.; Vickers, M.; West, A. R.; Jones, A. G., Crystal structure determination by combined synchrotron powder X-ray diffraction and crystal structure prediction: 1: 1 L-ephedrine D-tartrate. *CrystEngComm* **2013,** *15* (10), 1853-1859.
16. van de Streek, J.; Neumann, M. A., Validation of molecular crystal structures from powder diffraction data with dispersion-corrected density functional theory (DFT-D). *Acta Crystallographica Section B-Structural Science Crystal Engineering and Materials* **2014,** *70*, 1020-1032.
17. Habermehl, S.; Morschel, P.; Eisenbrandt, P.; Hammer, S. M.; Schmidt, M. U., Structure determination from powder data without prior indexing, using a similarity measure based on cross-correlation functions. *Acta Crystallographica Section B-Structural Science Crystal Engineering and Materials* **2014,** *70*, 347-359.
18. Baias, M.; Dumez, J. N.; Svensson, P. H.; Schantz, S.; Day, G. M.; Emsley, L., De Novo Determination of the Crystal Structure of a Large Drug Molecule by Crystal Structure Prediction-Based Powder NMR Crystallography. *Journal of the American Chemical Society* **2013,** *135* (46), 17501-17507.
19. Salager, E.; Day, G. M.; Stein, R. S.; Pickard, C. J.; Elena, B.; Emsley, L., Powder Crystallography by Combined Crystal Structure Prediction and High-Resolution H-1 Solid-state NMR Spectroscopy. *Journal of the American Chemical Society* **2010,** *132* (8), 2564-2566.
20. Baias, M.; Widdifield, C. M.; Dumez, J.-N.; Thompson, H. P. G.; Cooper, T. G.; Salager, E.; Bassil, S.; Stein, R. S.; Lesage, A.; Day, G. M.; Emsley, L., Powder crystallography of pharmaceutical materials by combined crystal structure prediction and solid-state 1H NMR spectroscopy. *Physical Chemistry Chemical Physics* **2013,** *15* (21), 8069-8080.
21. Eddleston, M. D.; Hejczyk, K. E.; Bithell, E. G.; Day, G. M.; Jones, W., Determination of the Crystal Structure of a New Polymorph of Theophylline. *Chemistry - A European Journal* **2013,** *19* (24), 7883-7888.
22. Popovic, D. J.; Posa, M.; Popovic, K. J.; Kolarovic, J.; Popovic, J. K.; Banovic, P. Z., Application of a widely-used tropical anti-worm agent, mebendazole, in modern oncology. *Tropical Journal of Pharmaceutical Research* **2017,** *16* (10), 2555-2562.
23. Kanagale, P. D., Novel age-appropriate formulation of mebendazole (Vermox) tablets 500 mg for who's children without worm donation programme. *International Journal of Pharmaceutics* **2016,** *511* (2), 1139-1139.
24. Ferreira, F. F.; Antoni, S. G.; Rosa, P. C. P.; Paiva-Santos, C. D., Crystal Structure Determination of Mebendazole Form A Using High-Resolution Synchrotron X-Ray Powder Diffraction Data. *Journal of Pharmaceutical Sciences* **2010,** *99* (4), 1734-1744.
25. Chen, J.-M.; Wang, Z.-Z.; Wu, C.-B.; Li, S.; Lu, T.-B., Crystal engineering approach to improve the solubility of mebendazole. *CrystEngComm* **2012,** *14* (19), 6221-6229.
26. Al-Badr, A.; Tariq, M., *50- Analytical Profile of Mebendazole*. 1987; p 291-326.
27. Brugmans, J. P.; Thienpont, D. C.; van Wijngaarden, I.; Vanparijs, O. F.; Schuermans, V. L.; Lauwers, H. L., Mebendazole in enterobiasis radiochemical and pilot clinical study in 1,278 subjects. *JAMA* **1971,** *217* (3), 313-316.
28. (WHO), W. H. O., WHO Model List of Essential Medicines. 2015.
29. Gutiérrez, E. L.; Souza, M. S.; Diniz, L. F.; Ellena, J., Synthesis, characterization and solubility of a new anthelmintic salt: Mebendazole nitrate. *Journal of Molecular Structure* **2018,** *1161*, 113-121.

30. Zimmermann, S. C.; Tichy, T.; Vavra, J.; Dash, R. P.; Slusher, C. E.; Gadiano, A. J.; Wu, Y.; Jancarik, A.; Tenora, L.; Monincova, L.; Prchalova, E.; Riggins, G. J.; Majer, P.; Slusher, B. S.; Rais, R., N-Substituted Prodrugs of Mebendazole Provide Improved Aqueous Solubility and Oral Bioavailability in Mice and Dogs. *Journal of Medicinal Chemistry* **2018,** *61* (9), 3918-3929.
31. Liu, C.-s.; Zhang, H.-b.; Jiang, B.; Yao, J.-m.; Tao, Y.; Xue, J.; Wen, A.-d., Enhanced bioavailability and cysticidal effect of three mebendazole-oil preparations in mice infected with secondary cysts of Echinococcus granulosus. *Parasitology Research* **2012,** *111* (3), 1205-1211.
32. Calvo, N. L.; Kaufman, T. S.; Maggio, R. M., Mebendazole crystal forms in tablet formulations. An ATR-FTIR/chemometrics approach to polymorph assignment. *Journal of Pharmaceutical and Biomedical Analysis* **2016,** *122*, 157-165.
33. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C., The Cambridge Structural Database. *Acta Crystallographica Section B-Structural Science Crystal Engineering and Materials* **2016,** *72*, 171-179.
34. Martins, F. T.; Neves, P. P.; Ellena, J.; Cami, G. E.; Brusau, E. V.; Narda, G. E., Intermolecular Contacts Influencing the Conformational and Geometric Features of the Pharmaceutically Preferred Mebendazole Polymorph C. *Journal of Pharmaceutical Sciences* **2009,** *98* (7), 2336-2344.
35. Bernstein, J., Polymorphism - A Perspective. *Crystal Growth & Design* **2011,** *11* (3), 632-650.
36. Brits, M.; Liebenberg, W.; De Villiers, M. M., Characterization of Polymorph Transformations That Decrease the Stability of Tablets Containing the WHO Essential Drug Mebendazole. *Journal of Pharmaceutical Sciences* **2010,** *99* (3), 1138-1151.
37. van de Streek, J., Searching the Cambridge Structural Database for the 'best' representative of each unique polymorph. *Acta Crystallographica Section B - Structural Science* **2006,** *62*, 567-579.
38. Salvi, S. T. B.; Antonio, S. G.; Ferreira, F. F.; Paiva-Santos, C. O., Rietveld Method in the Analysis of Polymorphism in Mebendazole Tablets Acquired in Brazil's Drugstores. *Journal of the Brazilian Chemical Society* **2015,** *26* (9), 1760-1768.
39. Swanepoel, E.; Liebenberg, W.; Devarakonda, B.; De Villiers, M. M., Developing a discriminating dissolution test for three mebendazole polymorphs based on solubility differences. *Pharmazie* **2003,** *58* (2), 117-121.
40. Charoenlarp, P.; Waikagul, J.; Muennoo, C.; Srinophakun, S.; Kitayaporn, D., Efficacy of single-dose mebendazole, polymorphic forms A and C, in the treatment of hookworm and Trichuris infections. *Southeast Asian Journal of Tropical Medicine and Public Health* **1993,** *24* (4), 712-716.
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. 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, D. J. *Gaussian 09, Revision D.01*, 2009.
42. Uzoh, O. G.; Galek, P. T. A.; Price, S. L., Analysis of the conformational profiles of fenamates shows route towards novel, higher accuracy, force-fields for pharmaceuticals. *Physical Chemistry Chemical Physics* **2015,** *17* (12), 7936-7948.
43. Vasileiadis, M.; Pantelides, C. C.; Adjiman, C. S., Prediction of the crystal structures of axitinib, a polymorphic pharmaceutical molecule. *Chemical Engineering Science* **2015,** *121*, 60-76.
44. Bhardwaj, R. M.; Price, L. S.; Price, S. L.; Reutzel-Edens, S. M.; Miller, G. J.; Oswald, I. D. H.; Johnston, B.; Florence, A. J., Exploring the Experimental and Computed Crystal Energy Landscape of Olanzapine. *Crystal Growth & Design* **2013,** *13* (4), 1602-1617.
45. Price, L. S.; McMahon, J. A.; Lingireddy, S. R.; Lau, S. F.; Diseroad, B. A.; Price, S. L.; Reutzel-Edens, S. M., A molecular picture of the problems in ensuring structural purity of tazofelone. *Journal of Molecular Structure* **2014,** *1078*, 26-42.
46. Karamertzanis, P. G.; Pantelides, C. C., Ab initio crystal structure prediction. II. Flexible molecules. *Molecular Physics* **2007,** *105* (2-3), 273-291.

47. 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. *Journal of Chemical Theory and Computation* **2015,** *11* (4), 1957-1969.
48. Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C.; Price, S. L.; Galek, P. T.; Day, G. M.; Cruz-Cabeza, A. J., Successful prediction of a model pharmaceutical in the fifth blind test of crystal structure prediction. *International Journal of Pharmaceutics* **2011,** *418* (2), 168-178.
49. Pantelides, C. C.; Adjiman, C. S.; Kazantsev, A. V., General Computational Algorithms for Ab Initio Crystal Structure Prediction for Organic Molecules. *Topics in Current Chemistry* **2014,** *345*, 25-58.
50. Coombes, D. S., Deriving intermolecular potentials for predicting the crystal structures of polar molecules. *Philosophical Magazine B-Physics of Condensed Matter Statistical Mechanics Electronic Optical and Magnetic Properties* **1996,** *73* (1), 117-125.
51. Steed, K.; Steed, J., Packing problems: High Z' crystal structures and theri relationship to co-crystals, inclusion compounds and polymorphism. *Chemical Reviews* **2015**.
52. Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C., CrystalOptimizer. An efficient Algorithm for Lattice Energy Minimisation of Organic Crystal using Isolated-Molecule Quantum Mechanical Calculations. In *Molecular System Engineering*, Adjiman, C. S.; Galindo, A., Eds. WILEY-VCH Verlag GmbH & Co.: Weinheim, 2010; Vol. 6, pp 1-42.
53. Price, S. L., Predicting crystal structures of organic compounds. *Chemical Society Reviews* **2014,** *43* (7), 2098-2111.
54. Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C., Efficient Handling of Molecular Flexibility in Lattice Energy Minimization of Organic Crystals. *Journal of Chemical Theory and Computation* **2011,** *7* (6), 1998-2016.
55. Price, S. L.; Leslie, M.; Welch, G. W. A.; Habgood, M.; Price, L. S.; Karamertzanis, P. G.; Day, G. M., Modelling Organic Crystal Structures using Distributed Multipole and Polarizability-Based Model Intermolecular Potentials. *Physical Chemistry Chemical Physics* **2010,** *12* (30), 8478-8490.
56. Stone, A. J., Distributed multipole analysis: Stability for large basis sets. *Journal of Chemical Theory and Computation* **2005,** *1* (6), 1128-1132.
57. Chisholm, J. A.; Motherwell, S., COMPACK: a program for identifying crystal structure similarity using distances. *Journal of Applied Crystallography* **2005,** *38*, 228-231.
58. Habgood, M.; Price, S. L.; Portalone, G.; Irrera, S., Testing a Variety of Electronic-Structure-Based Methods for the Relative Energies of 5-Formyluracil Crystals. *Journal of Chemical Theory and Computation* **2011,** *7* (9), 2685-2688.
59. Nyman, J.; Reutzel-Edens, S. M., Crystal structure prediction is changing from basic science to applied technology. *Faraday Discussions* **2018**.
60. Cruz-Cabeza, A. J.; Bernstein, J., Conformational Polymorphism. *Chemical Reviews* **2014,** *114* (4), 2170-2191.
61. Habgood, M., Form II Caffeine: A Case Study for Confirming and Predicting Disorder in Organic Crystals. *Crystal Growth & Design* **2011,** *11* (8), 3600-3608.
62. Thompson, H.; Day, G., Which conformations make stable crystal structures? Mapping crystalline molecular geometries to the conformational energy landscape. *Chemical Science* **2014,** *5* (8), 3173-3182.
63. Kumar, S.; Chawla, G.; Bansal, A. K., Spherical Crystallization of Mebendazole to Improve Processability. *Pharmaceutical Development and Technology* **2008,** *13* (6), 559-568.
64. Price, S. L., Is zeroth order crystal structure prediction (CSP_0) coming to maturity? What should we aim for in an ideal crystal structure prediction code? *Faraday Discussions* **2018,** *in press*.

4.7 Appendix

Appendix Table 4.1: Definition of the main torsion angles of the A- and C-tautomers of mebendazole shown in Figure 4.3.

Torsion angle	Torsion angle definition
Φ_1	C18-C17-C14-C7 (A), C18-C17-C14-C8 (C)
Φ_2	O15-C14-C7-C8 (A), O15-C14-C8-C7 (C)
Φ_3	N1-C2-N10-C11 (A and C)
Φ_4	C2-N10-C11-O12 (A and C)
Φ_5	N10-C11-O12-C13 (A and C)
Φ_6	C11-O12-C13-H13a (A and C)

Appendix Table 4.2: Values taken by the torsion angles in Figure 4.3 in the conformations of the two solved crystal structures of mebendazole.

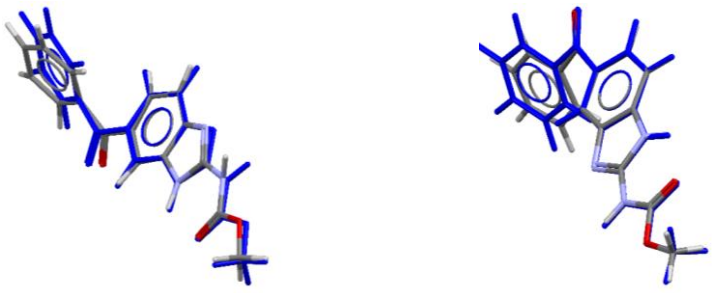
Torsion	Torsion angle values of form A conformation/°	Torsion angle values of form C conformation/°
Φ_1	336.73	329.67
Φ_2	148.40	147.69
Φ_3	359.57	0.17
Φ_4	180.76	181.35
Φ_5	180.27	180.32
Φ_6	42.89	65.00

Appendix Table 4.3: List of the 59 space groups considered in the crystal structure search performed with CrystalPredictor 1.8.

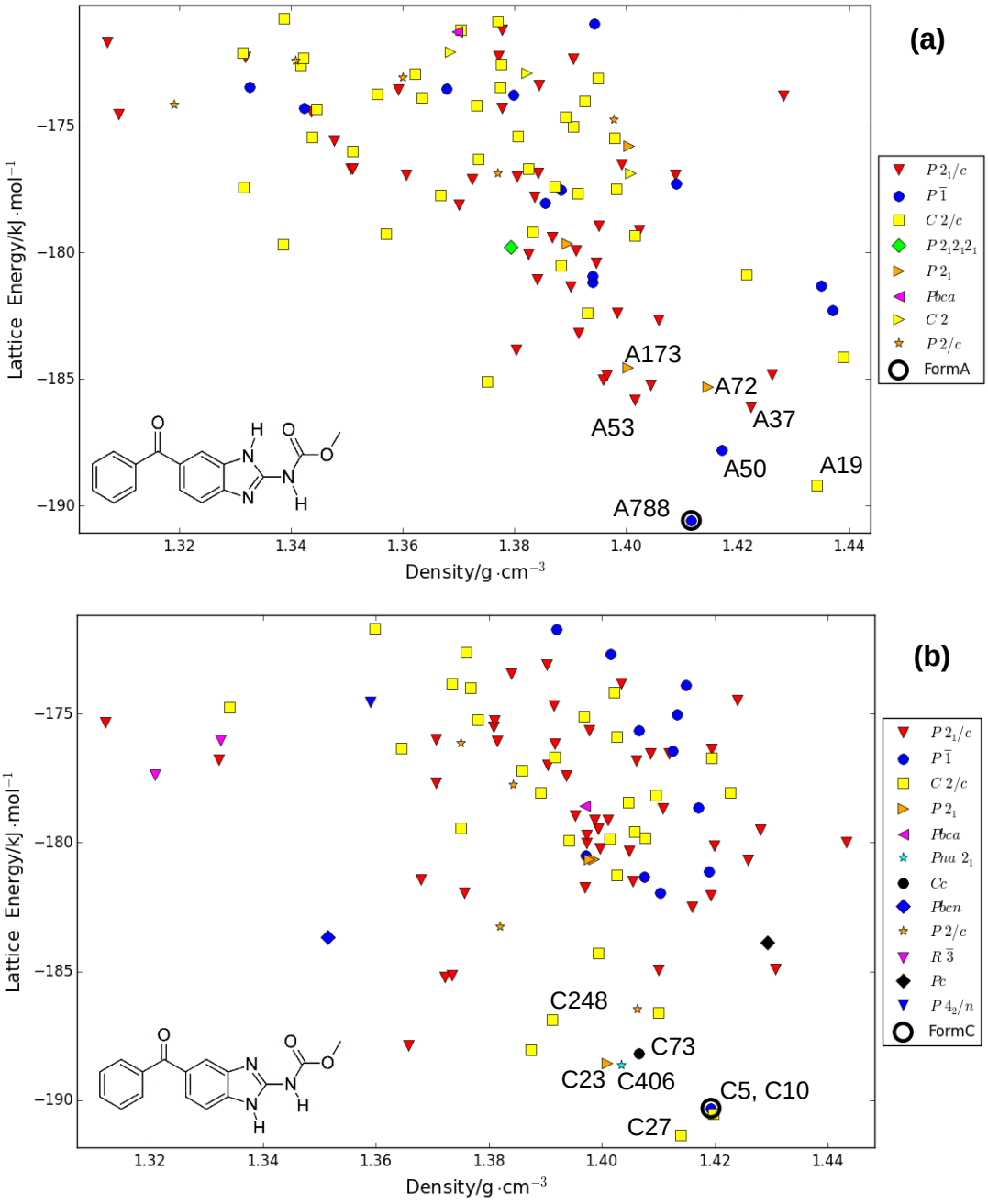
P1	$P\bar{1}$	$P2_1$	$P2_1/c$	$P2_12_12$	$P2_12_12_1$	$Pna2_1$	$Pca2_1$	Pbca	Pbcn
C2/c	Cc	C2	Pc	Cm	$P2_1/m$	C2/m	P2/c	$C222_1$	$Pmn2_1$
$Cmc2_1$	Aba2	Fdd2	Iba2	Pnna	Pccn	Pbcm	Pnnm	Pmmn	Pnma
Cmcm	Cmca	Fddd	Ibam	$P4_1$	$P4_3$	$I\bar{4}$	P4/n	$P4_2/n$	I4/m
$I4_1/a$	$P4_12_12$	$P4_32_12$	$P\bar{4}2_1c$	$I\bar{4}2d$	$P3_1$	$P3_2$	R3	$P\bar{3}$	$R\bar{3}$
$P3_121$	$P3_221$	R3c	$R\bar{3}C$	$P6_1$	$P6_3$	$P6_3/m$	$P2_13$	$PA\bar{3}$	

Appendix Table 4.4: Definition of the torsion and bond-angles of the two tautomers of mebendazole treated as independent CDFs in the final refinement stage with CrystalOptimizer. The atomic numbering of each tautomer is shown in Figure 4.6.

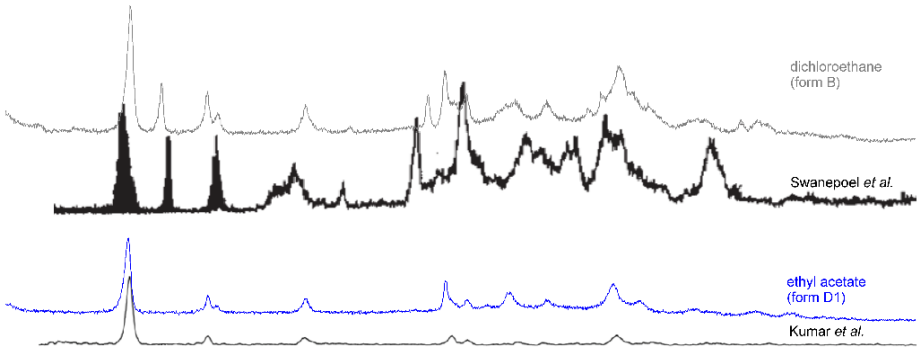
	Torsion angle definition	Bond-angle definition
A-tautomer	C18-C17-C14-C7, C21-C22-C17-C14, O15-C14-C7-C6, H1a-N1-C2-N10, N1-C2-N10-C11, O16-C11-O2-C13, C2-N10-C11-O12. H10a-N10-C11-O12	C22-C17-C14, C18-C17-C14, H1a-N1-C2, C2-N10-C11, H10a-N10-C11, C11-O12-C13
C-tautomer	N10-C11-O2-C13, O12-C11-N10-C2, N1-C2-N10-C11, C14-C8-C8-C4, C17-C14-C8-C7, C22-C17-C14-C8, O16-C11-O12-C13, H10a-N10-C11-O12, H13a-C13-O2-C11, H1a-N1-C2-N10	C22-C17-C14, C18-C17-C14, C17-C14-C7, H1a-N1-C2, H10a-N10-C11, C11-O12-C13



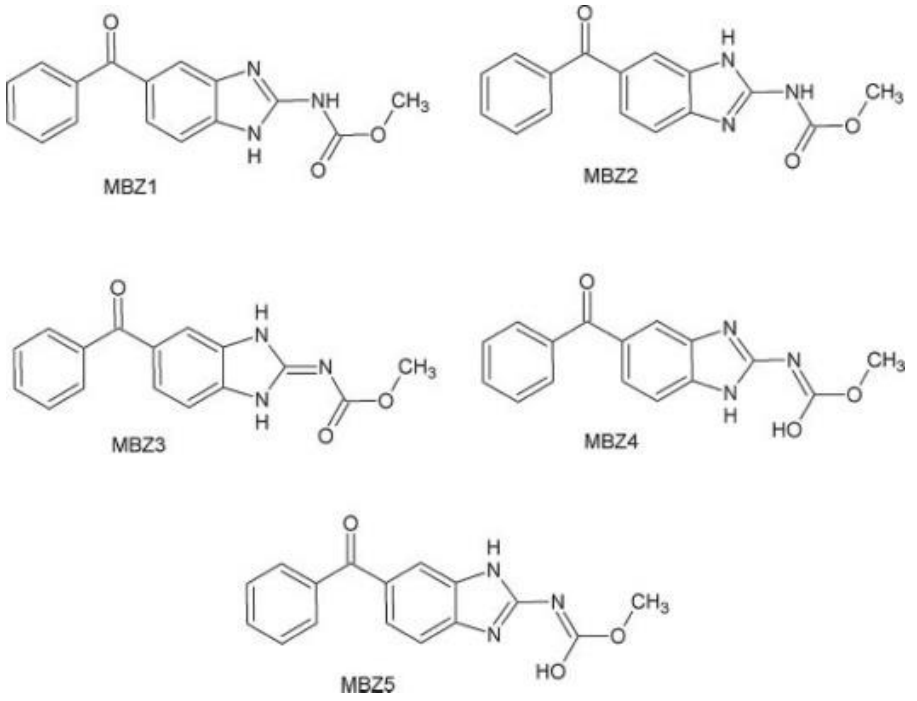
Appendix Figure 4.1: **Overlay of the conformation of the experimental crystal structure, coloured by elements, and the closest isolated-molecule local minimum in conformational energy, in blue, of (left) form A, with and $RMSD_1$ of 0.242 Å (right) form C, with and $RMSD_1$ of 0.096 Å.**



Appendix Figure 4.2: Summary plots of the crystal energy landscapes of (a) the A-tautomer (b) the C-tautomer of mebendazole after recalculating ΔE_{intra} and the charge density in a PCM with ε=3, showing all the crystal structures within 20 kJ·mol^{-1} of the global minimum in E_{latt}. Each point on the landscape corresponds to a separate crystal structure, labelled according to its space group. The structure matching the respective experimental forms are circled, and those labelled in Figure 4.10 and Figure 4.13 are indicated.



Appendix Figure 4.3: Diffractograms of form B, as reported by Kumar *et al.* and Swanepoel *et al.* Each of the two patterns (in black) is also compared with the diffractograms that were obtained from the crystallisation experiments using dichloroethane (grey) and ethyl acetate (blue) as solvents.



Appendix Figure 4.4: The five tautomers of mebendazole listed by Martins *et al*.[34] Note that MBZ1 and MBZ2 correspond to the C- and A-tautomer respectively. MBZ3, MBZ4 and MBZ5 were not considered in this CSP study.

Appendix Table 4.5: Lattice energies, structural and crystallographic parameters after the full optimisations of CrystalOptimizer of the 211 computer-generated crystal structures of both tautomers of mebendazole in the crystal energy landscape in Figure 4.7 and whose .res files were provided to the experimental collaborators. The energies obtained re-optimising the intermolecular interactions with the PCM (see Figure 4.9 for the plot) are also shown. The structure matching experimental form A is highlighted in orange, the one matching experimental form C in green.

Structure	E_{latt} /kJ·mol^{-1}	Density/g·cm^{3}	Packing coefficient/%	Space group	Z'	a/Å	b/Å	c/Å	α/°	β/°	γ/°	E_{latt} with PCM/kJ·mol^{-1}
A788	-183.02	1.408	72.03	$P\bar{1}$	1	5.3	11.3	12.7	66.97	80.4	77.24	-190.58
A19	-181.84	1.432	73.16	$C2/c$	1	5.1	13.1	20.9	94.11	90.0	101.24	-189.19
C27	-180.87	1.410	72.1	$C2/c$	1	5.1	16.0	17.8	106.01	98.2	90.00	-191.35
C5	-180.48	1.413	72.21	$P\bar{1}$	1	5.0	8.18	17.7	83.52	85.9	72.00	-190.32
C10	-180.39	1.413	72.37	$C2/c$	1	5.0	8.12	35.4	93.57	90.0	108.29	-190.53
A50	-179.88	1.418	72.33	$P\bar{1}$	1	5.4	10.7	12.2	91.93	101.	99.77	-187.81
A37	-178.94	1.418	72.44	$P2_1/c$	1	5.0	15.9	17.7	104.62	90.0	90.00	-186.09
C23	-178.69	1.397	71.4	$P2_1$	2	5.1	8.17	35.3	90.00	90.0	107.56	-188.58
C73	-178.69	1.400	71.59	Cc	2	5.0	8.16	35.5	89.52	88.4	71.86	-188.19
C406	-178.34	1.400	71.58	$Pna2_1$	2	5.0	16.1	34.2	90.00	90.0	90.00	-188.65
A53	-178.22	1.398	71.51	$P2_1/c$	1	4.1	17.8	20.8	113.79	90.0	90.00	-185.82
C53	-177.55	1.358	69.72	$P2_1/c$	1	5.1	8.14	36.2	90.00	90.0	108.08	-187.87
C25	-177.53	1.380	70.49	$C2/c$	1	5.1	8.15	35.9	86.59	85.9	71.68	-188.03
A173	-177.36	1.399	71.43	$P2_1$	2	5.3	12.5	21.4	90.00	90.0	101.04	-184.54
A72	-177.28	1.412	72.13	$P2_1$	1	4.0	10.6	16.3	97.04	90.0	90.00	-185.30
A49	-177.23	1.394	71.11	$P2_1/c$	1	4.7	14.7	20.2	90.45	90.0	90.00	-185.02
A78	-177.09	1.397	71.41	$P2_1/c$	1	4.0	16.6	21.0	99.74	90.0	90.00	-184.83
A90	-177.05	1.372	69.98	$C2/c$	1	5.3	11.4	24.3	84.57	83.7	76.46	-185.10
A291	-176.88	1.433	73.3	$C2/c$	1	7.4	8.48	24.5	83.22	81.2	64.03	-184.13
C248	-176.85	1.405	71.94	$C2/c$	1	5.2	14.5	19.5	108.31	90.0	100.48	-186.59
A306	-176.76	1.423	72.7	$P2_1/c$	1	4.5	10.2	29.7	90.00	93.4	90.00	-184.82
C46	-176.72	1.404	71.8	$P2/c$	1	5.0	16.0	17.4	101.36	90.0	90.00	-186.47
C24	-176.67	1.387	70.8	$C2/c$	1	5.0	16.1	18.1	106.04	98.0	90.00	-186.89
C115	-176.56	1.409	72.28	$P2_1/c$	1	4.5	16.2	19.5	106.93	90.0	90.00	-184.94
C509	-176.42	1.371	70.13	$P2_1/c$	1	5.4	14.8	18.6	109.11	90.0	90.00	-185.24
C583	-176.41	1.429	73.02	Pc	2	4.1	12.5	26.3	98.13	90.0	90.00	-183.88
C106	-176.30	1.430	72.88	$P2_1/c$	1	4.0	15.1	22.5	90.00	95.0	90.00	-184.92
A202	-176.30	1.378	70.53	$P2_1/c$	1	4.5	12.0	26.4	102.22	90.0	90.00	-183.85
A143	-175.83	1.435	72.96	$P\bar{1}$	1	7.1	8.60	12.8	109.05	90.2	113.30	-182.28
A89	-175.70	1.403	71.58	$P2_1/c$	1	7.8	8.40	21.1	94.23	90.0	90.00	-185.23
C908	-175.59	1.399	71.48	$C2/c$	1	5.3	13.9	19.8	105.63	90.0	101.03	-184.31
C244	-175.31	1.372	70.29	$P2_1/c$	1	5.2	15.2	19.5	112.57	90.0	90.00	-185.17
C132	-175.12	1.413	72.11	$P\bar{1}$	1	4.0	12.7	13.7	99.29	91.9	94.86	-181.95
A54	-174.94	1.430	72.61	$P\bar{1}$	1	7.5	10.3	10.5	116.61	92.9	107.47	-181.29
A109	-174.86	1.395	71.14	$C2/c$	1	5.4	12.4	21.9	79.40	82.9	77.45	-182.37
A71	-174.80	1.398	71.43	$P2_1/c$	1	4.8	10.4	27.8	90.55	90.0	90.00	-182.67
C111	-174.77	1.445	74.31	$P2_1/c$	1	3.9	17.4	19.6	90.00	90.0	95.63	-179.98
A604	-174.45	1.384	70.58	$C2/c$	1	3.8	17.4	21.3	98.15	90.0	96.40	-180.51
A75	-174.38	1.392	71.56	$P2_1/c$	1	4.1	11.9	28.2	90.00	91.1	90.00	-183.18
C199	-174.33	1.406	71.04	$P\bar{1}$	1	4.1	12.8	13.9	72.45	89.7	87.86	-181.32
C220	-174.27	1.408	72.22	$C2/c$	1	3.9	12.8	27.9	98.93	90.0	98.85	-181.25
A100	-174.22	1.392	71.04	$P2_1/c$	1	4.1	16.0	21.5	98.45	90.0	90.00	-182.36
A119	-174.11	1.390	70.85	$P2_1/c$	1	4.0	10.4	33.9	96.99	90.0	90.00	-181.34
A80	-174.07	1.392	70.43	$P\bar{1}$	1	4.0	10.5	16.9	101.28	94.7	90.65	-180.93
C394	-174.00	1.405	71.57	$P2_1/c$	1	4.9	12.8	22.0	97.93	90.0	90.00	-181.51
C114	-173.96	1.420	72.76	$P\bar{1}$	1	4.0	13.4	14.2	114.90	92.2	98.36	-181.12
C112	-173.91	1.421	72.72	$P2_1/c$	1	3.9	15.2	23.3	100.70	90.0	90.00	-180.12
C124	-173.90	1.411	72.31	$P2_1/c$	1	4.4	15.3	20.3	90.00	90.0	97.94	-182.51
A95	-173.76	1.416	72.41	$C2/c$	1	4.7	13.6	21.6	93.82	90.0	100.10	-180.85
C35	-173.72	1.378	70.13	$P2/c$	1	5.1	16.0	17.5	99.31	90.0	90.00	-183.28

C119	-173.70	1.427	72.58	$P2_1/c$	1	4.0	14.8	23.4	101.02	90.0	90.00	-180.67
C483	-173.70	1.412	71.91	$P2_1/c$	1	3.9	14.3	25.1	99.09	90.0	90.00	-180.33
C135	-173.69	1.346	68.86	*Pbca*	1	5.1	16.0	35.1	90.00	90.0	90.00	-183.67
C923	-173.41	1.418	72.18	$P2_1/c$	1	4.5	9.99	30.2	93.73	90.0	90.00	-182.05
C694	-173.38	1.430	72.7	$P2_1/c$	1	3.9	10.0	34.3	90.80	90.0	90.00	-179.50
C444	-173.29	1.398	71.43	$P2_1$	2	4.0	12.7	27.0	90.00	90.0	94.28	-180.66
C330	-173.20	1.374	69.8	$P2_1/c$	1	4.4	11.2	28.6	90.00	90.0	90.08	-181.96
C144	-173.13	1.403	71.98	$P2_1/c$	1	4.1	15.6	21.6	90.63	90.0	90.00	-179.48
C397	-173.08	1.402	71.69	$C2/c$	1	4.0	13.2	26.8	88.58	85.7	81.28	-179.84
A159	-173.07	1.395	70.84	$P2_1/c$	1	4.0	10.7	32.8	95.67	90.0	90.00	-180.41
A137	-173.07	1.380	70.36	$C2/c$	1	3.9	19.2	20.9	114.51	90.0	95.85	-179.18
A120	-172.89	1.385	70.57	$P2_1/c$	1	4.3	16.2	20.1	92.28	90.0	90.00	-180.07
C884	-172.88	1.409	71.77	$C2/c$	1	3.9	17.4	20.5	83.23	84.5	83.52	-179.59
A136	-172.71	1.378	70.35	$P2_1/c$	1	4.4	10.8	29.8	90.00	90.0	96.76	-181.05
C642	-172.65	1.397	71.56	$P2_1/c$	1	4.4	10.4	30.0	90.00	93.4	90.00	-181.74
A157	-172.55	1.399	71.42	$C2/c$	1	4.1	17.0	20.1	88.22	84.0	83.00	-179.33
C602	-172.54	1.405	71.95	$P2_1/c$	1	3.9	17.6	20.2	97.11	90.0	90.00	-179.11
A156	-172.24	1.389	70.72	$P2_1/c$	1	5.0	10.6	26.5	92.86	90.0	90.00	-179.93
C194	-172.13	1.424	72.96	$C2/c$	1	3.9	18.4	19.6	104.86	90.0	96.16	-178.08
C874	-171.99	1.397	71.53	$P2_1/c$	1	4.1	12.6	27.4	101.54	90.0	90.00	-178.95
A132	-171.97	1.379	70.43	$P2_1/c$	1	5.1	13.3	21.1	99.48	90.0	90.00	-179.40
C302	-171.88	1.415	72.59	$P\bar{1}$	1	5.0	10.8	14.0	109.60	96.8	99.23	-178.65
A76	-171.87	1.403	71.58	$P2_1/c$	1	3.9	10.6	33.5	96.91	90.0	90.00	-179.11
A85	-171.85	1.383	70.95	$P2_12_12$	1	4.3	10.4	31.5	90.00	90.0	90.00	-179.78
C644	-171.84	1.410	71.69	$P2_1/c$	1	4.0	10.0	34.1	90.00	92.8	90.00	-178.68
C657	-171.80	1.410	72.01	$C2/c$	1	3.9	15.0	23.7	88.58	85.2	82.51	-178.15
A175	-171.71	1.399	71.35	$C2$	1	3.8	10.6	17.4	97.30	96.3	90.00	-176.86
C59	-171.66	1.366	69.75	$P2_1/c$	1	5.1	13.6	20.4	90.00	92.5	90.00	-181.42
C541	-171.65	1.393	70.75	$P\bar{1}$	1	5.0	10.1	14.5	73.64	81.3	87.73	-180.50
A130	-171.62	1.392	71.19	$P2_1/c$	1	4.0	10.5	33.0	90.00	90.0	90.41	-178.93
C48	-171.60	1.388	70.8	$C2/c$	1	7.9	10.9	17.8	82.80	77.0	68.67	-179.92
C634	-171.54	1.395	71.32	$P2_1/c$	1	5.0	15.3	18.7	103.87	90.0	90.00	-180.24
A682	-171.53	1.360	69.43	$P2_1/c$	1	5.4	12.2	21.9	90.00	90.0	101.18	-176.93
A734	-171.38	1.339	68.41	$C2/c$	1	5.0	16.9	17.9	74.12	81.8	81.39	-179.66
A803	-171.34	1.390	71.27	$P\bar{1}$	1	4.9	8.25	17.9	97.05	94.6	104.47	-181.17
A497	-171.17	1.387	70.97	$C2/c$	1	4.5	13.5	23.8	104.26	90.0	99.75	-177.38
C190	-171.08	1.406	72.08	$C2/c$	1	4.2	14.1	24.2	75.04	84.9	81.36	-178.45
A307	-171.05	1.388	71.05	*P21*	2	5.6	8.31	30.7	90.00	90.0	101.10	-179.62
C295	-171.05	1.387	70.9	$C2/c$	1	4.9	15.8	18.3	85.01	82.1	80.91	-178.06
A218	-170.95	1.352	69.11	$C2/c$	1	5.3	11.3	24.9	85.65	83.8	76.45	-179.24
C761	-170.87	1.392	71.23	$P2_1/c$	1	4.6	10.0	30.0	90.00	90.0	94.17	-179.73
C75	-170.84	1.398	71.91	$P2_1/c$	1	3.9	15.3	23.3	90.00	90.0	96.79	-179.13
C584	-170.77	1.417	72.53	$C2/c$	1	4.1	18.2	20.1	112.43	90.0	96.50	-176.73
C838	-170.74	1.403	71.97	$C2/c$	1	5.1	14.7	18.5	96.39	98.0	90.00	-179.81
A429	-170.73	1.372	69.49	$P2_1/c$	1	4.3	10.7	30.9	90.00	92.2	90.00	-178.10
C777	-170.52	1.393	71.83	$P2/c$	1	4.1	16.5	20.7	97.99	90.0	90.00	-177.74
C439	-170.52	1.396	71.43	$P2_1/c$	1	3.9	19.0	19.4	107.04	90.0	90.00	-177.40
C465	-170.49	1.394	70.99	$C2/c$	1	3.9	15.3	23.9	77.80	85.2	82.57	-176.69
C275	-170.47	1.407	72.17	$P2_1/c$	1	4.1	17.0	19.7	90.00	94.6	90.00	-176.54
A179	-170.42	1.391	70.91	$C2/c$	1	5.2	10.8	25.6	96.18	90.0	104.08	-177.66
A706	-170.35	1.389	70.62	$P\bar{1}$	1	5.5	8.40	15.7	81.41	80.0	78.90	-178.03
C771	-170.29	1.390	70.82	$C2/c$	1	3.9	16.0	22.8	78.65	85.0	82.89	-177.21
C113	-170.22	1.368	69.82	$C2/c$	1	10.	10.9	12.4	94.87	94.8	104.48	-179.42
A182	-170.12	1.361	69.5	$C2/c$	1	4.2	11.1	31.8	80.39	86.2	79.11	-177.72
C174	-170.04	1.412	71.99	$P\bar{1}$	1	4.7	9.85	15.2	91.45	96.7	101.28	-175.02
C131	-170.03	1.393	71.23	$P2_1/c$	1	8.2	10.8	16.1	90.00	103.	90.00	-180.03
C432	-170.00	1.403	71.87	$C2/c$	1	3.9	18.5	19.7	105.54	90.0	96.17	-175.90
C893	-169.97	1.319	66.96	$R\bar{3}$	1	6.7	19.7	19.7	118.73	96.5	96.53	-177.38
A184	-169.96	1.389	70.95	$C2/c$	1	4.6	16.6	19.1	73.77	82.9	81.91	-177.47
A494	-169.83	1.389	70.81	$P\bar{1}$	1	5.6	8.29	15.9	75.18	80.6	78.21	-177.51
A435	-169.82	1.369	70.38	$P2_1/c$	1	7.4	7.86	24.6	90.00	98.3	90.00	-177.10
A88	-169.78	1.405	71.92	$P2_1/c$	1	4.9	16.8	17.6	109.20	90.0	90.00	-176.94
C173	-169.74	1.393	71.37	*Pbca*	1	7.4	11.2	33.5	90.00	90.0	90.00	-178.59

A482	-169.73	1.394	71.58	$P2/c$	1	3.8	10.6	35.0	96.78	90.0	90.00	-174.75
C396	-169.65	1.385	70.8	$P2_1/c$	1	4.2	13.2	25.9	101.63	90.0	90.00	-176.08
A457	-169.52	1.368	69.84	$C2/c$	1	4.1	18.4	20.5	67.93	84.2	83.58	-176.30
A142	-169.51	1.398	71.64	$P2_1/c$	1	3.9	19.0	20.8	115.13	90.0	90.00	-176.50
C660	-169.42	1.367	69.85	$C2/c$	1	3.9	13.1	28.0	84.46	85.9	81.29	-176.36
A167	-169.37	1.378	70.58	$P2_1/c$	1	4.2	18.0	19.9	111.77	90.0	90.00	-177.00
A129	-169.37	1.385	71.18	$P2_1/c$	1	4.6	11.6	26.2	90.00	90.7	90.00	-177.78
A147	-169.35	1.382	70.74	$P2_1/c$	1	4.4	17.7	19.0	109.62	90.0	90.00	-176.86
A588	-169.35	1.382	71.13	$P2/c$	1	3.9	16.9	22.0	104.33	90.0	90.00	-176.87
A194	-169.32	1.410	71.83	$P\bar{1}$	1	5.5	8.40	15.5	103.89	91.0	100.19	-177.26
C72	-169.31	1.403	71.98	$P2_1/c$	1	7.3	11.2	16.9	90.00	97.8	90.00	-176.83
A253	-169.29	1.356	69.29	$P2_1/c$	1	5.3	13.7	19.7	97.18	90.0	90.00	-176.70
C888	-169.29	1.379	71.06	$P2/c$	1	4.1	16.1	21.5	99.66	90.0	90.00	-176.12
C594	-168.94	1.396	71.47	$C2/c$	1	3.9	14.2	25.5	86.84	85.5	82.04	-175.09
C749	-168.93	1.367	69.94	$P2_1/c$	1	6.3	11.5	19.4	90.00	90.2	90.00	-177.67
A177	-168.85	1.387	70.82	$C2/c$	1	3.8	17.1	21.6	98.61	90.0	96.51	-174.01
A163	-168.78	1.374	70.13	$C2/c$	1	4.6	14.3	21.7	96.99	90.0	99.36	-176.70
C740	-168.69	1.382	70.93	$P2_1/c$	1	4.0	16.9	20.5	93.16	90.0	90.00	-175.52
A241	-168.63	1.389	70.98	$C2/c$	1	3.9	18.5	21.5	64.74	84.7	83.89	-174.65
A490	-168.62	1.382	70.52	$C2/c$	1	3.9	11.0	33.4	98.78	90.0	100.38	-175.38
A688	-168.57	1.347	68.92	$C2/c$	1	4.3	16.2	21.3	103.68	90.0	97.77	-175.99
A812	-168.55	1.329	67.55	$C2/c$	1	6.6	11.5	20.7	75.88	80.7	73.30	-177.42
A375	-168.51	1.397	71.19	$C2/c$	1	4.1	13.4	25.7	100.71	90.0	98.97	-175.48
A695	-168.48	1.392	71.05	$C2/c$	1	4.1	16.3	22.0	74.03	84.6	82.76	-175.03
A314	-168.43	1.398	71.35	$P2_1$	1	4.0	13.0	13.7	106.56	90.0	90.00	-175.80
C766	-168.32	1.394	71.44	$P2_1/c$	1	4.1	16.8	21.0	106.46	90.0	90.00	-175.26
C322	-168.31	1.423	72.84	$P2_1/c$	1	6.1	11.2	19.9	91.63	90.0	90.00	-174.49
C545	-168.29	1.378	70.69	$C2/c$	1	5.2	12.1	23.1	81.14	83.4	77.52	-175.25
A204	-168.28	1.351	68.76	$P2_1/c$	1	4.0	12.6	28.3	90.00	90.0	92.56	-176.68
C154	-168.22	1.407	71.96	$P2_1/c$	1	4.3	15.4	20.5	91.35	90.0	90.00	-173.83
C149	-168.17	1.403	71.77	$P2_1/c$	1	3.8	15.5	23.2	90.00	90.0	90.96	-176.54
C116	-168.15	1.393	71.75	$P2_1/c$	1	6.1	10.5	22.1	99.64	90.0	90.00	-175.67
A315	-168.13	1.341	68.07	$P2_1/c$	1	5.0	13.8	21.2	100.73	90.0	90.00	-175.59
A614	-168.12	1.383	70.34	$P2_1/c$	1	4.0	12.0	29.8	100.27	90.0	90.00	-174.29
C672	-168.06	1.407	72.06	$P\bar{1}$	1	7.0	9.33	10.6	88.44	84.0	86.74	-175.67
C797	-168.00	1.375	69.98	$C2/c$	1	4.0	15.9	22.6	88.77	84.9	82.79	-173.84
C811	-167.98	1.367	70.04	$P2_1/c$	1	4.9	14.2	20.5	94.00	90.0	90.00	-176.01
A658	-167.91	1.358	69.05	$P2_1/c$	1	3.8	17.5	21.6	96.63	90.0	90.00	-173.55
C93	-167.86	1.398	71.38	$C2/c$	1	8.7	9.16	19.9	92.29	90.0	118.45	-174.16
A616	-167.83	1.342	68.66	$C2/c$	1	4.3	17.9	20.9	65.72	84.1	83.13	-175.44
C165	-167.72	1.409	71.73	$P\bar{1}$	1	8.1	8.98	10.2	76.84	75.0	80.32	-176.45
C769	-167.66	1.397	71.55	$P2_1/c$	1	3.8	18.4	19.6	90.00	93.5	90.00	-173.10
C95	-167.64	1.309	66.47	$P2_1/c$	1	4.9	16.2	18.5	90.00	95.6	90.00	-175.35
A224	-167.50	1.309	66.62	$P2_1/c$	1	4.8	11.9	25.8	93.48	90.0	90.00	-174.54
C434	-167.50	1.324	67.81	$P2_1/c$	1	5.9	14.7	17.5	105.10	90.0	90.00	-176.79
A280	-167.46	1.375	70.07	$C2/c$	1	3.9	15.4	24.2	103.73	90.0	97.38	-174.18
C574	-167.44	1.412	72.08	$P\bar{1}$	1	5.8	10.6	11.7	99.74	103.	96.09	-173.88
A371	-167.24	1.377	70.26	$P\bar{1}$	1	7.5	7.63	12.7	102.90	94.7	91.46	-173.77
C529	-167.22	1.377	70.52	$C2/c$	1	4.8	10.0	30.5	91.03	90.0	103.90	-174.01
C270	-167.10	1.331	68.15	$R\bar{3}$	1	4.7	23.1	23.1	119.53	93.9	93.94	-176.02
A648	-166.88	1.319	66.92	$P2/c$	1	5.0	11.8	24.7	92.88	90.0	90.00	-174.13
A463	-166.86	1.397	71.48	$P\bar{1}$	1	3.9	13.1	13.6	100.04	95.0	93.01	-170.97
A758	-166.76	1.351	68.92	$C2/c$	1	4.1	16.3	21.6	96.52	90.0	97.33	-173.74
A226	-166.64	1.356	68.97	$C2/c$	1	5.1	13.5	21.9	76.68	83.3	79.14	-173.86
C546	-166.60	1.383	70.5	$P2_1/c$	1	6.5	14.5	15.1	90.00	90.0	98.40	-173.45
A396	-166.57	1.350	68.27	$P\bar{1}$	1	4.0	12.5	15.1	73.65	83.5	89.77	-174.27
A242	-166.56	1.422	72.27	$P2_1/c$	1	4.1	15.9	21.0	93.95	90.0	90.00	-173.81
C74	-166.51	1.386	71.02	$P2_1/c$	1	9.4	10.9	14.0	90.00	103.	90.00	-176.99
A415	-166.43	1.339	68.58	$P2_1/c$	1	7.3	9.49	20.9	91.71	90.0	90.00	-174.44
A205	-166.41	1.364	68.86	$P\bar{1}$	1	6.2	9.04	13.7	102.85	91.3	106.68	-173.51
C180	-166.38	1.389	70.9	$P2_1/c$	1	9.9	11.9	12.0	90.00	90.0	100.22	-176.17
A217	-166.34	1.379	70.17	$P2_1/c$	1	4.8	17.0	17.8	105.71	90.0	90.00	-173.38
A605	-166.25	1.357	69.4	$P2/c$	1	4.5	17.2	19.2	106.21	90.0	90.00	-173.06

C54	-166.23	1.397	71.57	$P\bar{1}$	1	7.5	10.2	10.8	109.90	102.	107.93	-172.68
A244	-166.21	1.373	70.37	$P2_1/c$	1	5.1	15.8	17.5	97.35	90.0	90.00	-171.20
C600	-166.15	1.335	68.03	$C2/c$	1	5.8	13.0	19.8	94.49	90.0	102.91	-174.76
A329	-166.02	1.360	69.8	$C2/c$	1	6.2	10.4	23.5	85.83	82.4	72.70	-172.93
A484	-166.01	1.371	69.87	$C2/c$	1	4.4	11.1	29.5	95.77	90.0	101.53	-173.44
A509	-165.98	1.374	70.39	$C2/c$	1	3.9	16.9	21.8	100.19	90.0	96.66	-172.57
A516	-165.94	1.389	70.66	$C2/c$	1	4.6	11.1	28.5	97.72	90.0	101.94	-173.09
A610	-165.93	1.374	70.15	$C2/c$	1	4.1	14.5	24.2	79.85	85.0	81.79	-170.36
A607	-165.79	1.337	68.28	$C2/c$	1	4.5	11.0	30.2	99.07	90.0	102.00	-174.33
A504	-165.64	1.376	70.33	$C2/c$	1	3.8	19.4	21.5	116.40	90.0	95.63	-170.86
C102	-165.61	1.387	70.61	$P2_1/c$	1	6.6	14.5	14.9	103.56	90.0	90.00	-174.67
C895	-165.60	1.377	69.99	$C2/c$	1	5.9	15.7	16.0	77.55	79.2	79.10	-172.63
A525	-165.55	1.332	67.8	$P2_1/c$	1	4.0	15.5	23.8	96.83	90.0	90.00	-172.29
A579	-165.50	1.341	68.4	$C2/c$	1	4.5	11.1	29.3	90.46	90.0	101.89	-172.59
A428	-165.50	1.415	72.53	$C2$	1	4.0	13.3	13.3	102.06	98.6	90.00	-169.75
A808	-165.33	1.334	68.17	$C2/c$	1	3.9	15.2	24.6	83.44	85.3	82.50	-172.08
A321	-165.22	1.378	70.36	$C2$	1	4.5	11.1	14.6	82.86	81.0	78.19	-172.91
C183	-165.17	1.358	69.67	$C2/c$	1	4.6	16.4	19.3	98.26	90.0	98.10	-171.69
A161	-164.96	1.393	71.41	$P2_1/c$	1	5.1	9.25	29.5	90.00	90.0	90.22	-170.57
A437	-164.90	1.392	71.03	$P2_1/c$	1	3.9	13.2	27.3	103.76	90.0	90.00	-172.34
A637	-164.89	1.338	67.77	$P2/c$	1	4.7	10.9	28.3	95.39	90.0	90.00	-172.42
A693	-164.88	1.335	67.94	$C2/c$	1	5.1	14.2	20.8	81.38	82.9	79.68	-172.30
C805	-164.86	1.362	69.89	$P4_2/n$	1	6.7	20.6	20.6	90.00	90.0	90.00	-174.55
A572	-164.82	1.364	69.52	$C2$	1	4.1	11.1	16.0	95.22	90.0	100.58	-172.05
A704	-164.78	1.369	69.93	$C2/c$	1	6.1	14.4	18.0	67.22	80.1	77.66	-171.18
A660	-164.68	1.328	67.84	$P\bar{1}$	1	6.0	10.4	12.6	100.97	98.6	106.29	-173.44
CisC32	-164.51	1.416	71.03	$P2_1/c$	1	4.0	16.0	21.3	90.00	93.2	90.00	-176.38
A243	-164.30	1.368	70.05	$Pbca$	1	5.4	15.9	32.8	90.00	90.0	90.00	-171.27
A330	-164.21	1.361	69.53	$P2_1/c$	1	5.0	17.3	17.6	111.35	90.0	90.00	-168.74
A511	-163.92	1.306	67.02	$P2_1/c$	1	4.0	16.7	22.2	95.09	90.0	90.00	-171.70
C100	-163.63	1.388	72.32	$P\bar{1}$	1	6.4	7.77	14.5	76.79	85.9	88.09	-171.72
A823	-163.42	1.371	69.84	$P2_1/c$	1	5.3	16.5	16.8	104.01	90.0	90.00	-172.24
A657	-163.39	1.404	72.21	$Pna2_1$	1	4.6	10.1	29.8	90.00	90.0	90.00	-170.41
A581	-163.14	1.397	71.69	$C2/c$	1	3.9	10.9	33.7	98.10	90.0	100.23	-167.82
A783	-163.09	1.358	69.49	$P2_1/c$	1	5.8	13.0	18.9	90.00	90.0	98.91	-170.09

Chapter 5: Crystal structure informatics for defining the conformational search space of large flexible molecules

5.1 Introduction

The previous two chapters have shown that computational cost is a big issue for performing CSP on pharmaceutical-like molecules, as it increases drastically with molecular size and flexibility (see also Chapter 1.2.3). Producing the crystal energy landscape of molecule XXVI[1] was twenty times more expensive than it was for two tautomers of mebendazole, even if an almost identical procedure was followed. The reason is that mebendazole is smaller than XXVI, and its main degrees of freedom can take fewer values; the latter was not obvious at the onset of these studies. Since most molecules in drug development are conformationally flexible,[2] this is one of the factors hindering a routine utilisation of computational methods in industrial solid form screening. Thus finding methods to reduce the computational expense and its scaling with molecular size and flexibility would be important.

All CSP methods start with a crystal structure search (see Chapter 2.4.1), which aims to generate plausible packing arrangements of a molecule by finding all the most important local minima on the potential energy surface.[1] In all successful CSP methodologies, the lattice energy (E_{latt}) surface is explored during the search.[1, 3, 4] The E_{latt} search space is immense, and its size increases drastically with the number of flexible conformational degrees of freedom (CDFs).[2, 5] It is fundamental that the conformations that occur in the most important E_{latt} minima are included in the search to guarantee completeness. These cannot be limited to the isolated-molecule local minima in conformational energy, since intermolecular interactions can significantly distort conformers[6, 7] (as it was the case for molecule XXVI, see Chapter 3.3.1).

One possible solution is to perform a set of rigid searches for all the conformations that have an energy penalty small enough to be counterbalanced by improvements in intermolecular interactions.[8] This approach can be effective in some cases,[9, 10] but for a molecule like XXVI (or ritonavir) this would require an unfeasible number of searches.[11] An alternative approach is to treat some CDFs as explicit search variables.[8, 12] In general, only low-barrier torsion angles need to be treated as flexible in the search, as semi-rigid dihedrals (such as those in phenyl rings), bond-angles and bond-lengths do not separate structures belonging to different E_{latt} minima (but they are important for performing accurate optimisations of the generated structures, see Chapter 6).[12, 13] On the other hand if a torsion angle can only take a limited set of values within very limited ranges, such as an amide group that can only be in a *cis* or a *trans* configuration,[14] a more

efficient approach is to perform separate searches in for all its possible values, which are deemed to define separate conformational regions (CRs).[13, 15]

Treating torsion angles as explicitly flexible in a search requires an accurate calculation of the energy penalty for their variation from the isolated molecule global minimum in conformational energy (*i.e.* ΔE_{intra}).[8, 12] This is because the use of cheap transferable force-fields can be inaccurate and limit the ability of a CSP method to locate all the most important local minima in E_{latt}[16] (this occurred in an early study on aspirin).[17] Hence, in the most effective CSP methodologies electronic structure calculations are performed to estimate ΔE_{intra} for varying the flexible torsion angles.[1] This can be done implicitly through the use of *ab initio* calculations to establish suitable tailor-made force-fields,[18] or explicitly by performing 'scans' on the isolated molecule with electronic structure methods to create grids of ΔE_{intra} values;[8] the latter method is used in CrystalPredictor[19] (see Chapter 2.4.1.2). However, both approaches are computationally expensive, in particular for very flexible molecules, and they often require a large amount of human time to set up the calculations.[1, 20]

This chapter seeks to simplify the definition of the conformational search space of flexible molecules using information retrieved from the Cambridge Structural Database (CSD),[21] with the goal of reducing the computational cost of CSP searches without reducing their effectiveness in identifying potential polymorphs. After exploring what sort of data on conformational preferences can be retrieved from the CSD on small molecules, as an essential validation of the approach, this chapter illustrates the development and testing of a workflow that uses this information to set up CSP searches. This workflow reduces the number of *ab initio* calculations that are required to determine the flexibility ranges of the most important torsion angles and to generate plausible crystal structures. This approach was developed from the five pharmaceutical-like flexible molecules shown in Figure 5.4, which had been subject to previous CSP studies in the Price group performed with a standard methodology: molecule XXVI (see Chapter 3),[1] mebendazole (Chapter 4), GSK269984B,[20] molecule XXIII[1] and molecule XX.[15, 22] It was then tested for its ability to generate the most important low-energy crystal structures of the same five molecules. Finally, the workflow was applied to succinic acid to test whether it could reproduce a newly discovered conformational polymorph (the γ form).[23]

5.2 Methods

5.2.1 Preliminary analysis of CSD conformational information on small molecules

5.2.1.1 Rotamer distributions

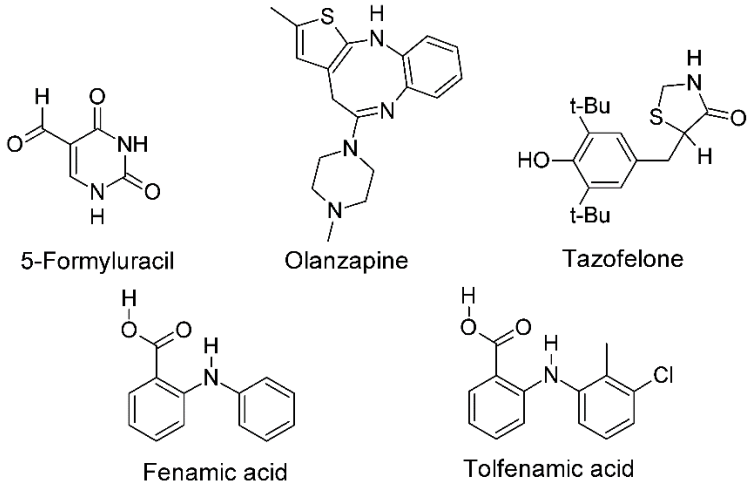


Figure 5.1: Chemical diagrams of the small molecules used to investigate the ability of CSD information on geometric preferences to define the conformational search space.

An initial test was performed on five small molecules (Figure 5.1), all of which had been subject to full CSP studies,[24-27] to illustrate some types of conformational behaviours and how CSD information on torsion angle distributions could be applied in CSP. The aim was to verify whether CSD informatics tools provided information on the conformational space of some relatively simple molecules that is consistent with what had been determined using accurate *ab initio* methods, before testing them on some more challenging large and flexible targets. For each molecule, the rotamer distributions were retrieved from the CSD conformational libraries (see Chapter 2.6.3),[28] using a stand-alone programme. The retrieved distributions were also analysed via kernel density estimation (KDE) with the Von Mises kernel, using the method described in Chapter 2.6.3.1.[29] Figure 5.2 shows three examples of possible histograms and PDFs derived from the rotamer distributions.

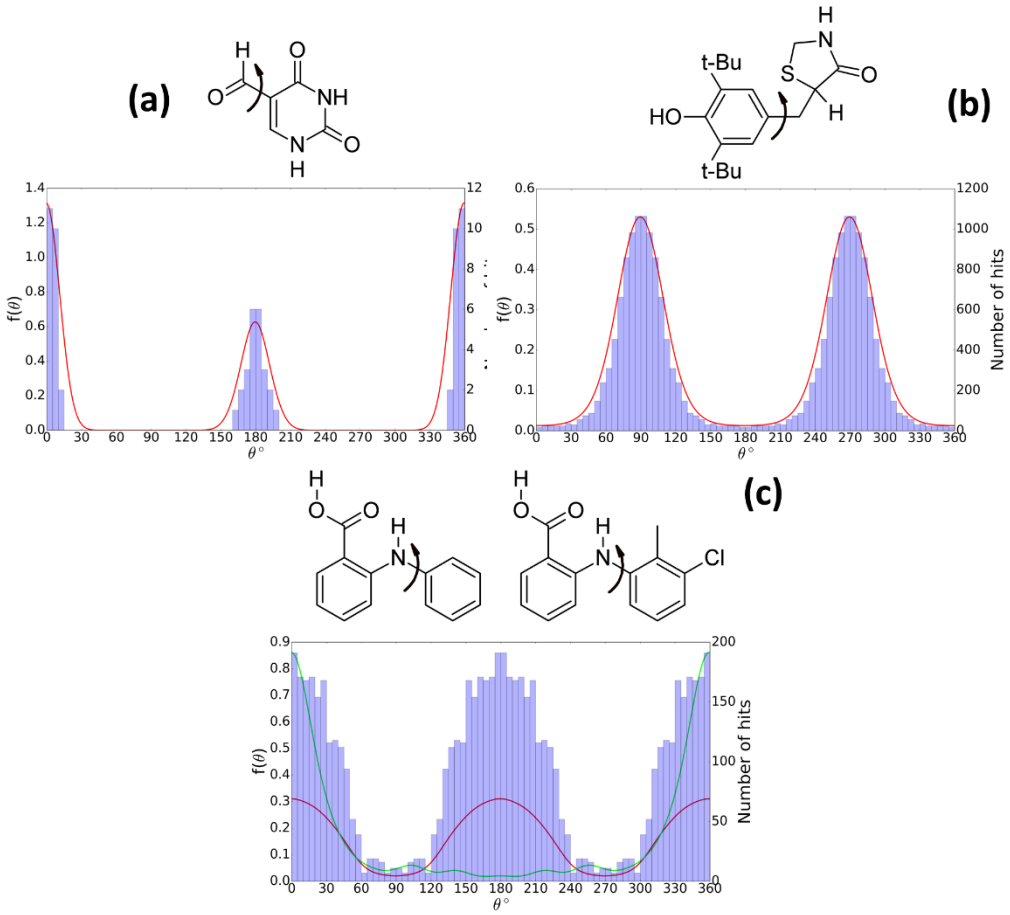


Figure 5.2: Histograms (light purple bars) and Von Mises KDE PDFs (red lines) describing the torsion angle distributions of the dihedral angles indicated on each molecular diagram of (a) 5-Formyluracil (0° in the diagram) (b) Tazofelone (0° in the diagram) (c) Fenamic acid (0° in the diagram), with an overlay of the PDF for tolfenamic acid in green, showing the effect of the additional methyl and Cl substituents.

In all three cases, the insights from the rotamer distributions are similar to those derived from the *ab initio* scans performed on these molecules in the original CSP studies. For 5-formyluracil (Figure 5.2a), the distributions indicate that only two values are possible within very narrow ranges (i.e. 0 and 180°).[24] In tazofelone, there is quite a wide spread of possible values around both 90 and 270° (Figure 5.2b), while the remainder have very low probabilities.[26] In the final example of the fenamates, the only low-barrier torsion angle (Figure 5.2c) can adopt any value in fenamic acid, but the methyl substitution in tolfenamic acid significantly reduces the probability of a crystalline conformation between 60 and 300° because of steric interactions.[27]

In summary, this initial test has shown that CSD torsion angle distributions can capture the conformational behaviour of these molecules consistently with their *ab initio* energy profiles.

5.2.1.2 The CSD Conformer Generator (CG)

The CSD Conformer Generator (CG) uses the information from the rotamer distributions to generate plausible molecular conformations;[30] its functioning is explained in Chapter 2.6.4. The effectiveness of the CG was tested by verifying whether it could reproduce the molecular conformations of the experimentally-known single-component crystal structures of the molecules in Figure 5.1. Hence the CG, with its default settings for molecular clustering and maximum number of unusual torsion angles and without any limit in number of conformations and probability scores, was used on each of these molecules. It produced two conformations for 5-formyluracil, 37 for olanzapine, 10 for tazofelone, 67 for fenamic acid and 83 for tolfenamic acid. The full set of generated conformations was analysed to verify whether they contained matches to the experimental targets. The analysis was performed with the Crystal Packing Similarity tool[31] (see Chapter 2.5.1) available through the CSD Python API (Chapter 2.6.5).[21] All experimental crystalline conformations of the five small molecules were reproduced very well by the CG, with $RMSD_1$ < 0.35 Å, as shown in Figure 5.3 and Table 5.1.

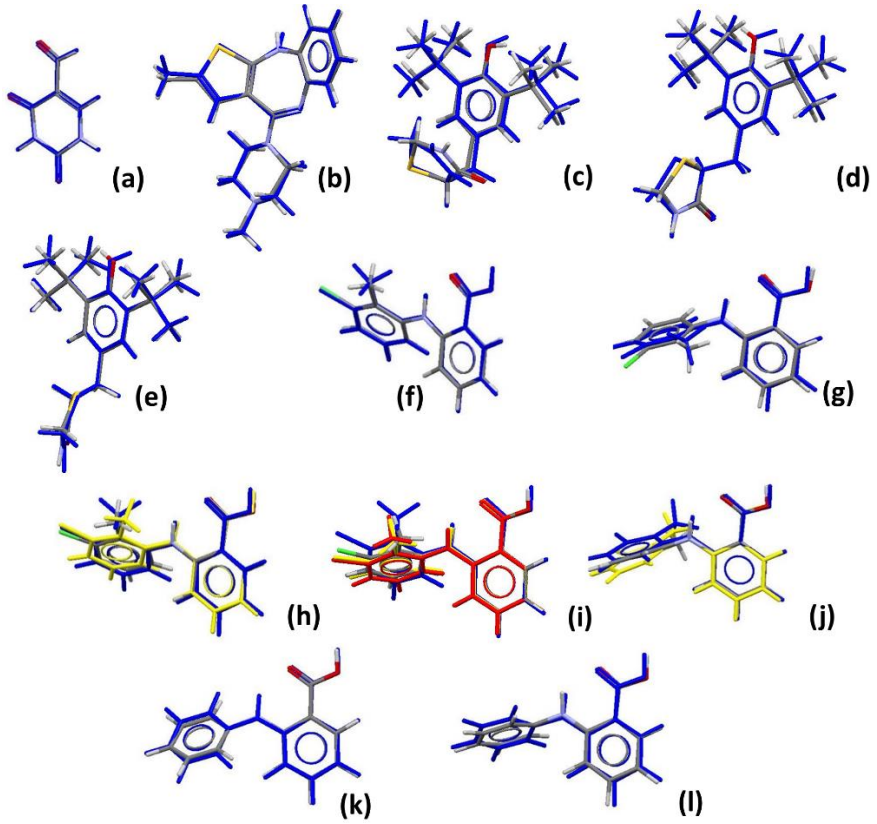


Figure 5.3: Overlays of the experimental conformations of the molecules in Figure 5.1 (coloured by elements) with their best matches produced by the CG (in blue). If the same CG conformation was the closest match of each molecule in the asymmetric unit of Z'>1 crystal structures, the other experimental conformations are coloured in red or in yellow. Polymorphs with very similar conformations are shown only once.

Table 5.1: Quantification of the ability of CG to reproduce the experimental conformations, as shown in Figure 5.3.

Label on Figure 5.3	Crystal structure of molecule	Generated conformations	Ranking of most similar conformation	$RMSD_1$(Å)
a	5-Formyluracil	2	2	0.019
b	Olanzapine forms I and II	37	1	0.151
c	Tazofelone forms I and II	10	5	0.337
d, e	Tazofelone form III (solid solution)	"	4, 1	0.260, 0.123
f	Tolfenamic acid form 1	83	2	0.087
g	Tolfenamic acid form 2	"	16	0.258
h	Tolfenamic acid form 3 (Z` = 2)	"	2, 2	0.219, 0.175
i	Tolfenamic acid form 4 (Z` = 3)	"	2, 2, 2	0.277, 0.314, 0.287
j	Tolfenamic acid form 5 (disordered)	"	15, 15	0.301, 0.228
k, l	Fenamic acid (Z` = 2)	67	2, 6	0.169, 0.185

This analysis revealed that the range of CG conformations covers, or even exceeds, the conformational search space considered in the original searches. Hence, a set of rigid searches would have captured the entire flexibility ranges of these small molecules. Furthermore an analysis of the rotamer distributions, *e.g.* those in Figure 5.2, would make the choice of which torsion angle/s should be treated as explicitly flexible to make searches more efficient straightforward.

5.2.2 Development of a workflow to generate the crystal structures of the five flexible molecules

5.2.2.1 Extension of the use of conformational information on larger molecules

Although the analysis in Chapter 5.1.1 showed that CSD conformational information could straightforwardly be applied to the small molecules in Figure 5.1, for larger and more flexible targets more significant challenges can be anticipated. In this chapter, the five molecules shown in Figure 5.4 were used to test the applicability of CSD conformational information to molecules of pharmaceutical interest.

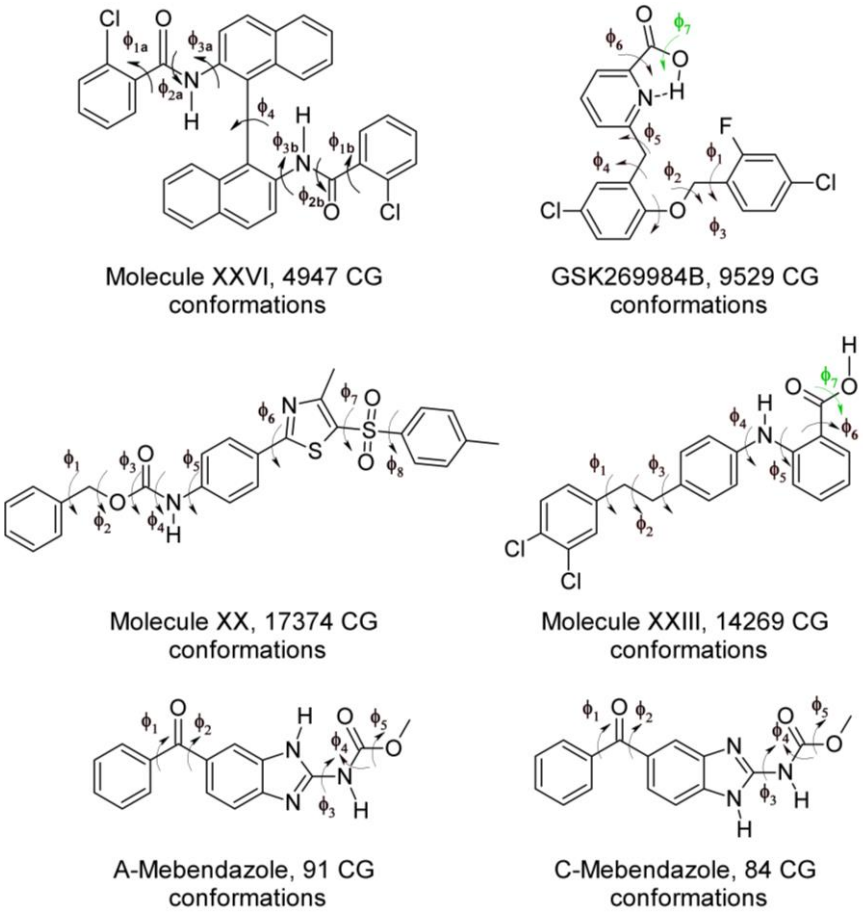


Figure 5.4: Chemical diagrams of the molecules used to test the applicability of CSD information to large and flexible targets, showing the torsion angles that are identified as flexible by the rotamer libraries and the number of distinct conformations generated by the CG. The additional angles not identified by the rotamers libraries are in green and define the position of polar hydrogen atoms. Atomic numbering can be found in Appendix Figure 5.1, and the definition of the torsion angles in Appendix Table 5.1.

The CG generated a significantly higher number of possible conformations than for the smaller molecules, as shown in Figure 5.4. The Crystal Packing Similarity tool was used to find the closest matches to the experimental conformations. The results can be found in Figure 5.5 and Table 5.2.

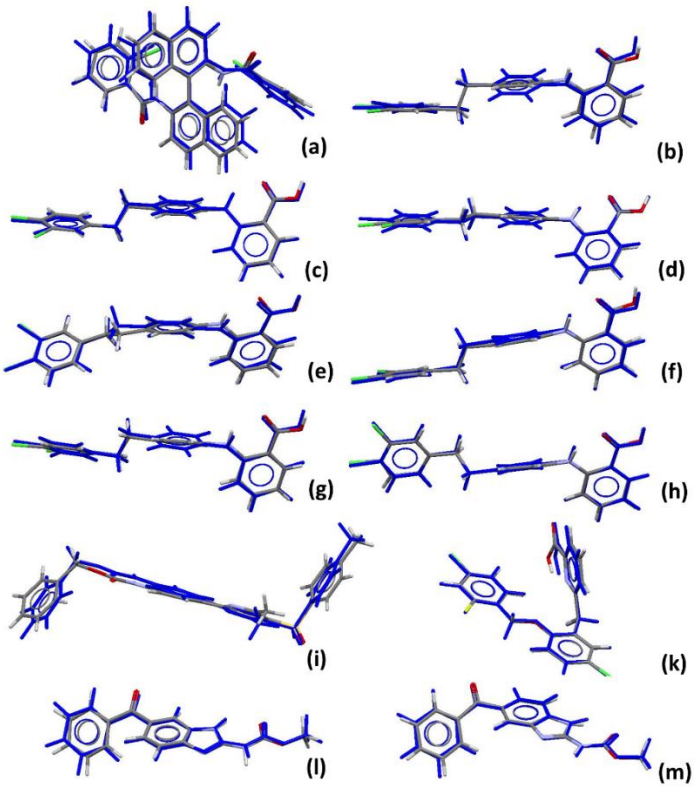


Figure 5.5: Overlays of the experimental conformations of the molecules in Figure 5.4 (coloured by elements) with their best matches produced by the CG (in blue).

Table 5.2: Quantification of the ability of CG to reproduce the experimental conformations, as shown in Figure 5.5.

Label on Figure 5.5	Crystal structure of molecule	Generated conformations	Ranking of most similar conformation	$RMSD_1$(Å)
a	XXVI	4947	685	0.386
b	XXIII form a	14269	440	0.683
c	XXIII form b	"	406	0.65
d, e	XXIII form c (Z` = 2)	"	657, 3160	0.66, 0.692
f	XXIII form d	"	406	0.192
g, h	XXIII form e (Z` = 2)	"	491, 411	0.235, 0.627
l	XX	17374	15	0.419
j	GSK269984B	9529	166	0.129
k	Mebendazole form A	91	2	0.157
l	Mebendazole form C	84	1	0.133

Compared to the results in Table 5.1, Table 5.2 shows that the reproductions of the conformations of the larger molecules were of poorer quality. Even when the $RMSD_1$ values are small, there are visual differences in the orientations of key peripheral bulky groups. This suggests that performing CSP searches with the closest CG-generated conformations being treated as rigid may not find some experimental structures because the exact E_{latt} local minima would not be reachable upon optimisation.[2] Furthermore the large number of generated conformations means that many *ab initio* calculations would be required to select those that should be searched, given how poorly the CG probability-based score correlates with conformational energy (see Appendix Figure 5.2).

Performing *ab initio* calculations and possibly rigid searches on thousands of conformations would be unfeasibly expensive.

A related issue is that some torsion angles affect the overall shape of the molecule more than others, as illustrated in Figure 5.6.[11] Hence the effect of the variation of a torsion angle on the overall molecular shape, and so on its packing capability, should be considered in CSP.

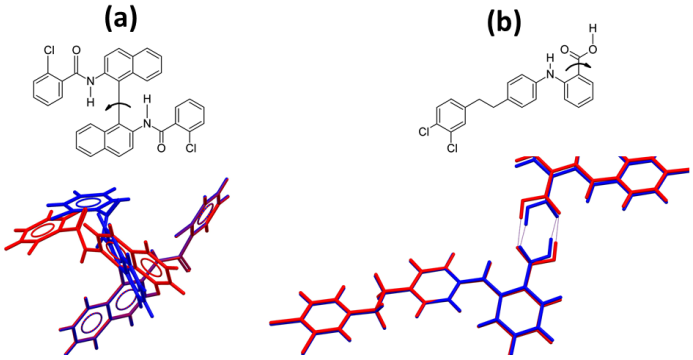


Figure 5.6: Visual comparison of the effects of changing by 30° (a) torsion angle Φ_4 in molecule XXVI, which strongly affects molecular shape and (b) torsion angle Φ_6 in molecule XXIII, which has a very small effect on the shape of the hydrogen bonded dimer, except in the proximity of the carboxylic acid functional groups.

In summary, for large and flexible molecules CSD conformational information cannot be used as it is to perform CSP searches, but a more complex workflow is required.

5.2.2.2 The workflow

The proposed workflow was developed from the molecules in Figure 5.4.[11] The specific parameters of this workflow were heuristically determined, and may need adaption to different molecules, in particular if they are larger and more flexible. However, the aim is to propose a general approach to use information retrieved from the CSD to establish the most efficient treatment of conformational flexibility in the searches.

More specifically, the individual rotamer distributions fall into two classes. On the one hand the rotamer distributions can be tightly clustered around a few specific values (*e.g.* in 5-formyluracil, see Figure 5.2a), which will be picked by the CG. These angles can be treated as fixed in the searches, with different values defining separate CRs.[13, 15] Alternatively, there can be a wide range of possible values (*e.g.* in fenamic acid, see Figure 5.2c): these torsion angles must be considered as search variables.[13] Furthermore, the effect on molecular shape (Figure 5.6) needs to be accounted for in deciding how to treat a torsion angle: dihedrals that have a large shape impact should only be constrained if they can take few values within very narrow intervals, as even a small variation within a range could affect the packing possibilities of the molecule.

The workflow developed in this chapter includes several steps:

1) *For each torsion angle, analyse the rotamer distributions and the effect on molecular shape.* After extracting the rotamer distributions from the knowledge-based libraries, histograms of the torsion angles in Figure 5.4 were plotted and superimposed with their PDFs calculated via Von Mises KDE using the Matplotlib[32] Python[33] package. The effect of each torsion angle on molecular shape was then analysed by ultra-fast shape recognition (USR, see Chapter 2.5.3),[34] using the RDKit[35] and USRCAT[36] Python packages. The results of these analyses can be found in Appendix Figures 5.3-5.7 and Appendix Tables 5.2-5.6.

For torsion angles defining the positions of polar hydrogen atoms, which are very important in determining hydrogen bonding patterns,[37] or any other angle identified by CG as potentially flexible but with insufficient data in the CSD, explicit *ab initio* conformational energy scans must be performed. This was the case for angles Φ_7 in GSK269984B and Φ_7 in XXIII, both of which define the position of polar hydrogen atoms. They were scanned with Gaussian 09[38] at the PBE0 6-31G(d,p) level of theory.

2) Assess the nature of each torsion angle. This analysis aims to discriminate between torsion angles to be treated as search variables and those that can be constrained to a set of CG values, defining separate CRs. The decision tree in Figure 5.7 was used for this purpose.

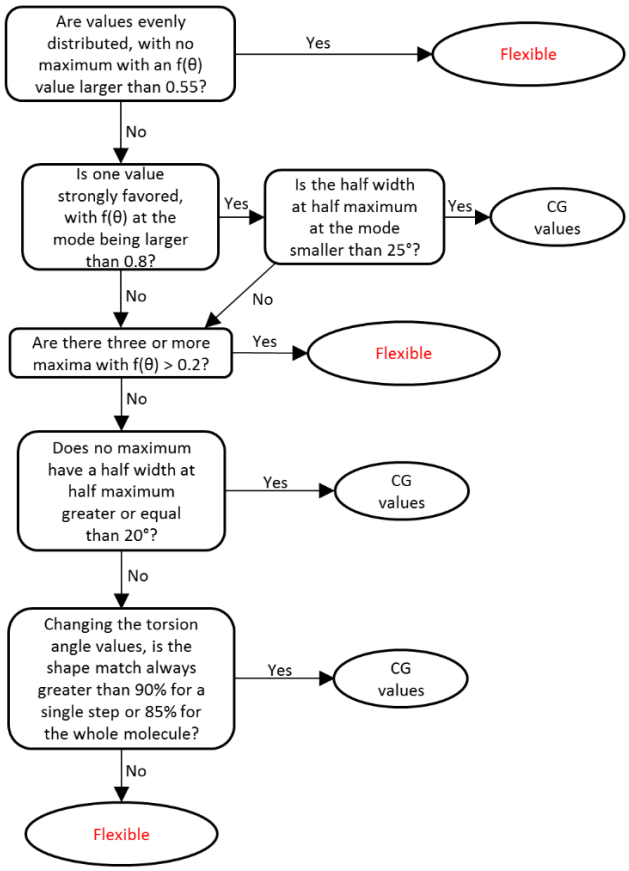


Figure 5.7: Decision tree used to discriminate between constrained and explicitly flexible torsion angles on the basis of the PDF *f(θ)* values and the changes in shape associated with their variation.

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. A set of CG-generated conformations with different combinations of the constrained torsion angles needs to be selected, which define separate CRs. Each CR requires a separate sub-search. Since the CG has a complex algorithm for picking the values of each torsion angle, aimed at maximising diversity,[30] many CG conformations do not have angles that are at the peaks of distributions. This diversity is important, but the number of conformations (and associated CRs) to be searched needs to be limited to keep the computational cost manageable. Thus, the decision tree shown in Figure 5.8 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 Python script to select those CG-generated conformations that were worth considering as distinct CRs.

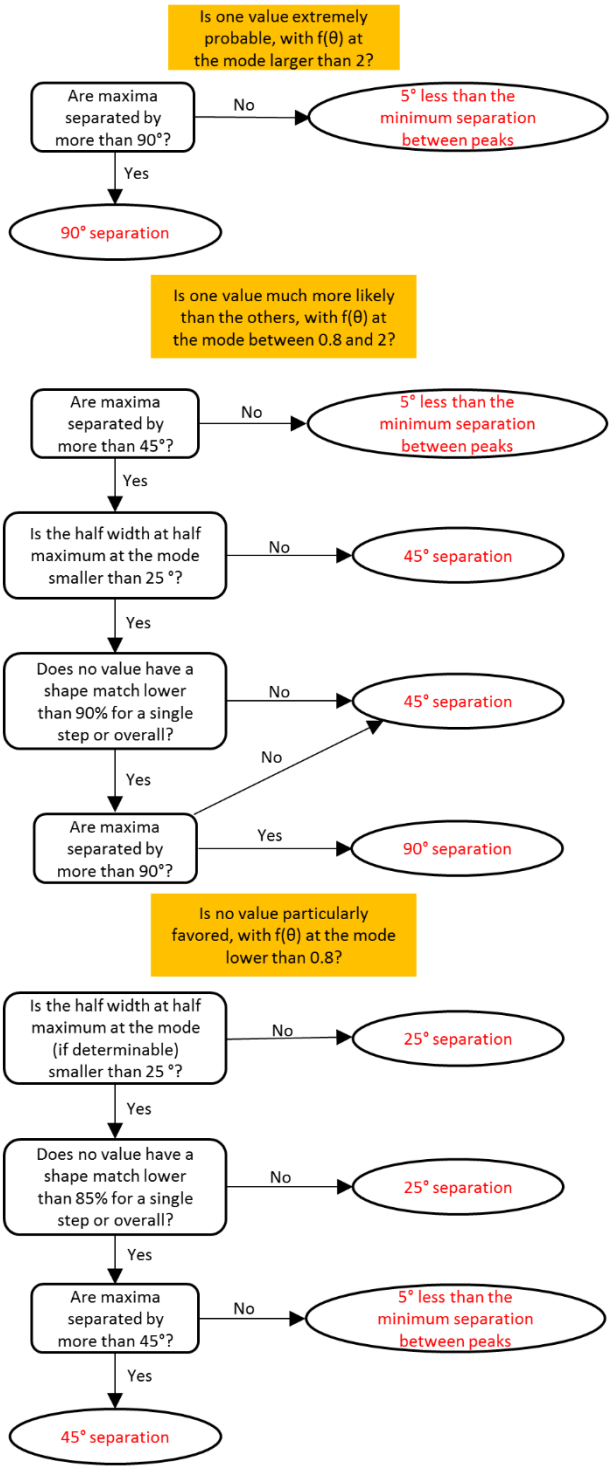


Figure 5.8: Decision tree used to choose the separation threshold of each torsion angle defining a separate CR from the PDF *f(θ)* and shape-matching characteristics.

b) Specification of the ranges for the flexible torsion angles. The range of the torsion angles treated as search variables, and a suitable grid spacing for representing their energy behaviour, were selected using Mogul,[39] as it produces more chemistry-specific distributions than the rotamer libraries.[28] Since the CG removes any gross steric clash that can occur in large molecules,[30] the torsion angle ranges produced by the CG can sometimes further reduce those observed in the Mogul distributions. The Mogul

distributions for the explicitly flexible torsion angles of the five molecules are shown in Appendix Figures 5.8-5.12.

c) Add polar hydrogen torsion angles. The same principles of analysing the PDFs were applied to the *ab initio* torsion energy surfaces. Since the scans showed that torsion angles defining the position of polar hydrogren atoms (*i.e.* Φ_7 in GSK269984B and in molecule XXIII, see Appendix Figures 5.13-5.17) had sharp energy minima, they were constrained to these values, which were added to the selected conformations.

4) Eliminate conformational regions that are energetically implausible. The CG conformations representative of distinct CRs, with all the angles to be fixed during the search constrained at their initial, CG values, were optimised with Gaussian 09 at the PBE0 6-31G(d,p) level of theory. This optimisation of the flexible torsions and all the other bond-lengths, bond-angles and dihedrals not identified as flexible by CG was performed to calculate the energy of the nearest local minimum in conformational energy. Only CRs whose optimised conformational energy penalty (ΔE_{intra}), relative to the most stable fully optimised gas-phase conformer, was plausible for solid-state conformations were kept after this stage. A recently found threshold of $\Delta E_{\mathrm{Intra}}^{\mathrm{CR}}$ ≤ 26 kJ·mol^{-1} was utilised.[40]

In summary, the workflow steps outlined thus far define the conformational space of the molecule as a set of CRs and angles that can adopt a specific range of possible values.

5) Perform CSP searches using the remaining conformational regions. The searches were performed with CrystalPredictor 1.8[19] for each CR. Grids of ΔE_{intra} values were calculated for the ranges and grid spacings determined in step 3b with Gaussian 09 at the PBE0 6-31G(d,p) level of theory. The dimensionality of the grids was reduced by identifying smaller surrogate molecules containing subsets of torsion angles when it was reasonable to assume that their ΔE_{intra} values were not affected by the rest of the molecule.[8, 12, 15] The chosen surrogate molecules are shown in Appendix Figures 5.18-5.22. They had hydrogen atoms or methyl groups added at their edges to avoid the presence of unphysical free bonds, and were large enough to represent the influence of the bonding environment on ΔE_{intra} for the subset of torsion angles.

In the searches the ΔE_{intra} values estimated from the grids were combined with the intermolecular energy (U_{inter}) calculated as a sum of an electrostatic component derived from fixed point charges (fitted from the charge densities of the optimised CRs) and a repulsion-dispersion component derived from the empirically fitted FIT potential.[41] Z'=1 Crystal structures were generated in the 59 most-common space groups in the CSD, which are listed in Appendix Table 5.7. The extent of the search was weighted according to the lowest energy CR. A total of 300,000 structures were generated in CRs with $\Delta E_{\mathrm{Intra}}^{\mathrm{CR}}$ ≤ 4 kJ·mol^{-1}, 150,000 for those with 4 < $\Delta E_{\mathrm{Intra}}^{\mathrm{CR}}$ ≤ 17 kJ·mol^{-1}and 50,000 for those

with 17< ΔE_{Intra}^{CR} ≤ 26 kJ·mol^{-1}. By running the sub-searches in order of increasing ΔE_{Intra}^{CR}, it was possible to track the current global E_{latt} minimum for the overall search. Sub-searches that after generating at least 10,000 crystal structures had not produced any with a lattice energy within 25 kJ·mol^{-1} of the global minimum were terminated to save computational resources, with a high degree of confidence that the U_{inter} was unlikely to compensate for the ΔE_{intra} penalty and generate thermodynamically competitive structures. This contributed to the saving in computational resources.

5.2.2.3 Assessing the adequate coverage of conformational search space

After completing the workflow, its effectiveness was assessed. The workflow could only be regarded as successful if all the "significant" crystal structures that would be analysed in the interpretation of the CSP study had been generated.[11] These were selected as those matching the experimentally observed crystal structure/s of each molecule and the most competitive unobserved putative polymorphs (PPMs) found in the original CSP studies. For some molecules a few higher energy conformationally diverse computer-generated crystal structures that were examined in the original CSP studies to exclude possible types of conformational polymorphism were also considered.

For molecule XXVI, the set of significant structures included all the 35 crystal structures within 10 kJ·mol^{-1} of the global minimum in the first list of predictions submitted by the Price group for the 6th Blind Test, including a match to the experimental structure (CSD refcode XAFQIH, see chapter 3), as well as one extra structure that was made highly competitive by polarisation in the second list.[1] For molecule XXIII, it consisted of the 49 crystal structures within 10 kJ·mol^{-1} of the global minimum in the second list of predictions, including matches to experimental forms B (XAFPAY01) and D (XAFPAY03), three further structures in a different conformational region present in the second list and experimental form A (XAFPAY), which was not present in either submitted list; Z'=2 experimental structures were ignored.[1] For GSK269984B, the 38 crystal structures within 10 kJ·mol^{-1} of the global minimum, including a match to the experimental form (BIFHOP),[20] which are listed in the publication were considered.[20] For molecule XX, the 21 unique crystal structures within 10 kJ·mol^{-1} of the global minimum submitted in the extended list of predictions for the 5th Blind Test, which included a match to the experimentally known crystal structure (OBEQIX), were deemed to be significant, as well as a structure with a *cis* amide, which was submitted as one of the three main predictions.[15, 22] Finally for mebendazole the significant structures were the 31 lowest energy crystal structures of both tautomers, including matches to the solved forms A (TUXPEJ)[42] and C (YULGIW),[43] together with the only competitive crystal structure with a *cis* amide configuration (see Chapter 4 for details).

For each molecule, all the crystal structures generated by the workflow with E_{latt} values (as calculated by CrystalPredictor) within 40 kJ·mol^{-1} of the global minimum were checked to verify whether the significant crystal structures were present among them. This is on the conservative side of the spectra of ranges used to determine which CrystalPredictor-generated structures should be taken to the final refinement stage.[1]

The generated crystal structures were compared with the significant ones using the Crystal Packing Similarity tool available through the CSD Python API. Clusters of 15 molecules were considered, with a 50% distance and a 40° angle tolerances. These values are larger than the default tolerances to account for the differences in the quality of the models used for generating the crystal structures and for the full optimisations in the original CSP studies. A fully optimised significant structure was considered as 'found' when a 15/15 molecule overlay was possible with at least one crystal structure generated with the new methodology. If the $RMSD_{15}$ was smaller than 0.8 Å, which is the criterion used to test the success of the searches for the latest Blind Test,[1] the structure was considered to have been 'certainly found' (*i.e.* a full optimisation was deemed to lead to an almost exact match).[11] If the $RMSD_{15}$ was larger than 0.8 Å, the structure was considered to have been 'probably found': it was deemed highly likely, but not certain, that the generated structure would optimise to the corresponding significant one.[11] Optimisations of these crystal structures are performed in Chapter 7.

5.2.3 Testing the workflow on succinic acid

After the development of the workflow, the Price group was challenged by the Blagden team at the University of Lincoln to predict a newly discovered conformational polymorph (γ form) of succinic acid they had produced during a co-crystallisation experiment aimed at purifying a peptide from a set of impurities.[23] Before that discovery, two single-components forms of succinic acid were known: stable β form (CSD refcode SUCACB03[44] is the best experimental determination), and metastable α form (the highest-quality determination being CSD refcode SUCACB07).[23, 45] The α and β forms both contain a planar conformation, while the new γ conformational polymorph contains a folded molecular geometry.[23] Succinic acid is smaller and less flexible than the molecules in Figure 5.4 from which the workflow was developed. It had already been subject to a partial CSP study, which had however dismissed the possibility of a conformational polymorph existing.[46] Its chemical diagram with the main torsion angles indicated is shown in Figure 5.9.

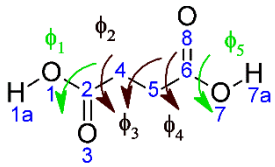


Figure 5.9: Chemical diagram and atomic numbering of succinic acid. The torsion angles identified as flexible by the CG are in black, while those in green define the position of polar hydrogen atoms. Their definition can be found in Appendix Table 5.8.

In order to verify whether the methodology developed in this chapter could generate form γ, the CG was used on succinic acid, generating 46 conformations, and the workflow steps outlined in Chapter 5.2.2.3 were performed. The results of the various steps of the workflow for succinic acid can be found in Appendix Table 5.9 and Appendix Figures 5.23-5.24.

Since succinic acid is not a very flexible molecule, with no torsion angle selected by the workflow for an explicitly flexible treatment, the number of structures generated with CrystalPredictor was halved compared to what had been done for the larger and more flexible molecules in Figure 5.4: for CRs with ΔE_{Intra}^{CR} ≤ 4 kJ·mol^{-1} 150,000 structures were generated; 75,000 for those with 4 < ΔE_{Intra}^{CR} ≤ 17 kJ·mol^{-1}and 25,000 for those with 17 < ΔE_{Intra}^{CR} ≤ 26 kJ·mol^{-1}..

5.3 Results and discussion

5.3.1 Test of the workflow for the five large flexible molecules

A summary of the results of the workflow for the molecules in Figure 5.4 is shown in Figure 5.10 and Table 5.3.

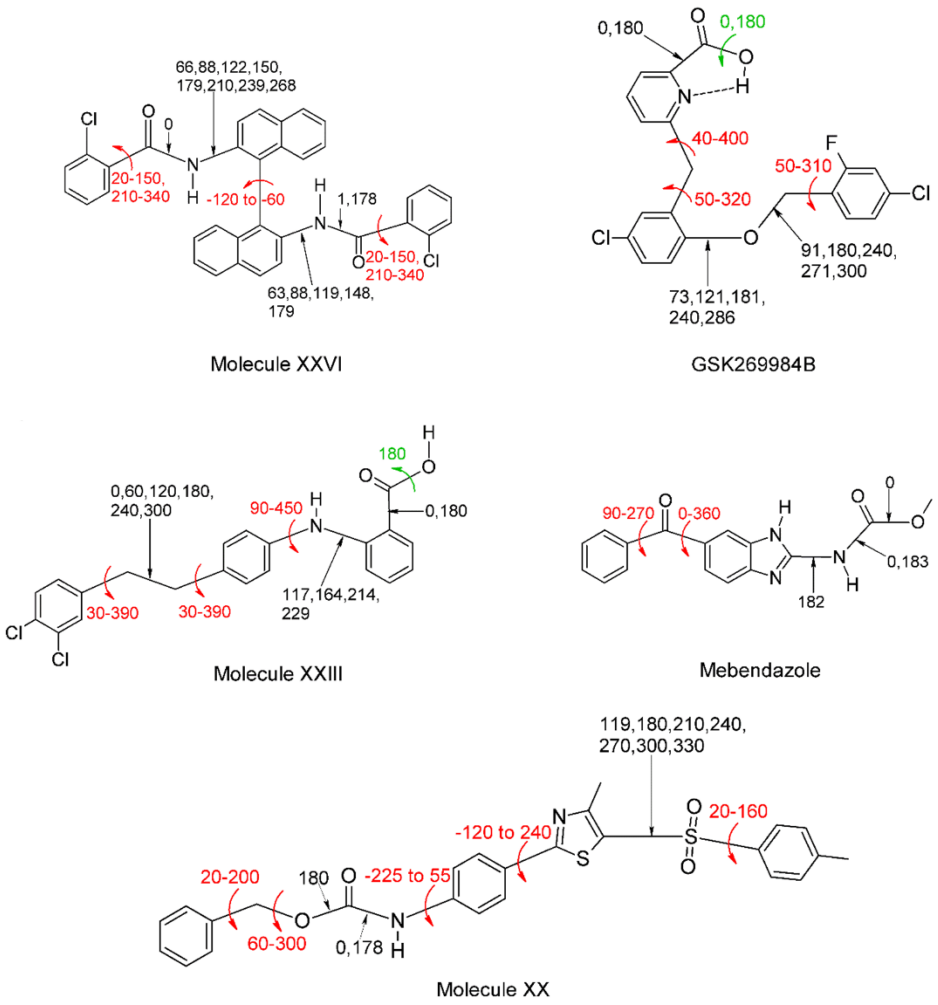


Figure 5.10: Summary of the application of the workflow on the molecules in Figure 5.4. The torsion angles in red were treated as flexible in the searches, covering the ranges given in degrees. Torsion angles in black were constrained to a set of CG values, with the values used in at least one search indicated; note that many combinations of these values were eliminated as energetically unfeasible (see Appendix Tables 5.10-5.15). 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 (the motivation for this assumption is illustrated in Chapter 4).

Table 5.3: Summary of the CSP studies and their results.

Molecule	# Selected CRs	# ΔE_{intra} (CR)			# ~ crystal structures / $\cdot 10^6$	Found?			Saving in CPU hours, relative to original CSP studies
		0-4	4-17	17-26		Yes	Probably	No	
XXVI[1]	138	1	8	14	2.2	31	4	1	~14,000 (-50%)
GSK269984B[20]	51	1	7	10	1.9	28	6	4	not recorded
XX[15, 22]	21	2	8	3	2	20	2	0	~6,000 (-30%)
XXIII[1]	127	3	6	7	2.1	37	12	4	~14,000 (-70%)
Mebendazole A	4	1	1	0	0.45	28	4	0	~5,000 (-70%)
Mebendazole C	4	1	0	1	0.35				

The workflow defined a number of CRs and explicitly flexible torsion angles. This breakdown is very dependent on the specific molecule, reflecting the level of flexibility and the shape effect of different torsion angles. For example, there are eight possible angles between the aromatic ring and amide groups in XXVI (*i.e.* angle Φ_{3a}), but only one for mebendazole (*i.e.* Φ_3), because of an internal hydrogen bond.

The reproduction of the significant crystal structures listed in Chapter 5.2.2.3 is summarised in Figure 5.11. More details can also be found in Appendix Tables 5.16-5.20.

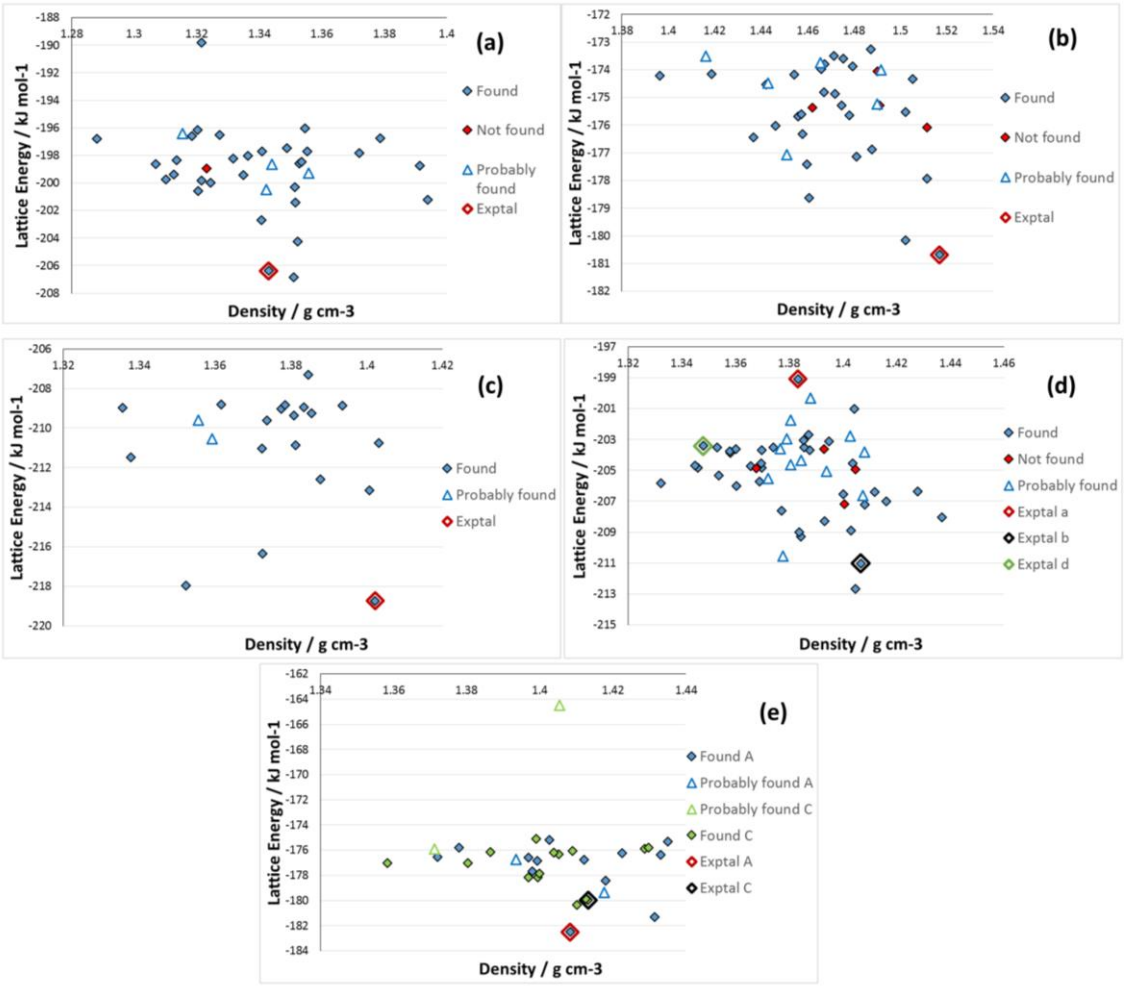


Figure 5.11. Plots of the significant CSP-generated crystal structures found in previous studies, classified as to whether they were found by the search workflow, of (a) molecule XXVI, (b) GSK269984B, (c) molecule XX, (d) molecule XXIII and (e) mebendazole. Higher energy structures were included for molecule XXVI (a structure whose stability was very dependent on the energy model), molecule XX and mebendazole (a competitive crystal structure with a *cis* amide for both molecules). Structures matching the experimentally known forms are indicated by open diamonds.

The workflow generated millions of crystal structures at a considerably reduced computational cost, as shown in Table 5.3.

Figure 5.11 evaluates whether the searches managed to reproduce the significant crystal structures, classifying the output of the original CSP studies as to whether the

workflow generated similar enough matches. The vast majority of the significant crystal structures, including those matching all the experimentally-characterised forms, were successfully generated. For molecule XX and mebendazole no significant crystal structure was missed, and only a few higher-energy PPMs were not generated for molecule XXVI, GSK269984B and molecule XXIII. This was not due to an inadequacy in the workflow, since the conformations contained in the missed structures were found in other structures that were generated in the searches. These few structures were probably missed because they are located in particularly narrow E_{latt} minima, which are rarely found[47] by the pseudo-random element of the CrystalPredictor searches. The accuracy of the conformational model was also evidenced by most of the generated structures being good matches with the corresponding fully optimised significant structures (see Appendix Tables 5.16-5.20 for further details).

The use of the workflow decreased the computational cost by 30-70%, as shown in Table 5.3, although some of these savings would be due to the use of different computer clusters. This reduction came from replacing the initial *ab initio* conformational analyses on individual torsion angles with a fast assessment based on CSD informatics, from requiring fewer calculations to produce the ΔE_{intra} grids, and from terminating some Crystal Predictor runs early in a systematic manner. Furthermore, this workflow takes less than an hour to set up, which is a considerable saving in terms of human time compared to other similar methods.

5.3.2. Testing the workflow on succinic acid

The results of the application of the workflow to succinic acid are summarised in Figure 5.12 and Table 5.4.

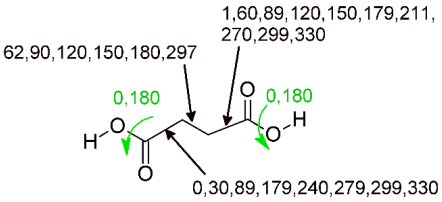


Figure 5.12: Summary of the application of the workflow to succinic acid. Torsion angles in black were constrained to a set of CG values, with the values used in at least one search indicated; many combinations of these values were eliminated as energetically unfeasible (see Appendix Table 5.21). Torsion angles in green were constrained to the indicated values having been determined from an *ab initio* conformational-energy scan. Since the two halves of the molecule are symmetric, only one combination 180°-0° values for the green torsion angles Φ_1 and Φ_5 was considered, as they would represent identical molecules upon switching.

Table 5.4: Summary of the succinic acid searches.

# CG conformations	# Selected CRs	# CRs ΔE_{intra} (CR) ≤ 26 kJ·mol^{-1}			# ~ generated structures/ ·10^6	CPU cost/hours
		0-4	4-17	17- 26		
46	72	1	10	15	0.42	~600

A match to new conformational polymorph (γ form) was successfully found by the workflow. This is significant because the γ form contains a folded conformation, with a Φ_3 value of 75.43° that is rare in the CSD (see Appendix Figure 5.23). A CSD survey of all the salts, co-crystals and solvates that contain succinic acid showed the folded conformations are indeed unusual, as they are found in only ~11% of the retrieved entries.[23] It is promising that the workflow generated a match to a new polymorph that contains such an uncommon molecular geometry. The most stable β form, which contains the more common planar conformation (*i.e.* with Φ_3=180.00°), was also successfully identified, while the α form was missed. However, the α form has a Z' of two halves molecules,[23, 46] and so it was outside the scope of this search. The results are summarised in Table 5.5.

Table 5.5: Comparison of the crystal structures of succinic acid generated in the workflow and the two Z'=1 experimentally known polymorphs.

Experimental form	Found?	Conformation number	Ranking after search	$RMSD_{15}$
β form	YES	1	1	0.272
γ form	YES	2	834	0.456

Although a parallel CSP study performed with a traditional CSP workflow was published in the paper describing the crystallisation of the γ form,[23] this search showed that the workflow developed and tested in Chapter 5.2.2.2 can be effective even for a molecule that was not involved in its development. Succinic acid is a relatively small molecule, and it is important that a search workflow developed heuristically from larger targets was successful. Nonetheless, more testing on larger molecules that differ more from those that were considered in this work is needed.

5.3.3 Discussion

The results illustrated in this chapter show that the conformational space that needs to be covered is very molecule-dependent. The interplay between the different torsion angles within a large and flexible molecule is complex: some combinations of values are restricted by steric clashes, or there can be specific correlations that allow large variations with little overall effect on the molecular shape.

The CSD Conformer Generator is effective at describing the likely conformational space of pharmaceutical-like molecules, but the vast number of conformations it generates makes a detailed analysis of both the individual torsion angle distributions and their effect on the overall molecular shape necessary. These analyses allow a setup of

a CSP study as a set of searches with some fixed torsion angles and/or some that are allowed to vary over defined ranges, vastly reducing the number of costly *ab initio* energy calculations required. This approach is effective: it generated most of the significant crystal structures of the five large and flexible molecules in Figure 5.4, including matches to all Z'=1 experimentally-characterised forms, with a large reduction in computational cost and minimal human effort. It also reproduced both Z'=1 polymorphs of succinic acid, including the newly discovered γ form with an unusual conformation.[23] Very few significant structures were missed for the five test molecules (see Table 5.3 and Appendix Tables 5.16-5.20), and this was not due to an incomplete coverage of the conformational space. However the significant crystal structures were often poorly ranked among a plethora of alternatives because of the approximations in the CrystalPredictor E_{latt} model, and this poses important challenges for the final refinement stage of CSP.

Although the results presented in this chapter are promising, the workflow was only tested for its ability to reproduce the results of CSP studies based on the assumption that a conformation can plausibly occur in a stable crystal structure only if it has a low ΔE_{intra} value as determined by *ab initio* calculations (see Chapter 2.4.1.1). However in larger molecules intermolecular dispersion effects, which favour planar conformations,[40] and differences in electrostatic contributions, such as switching between inter- and intramolecular hydrogen bonds,[6] can stabilise molecular geometries with very high ΔE_{intra} values. Hence as molecules get larger and more flexible the conformational energy cut-off values, such as the 26 $kJ \cdot mol^{-1}$ used here to determine whether a CR is likely to be observed in a crystal structure, might require adaption. It may also be that CSD information, capturing solid-state rather than isolated-molecule conformational preferences, reflects more than the relative thermodynamic stability of gas-phase conformations, *e.g.* which molecular geometries optimise intermolecular interactions or crystallisation kinetics.[48, 49]

Some further issues related to a large-scale application of this specific workflow can be anticipated. First of all, since the workflow is based on conformational information derived from molecular fragments, it may miss some unusual but potentially important conformations involving torsion angles that have few CSD entries. Moreover, the workflow could only be developed and validated from a limited number of CSP studies, as relatively few have been performed on molecules of the size of small pharmaceuticals. Thus it is doubtful whether this methodology could be extended to molecules like ritonavir: the current version of the CG on standard settings does not produce the molecular conformation of the most stable observed polymorph (form II) because of the unusual *cis* conformation of the carbamate group in combination with the other 20 flexible torsion angles.[50] However, this does not necessarily mean that the workflow would lead

to an incomplete search, as the combination of constrained and explicitly flexible torsion angles may still include form II within its boundaries. Finally, all the assumptions made in this chapter to limit the use of human and computer resources (such as assuming Z'=1) risk missing some thermodynamically stable crystal structures.[5, 51]

Nonetheless, the results show that using CSD-derived information on geometrical preferences to define the conformational search space can be an efficient route to performing CSP searches on molecules of pharmaceutical interest. The use of shape matching criteria to determine which angles have a large effect on molecular 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 CRs, are likely to be applicable to other CSP workflows and algorithms. Finally, the poor ranking of the significant crystal structures among many others that are of little practical interest shows how cost effective methods to optimise and re-rank the generated crystal structures are needed. This topic will be discussed in Chapter 7.

5.4 Conclusion

A workflow for using CSD conformational information to set up CSP searches was developed, which uses 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 also which torsion angles should be treated as search variables within specified ranges. Using this workflow in conjunction with CrystalPredictor successfully generated most of the significant crystal structures of five large flexible molecules of pharmaceutical interest, including all the experimentally- characterised Z'=1 forms, at a 30-70% lower computational cost. It must be noted that some of the calculations were not performed on the same machines. A newly discovered conformational polymorph of succinic acid was also found, showing that the workflow can be successful for molecules not considered in its development.

The results illustrated in this chapter are very important in the context of this thesis. First of all, they show that the use of CSD conformational information is a viable way to limit the computational cost of setting up and performing CSP searches on large flexible molecules of pharmaceutical interest. This approach does not seem to limit the coverage of the conformational search space, as crystal structures with unusual molecular geometries were successfully generated, *e.g.* the γ form of succinic acid. Another important point is that CSD information cannot be used on its own in CSP for molecules of pharmaceutical interest, since the informatics tools generate an unmanageable number of conformational possibilities, and other factors, such as *ab initio* energies and shape effects, must be accounted for. Although the workflow is tailored to

CrystalPredictor, the ideas behind it are applicable to other CSP methodologies, as long as the necessary modifications to the parameters are brought. Finally, this chapter shows that speeding up searches is by no means sufficient to perform an effective CSP study: although the most important crystal structures were generated, they were often poorly ranked among a plethora of unimportant alternatives. This is due to the approximate E_{latt} model used to perform the CSP searches. Hence, an effective CSP methodology requires cost effective yet accurate methods to optimise and re-rank search-generated structures. The development of a methodology to achieve a fast but high-quality refinement of the thousands of crystal structures generated by the workflow for the molecules in Figure 5.4 is the main goal of Chapter 7.

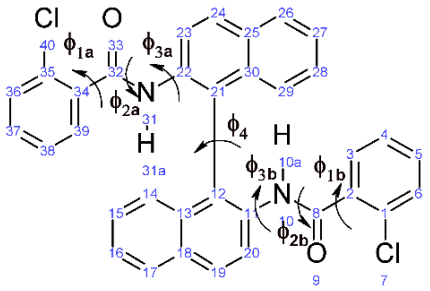
5.5 References

1. 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.; Vacarro, F.; Vasileiadis, M.; Vazquez-Mayagoitia, A.; Vogt, L.; Wang, Y.; Watson, R. E.; de Wijs, G. A.; Yang, J. Z.; Zhu, Q.; Groom, C. R., Report on the sixth blind test of organic crystal-structure prediction methods. *Acta Crystallographica Section B - Structural Science* **2016**.
2. Price, S. L.; Braun, D. E.; Reutzel-Edens, S. M., Can computed crystal energy landscapes help understand pharmaceutical solids? *Chemical Communications* **2016,** *52*, 7065-7077.
3. Price, S. L., Is zeroth order crystal structure prediction (CSP_0) coming to maturity? What should we aim for in an ideal crystal structure prediction code? *Faraday Discussions* **2018,** *in press*.
4. Neumann, M.; van de Streek, J., How many Ritonavir cases are there still out there? *Faraday Discussions* **2018,** *Advance article*.
5. Price, S. L.; Reutzel-Edens, S. M., The potential of computed crystal energy landscapes to aid solid form development. *Drug Discovery Today* **2016,** *21* (6), 912-923.
6. Cruz-Cabeza, A. J.; Bernstein, J., Conformational Polymorphism. *Chemical Reviews* **2014,** *114* (4), 2170-2191.
7. Thompson, H.; Day, G., Which conformations make stable crystal structures? Mapping crystalline molecular geometries to the conformational energy landscape. *Chemical Science* **2014,** *5* (8), 3173-3182.
8. Pantelides, C. C.; Adjiman, C. S.; Kazantsev, A. V., General Computational Algorithms for Ab Initio Crystal Structure Prediction for Organic Molecules. *Topics in Current Chemistry* **2014,** *345*, 25-58.
9. Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C.; Price, S. L.; Galek, P. T.; Day, G. M.; Cruz-Cabeza, A. J., Successful prediction of a model pharmaceutical in the fifth blind test of crystal structure prediction. *International Journal of Pharmaceutics* **2011,** *418* (2), 168-178.
10. Abramov, Y. A., Current Computational Approaches to Support Pharmaceutical Solid Form Selection. *Organic Process Research & Development* **2012,** *17* (3), 472-485.
11. Iuzzolino, L.; Reilly, A. M.; McCabe, P.; Price, S. L., Use of Crystal Structure Informatics for Defining the Conformational Space Needed for Predicting Crystal Structures of

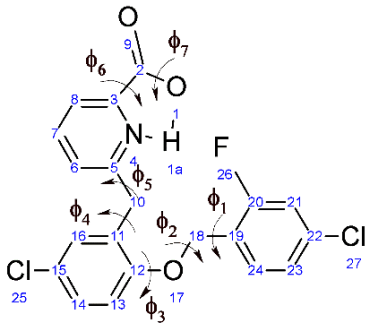
Pharmaceutical Molecules. *Journal of Chemical Theory and Computation* **2017,** *13* (10), 5163-5171.
12. Vasileiadis, M.; Pantelides, C. C.; Adjiman, C. S., Prediction of the crystal structures of axitinib, a polymorphic pharmaceutical molecule. *Chemical Engineering Science* **2015,** *121*, 60-76.
13. Price, S. L., Predicting crystal structures of organic compounds. *Chemical Society Reviews* **2014,** *43* (7), 2098-2111.
14. Nguyen, K.; Iskandar, M.; Rabenstein, D. L., Kinetics and Equilibria of Cis/Trans Isomerization of Secondary Amide Peptide Bonds in Linear and Cyclic Peptides. *The Journal of Physical Chemistry B* **2010,** *114* (9), 3387-3392.
15. 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. *International Journal of Pharmaceutics* **2011,** *418* (2), 168-178.
16. Day, G. M.; Motherwell, W. D. S.; Jones, W., A strategy for predicting the crystal structures of flexible molecules: the polymorphism of phenobarbital. *Physical Chemistry Chemical Physics* **2007,** *9* (14), 1693-1704.
17. Payne, R. S.; Rowe, R. C.; Roberts, R. J.; Charlton, M. H.; Docherty, R., Potential polymorphs of aspirin. *Journal of Computational Chemistry* **1999,** *20* (2), 262-273.
18. Neumann, M. A. *GRACE (the Generation, Ranking and Characterisation Engine)*, 1.0; Avant-garde Materials Simulation Deutschland GmbH: 2007.
19. Karamertzanis, P. G.; Pantelides, C. C., Ab initio crystal structure prediction. II. Flexible molecules. *Molecular Physics* **2007,** *105* (2-3), 273-291.
20. 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. *Crystal Growth & Design* **2013,** *13* (6), 2396-2406.
21. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C., The Cambridge Structural Database. *Acta Crystallographica Section B-Structural Science Crystal Engineering and Materials* **2016,** *72*, 171-179.
22. 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 Crystallographica Section B-Structural Science* **2011,** *67*, 535-551.
23. Lucaioli, P.; Nauha, E.; Gimondi, I.; Price, L. S.; Guo, R.; Iuzzolino, L.; Singh, I.; Salvalaglio, M.; Price, S. L.; Blagden, N., Serendipitous isolation of a disappearing conformational polymorph of succinic acid challenges computational polymorph prediction. *CrystEngComm* **2018,** *20* (28), 3971-3977.
24. Habgood, M.; Price, S. L.; Portalone, G.; Irrera, S., Testing a Variety of Electronic-Structure-Based Methods for the Relative Energies of 5-Formyluracil Crystals. *Journal of Chemical Theory and Computation* **2011,** *7* (9), 2685-2688.
25. Bhardwaj, R. M.; Price, L. S.; Price, S. L.; Reutzel-Edens, S. M.; Miller, G. J.; Oswald, I. D. H.; Johnston, B.; Florence, A. J., Exploring the Experimental and Computed Crystal Energy Landscape of Olanzapine. *Crystal Growth & Design* **2013,** *13* (4), 1602-1617.
26. Price, L. S.; McMahon, J. A.; Lingireddy, S. R.; Lau, S. F.; Diseroad, B. A.; Price, S. L.; Reutzel-Edens, S. M., A molecular picture of the problems in ensuring structural purity of tazofelone. *Journal of Molecular Structure* **2014,** *1078*, 26-42.
27. Uzoh, O. G.; Cruz-Cabeza, A. J.; Price, S. L., Is the Fenamate Group a Polymorphophore? Contrasting the Crystal Energy Landscapes of Fenamic and Tolfenamic Acids. *Crystal Growth & Design* **2012,** *12* (8), 4230-4239.
28. Taylor, R.; Cole, J.; Korb, O.; McCabe, P., Knowledge-Based Libraries for Predicting the Geometric Preferences of Druglike Molecules. *Journal of Chemical Information and Modeling* **2014,** *54* (9), 2500-2514.
29. McCabe, P.; Korb, O.; Cole, J., Kernel Density Estimation Applied to Bond Length, Bond-angle, and Torsion Angle Distributions. *Journal of Chemical Information and Modeling* **2014,** *54* (5), 1284-1288.
30. Cole, J. C.; Korb, O.; McCabe, P.; Read, M. G.; Taylor, R., Knowledge-Based Conformer Generation Using the Cambridge Structural Database. *Journal of Chemical Information and Modeling* **2018,** *58* (3), 615-629.

31. Chisholm, J. A.; Motherwell, S., COMPACK: a program for identifying crystal structure similarity using distances. *Journal of Applied Crystallography* **2005,** *38*, 228-231.
32. Hunter, J. D., Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **2007,** *9* (3), 90-95.
33. van Rossum, G. *Python: a computer language*, 1.5.1; 1998.
34. Ballester, P. J.; Richards, W. G., Ultrafast shape recognition to search compound databases for similar molecular shapes. *J Comput Chem* **2007,** *28*.
35. RDKit: Cheminformatics and Machine Learning Software. http://www.rdkit.org.
36. Schreyer, A. M.; Blundell, T., USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. *Journal of Cheminformatics* **2012,** *4*, 12.
37. Nyman, J.; Day, G. M., Static and lattice vibrational energy differences between polymorphs. *Crystengcomm* **2015,** *17* (28), 5154-5165.
38. 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. 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.
39. 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. *Journal of Chemical Information and Computer Sciences* **2004,** *44* (6), 2133-2144.
40. Thompson, H. P. G.; Day, G. M., Which conformations make stable crystal structures? Mapping crystalline molecular geometries to the conformational energy landscape. *Chemical Science* **2014,** *5* (8), 3173-3182.
41. Coombes, D. S., Deriving intermolecular potentials for predicting the crystal structures of polar molecules. *Philosophical Magazine B-Physics of Condensed Matter Statistical Mechanics Electronic Optical and Magnetic Properties* **1996,** *73* (1), 117-125.
42. Ferreira, F. F.; Antoni, S. G.; Rosa, P. C. P.; Paiva-Santos, C. D., Crystal Structure Determination of Mebendazole Form A Using High-Resolution Synchrotron X-Ray Powder Diffraction Data. *Journal of Pharmaceutical Sciences* **2010,** *99* (4), 1734-1744.
43. Martins, F. T.; Neves, P. P.; Ellena, J.; Cami, G. E.; Brusau, E. V.; Narda, G. E., Intermolecular Contacts Influencing the Conformational and Geometric Features of the Pharmaceutically Preferred Mebendazole Polymorph C. *Journal of Pharmaceutical Sciences* **2009,** *98* (7), 2336-2344.
44. Leviel, J. L.; Auvert, G.; Savariault, J. M., Hydrogen bond studies. A neutron diffraction study of the structures of succinic acid at 300 and 77 K. *Acta Crystallographica Section B - Structural Crystallography and Crystal Chemistry* **1981,** *37* (12), 2185-2189.
45. Dodd, I. M.; Maginn, S. J.; Harding, M. M.; Davey, R. J., 1998.
46. Issa, N.; Barnett, S. A.; Mohamed, S.; Braun, D. E.; Copley, R. C. B.; Tocher, D. A.; Price, S. L., Screening for cocrystals of succinic acid and 4-aminobenzoic acid. *CrystEngComm* **2012,** *14* (7), 2454-2464.
47. Case, D. H.; Campbell, J. E.; Bygrave, P. J.; Day, G. M., Convergence Properties of Crystal Structure Prediction by Quasi-Random Sampling. *Journal of Chemical Theory and Computation* **2016,** *12* (2), 910-924.
48. 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. *Journal of the American Chemical Society* **2005,** *127* (30), 10545-10559.
49. 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? *Journal of the American Chemical Society* **2015**.
50. Bauer, J.; Spanton, S.; Henry, R.; Quick, J.; Dziki, W.; Porter, W.; Morris, J., Ritonavir: An Extraordinary Example of Conformational Polymorphism. *Pharmaceutical Research* **2001,** *18* (6), 859-866.
51. Nyman, J.; Reutzel-Edens, S. M., Crystal structure prediction is changing from basic science to applied technology. *Faraday Discussions* **2018,** *Advance article*.

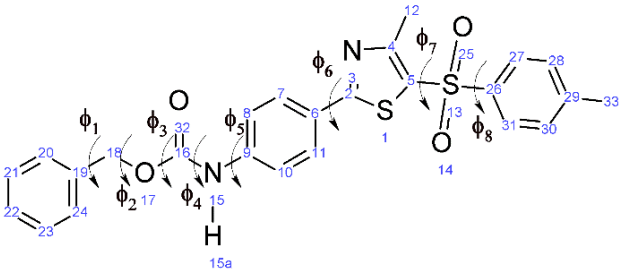
5.6 Appendix



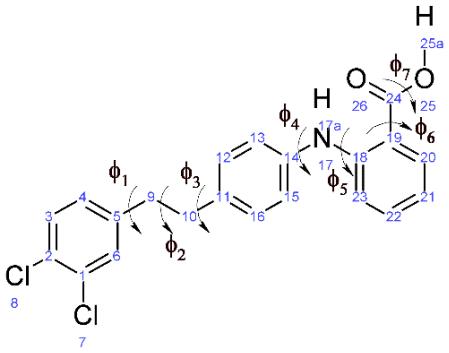
Molecule XXVI



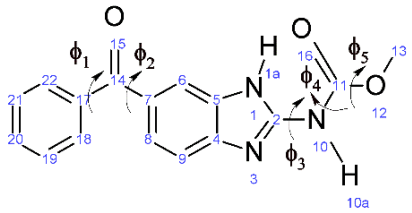
GSK269984B



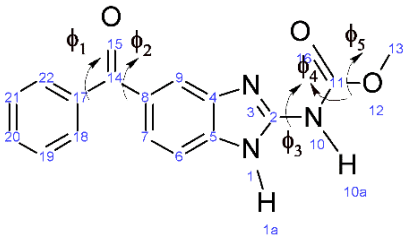
Molecule XX



Molecule XXIII



A-Mebendazole

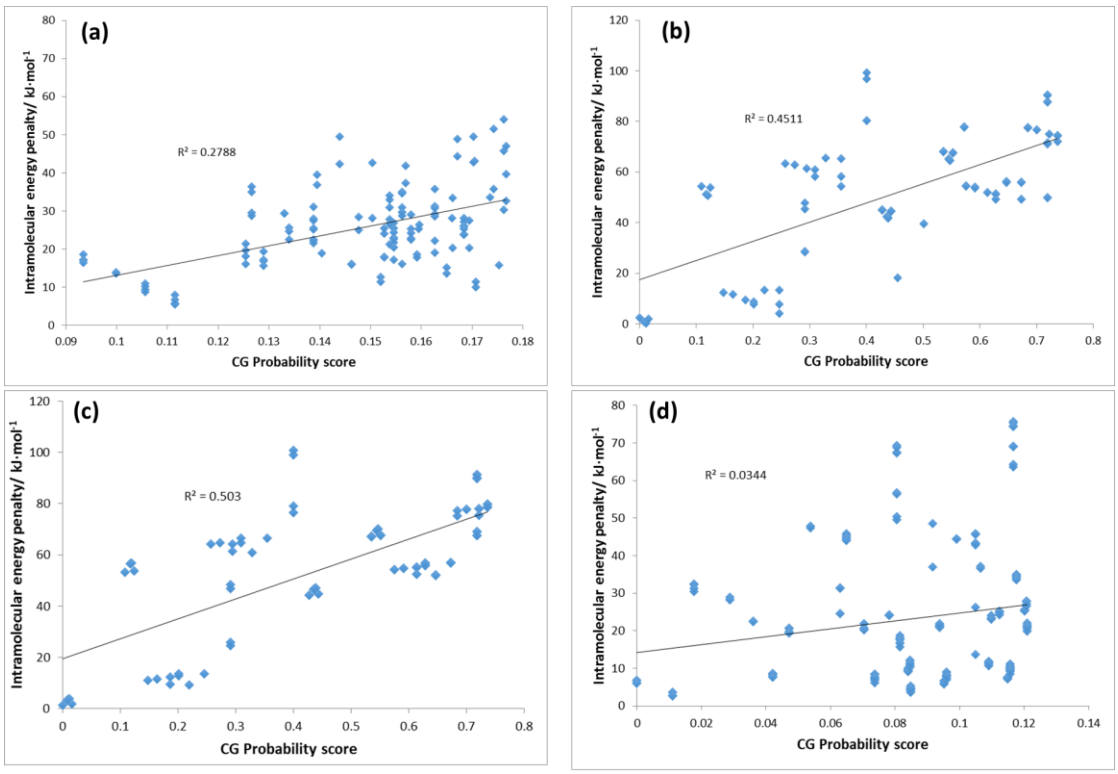


C-Mebendazole

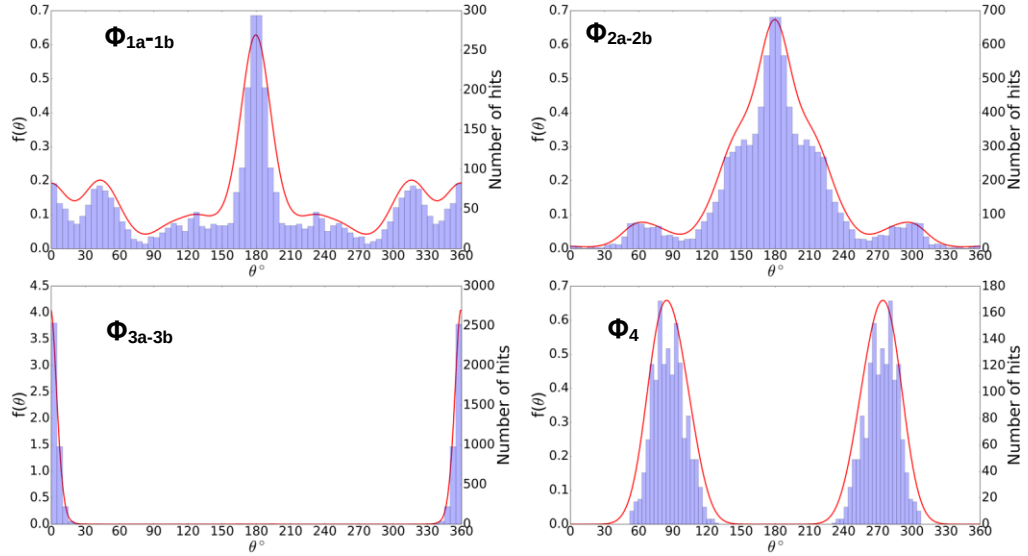
Appendix Figure 5.1: Chemical diagrams of the molecules in Figure 5.4, with the atomic numbering that precisely identifies the rotatable torsion angles in Appendix Table 5.1.

Appendix Table 5.1: Atomic numbering definition of the key torsion angles shown in Appendix Figure 5.1.

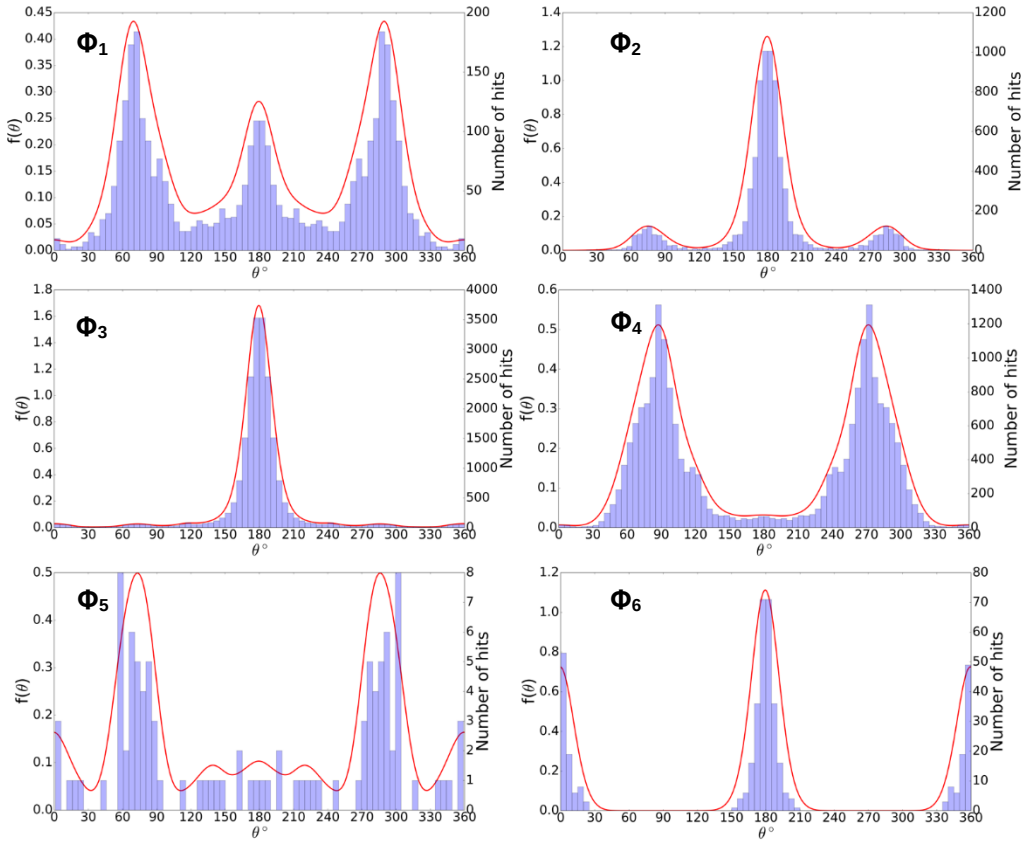
Label for XXVI	Torsion Angle Definition
Φ_{1a}	35-34-32-33
Φ_{2a}	33-32-31-22
Φ_{3a}	32-21-22-21
Φ_{1b}	1-2-8-9
Φ_{2b}	9-8-10-11
Φ_{3b}	8-10-11-12
Φ_{4}	30-21-12-13
Label for GSK269984B	**Torsion Angle Definition**
Φ_{1}	17-18-19-20
Φ_{2}	12 -17-18-19
Φ_{3}	11-12-17-18
Φ_{4}	12-11-10-5
Φ_{5}	11-10-5-4
Φ_{6}	4-3-2-9
Φ_{7}	3-2-1-1a
Label for XX	**Torsion Angle Definition**
Φ_{1}	20-19-18-17
Φ_{2}	19-18-17-16
Φ_{3}	18-17-16-16
Φ_{4}	32-16-15-9
Φ_{5}	16-15-9-8
Φ_{6}	7-6-2-3
Φ_{7}	4-5-13-26
Φ_{8}	5-13-26-31
Label for XXIII	**Torsion Angle Definition**
Φ_{1}	6-5-9-10
Φ_{2}	5-9-10-11
Φ_{3}	9-10-11-12
Φ_{4}	13-14-17-18
Φ_{5}	14-17-18-19
Φ_{6}	18-19-26-26
Φ_{7}	19-24-25-25a
Label for Mebendazole A, C	**Torsion Angle Definition**
Φ_{1}	22-17-14-7 (A), 22-17-14-8 (C)
Φ_{2}	17-14-7-8 (A), 17-14-8-9 (C)
Φ_{3}	3-2-10-11
Φ_{4}	2-10-11-16
Φ_{5}	16-11-12-13



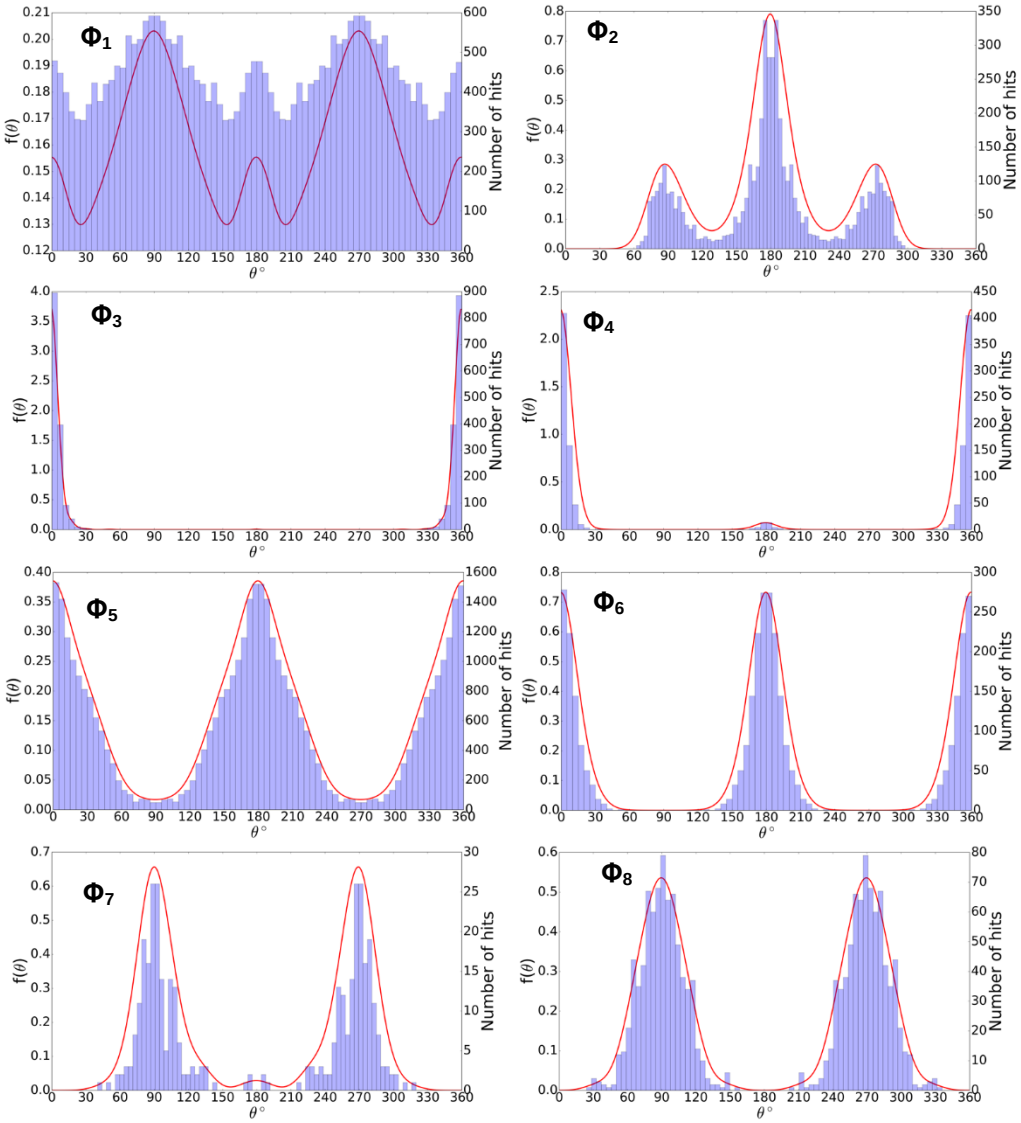
Appendix Figure 5.2: Plots of ΔE_{intra} versus CG probability scores (explained in Chapter 2.6.4) for the 200 most probably conformations of (a) Molecule XXVI (b) the A-tautomer of Mebendazole (c) the C-tautomer of Mebendazole (d) Molecule XXIII. This test shows that the CG probability score does not correlate well with *ab initio* ΔE_{intra} values, and that some high-probability conformations have unfeasible conformational energies for solid-state candidates. This is probably because the CG does not capture the interactions between different fragments.[30]



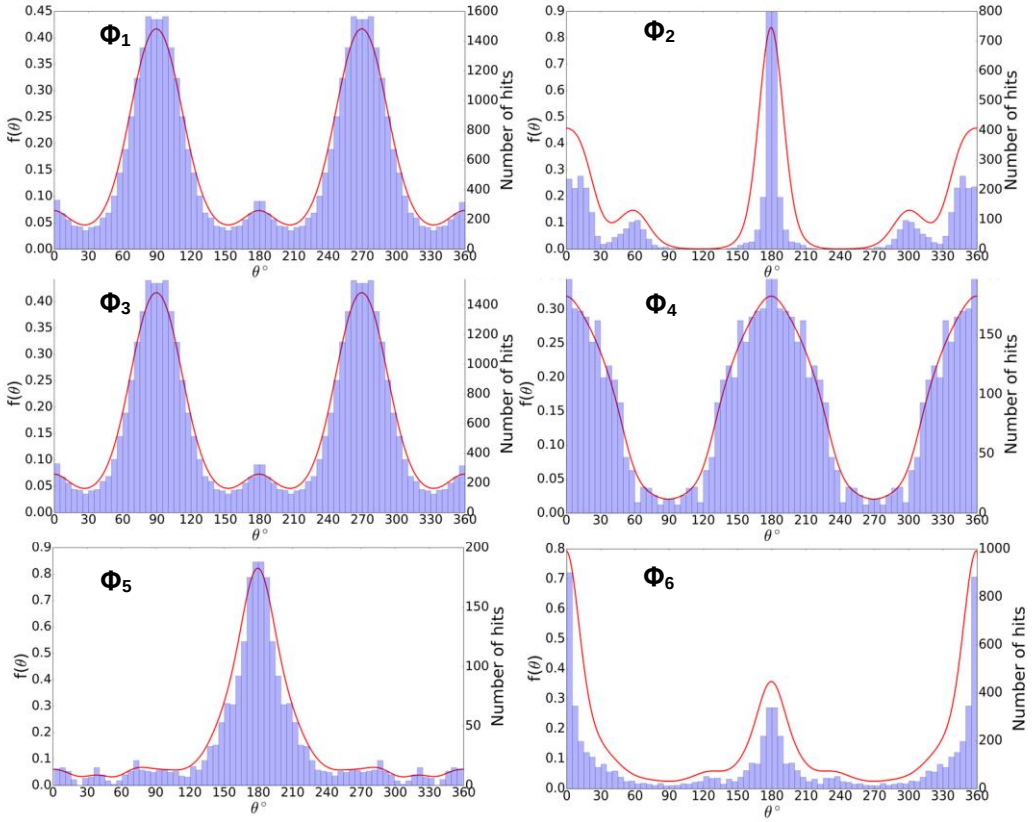
Appendix Figure 5.3: **Histograms (light purple bars) and PDFs (red lines) of the torsion angle distributions of molecule XXVI.**



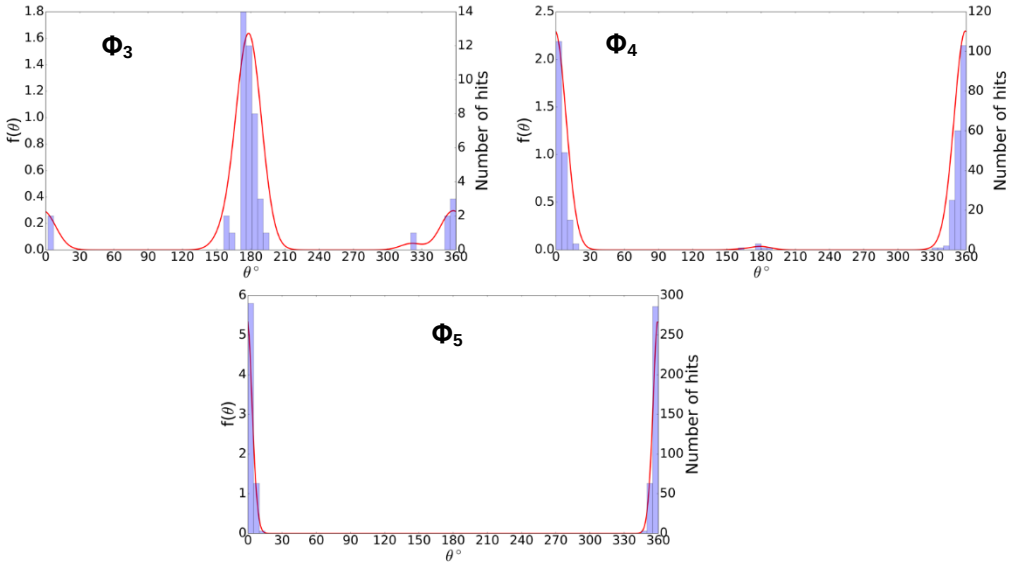
Appendix Figure 5.4: **Histograms (light purple bars) and PDFs (red lines) of the torsion angle distributions of GSK269984B.**



Appendix Figure 5.5: Histograms (light purple bars) and PDFs (red lines) of the torsion angle distributions of molecule XX.



Appendix Figure 5.6: Histograms (light purple bars) and PDFs (red lines) of the torsion angle distributions of molecule XXIII.



Appendix Figure 5.7: Histograms (light purple bars) and PDFs (red lines) of the torsion angle distributions of both tautomers of mebendazole. For Φ_1 and Φ_2, the rotamer libraries failed to generate individual distributions since they are coupled around C14 (see Appendix Figure 5.1). Hence the Mogul distributions in Appendix Figure 12 were used as a replacement, and they suggested that both Φ_1 and Φ_2 should be treated as explicitly flexible in the searches.

Appendix Table 5.2: Shape matches for varying each torsion angle of molecule XXVI with 30° steps with both the previous step and the starting conformation. The angles are normalised to a 0-360° range. The first value for each torsion angle corresponds to the initial value.

$\Phi_{1a,2a}$/°	% shape match with previous step	% shape match with original conformation	$\Phi_{2a,2b}$/°	% shape match with previous step	% shape match with original conformation
214.37	/	/	**4.14**	/	/
244.37	99.39	99.39	**34.14**	95.47	95.47
274.37	99.60	99.04	**64.14**	96.53	92.30
304.37	99.51	98.90	**94.14**	98.30	91.09
334.37	96.26	96.39	**124.14**	98.75	92.06
4.37	97.31	97.77	**154.14**	96.72	94.06
34.37	97.48	98.55	**184.14**	95.70	96.18
64.37	99.46	98.37	**214.14**	95.61	96.16
94.37	99.31	98.23	**244.14**	96.42	93.00
124.37	99.28	98.28	**274.14**	98.07	91.53
154.37	99.50	98.62	**304.14**	98.86	92.41
184.37	99.32	99.24	**334.14**	96.63	95.47
$\Phi_{3a,3b}$/°	% shape match with previous step	% shape match with original conformation	Φ_4/°	% shape match with previous step	% shape match with original conformation
222.46	/	/	**258**	/	/
252.46	98.71	98.71	**287.97**	91.41	91.41
282.46	98.41	97.31	**317.97**	93.60	86.04
312.46	98.14	95.83	**347.97**	93.32	82.99
342.46	98.19	94.40	**17.97**	96.53	80.93
12.46	98.65	93.58	**47.97**	88.74	81.81
42.46	98.90	93.54	**77.97**	88.27	82.76
72.46	96.24	95.21	**107.97**	92.50	85.50
102.46	98.39	94.86	**137.97**	76.01	77.20
132.46	98.70	95.15	**167.97**	77.46	71.13
162.46	96.79	97.82	**197.97**	93.17	73.82
192.46	98.75	98.98	**227.97**	83.49	85.93

Appendix Table 5.3: Shape matches for varying each torsion angle of GSK269984B with 30° steps with both the previous step and the starting conformation. The angles are normalised to a 0-360° range. The first value for each torsion angle corresponds to the initial value.

Φ_1/°	% shape match with previous step	% shape match with original conformation	Φ_2/°	% shape match with previous step	% shape match with original conformation
181	/	/	**178.55**	/	/
211	99.13	99.13	**208.55**	96.87	96.87
241	98.94	98.31	**238.55**	92.07	92.06
271	98.47	97.11	**268.55**	92.11	85.33
301	98.09	95.98	**298.55**	95.07	82.12
331	99.00	95.29	**328.55**	95.24	81.17
1	99.01	95.88	**358.55**	97.28	80.20
31	98.16	96.86	**28.55**	96.15	79.86
61	98.51	97.67	**58.55**	94.11	79.63
91	98.87	98.76	**88.55**	92.84	80.79
121	98.90	99.53	**118.55**	93.03	83.86
151	99.49	99.48	**148.55**	88.87	91.13
Φ_3/°	% shape match with previous step	% shape match with original conformation	Φ_4/°	% shape match with previous step	% shape match with original conformation
182.96	/	/	**64.96**	/	/
212.96	98.14	98.14	**94.96**	89.55	89.55
242.96	97.44	95.70	**124.96**	90.89	82.19
272.96	96.38	92.87	**154.96**	91.70	77.69
302.96	96.07	89.60	**184.96**	92.79	75.11
332.96	98.24	88.20	**214.96**	89.34	72.58
2.96	97.20	90.34	**244.96**	94.72	74.44
32.96	95.98	93.41	**274.96**	90.19	80.27
62.96	96.61	96.58	**304.96**	86.80	89.57
92.96	97.53	98.85	**334.96**	88.04	89.18
122.96	98.49	99.42	**4.96**	93.68	84.48
152.96	99.58	99.20	**34.96**	88.78	92.03
Φ_5/°	% shape match with previous step	% shape match with original conformation	Φ_6/°	% shape match with previous step	% shape match with original conformation
214.27	/	/	**184.8**	/	/
244.27	92.15	92.15	**214.8**	99.83	99.83
274.27	96.75	90.94	**244.8**	96.88	96.72
304.27	93.97	92.95	**274.78**	98.31	95.16
334.27	92.08	90.42	**304.78**	99.04	95.14
4.27	93.22	87.38	**334.78**	98.23	96.68
34.27	95.29	87.47	**4.78**	96.68	99.70
64.27	95.39	84.64	**34.78**	99.91	99.66
94.27	96.64	83.85	**64.78**	97.22	97.28
124.27	93.86	84.70	**94.78**	98.61	96.03
154.27	90.66	89.01	**124.78**	99.35	95.99
184.27	94.18	90.20	**154.78**	98.73	97.03

Appendix Table 5.4: Shape matches for varying each torsion angle of molecule XX with 30° steps with both the previous step and the starting conformation. The angles are normalised to a 0-360° range. The first value for each torsion angle corresponds to the initial value.

Φ_1/°	% shape match with previous step	% shape match with original conformation	Φ_2/°	% shape match with previous step	% shape match with original conformation
82.16	/	/	-105.82	/	/
112.16	96.73	96.73	-75.82	92.21	92.21
142.16	99.00	95.80	-45.82	89.91	87.53
172.16	98.38	96.51	-15.82	89.36	80.07
202.16	96.84	99.38	14.18	88.41	73.23
232.16	98.77	98.77	44.18	90.57	70.52
262.16	98.54	99.74	74.18	91.14	69.33
292.16	96.81	96.58	104.18	83.40	72.75
322.16	99.16	95.82	134.18	87.45	72.05
352.16	98.34	96.59	164.18	87.44	79.87
22.16	96.77	99.21	194.18	90.05	84.41
52.16	98.74	98.52	224.18	85.74	91.60
Φ_3/°	% shape match with previous step	% shape match with original conformation	Φ_4/°	% shape match with previous step	% shape match with original conformation
6.3	/	/	355.83	/	/
36.3	97.13	97.13	25.83	79.10	79.10
66.3	96.04	95.35	55.83	77.91	75.17
96.3	92.59	89.79	85.83	80.14	67.42
126.3	91.09	82.54	115.83	84.26	60.41
156.3	91.58	76.72	145.83	91.36	61.17
186.3	90.58	71.42	175.83	88.85	65.04
216.3	85.41	71.05	205.83	84.93	70.56
246.3	94.06	74.17	235.83	77.86	72.66
276.3	88.51	81.72	265.83	77.89	90.07
306.3	91.30	88.15	295.83	91.79	91.39
336.3	92.42	94.76	325.83	95.81	94.03
Φ_5/°	% shape match with previous step	% shape match with original conformation	Φ_6/°	% shape match with previous step	% shape match with original conformation
0.09	/	/	348.21	/	/
31.09	96.13	96.13	18.21	92.24	92.24
61.09	94.90	91.49	48.21	92.39	85.74
91.09	94.36	86.91	78.21	94.67	81.96
121.09	94.09	82.95	108.21	97.81	81.01
151.09	86.96	76.37	138.21	96.05	82.84
181.09	93.73	75.67	168.21	94.17	85.94
211.09	95.29	78.21	198.21	93.43	88.28
241.09	91.33	83.22	228.21	93.73	89.06
271.09	88.43	91.89	258.21	94.58	89.85
301.09	95.82	94.58	288.21	95.07	91.18
331.09	97.20	97.05	318.21	95.92	94.13
Φ_7/°	% shape match with previous step	% shape match with original conformation	Φ_8/°	% shape match with previous step	% shape match with original conformation
286.42	/	/	107.04	/	/
316.42	94.13	94.13	137.04	99.53	99.53
346.42	95.92	91.18	167.04	99.39	98.95
16.42	95.07	89.85	197.04	99.26	98.41
46.42	94.58	89.06	227.04	98.65	97.32
76.42	93.73	88.28	257.04	99.59	97.09
106.42	93.43	85.94	287.04	99.62	97.42
136.42	94.17	82.84	317.04	99.14	98.16
166.42	96.05	81.01	347.04	99.02	98.85

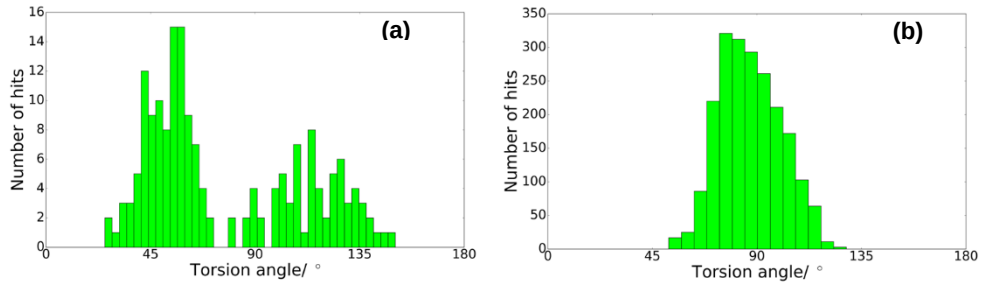
196.42	97.81	81.96	**17.04**	99.24	99.37
226.42	94.67	85.74	**47.04**	99.63	99.52
256.42	92.39	92.24	**77.04**	99.56	99.85

Appendix Table 5.5: Shape matches for varying each torsion angle of molecule XXIII with 30° steps with both the previous step and the starting conformation. The angles are normalised to a 0-360° range. The first value for each torsion angle corresponds to the initial value.

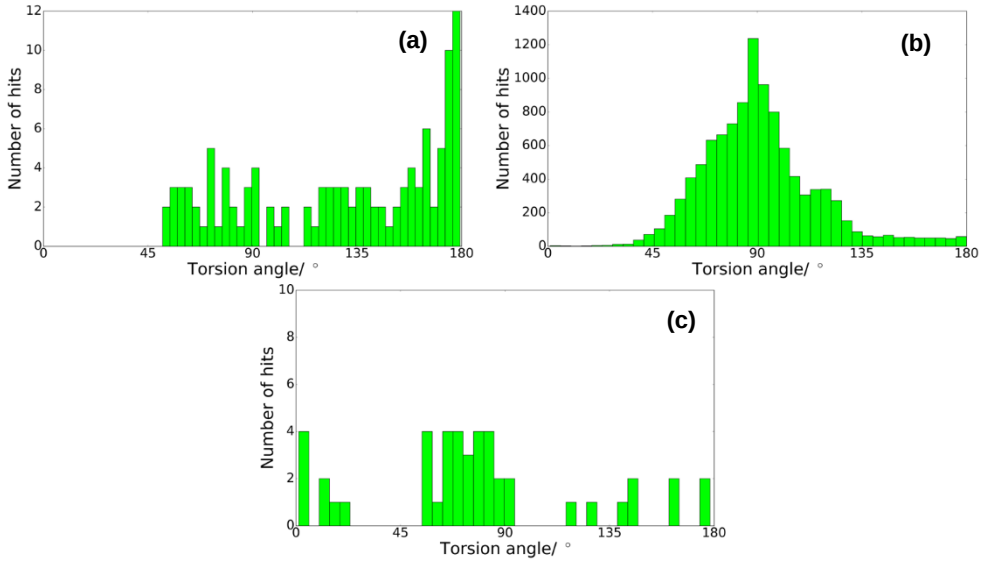
Φ_1/°	**% shape match with previous step**	**% shape match with original conformation**	**Φ_2/°**	**% shape match with previous step**	**% shape match with original conformation**
269.87	/	/	**188.84**	/	/
299.87	98.29	98.29	**218.84**	92.11	92.11
329.87	99.00	97.52	**248.84**	91.00	87.65
359.87	98.03	97.24	**278.84**	90.13	80.41
29.87	94.35	92.17	**308.84**	89.06	73.85
59.87	98.59	93.13	**338.84**	90.73	68.96
89.87	91.96	96.75	**8.84**	89.68	69.65
119.87	97.26	94.30	**38.84**	86.73	72.40
149.87	97.00	95.95	**68.84**	91.64	76.79
179.87	95.29	95.52	**98.84**	89.29	84.32
209.87	98.43	94.93	**128.84**	86.77	87.08
239.87	96.75	97.57	**158.84**	88.76	96.55
Φ_3/°	**% shape match with previous step**	**% shape match with original conformation**	**Φ_4/°**	**% shape match with previous step**	**% shape match with original conformation**
87.22	/	/	**132.45**	/	/
117.22	97.54	97.54	**162.45**	97.83	97.83
147.22	93.52	92.18	**192.45**	98.58	96.47
177.22	98.58	92.72	**222.45**	93.77	92.15
207.22	96.14	90.77	**252.45**	96.17	90.65
237.22	97.74	92.21	**282.45**	95.55	92.97
267.22	94.80	90.36	**312.45**	91.32	94.18
297.22	93.37	96.45	**342.45**	98.04	95.97
327.22	94.51	97.61	**12.45**	95.12	98.73
357.22	92.81	92.50	**42.45**	91.06	91.07
27.22	99.65	92.58	**72.45**	98.83	91.97
57.22	98.67	93.28	**102.45**	93.00	97.35
Φ_5/°	**% shape match with previous step**	**% shape match with original conformation**	**Φ_6/°**	**% shape match with previous step**	**% shape match with original conformation**
166.75	/	/	**4.47**	/	/
196.75	97.29	97.29	**34.47**	98.40	98.40
226.75	91.32	90.47	**64.47**	94.44	93.41
256.75	97.13	88.35	**94.47**	99.77	93.23
286.75	97.73	87.06	**124.47**	99.84	93.29
316.75	97.72	86.89	**154.47**	93.84	97.17
346.75	97.39	87.48	**184.47**	98.67	98.42
16.75	97.57	88.88	**214.47**	98.66	97.19
46.75	98.37	89.96	**244.47**	94.56	93.42
76.75	98.33	90.68	**274.47**	99.88	93.41
106.75	98.07	91.71	**304.47**	99.76	93.56
136.75	97.81	92.99	**334.47**	93.68	98.59

Appendix Table 5.6: Shape matches for varying each torsion angle of mebendazole with 30° steps with both the previous step and the starting conformation. The angles are normalised to a 0-360° range. The first value for each torsion angle corresponds to the initial value.

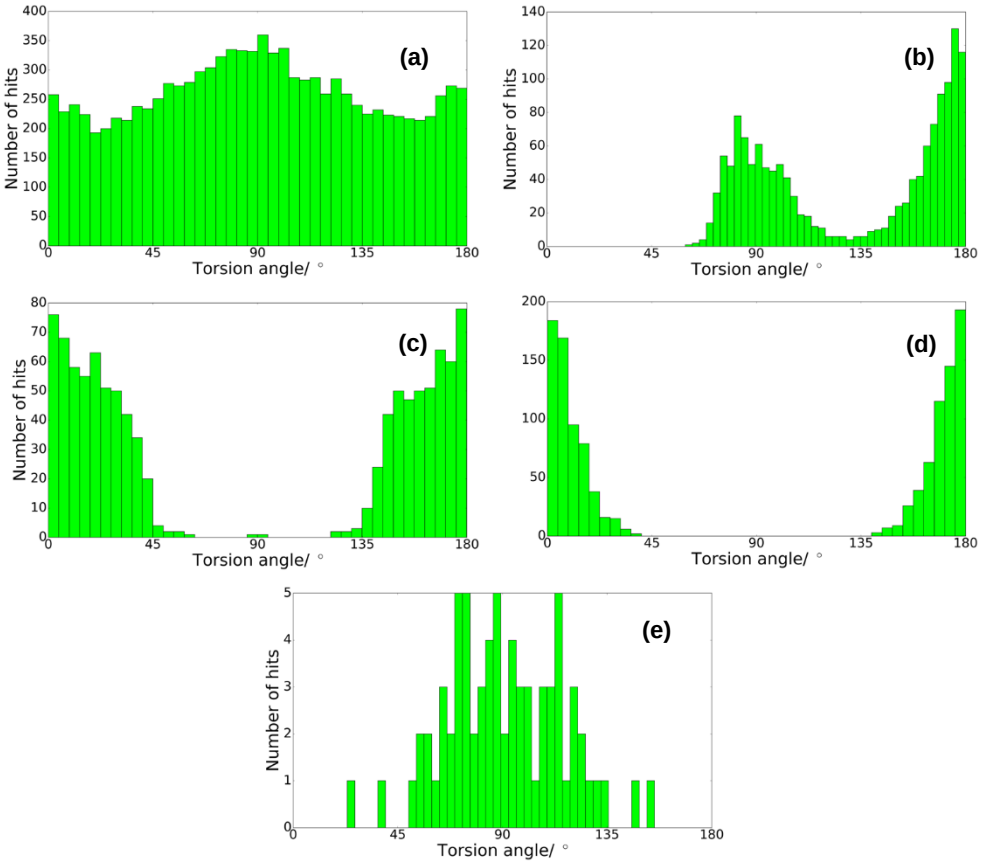
Φ_1/°	% shape match with previous step	% shape match with original conformation	Φ_2/°	% shape match with previous step	% shape match with original conformation
166.19	/	/	**327.79**	/	/
199.19	97.62	97.62	**357.79**	93.86	93.86
229.19	97.59	95.35	**27.79**	88.55	85.12
259.19	90.64	87.23	**57.79**	91.60	86.59
289.19	88.47	95.96	**87.79**	91.08	90.66
319.19	98.10	96.42	**117.79**	92.04	88.46
349.19	97.77	96.20	**147.79**	95.51	85.08
19.19	97.79	95.56	**177.79**	96.40	82.72
49.19	98.32	94.59	**207.79**	97.04	81.73
79.19	94.43	95.84	**237.79**	93.19	86.22
109.19	99.32	96.47	**267.79**	94.90	90.05
139.19	98.47	97.94	**297.79**	93.66	93.47
Φ_3/°	% shape match with previous step	% shape match with original conformation	Φ_4/°	% shape match with previous step	% shape match with original conformation
176.61	/	/	**0.42**	/	/
206.61	98.20	98.20	**30.42**	98.21	98.21
236.61	97.81	96.08	**60.42**	96.89	95.61
266.61	97.75	94.12	**90.42**	95.70	91.68
296.61	97.00	91.71	**120.42**	94.81	88.35
326.61	98.21	90.93	**150.42**	96.44	85.77
356.61	97.14	91.57	**180.42**	95.59	86.03
26.61	97.66	93.26	**210.42**	98.86	86.59
56.61	97.49	94.45	**240.42**	96.88	86.81
86.61	97.27	96.24	**270.42**	95.72	90.21
116.61	99.28	96.63	**300.42**	95.32	94.31
146.61	98.31	97.75	**330.42**	96.15	97.87
Φ_5/°	% shape match with previous step	% shape match with original conformation			
0.62	/	/			
30.62	98.44	98.44			
60.62	97.42	95.95			
90.62	96.96	93.22			
120.62	96.30	90.10			
150.62	98.40	89.01			
180.62	99.06	89.22			
210.62	98.23	90.57			
240.62	97.40	92.74			
270.62	97.20	95.21			
300.62	97.57	97.18			
330.62	98.49	98.27			



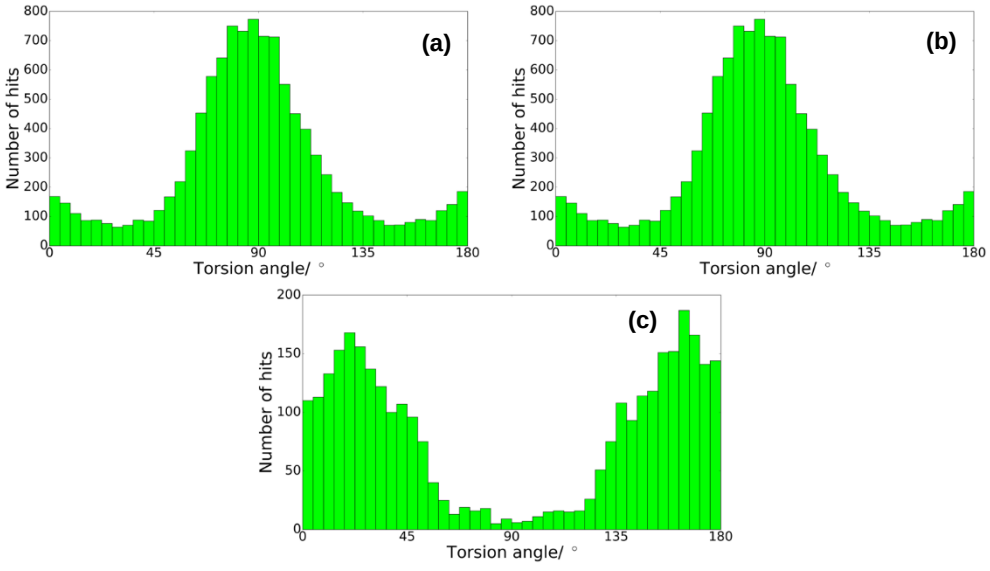
Appendix Figure 5.8: Mogul distributions used to select the ΔE_{intra} grid dimensions for (a) $\Phi_{1a\text{-}1b}$ and (b) Φ_4 of molecule XXVI.



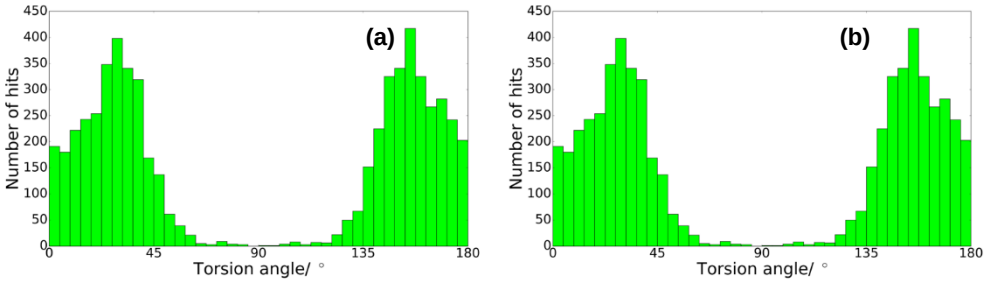
Appendix Figure 5.9: Mogul distributions used to select the ΔE_{intra} grid dimensions for (a) Φ_1 (b) Φ_4 and (c) Φ_5 of GSK269984B



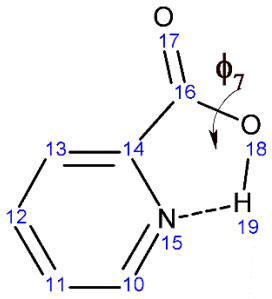
Appendix Figure 5.10: Mogul distributions used to select the ΔE_{intra} grid dimensions for (a) Φ_1 (b) Φ_2 (c) Φ_5 (d) Φ_6 and (e) Φ_8 of molecule XX.



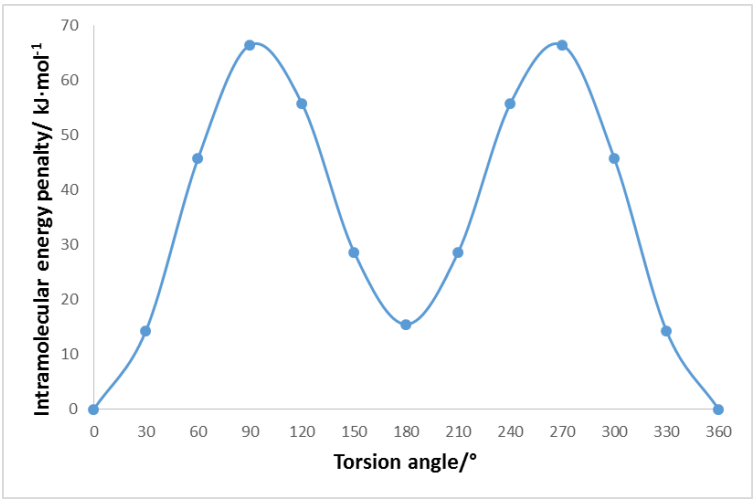
Appendix Figure 5.11: Mogul distributions used to select the ΔE_{intra} grid dimensions for (a) Φ_1 (b) Φ_3 (c) Φ_4 of molecule XXIII.



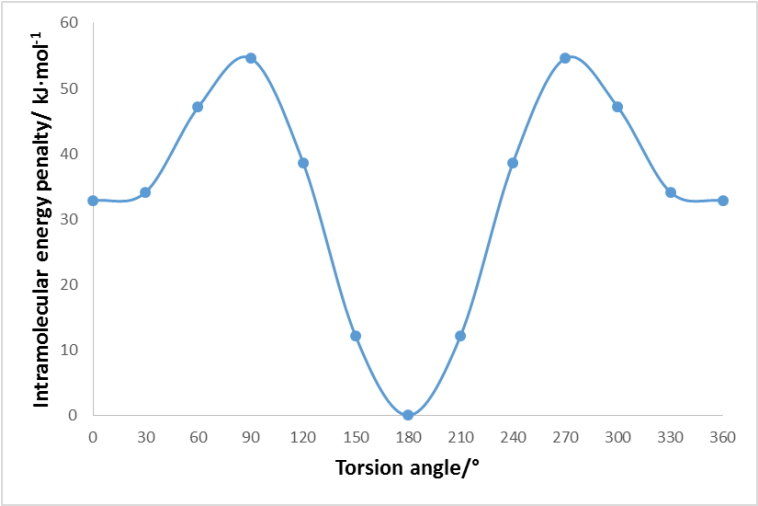
Appendix Figure 5.12: Mogul distributions used to decide the flexible treatment and select the ΔE_{intra} grid dimensions for (a) Φ_1 (b) Φ_2 of mebendazole.



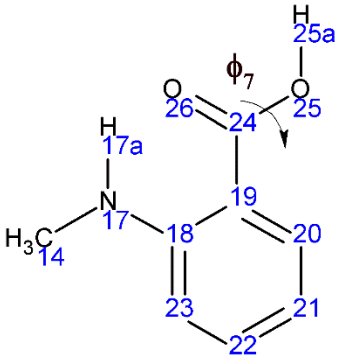
Appendix Figure 5.13: Fragment used to scan angle Φ_7 in GSK269984B.



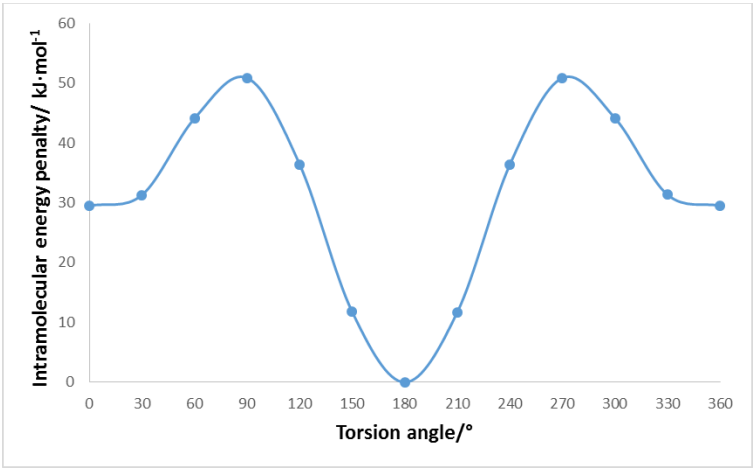
Appendix Figure 5.14: Constrained angle scan of Φ_7 in GSK269984B, with 30° steps, when the OH group is on the same side as the N atom in the fragment shown in Appendix Figure 5.13.



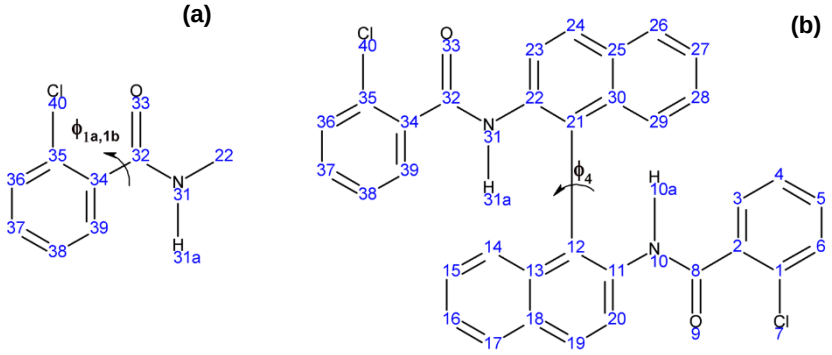
Appendix Figure 5.15: Constrained angle scan of Φ_7 in GSK269984B, with 30° steps, when the OH group is on the opposite side of the N atom in the fragment shown in Appendix Figure 5.13.



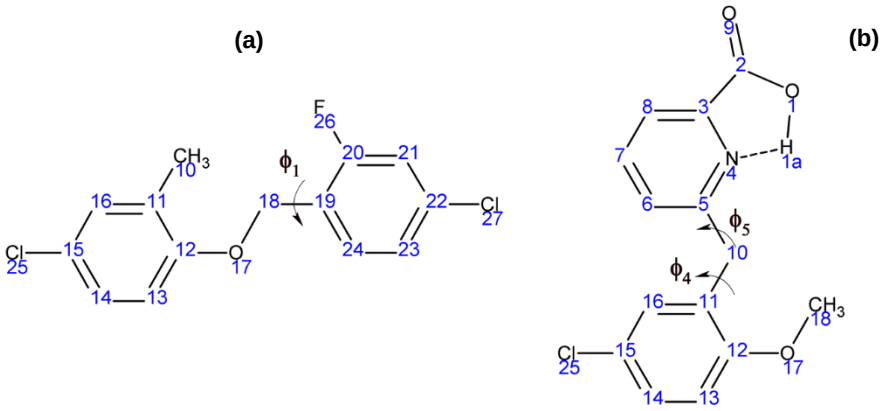
Appendix Figure 5.16: Fragment used to scan angle Φ_7 in molecule XXIII.



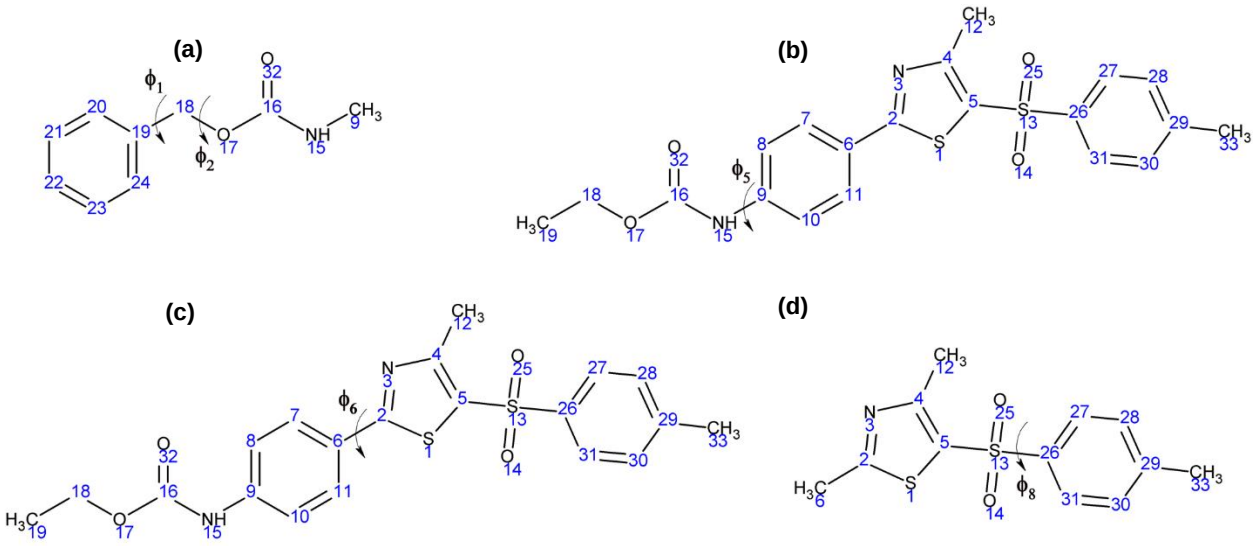
Appendix Figure 5.17: Constrained angle scan of Φ_7 in molecule XXIII, with 30° steps, using the fragment shown in Appendix Figure 5.16.



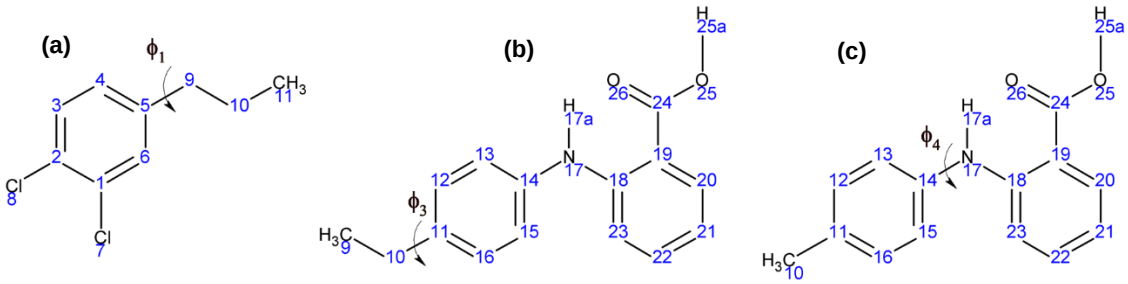
Appendix Figure 5.18: (a) Surrogate molecule used to calculate the ΔE_{intra} grids of Φ_{1a} and Φ_{1b} of molecule XXVI. (b) Whole molecule XXVI used to calculate the ΔE_{intra} grid of sterically congested Φ_4.



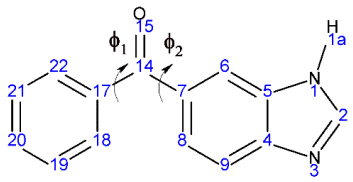
Appendix Figure 5.19: Surrogate molecules used to calculate the ΔE_{intra} grids of (a) Φ_1 (b) Φ_4 and Φ_5 of GSK269984B.



Appendix Figure 5.20: Surrogate molecules used to calculate the ΔE_{intra} grids of (a) Φ_1 and Φ_2 (b) Φ_5 (c) Φ_6 and (d) Φ_8 of molecule XX.



Appendix Figure 5.21: Surrogate molecules used to calculate the ΔE_{intra} grids of (a) Φ_1 (b) Φ_3 (c) Φ_4 of molecule XXIII.



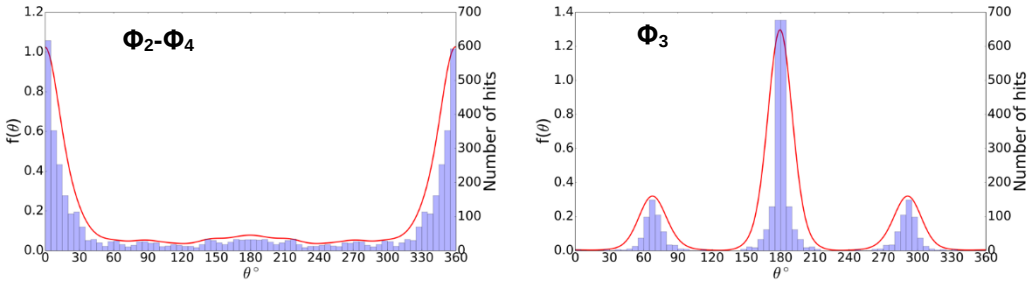
Appendix Figure 5.22: Surrogate molecule used to calculate the ΔE_{intra} grid of Φ_1 and Φ_2 of both tautomers of mebendazole.

Appendix Table 5.7: List of the 59 space groups considered in the searches.

P1	$P\bar{1}$	$P2_1$	$P2_1/c$	$P2_12_12$	$P2_12_12_1$	$Pna2_1$	$Pca2_1$	Pbca	Pbcn
C2/c	Cc	C2	Pc	Cm	$P2_1/m$	C2/m	P2/c	$C222_1$	$Pmn2_1$
$Cmc2_1$	Aba2	Fdd2	Iba2	Pnna	Pccn	Pbcm	Pnnm	Pmmn	Pnma
Cmcm	Cmca	Fddd	Ibam	$P4_1$	$P4_3$	$I\bar{4}$	P4/n	$P4_2/n$	I4/m
$I4_1/a$	$P4_12_12$	$P4_32_12$	$P\bar{4}2_1c$	$I\bar{4}2d$	$P3_1$	$P3_2$	R3	$P\bar{3}$	$R\bar{3}$
$P3_121$	$P3_221$	R3c	$R\bar{3}C$	$P6_1$	$P6_3$	$P6_3/m$	$P2_13$	$PA\bar{3}$	

Appendix Table 5.8: Atomic numbering definition of the key torsion angles of succinic acid shown in Figure 5.9.

Label for XXVI	Torsion Angle Definition
Φ_1	1a-1-2-4
Φ_2	3-2-4-5
Φ_3	2-4-5-6
Φ_4	4-5-6-8
Φ_5	5-6-7-7a

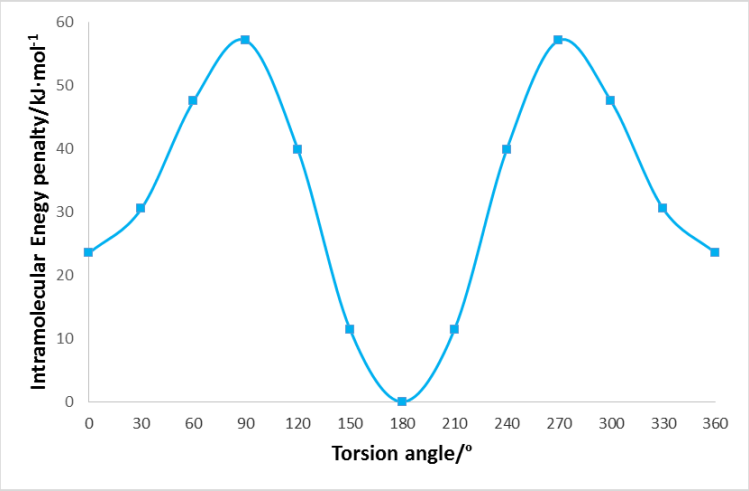


Appendix Figure 5.23: Histograms (light purple bars) and Von Mises kernel density approximations (red lines) for torsion angle distributions of angles succinic acid.

Appendix Table 5.9: Shape matches for varying each torsion angle of succinic acid with 30° steps with both the previous step and the starting conformation. The angles are normalised to a 0-360° range. The first value for each torsion angle corresponds to the initial value.

Φ_2/°	% shape match with previous step	% shape match with original conformation	Φ_3/°	% shape match with previous step	% shape match with original conformation
10.76	/	/	**180**	/	/
40.76	94.90	94.90	**210**	94.99	94.99
70.76	96.84	92.05	**240**	92.15	87.87
100.76	98.53	92.71	**270**	93.80	83.05
130.76	94.09	91.83	**300**	94.43	79.41
160.76	95.54	95.10	**330**	95.46	77.37
190.76	99.13	94.52	**0**	97.72	76.31
220.76	94.81	93.21	**30**	97.72	77.37
250.76	93.66	93.15	**60**	95.46	79.41
280.76	97.63	91.35	**90**	94.43	83.05
310.76	97.00	94.00	**120**	93.80	87.88
340.76	96.62	96.63	**150**	92.15	94.99

Φ_4/°	% shape match with previous step	% shape match with original conformation
349.24	/	/
19.24	96.63	96.63
49.24	96.62	94.01
79.24	97.00	91.35
109.24	97.63	93.15
139.24	93.66	93.21
169.24	94.81	94.52
199.24	99.13	95.10
229.24	95.54	91.83
259.24	94.09	92.71
289.24	98.53	92.05
319.24	96.84	94.90



Appendix Figure 5.24: Constrained angle scan of Φ_1 in succinic acid, with 30° steps, using the whole molecule (Figure 5.9). This scan was considered to be valid for torsion angle Φ_5 because of symmetry.

Appendix Table 5.10: Selected conformational regions and calculation of the associated ΔE_{intra} value of molecule XXVI. The conformational regions highlighted in blue were taken to the search stage; some with ΔE_{Intra}^{CR} < 26 kJ·mol^{-1} were excluded because of an approximate molecular symmetry relationship; in those cases, the lower energy one was always chosen.

CG conformation number	Φ_{2a}/°	Φ_{3a}/°	Φ_{2b}/°	Φ_{3b}/°	ΔE_{intra}
1	-4	178	9	179	1.92
25	0	212	-8	212	11.7
26	-4	151	-8	151	11.63
29	2	148	4	178	4.35
30 ~ 41 for symmetry	-2	212	-4	182	6.54
31 ~ 29 for symmetry	-2	178	2	148	4.47
37	2	209	8	148	7.93
38 ~ 37 for symmetry	-2	151	-8	212	8.67
41	2	178	2	209	5.54
60 ~ 124 for symmetry	0	178	8	238	15.77
61	0	182	-8	122	12.86
89	-3	66	-9	181	20.98
93	-4	209	8	118	16.42
94 ~ 198 for symmetry	4	151	-8	242	21.14
122 ~ 61 for symmetry	0	122	-4	182	12.93
124	-1	238	0	178	15.76
174 ~ 176 for symmetry	3	148	8	119	22.11
175	2	238	4	209	20.27
176	-2	122	-4	151	19.03
186 ~ 1502 for symmetry	3	177	5	89	25.78
187	-4	182	-4	272	26.34
190	2	268	9	179	25.54
198	2	238	2	148	19.63
199	-2	122	-2	212	16.07
218	-4	268	8	209	29.11
219	-4	148	8	88	31.66
224 ~ 175 for symmetry	-2	212	-4	242	22.61
232	0	65	0	212	23.61
241	-4	177	8	297	30.71
275	0	148	0	268	29.59
277	2	209	8	88	18.32
279	-2	212	-8	272	30.9
294	0	298	3	177	31.31
322	0	242	0	242	32.23
323	-4	122	-4	122	32.72
338	2	268	8	148	29.33
353 ~ 89 for symmetry	-4	179	8	59	22.16
364 ~ 396 for symmetry	-4	122	-8	242	25.54
375 ~ 805 for symmetry	-3	66	-8	242	21.01
396	-2	242	-8	122	25.53
407	0	148	8	297	33.26
417	-1	88	2	148	22.96
465	2	298	1	209	33.73
466	-2	212	-8	301	34.75
471	4	122	-8	272	36.93
472	-4	238	8	88	21.61
510	2	298	4	148	33.49
516	4	93	0	123	34.34
517	-4	272	-8	242	40.83
556	0	238	4	268	38.7
562	-2	272	-8	122	37.02
601	-4	58	2	148	38.17
609	4	123	-3	93	34.26

617	1	65	-8	272	31.18
646	0	152	-2	62	36.58
648	-4	239	8	297	43.3
696	-4	268	8	268	47.07
698 ~ 472 for symmetry	-2	92	-4	242	22.23
723 ~ 232 for symmetry	2	209	0	59	25.09
730 ~ 277 for symmetry	4	91	-2	212	23.57
741	1	298	4	239	43.27
742	0	122	0	301	41.14
754	-4	298	8	297	51.74
796	4	298	8	118	42.16
803	-4	117	8	59	37.67
805	4	241	-8	63	20.96
852	8	268	4	88	30.19
886	0	58	4	117	37.75
905	-10	61	-10	62	84.61
1160	2	298	4	268	51.42
1194	-2	272	-2	301	51.96
1214	4	91	-8	91	38.87
1502	-2	92	-9	181	23.87
1534	-4	178	178	143	23.2
1535	4	182	-178	217	30.07
1767	-178	178	9	179	31.86
1773	176	148	-5	182	27.63
1826	-179	216	-4	182	30.17
2000	-2	182	178	182	30.64
2099	179	144	2	148	35.22
2100	-179	216	-2	212	35.85
2101	-175	213	2	148	33.17
2102	174	147	-2	212	40.4
2380	-179	177	2	209	36.19
2381	179	183	-2	151	34.79
2533	-179	216	-8	242	47.54
2534	179	144	9	239	41.28
2535	179	144	8	118	38.82
2537	-174	213	8	118	40.85
2776	-2	242	174	146	49.44
2813	178	182	-8	242	46.44
2814	179	183	-8	122	45.51
2875	-4	212	175	236	43.66
2876	0	151	175	236	38.68
2954	-2	122	-180	213	40.05
2972	4	212	174	147	40.27
3058	2	238	-178	178	46.43
3060	2	118	-178	178	48.81
3188	2	209	-176	209	33.09
3189	-2	151	174	146	33.95
3192	1	123	179	147	45.88
3213	-179	216	0	92	41.97
3214	-179	216	0	272	54.2
3308	-180	146	5	267	49.48
3403	-180	146	5	91	43.72
3421	2	238	179	233	53.41
3445	-178	268	9	178	56.52
3489	-174	213	8	297	53.99
3518	-2	68	174	234	44.9
3596	2	177	-178	268	56.33
3682	-178	88	9	180	43.28
3716	-179	149	11	63	44.79
3717	-180	213	-8	63	43.97
3833	2	180	-177	89	43.14
3840	2	207	-178	270	64.2
3841	177	257	2	205	53.53

3842	178	272	-2	151	60.4
3912	-179	100	-8	211	39.19
4007	0	150	-177	121	34.63
4097	-175	92	2	148	41.38
4098	-178	270	8	121	67.03
4099	-178	271	8	239	78.72
4109	5	300	-178	236	59.1
4138	-4	122	-178	238	45.2
4228	-2	122	178	272	50.67
4229	-2	242	178	273	68.4
4310	178	91	-8	241	53.02
4395	2	238	-178	88	51.68
4396	5	95	-172	124	46.35
4774	2	148	-176	210	32.72
4819	179	144	178	143	40.77
4820	-179	216	-178	217	56.76
4821	-174	213	178	143	46
4822	174	147	-178	217	48.01
4872	6	237	-178	119	43.25
4899	2	209	-178	117	38.59
4928	-178	238	2	122	45.03
4941	-178	215	178	242	66.95

Appendix Table 5.11: Selected conformational regions and calculation of the associated ΔE_{intra} value of GSK269984B. The values of the polar hydrogen torsion angle highlighted in yellow were added manually, and those conformations that differ from the CG-generated ones only in the values of this dihedral are indicated using a '_N' notation. The conformational regions highlighted in blue were taken to the search stage; some with ΔE_{Intra}^{CR} < 26 kJ·mol^{-1} were excluded because of an approximate molecular symmetry relationship; in those cases, the lower energy one was always chosen.

CG conformation number	Φ_2/°	Φ_3/°	Φ_6/°	Φ_7/°	ΔE_{intra}
1	180	180	180	0	0.07
1_2	180	180	180	180	12.89
11	0	180	180	180	14.02
184	180	181	272	0	5.84
184_2	180	181	272	180	21.64
185 ~ 184 for symmetry	180	178	88	0	10.17
185_2 ~ 184_2 for symmetry	180	178	88	180	25.75
306	0	181	272	180	19.42
307 ~ 306 for symmetry	0	178	88	180	19.56
475	180	120	180	0	16.2
475_2	180	120	180	180	22.57
477	180	240	180	0	13.97
477_2	180	240	180	180	22.39
666	360	120	180	180	26.74
668	0	240	182	180	26.92
1694	180	75	181	0	23.31
1694_2	180	75	181	180	37.41
2018	180	285	179	0	20.48
2018_2	180	285	179	180	34.7
2302	360	75	181	180	43.91
2655	0	285	179	180	39.89
2730 ~ 2731 for symmetry	180	240	270	0	20.3
2730_2	180	240	270	180	35.67
2731	180	239	91	0	16.89
2731_2	180	239	91	180	28.31
2732	180	120	90	0	18.44
2732_2	180	120	90	180	33.64
3487	0	240	270	180	41.6

3488	0	121	269	180	42.66
3489	0	120	90	180	39.97
3491	360	238	91	180	31.57
5036	180	124	300	0	16.92
5036_2	180	124	300	180	32.57
5162	180	287	267	0	25.91
5162_2	180	287	267	180	37.21
5974	180	71	91	0	24.6
5974_2	180	71	91	180	39.48
6136	0	287	267	180	41.77
6810	180	120	240	0	24.51
6810_2	180	71	91	180	38.58
6875	0	71	91	180	45.49
9131	180	91	33	0	27.33
9131_2	180	91	33	180	34.39
9132	180	271	327	0	29.88
9132_2	180	271	327	180	42.18
9267	180	242	32	0	26.04
9267_2	180	242	32	180	38.83
9327	0	91	33	180	34.81
9329	360	271	327	180	46.87
9395	360	242	32	180	41.9
9458	0	123	327	180	60.49

Appendix Table 5.12: Selected conformational regions and calculation of the associated ΔE_{intra} value of molecule XX. The conformational regions highlighted in blue were taken to the search stage; some with ΔE_{Intra}^{CR} < 26 kJ·mol^{-1} were excluded because of an approximate molecular symmetry relationship; in those cases, the lower energy one was always chosen.

CG conformation number	Φ_3/°	Φ_4/°	Φ_7/°	ΔE_{intra}
1	180	180	270	1.45
25	180	180	240	5.25
78	180	180	300	3.34
92	180	180	180	12.33
245	180	357	270	11.68
273	180	180	210	8.48
413	180	180	330	15.50
515	180	3	240	15.50
745	180	357	300	13.79
826 ~ 273 for symmetry	179	180	150	8.46
1290	180	357	180	22.48
2911	180	357	330	25.84
3276 ~ 3739 for symmetry	180	357	150	18.56
3739	180	357	210	18.55
4480 ~ 245 for symmetry	180	357	90	11.69
4974 ~ 2911 for symmetry	180	357	30	25.87
7122	358	140	270	34.18
9765	358	140	240	38.30
10397	180	330	119	15.37
14548	233	177	120	26.23
16943	126	180	120	32.05

Appendix Table 5.13: Selected conformational regions and calculation of the associated ΔE_{intra} value of molecule XXIII. The values of the polar hydrogen torsion angle highlighted in yellow were added manually. The conformational regions highlighted in blue were taken to the search stage; some with ΔE_{Intra}^{CR} < 26 kJ·mol^{-1} were excluded because of an approximate molecular symmetry relationship; in those cases the lower energy one was always chosen.

CG conformation number	Φ_2/°	Φ_5/°	Φ_6/°	Φ_7/°	ΔE_{intra}
1	180	164	360	180	0.57
9	360	164	0	180	25.13
17	180	164	180	180	16.66
33	0	164	180	180	41.2
49	60	164	0	180	1.15
51 ~ 49 for symmetry	300	164	0	180	2.17
65	0	229	360	180	32.68
103	180	214	0	180	3.34
145 ~ 147 for symmetry	60	164	180	180	17.56
147	300	164	180	180	17.02
172	0	229	180	180	45.8
231	180	161	240	180	35.84
232	180	161	120	180	43.69
235	180	214	180	180	18.11
256	180	117	360	180	16.07
290	60	229	360	180	9.3
309	180	270	0	180	32.67
328	0	199	120	180	59.34
329	0	199	240	180	67.29
402	0	199	60	180	63.72
403	0	199	300	180	54.51
434	360	90	360	180	53.8
464	180	64	2	180	47.28
500	300	214	0	180	4.94
512	180	117	180	180	27.25
563	60	229	180	180	22.58
613	180	269	180	180	38.59
645	180	199	60	180	37.47
646	180	199	300	180	44.14
658	0	296	358	180	71.63
712	0	117	180	180	51.65
831	0	143	240	180	62.23
893	60	199	120	180	36.3
894	60	199	240	180	43.95
897	300	199	120	180	48.44
919	300	214	180	180	19.44
949	60	117	360	180	16.42
950 ~ 949 for symmetry	300	117	360	180	17.91
1055	300	199	60	180	38.75
1056	300	199	300	180	32.54
1177	300	270	360	180	34.13
1580	300	64	2	180	47.64
1581	60	64	2	180	42.06
1640	0	138	90	180	76.89
1677	60	117	180	180	27.94
1678	300	117	180	180	27.17
1720	300	164	240	180	36.33
1957	300	270	180	180	39.94
2016	60	199	60	180	40.49
2017	60	199	300	180	31.37
2279	300	229	240	180	49.39
2334	180	61	213	180	45.36
2511	0	61	213	180	62.18

2512	0	299	147	180	69.56
2527	180	207	240	180	44.01
2530	180	229	120	180	39.45
2661	60	296	358	180	48.52
2915	180	131	300	180	46.17
2916	180	131	60	180	35.38
2926	300	146	90	180	52.46
2932	60	143	90	180	53.14
2933	60	143	270	180	55.14
3001	0	270	240	180	83.92
3020	0	117	300	180	76.46
3098	180	300	300	180	54.68
3099	180	61	60	180	54.68
3169	180	300	120	180	52.81
3170	180	63	118	180	66.95
3468	0	90	60	180	64.76
3469	0	270	300	180	72.65
3489	0	90	120	180	77.15
3882	0	270	60	180	85.38
3948 ~ 3950 for symmetry	120	164	0	180	15.94
3950	240	164	360	180	15.9
4283	359	36	60	180	70.45
4484	180	267	241	180	58.81
4810	360	63	269	180	82.6
4889	60	61	213	180	44.03
4891	60	299	147	180	46.9
4991	300	120	300	180	51.82
5452	300	61	213	180	39.86
5580	180	61	299	180	64.73
5590	180	299	61	180	65.36
5609	60	90	120	180	61.79
5741	120	164	180	180	32.03
5743	240	164	180	180	31.73
6049	300	90	60	180	44.92
6050	60	90	300	180	57.91
6155	300	90	120	180	53.02
6560	60	60	60	180	46.32
6563	60	299	300	180	55.85
6586	300	270	60	180	62.25
6587	300	270	300	180	49.82
6657	300	292	121	180	53.91
7010	330	325	300	180	67.45
7199	120	229	360	180	24.24
7242	300	35	60	180	55.32
7811	60	270	240	180	60.15
8717	240	214	0	180	18.84
8819	60	270	60	180	61.7
9063	120	229	180	180	37.37
9528	299	61	299	180	60.17
10475	240	161	120	180	58.81
10476	240	199	240	180	58.74
10557	240	214	180	180	33.5
10643	240	117	360	180	31.44
10644	120	117	0	180	31.05
10888	120	199	60	180	55.41
10889	240	161	60	180	46.51
10890	240	161	300	180	55.45
11600	240	270	0	180	48.03
11777	120	164	270	180	68.83
11924	120	64	2	180	62.08
11926	120	296	358	180	62.73
11928	240	64	2	180	62.18
12019	120	117	180	180	42.43

12024	240	117	180	180	42.54
12045	120	164	120	180	58.88
12267	240	270	180	180	53.82
12351	293	297	242	180	66.85
12628	240	229	120	180	54.52
13511	120	210	120	180	51.8
13512	120	210	240	180	59.61
13514	240	150	240	180	51.53
13625	120	150	60	180	47.08
13628	240	229	300	180	50.56
14103	240	214	60	180	57.34

Appendix Table 5.14: Selected conformational regions and calculation of the associated ΔE_{intra} value of the A-tautomer of mebendazole. The conformational regions highlighted in blue were taken to the search stage.

CG conformation number	Φ_3/°	Φ_4/°	Φ_5/°	ΔE_{intra}
1	180	360	0	0.00
5	0	0.01	360	50.49
35	184	186	360	16.25
46	2	183	0	48.17

Appendix Table 5.15: Selected conformational regions and calculation of the associated ΔE_{intra} value of the C-tautomer of mebendazole. The conformational regions highlighted in blue were taken to the search stage.

CG conformation number	Φ_3/°	Φ_4/°	Φ_5/°	ΔE_{intra}
1	180	0	0	3.27
5	0	0	0	56.54
33	184	186	0	21.08
41	2	183	360	52.89

Appendix Table 5.16: Comparison of the crystal structure search with the new workflow and the previous CSP results for molecule XXVI (Figure 5.11a). The structure highlighted in yellow is a match to the experimental form, the one in red was not found with the new method. When $RMSD_{15}$ values are highlighted in blue, this indicates that the structure had been probably found in the search (*i.e.* $RMSD_{15} > 0.8$ Å).

Structure name	Found?	Conformation number	Previous CSP ranking	Previous CSP lattice energy/kJ mol^{-1}	New method ranking after search	$RMSD_{15}$
3525	YES	29	1	-206.86	1	0.506
1600	YES	124	2	-206.37	251	0.695
675	YES	124	3	-204.25	61	0.257
38	YES	1	4	-202.71	122	0.597
421	YES	124	5	-201.43	61	0.295
3104	YES	805	6	-201.20	164	0.297
615	YES	41	7	-200.58	72	0.426
239	PROBABLY	1	8	-200.48	233	0.805
2930	YES	29	9	-200.30	51	0.37
354	YES	29	10	-199.98	71	0.377
851	YES	29	11	-199.82	1071	0.442
6460	YES	29	12	-199.74	7	0.666
6335	YES	29	13	-199.41	10	0.579
221	YES	29	14	-199.39	7	0.648
2231	PROBABLY	29	15	-199.29	56	0.809
2496	NO	/	16	-198.93	/	/
185	YES	805	17	-198.75	159	0.325
4201	PROBABLY	41	18	-198.65	670	1.223
314	YES	29	19	-198.63	13	0.406
508	YES	1	20	-198.57	31	0.505
4946	YES	29	21	-198.48	56	0.342
6879	YES	29	22	-198.35	632	0.43
506	YES	29	23	-198.23	20	0.42
4842	YES	29	24	-198.02	39	0.478
43	YES	41	25	-197.84	26	0.467
1236	YES	41	26	-197.71	33	0.387
1537	YES	1	27	-197.69	21	0.43
188	YES	61	28	-197.45	116	0.661
5126	YES	1	29	-196.81	855	0.504
444	YES	25	30	-196.74	42	0.683
544	YES	29	31	-196.57	1071	0.408
686	YES	29	32	-196.52	287	0.406
89	PROBABLY	41	33	-196.42	614	0.841
20	YES	29	34	-196.16	138	0.597
83	YES	29	35	-196.04	3	0.546
2591	YES	805	132	-189.83	1699	0.304

Appendix Table 5.17: Comparison of the crystal structure search with this new workflow and the previous CSP results for GSK269984B (Figure 5.11b in the main paper). The structure highlighted in yellow is a match to the experimental form, the ones in red were not found with the new method. When $RMSD_{15}$ values are highlighted in blue, this indicates that the structure had been probably found in the search (*i.e.* $RMSD_{15}$ > 0.8 Å).

Structure name	Found?	Conformation number	Previous CSP ranking	Previous CSP lattice energy/kJ mol^{-1}	New method ranking after search	$RMSD_{15}$
180Intra1	YES	1	1	-180.68	46	0.129
90InterB3	YES	306	2	-180.15	5072	0.626
180Inter	YES	1_2	3	-178.62	776	0.374
180Inter	YES	1_2	4	-177.92	225	0.543
180InterB	YES	11	5	-177.42	354	0.32
180Intra8	YES	1	6	-177.13	31	0.142
180Intra3	PROBABLY	1	7	-177.06	68	1.175
180InterB	YES	11	8	-176.88	942	0.282
180Intra7	YES	1	9	-176.44	1022	0.486
180Inter	YES	1_2	10	-176.32	352	0.722
90InterB6	NO	/	11	-176.07	/	/
180Intra1	YES	1	12	-176.02	87	0.211
180Intra7	YES	1	13	-175.68	2217	0.341
180Intra4	YES	1	14	-175.64	38	0.355
180Inter	YES	1_2	15	-175.60	1104	0.182
180Intra2	YES	1	16	-175.52	1	0.344
180Inter	NO	/	17	-175.36	/	/
180Inter	YES	1_2	18	-175.29	1320	0.15
180Intra8	NO	/	19	-175.28	/	/
180Intra5	PROBABLY	1	20	-175.24	53	0.977
180Inter	YES	1_2	21	-174.87	180	0.298
90Intra31	YES	184	22	-174.80	7969	0.349
180Intra3	YES	1	23	-174.53	117	0.177
180Inter	PROBABLY	1_2	24	-174.49	325	2.052
180Intra9	YES	1	25	-174.33	53	0.631
180Inter	YES	1_2	26	-174.21	325	0.505
180Inter	YES	1_2	27	-174.17	1499	0.172
180InterB	YES	11	28	-174.15	489	0.507
90InterA	NO	/	29	-174.06	/	/
180Intra8	PROBABLY	1	30	-174.01	926	1.437
180Intra4	YES	1	31	-173.96	125	0.294
180Intra6	YES	1	32	-173.87	50	0.118
90InterA	YES	184_2	33	-173.79	14162	0.362
180Intra5	PROBABLY	1	34	-173.74	49	0.905
180Intra2	YES	1	35	-173.60	68	0.241
180Inter	PROBABLY	1_2	36	-173.52	325	1.884
180InterB	YES	11	37	-173.50	1749	0.192
180Intra5	YES	1	38	-173.25	114	0.156

Appendix Table 5.18: Comparison of the crystal structure search with this new workflow and the previous CSP results for molecule XX (Figure 5.11c in the main paper). The structure highlighted in yellow is a match to the experimental form. When $RMSD_{15}$ values are highlighted in blue, this indicates that the structure had been probably found in the search (i.e. $RMSD_{15} > 0.8$ Å).

Structure name	Found?	Conformation number	Previous CSP ranking	Previous CSP lattice energy/kJ mol^{-1}	New method ranking after search	$RMSD_{15}$
dfAa132	YES	1	1	-218.73	10	0.386
dfAc102	YES	1	2	-217.95	161	0.264
dfAa180	YES	1	3	-216.35	61	0.618
dfAc14	YES	1	5	-213.14	70	0.308
dfAc48	YES	1	10	-212.58	278	0.46
dfAc19	YES	1	11	-212.30	223	0.688
dfAc7	YES	78	12	-211.47	14	0.365
dfAc43	YES	78	14	-211.04	1740	0.778
dfAc17	YES	1	15	-210.87	40	0.449
dfAc172	YES	1	16	-210.76	58	0.572
dfAc29	PROBABLY	1	17	-210.54	2154	1
dfAb181	YES	78	22	-209.62	2428	0.79
dfAd152	PROBABLY	78	23	-209.60	1596	1.012
dfAc86	YES	1	24	-209.37	13	0.532
dfAc67	YES	1	25	-209.25	94	0.167
dfAa277	YES	1	27	-209.03	134	0.546
dfAa4	YES	1	28	-208.97	2	0.376
dfAa1	YES	1	29	-208.95	7	0.376
dfAb161	YES	1	31	-208.86	49	0.26
dfAb1	YES	1	32	-208.83	1	0.123
dfAd79	YES	1	33	-208.80	602	0.444
dfBa28	YES	245	47	-207.29	302	0.65

Appendix Table 5.19: Comparison of the crystal structure search with this new workflow and the previous CSP results for molecule XXIII (Figure 5.11d in the main paper). T

he structures highlighted in yellow are matches to the experimental Z'=1 forms, the ones in red were not found with the new method. When $RMSD_{15}$ values are highlighted in blue, this indicates that the structure had been probably found in the search (*i.e.* $RMSD_{15}$ > 0.8 Å).

Structure name	Found?	Conformation number	Previous CSP ranking	Previous CSP lattice energy/kJ mol^{-1}	New method ranking after search	$RMSD_{15}$
A1361	YES	1	1	-212.68	11	0.222
A70	YES	1	2	-211.02	10	0.309
A6494	PROBABLY	1	3	-210.55	5375	1.394
A691	YES	1	4	-209.30	46	0.767
A3457	YES	1	5	-209.00	47	0.433
A72	YES	1	6	-208.87	163	0.384
A424	YES	1	7	-208.27	3	0.483
A771	YES	1	8	-208.04	1	0.237
A191	YES	103	9	-207.61	514	0.719
A4890	YES	1	10	-207.22	24	0.746
A5191	NO	/	11	-207.16	/	/
A272	YES	1	12	-207.00	778	0.749
A63	PROBABLY	1	13	-206.63	678	0.894
A118	YES	1	14	-206.55	70	0.449
A75	YES	1	15	-206.39	358	0.485
A1413	YES	1	16	-206.35	13	0.375
A2457	YES	1	17	-206.02	951	0.671
A587	YES	1	18	-205.83	2821	0.422
A2417	YES	1	19	-205.71	399	0.407
A138	PROBABLY	1	20	-205.51	782	1.11
A227	YES	1	21	-205.34	277	0.514
A1949	PROBABLY	1	22	-205.07	497	1.114
A3174	NO	/	23	-204.92	/	/
A2054	NO	/	24	-204.87	/	/
A3023	YES	103	25	-204.83	106	0.481
A2311	YES	1	26	-204.82	5	0.343
A3513	YES	1	27	-204.71	475	0.686
A1109	YES	1	28	-204.69	2	0.447
A894	PROBABLY	1	29	-204.61	1279	1.259
A1422	YES	1	30	-204.53	377	0.488
A1127	YES	1	31	-204.53	7	0.276
A6634	PROBABLY	1	32	-204.34	3394	1.474
A282	YES	1	33	-203.87	3838	0.322
A323	PROBABLY	1	34	-203.83	199	0.807
A2715	YES	1	35	-203.76	2489	0.537
A24995	YES	1	36	-203.70	2983	0.42
A3746	YES	1	37	-203.69	735	0.615
A368	YES	1	38	-203.62	82	0.606
A6738	NO	/	39	-203.61	/	/
A4228	PROBABLY	1	40	-203.60	2304	1.073
A1752	YES	1	41	-203.52	511	0.471
A113	YES	1	42	-203.51	125	0.275
A3750	YES	1	43	-203.49	584	0.256
A505	YES	1	44	-203.41	1262	0.37
A12658	YES	1	45	-203.12	626	0.31
A1918	YES	1	46	-203.04	802	0.757
A1411	PROBABLY	1	47	-202.96	350	0.855
A5145	PROBABLY	1	48	-202.76	262	0.872
A710	YES	1	49	-202.70	155	0.338
B204	PROBABLY	49	66	-201.75	465	1.543

B60	YES	49	83	-201.03	2472	0.427
B184	PROBABLY	49	100	-200.32	846	1.284
Exptal A	YES	103	(167)	-199.08	4218	0.377

Appendix Table 5.20: Comparison of the crystal structure search with this new workflow and the previous CSP results for mebendazole (Figure 5.11e in the main paper). The structures highlighted in yellow are matches to the solved forms. When $RMSD_{15}$ values are highlighted in blue, this indicates that the structure had been probably found in the search (*i.e.* $RMSD_{15} > 0.8$ Å).

Structure name	Found?	Conformation number	Previous CSP ranking	Previous CSP lattice energy/kJ mol^{-1}	New method ranking after search	$RMSD_{15}$
A788	YES	A1	1	-182.51	1	0.302
A19	YES	A1	2	-180.35	3	0.160
C27	YES	C1	3	-179.96	111	0.260
C5	YES	C1	4	-179.96	22	0.247
C10	YES	C1	5	-179.88	30	0.265
A50	PROBABLY	A1	6	-179.36	8	1.031
A37	YES	A1	7	-178.43	5	0.357
C23	YES	C1	8	-178.17	66	0.124
C73	YES	C1	9	-178.17	223	0.101
C406	YES	C1	10	-177.83	1313	0.159
A53	YES	A1	11	-177.70	12	0.590
C53	YES	C1	12	-177.03	123	0.242
C25	YES	C1	13	-177.01	60	0.241
A173	YES	A1	14	-176.84	119	0.418
A72	YES	A1	15	-176.76	6	0.312
A49	PROBABLY	A1	16	-176.72	4	1.066
A78	YES	A1	17	-176.58	46	0.310
A90	YES	A1	18	-176.54	23	0.326
A291	YES	A1	19	-176.37	284	0.289
C248	YES	C1	20	-176.33	129	0.210
A306	YES	A1	21	-176.24	87	0.311
C46	YES	C1	22	-176.21	201	0.243
C24	YES	C1	23	-176.15	111	0.266
C115	YES	C1	24	-176.04	105	0.440
C509	PROBABLY	C1	25	-175.90	128	1.857
C583	YES	C1	26	-175.89	360	0.353
A202	YES	A1	27	-175.79	51	0.603
C106	YES	C1	28	-175.78	56	0.235
A143	YES	A1	29	-175.31	178	0.429
A89	YES	A1	30	-175.18	29	0.247
C908	YES	C1	31	-175.08	206	0.513
CCis32	PROBABLY	C33	67	-164.51	1486	0.855

Appendix Table 5.21: Selected conformational regions and calculation of the associated intramolecular energy penalty of succinic acid. The values of the polar hydrogen torsion angles highlighted in yellow were added manually, and those conformations that differ from the CG-generated ones only in the values of these dihedrals are indicated using a '_N' notation. The conformational regions highlighted in blue were taken to the search stage.

CG conformation number	Φ_1/°	Φ_2/°	Φ_3/°	Φ_4/°	Φ_5/°	ΔE_{intra}/kJ·mol^{-1}
1	180	0	180	0	180	1.92
1_2	180	0	180	0	0	25.42
1_3	0	0	180	0	0	48.45
2	180	0	62	0	180	5.93
2_2	180	0	62	0	0	32.49
2_3	0	0	62	0	0	59.86
4	180	0	180	180	180	9.12
4_2	180	0	180	180	0	33.92
4_3	0	0	180	180	0	69.42
5	180	0	180	270	180	10.16
5_2	180	0	180	270	0	35.20
5_3	0	0	180	270	0	62.64
7	180	178	295	359	180	9.13
7_2	180	178	295	359	0	42.33
7_3	0	178	295	359	0	69.74
8	180	178	295	178	180	9.11
8_2	180	359	295	178	0	34.65
8_3	0	359	295	178	0	69.71
9	180	269	63	1	180	24.27
9_2	180	269	63	1	0	54.45
9_3	0	269	63	1	0	84.13
11	180	87	64	2	180	21.57
11_2	180	87	64	2	0	21.48
11_3	0	87	64	2	0	46.21
13	180	180	180	180	180	15.97
13_2	180	180	180	180	0	46.78
13_3	0	180	180	180	0	76.27
14	180	180	180	90	180	16.84
14_2	180	180	180	90	0	46.45
14_3	0	180	180	90	0	76.23
16	180	270	180	270	180	18.74
16_2	180	270	180	270	0	45.49
16_3	0	270	180	270	0	78.39
17	180	270	180	90	180	18.37
17_2	180	270	180	90	0	44.19
17_3	0	270	180	90	0	69.75
23	180	178	296	150	180	13.24
23_2	180	178	296	150	0	42.45
23_3	0	178	296	150	0	72.68
25	180	90	300	150	180	20.31
25_2	180	90	300	150	0	48.85
25_3	0	90	300	150	0	73.80
27	180	89	62	211	180	21.63
27_2	180	89	62	211	0	32.05
27_3	0	89	62	211	0	80.92
29	180	90	60	90	180	19.46
29_2	180	90	60	90	0	44.81
29_3	0	90	60	90	0	64.30
30	180	90	300	90	180	20.08
30_2	180	90	300	90	0	50.28
30_3	0	90	300	90	0	82.27
33	180	89	61	299	180	21.79
33_2	180	89	61	299	0	47.42

33_3	0	89	61	299	0	75.11
34	180	299	61	89	180	21.79
34_2	180	299	61	89	0	50.13
34_3	0	299	61	89	0	75.11
37	180	30	90	120	180	13.25
37_2	180	30	90	120	0	39.52
37_3	0	30	90	120	0	59.04
39	180	330	150	60	180	16.56
39_2	180	330	150	60	0	41.39
39_3	0	330	150	60	0	64.03
41	180	240	150	330	180	15.88
41_2	180	240	150	330	0	43.45
41_3	0	240	150	330	0	72.22
43	180	30	120	180	180	20.02
43_2	180	30	120	180	0	44.70
43_3	0	30	120	180	0	86.30
45	180	30	120	270	180	24.10
45_2	180	30	120	270	0	50.61
45_3	0	30	120	270	0	78.78

Chapter 6: Molecular flexibility in crystal structures

6.1 Introduction

The previous chapter was focused on the treatment of molecular flexibility in crystal structure searches. In this chapter the importance of treating flexibility in the final refinement stage of CSP studies (see Chapter 2.4.2 for background), a related but different problem, is assessed.

Upon crystallisation, the conformation of a flexible molecule always undergoes some adjustment from its closest isolated-molecule local minimum in conformational energy to improve the intermolecular interactions, U_{inter} (see Chapter 2.3.1.1 for a detailed discussion).[1, 2] These improvements compensate for the increase in intramolecular energy (ΔE_{intra}) associated with conformational adjustment.[1] Hence, the accurate modelling of molecular flexibility is fundamental to guarantee a full coverage of the potential energy surface in CSP.[3] However, while crystal structure searches aim only to locate all possible local minima in lattice energy (E_{latt}), the final refinement has the purpose of finding the most accurate geometry and energy for each local minimum.[4-7] Thus in a search only some very flexible torsion angles need to be treated as independent conformational degrees of freedom (CDFs).[3, 6-10] Nonetheless, this approximation is too gross for the final refinement stage, where accuracy is of utmost importance: even minor changes in CDFs of limited flexibility, with a minor influence on molecular shape, can affect the balance between intra- and intermolecular interactions, in particular if they define the position of donors and acceptors in hydrogen bonded crystal structures.[3, 6, 8]

If the final refinement is performed with modelling approaches based on optimising the whole crystal structure (the Ψ_{crys} method, see Chapter 2.3.2), then all the degrees of freedom are minimised, without any selection required.[3, 11] On the other hand, Ψ_{mol} approaches (see Chapter 2.3.1) that calculate the wave-function of each molecular conformation require a selection of the CDFs to be treated as independent.[3, 6, 8, 11] The remaining CDFs are considered not affected by intermolecular interactions.[6,12] CrystalOptimizer[13] is an example of E_{latt} minimisation algorithm utilising the Ψ_{mol} approach; its functioning in described in Chapter 2.4.2.2. Its computational efficiency is due to the use of local approximate models (LAMs, see Chapter 2.4.1.2 for details), and this allows the optimisation of several independent CDFs, in contrast with earlier applications of the Ψ_{mol} approach where only a few flexible torsion angles around acyclic bonds, similar to those considered in the search,[6, 8] could be explicitly optimised.[6, 13]

Ideally, all CDFs should be considered as independent (*i.e.* a fully atomistic optimisation),[6, 13] similarly to the Ψ_{crys} approach. However, the computational cost increases with the number of CDFs treated as independent,[6, 13] and fully atomistic optimisations are often unfeasibly expensive, in particular for large and flexible molecules. Hence the independent CDFs are generally limited to acyclic torsion angles (including methyl rotations that are ignored in the searches, see Chapter 4.2.2) and some bond-angles that are deemed to be affected by the solid-state environment; bond-lengths are usually treated as dependent CDFs (*i.e.* they take values that minimise the conformational energy of the isolated molecule for a set of values of the independent CDFs) [4, 6, 8, 13, 14]

The importance of selecting a set of degrees of freedom that combines accuracy and cost efficiency is fundamental when the Ψ_{mol} method is used.[6, 8, 13] However, there is no well-defined approach to perform this selection, and the user has to choose the independent CDFs for the final refinement stage of CSP. This choice is generally based on experience and/or chemical intuition, or on analysing the isolated-molecule behaviour of the CDFs, although in some cases fully-atomistic calculations on a set of crystal structures are performed to verify which torsion and bond-angles vary the most (this method was utilised in Chapter 4.2.3.1).[4, 8, 15-18] The lack of clear rules for selecting the independent CDFs for the final refinement stage of CSP can be problematic. The set of independent CDFs could be sub-optimal and lead to inaccurate energies and geometries, possibly missing experimental forms, or to unnecessarily high computational costs.

In this chapter, three criteria for selecting the CDFs that can be the most affected by the crystalline environment, and that should be treated as independent in the final refinement stage of a CSP study, were tested on sets of 20 computer-generated crystal structures (some matching known experimental forms) of five flexible molecules. The results from using these three criteria were compared in terms of absolute and relative energies, reproduction of the experimental forms and computational cost with those obtained optimising the same crystal structures with the more accurate but expensive treatment of all torsion and bond-angles as independent CDFs. The first criterion was purely based on chemical intuition and consisted in the independent treatment in the optimisations of only the acyclic torsion angles. The second criterion was based on a set of rules derived from combining chemical intuition with experience from modelling crystal structures of small molecules; these rules are embedded in the AUTODOF[19] programme and lead to the selection of the several torsion angles and a few bond-angles as independent CDFs.[14, 20] Finally, the third criterion integrated the independent CDFs from AUTODOF with some intuitively rigid cyclic torsion angles, as well as some bond-angles,

which were found to vary significantly in the optimisations where they were all treated as independent.

6.2 Methods

6.2.1 Choice of molecules and sample crystal structures

The five molecules shown in Figure 6.1 were used to perform the analyses in this chapter: molecules XXIII and XXVI (whose CSP study is outlined in Chapter 3) from the sixth Blind Test,[4] the two tautomers of mebendazole[21] (see Chapter 4 for details of the CSP study), and naproxen, whose CSP study was undertaken in a work aimed at contrasting its racemic and enantiopure forms.[18] Naproxen was chosen because of the presence of a rare non-planar naphthalene group in the known enantiopure crystal structure, which exemplifies how the solid-state environment can sometimes distort CDFs that are intuitively rigid.[18]

For each molecule, a set of 20 crystal structures generated in the respective CSP studies were used for this test. For molecules XXVI and XXIII, the 20 lowest energy crystal structures submitted in the first list of predictions for the sixth Blind Test (*i.e.* the 20 most stable in E_{latt} after the final optimisations with CrystalOptimizer, without considering polarisation, temperature effects and diversity, see Chapter 3 and the publication[4] for details) were considered; structure 1600 of molecule XXVI matches the only known experimental form (CSD[22] refcode XAFQIH),[4] while structure A70 of molecule XXIII matches experimental form B (CSD refcode XAFPAY01).[4] For both tautomers of mebendazole, the 20-lowest energy crystal structures predicted in the CSP study (see Chapter 4 for details) were considered; structure A788 is a match to experimental form A (CSD refcode TUXPEJ),[23] while structure C5 matches experimental form C (CSD refcode YULGIW).[24] Finally for naproxen the 20-lowest energy crystal structures in the crystal energy landscape of the published CSP study were considered (the Z'=2 version of the global minimum was ignored);[18] this set includes structure CO_1, which matches the known racemic form (CSD refcode PAPTUX is the best experimental determination),[18] and structure af92, a match to the enantiopure form (CSD refcode COYRUD11 is the highest-quality determination).[25] The starting points for all CrystalOptimizer minimisations were the same from which the final optimisations in the original CSP studies had been started, in order to avoid to begin from structures already completely converged. For molecule XXVI[4] and the two tautomers of mebendazole[21] the starting points were the output of an intermediate optimisation with a single-iteration of CrystalOptimizer, while for molecule XXIII[4] and naproxen[18] of a rigid-body minimisation of U_{inter} with DMACRYS.[26]

6.2.2 Benchmark calculations with all torsion and bond-angles as independent CDFs

In order to perform an accurate analysis, a benchmark was needed against which the effectiveness of various approaches for selecting the independent CDFs could be tested in terms of absolute and relative lattice energies, computational cost and reproduction of the experimentally known forms. Data from original CSP studies were not considered appropriate for this test because the independent CDFs for the CrystalOptimizer optimisations had not been selected with consistent criteria. Hence, in order to produce an accurate and feasible benchmark, all the 100 crystal structures of the five molecules were optimised with CrystalOptimizer at the PBE0 6-31G(d,p) level of theory for both intramolecular interactions and charge density calculations treating all torsion and bond-angles (from now on the $CDFs_{all}$) as independent CDFs. Bond-lengths were treated as dependent CDFs: their values are generally not affected by the solid-state environment because their distortion requires very high energies.[13]

6.2.3 Treating only the acyclic torsion angles as independent CDFs

The first selection criterion that was tested was the treatment of only the intuitively flexible torsion angles around acyclic bonds as independent CDFs (from now on the $CDFs_{torsion}$); this corresponds with what was achievable in early optimisations with the Ψ_{mol} method.[6, 13] Acyclic torsion angles are also generally used to define molecular flexibility in crystal structure searches,[6, 8, 9] with the exception of methyl rotations. Methyl rotations are ignored in searches because of their negligible effect on ΔE_{intra} and molecular shape, but refining their position can be important when more accurate optimisations are performed.[3, 8] The $CDFs_{torsion}$ are shown in Figure 6.1 for each molecule, and their full definition is given in Appendix Figure 6.1 and Appendix Tables 6.1-6.5.

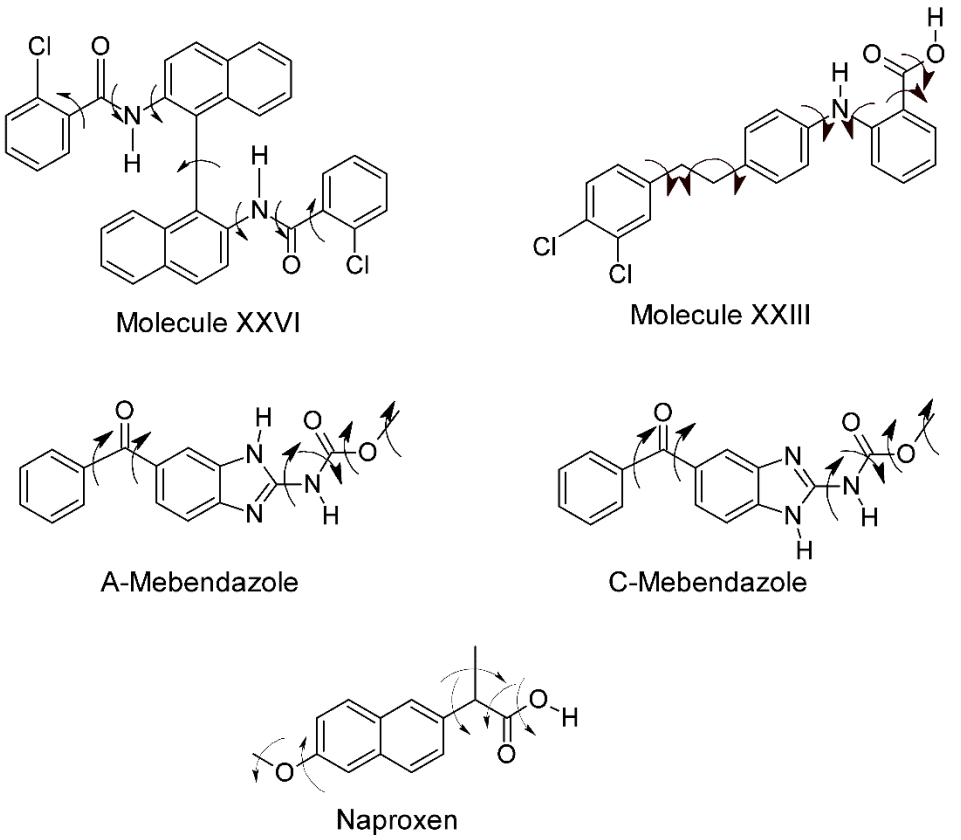


Figure 6.1: Chemical diagrams of the five molecules considered for this test. The black arrows indicate the $CDFs_{torsion}$.

Note that the $CDFs_{torsion}$ of molecule XXVI correspond to the independent CDFs used in the final refinement stage of the original CSP study (see Chapter 3).[4]

All 100 crystal structures were optimised with CrystalOptimizer at the PBE0 6-31G(d,p) level of theory for both intramolecular interactions and charge density calculations, as a function of the $CDFs_{torsion}$.

6.2.4 Treating the torsion and bond-angles selected with the AUTODOF programme as independent CDFs

In a recent paper Nyman and Day proposed a set of rules to select the independent CDFs, which they used in a study on the importance of temperature effects on the properties of different polymorphs of the same molecules.[20] These are:

- "Covalent bond lengths are optimised without considering packing forces;
- All angles and dihedrals containing a polar hydrogen atom (–OH, –NH, –SH) are optimised under the influence of packing forces;
- All exocyclic bonds are considered rotatable and are optimised under the influence of crystal packing forces;
- Dihedrals and angles in 3- and 4-membered rings are optimised without considering packing forces;

- Dihedrals and angles in 5- and 6-membered rings consisting of 3-coordinated carbon atoms and nitrogen in any combination are unaffected by packing forces, except dihedrals and angles that contain a polar hydrogen atom;
- Dihedrals and angles in 5-membered rings containing 2-coordinated sulphur or oxygen atoms bonded to two 3-coordinated carbon atoms are optimised without considering packing forces;
- Any remaining dihedrals are optimised with respect to packing forces."

Note that the CDFs optimised "with respect to packing forces" are independent and those optimised "without considering packing forces" are dependent. These are the only criteria to select the independent CDFs in the final refinement stage of CSP studies with the Ψ_{mol} approach that are clearly codified in the literature, although they have never been tested against possible alternatives. These rules are based on a combination of chemical intuition, as acyclic torsion angles are treated as independent CDFs and torsion angles in rigid rings as dependent, and experience derived from modelling small crystal structures, as all torsion and bond-angles that contain polar hydrogen atoms, whose position is key to accurately describe hydrogen bond geometries,[14] are explicitly optimised. AUTODOF[19, 20] is a Python[27] programme that can be used to automatically apply these rules to a given molecule. The independent CDFs chosen by AUTODOF (from now on the $CDFs_{AUTODOF}$) for each of the five molecules are shown in Figure 6.2, and their full definition is given in Appendix Figure 6.1 and Appendix Tables 6.1-6.5.

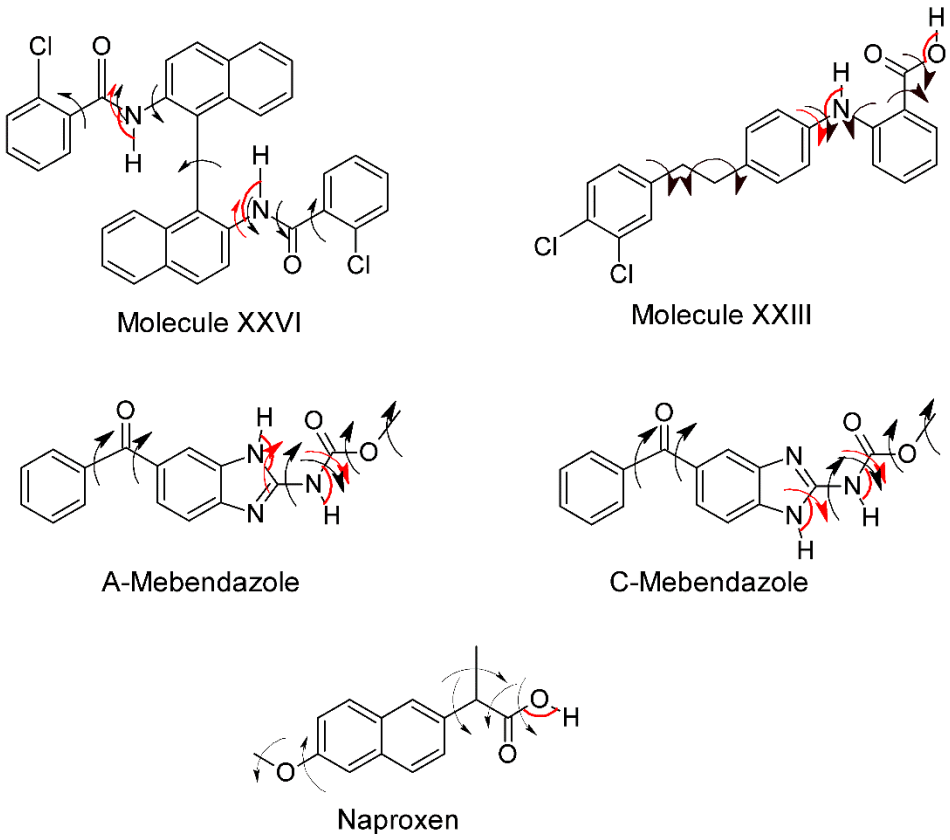


Figure 6.2: Chemical diagram of the five molecules considered for this test, showing the torsion angles and bond-angles treated as independent in the CrystalOptimizer optimisation with the $CDFs_{AUTODOF}$. Black arrows indicate the torsion angles that are also present in the $CDFs_{torsion}$ and that are selected according to chemical intuition, while red arrows and red arcs indicate the torsion and bond-angles added by AUTODOF as they include polar hydrogen atoms.

Comparing the independent CDFs in Figure 6.1 and Figure 6.2, it is clear that all the torsion angles in the $CDF_{torsion}$ are present in the $CDFs_{AUTODOF}$, which indicates how AUTODOF chooses the independent CDFs in line with what suggested by chemical intuition. However, there are a few additions. In particular, the $CDFs_{AUTODOF}$ include a handful of bond-angles around polar hydrogen atoms, as well as some extra torsion angles, most of which have the same central atoms as others intuitively flexible (this is indicated in Figure 6.2 by double arrows around the same bonds), with the exception of one torsion angle around a cyclic bond in the two tautomers of mebendazole.

All 100 crystal structures were optimised with CrystalOptimizer at the PBE0 6-31G(d,p) level of theory for both intramolecular interactions and charge density calculations, as a function of $CDFs_{AUTODOF}$.

6.2.5 Integrating the CDFs$_{AUTODOF}$ with extra torsion and bond-angles.

6.2.5.1 Determining which CDFs can vary to optimise intermolecular interactions

The results of the benchmark optimisations with the CDFs$_{all}$ were analysed to determine which CDFs had their values significantly distorted to improve the balance between intra- and intermolecular interactions. In particular, for each molecule all the torsion angles that varied by at least 5° in one or more crystal structures, and all the bond-angles that varied at least by 1° were recorded. Bond-angles including non-polar hydrogen atoms were ignored, and only one instance of torsion angles with the same central atoms was considered. The results of this analysis are shown in Figure 6.3.

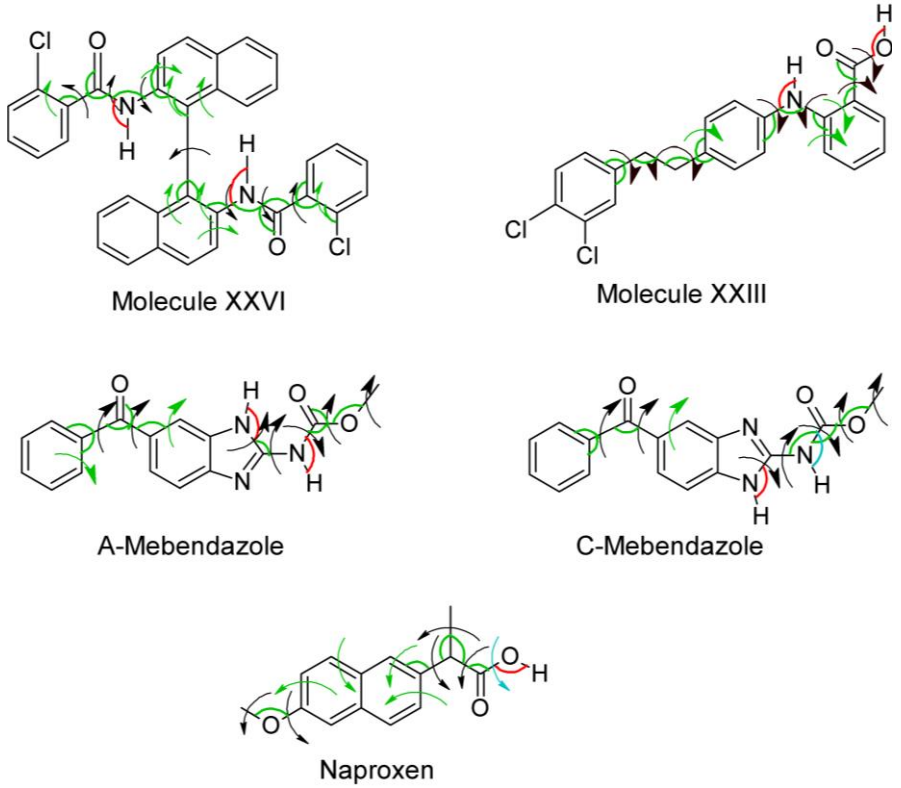


Figure 6.3: Chemical diagrams of the five molecules considered for this test. The torsion angles indicated by black arrows and the bond-angles indicated by red arcs were present in the CDFs$_{AUTODOF}$ and varied significantly in at least one crystal structure optimised with CrystalOptimizer with the CDFs$_{all}$. Green arrows and arcs indicate torsion and bond-angles not included in the CDFs$_{AUTODOF}$ but that varied significantly in at least one crystal structure, while those in turquoise are included in the CDFs$_{AUTODOF}$ but did not vary significantly in any crystal structure.

6.2.5.2 Comparison with CDFs$_{AUTODOF}$ and integration with extra CDFs

From Figure 6.3 it is clear that most torsion and bond-angles in the CDFs$_{AUTODOF}$ do respond significantly to packing forces in the sample of crystal structures that was considered. The only exceptions (indicated in turquoise in Figure 6.3) are the torsion angle in the carboxyl group of naproxen and the bond-angle around the amine group in

the C-tautomer of mebendazole (although another bond-angle with the same central N atom is present in Figure 6.3). With the exception of that bond-angle, all the other CDFs that are included in the $CDFs_{AUTODOF}$ but that are absent in the $CDFs_{torsion}$ varied significantly in the benchmark optimisations with the $CDFs_{all}$. This shows how for these molecules the combination of chemical intuition and experience on modelling crystal structures of small molecules forms a solid base for selecting the independent CDFs. However, the $CDFs_{AUTODOF}$ do not include several intuitively rigid torsion and bond-angles (indicated in green in Figure 6.3) that can undergo large distortions when optimised under the influence of packing forces. Analysing Figure 6.3, they can be grouped into three main categories:

1) Torsion angles in 5 or 6-membered aromatic rings or 5-membered rings consisting of 3-coordinated carbon atoms and nitrogen in any combination, in which one of the central atoms is bonded to a cyclic or acyclic substituent containing more than one non-H atom.
2) Bond-angles between atoms in 5 or 6-membered aromatic rings or 5-membered rings consisting of 3-coordinated carbon atoms and nitrogen in any combination and an acyclic substituent containing more than one non-H atom.
3) Bond-angles between atoms in two consecutive acyclic flexible bonds.

CDFs belonging to these three categories were manually added to the set of independent CDFs selected by AUTODOF, forming the $CDFs_{AUTODOF+}$ that are shown in Figure 6.4. Only one torsion angle around the same bond was added to the $CDFs_{AUTODOF}$, and only one bond-angle per central atom; the full definition of the $CDFs_{AUTODOF+}$ is given in Appendix Figure 6.1 and Appendix Tables 6.1-6.5.

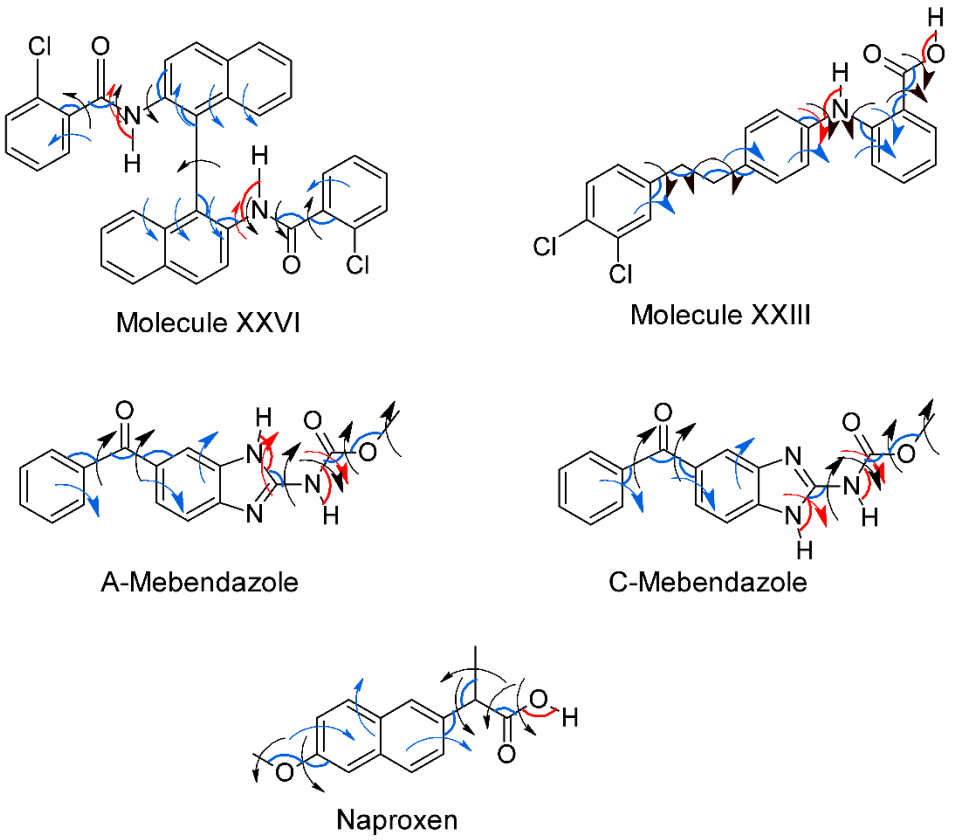


Figure 6.4: Chemical diagrams of the five molecules considered for this test, showing the torsion and bond-angles treated as independent in the CrystalOptimizer optimisation with the $CDFs_{AUTODOF+}$. The torsion angles indicated by red and black arrows and bond-angles indicate by red arcs were present in the $CDFs_{AUTODOF}$. Blue arrows and arcs indicate the torsion and bond-angles that were manually added as they belong to one of the three categories of intuitively rigid CDFs that varied significantly in the CrystalOptimizer minimisations performed with the $CDFs_{all}$.

All 100 crystal structures were optimised with CrystalOptimizer at the PBE0 6-31G(d,p) level of theory for both intramolecular interactions and charge density calculations as a function of the $CDFs_{AUTODOF+}$.

6.3 Results and Discussion

6.3.1 Absolute lattice energies

The quality of the various sets of independent CDFs, and of the approaches for selecting them, was firstly assessed by their ability to reproduce the E_{latt} values obtained in the benchmark optimisations with the $CDFs_{all}$. Comparisons between the lattice energies obtained with each selection approach and in the benchmark optimisations for the 20 crystal structures of the five molecules are shown in Figure 6.5 and Figure 6.6; these data are also summarised in Table 6.1 and shown more extensively in Appendix Tables 6.6-6.10.

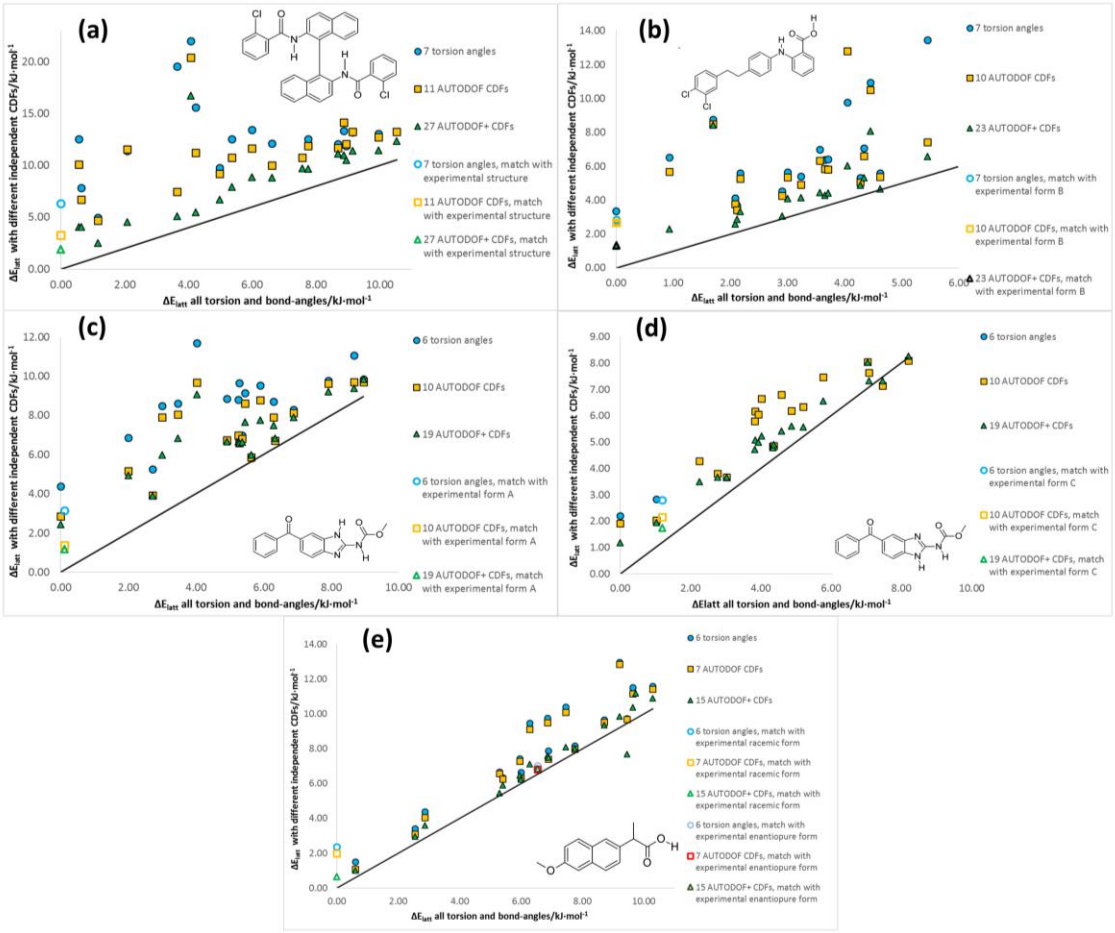


Figure 6.5: Comparison between the lattice energies obtained after full optimisations with CrystalOptimizer with the various sets of independent CDFs and those with the CDFsall for the 20 crystal structures of (a) molecule XXVI (b) molecule XXIII (c) the A-tautomer of mebendazole (d) the C-tautomer of mebendazole and (e) naproxen. For each molecule, the lattice energies are plotted as the difference with the Elatt value of the global minimum in the benchmark optimisations with the CDFsall. The black lines indicate the lattice energies obtained in the optimisations with the CDFsall, and the structures matching the experimentally-characterised forms are indicated. See Appendix Tables 6.6-6.10 for more detailed results. The presence of crystal structures with energies lower than those obtained with the CDFsall is probably due to the tolerances in the convergence criteria of the optimisations.

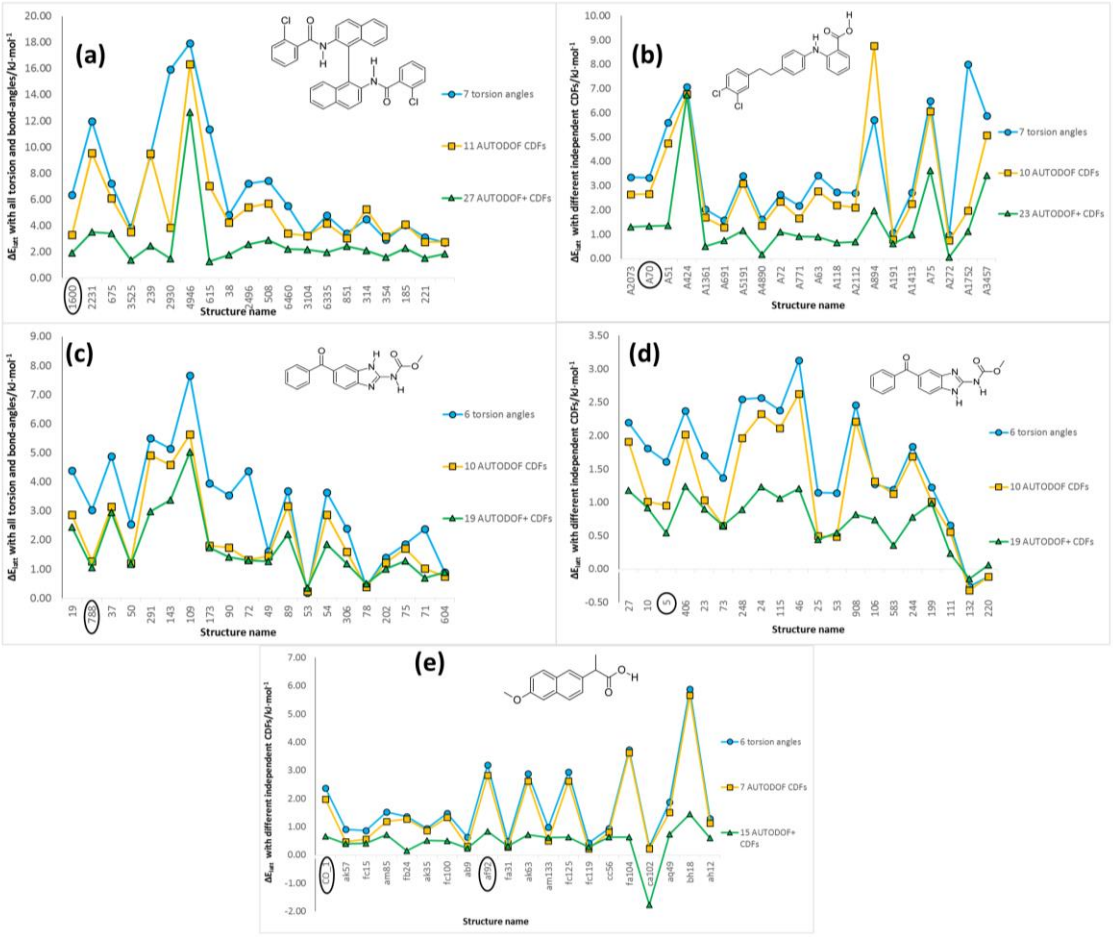


Figure 6.6: Comparison between the lattice energies obtained after full optimisations with CrystalOptimizer with the various sets of independent CDFs and those with the $CDFs_{all}$ for the 20 crystal structures of (a) molecule XXVI (b) molecule XXIII (c) the A-tautomer of mebendazole (d) the C-tautomer of mebendazole and (e) naproxen. For each crystal structure, the energy is shown as the difference with that obtained in the benchmark optimisations with the $CDFs_{all}$ for the same structure. The crystal structures that match experimentally-characterised forms are circled. See Appendix Tables 6.6-6.10 for more detailed results. The presence of crystal structures with energies lower than those obtained with the $CDFs_{all}$ is probably due to the tolerances in the convergence criteria of the optimisations.

Table 6.1: Summary of the results shown in Figure 6.5, Figure 6.6 and Appendix Tables 6.6-6.10, showing for each molecule the number of torsion and bond-angles treated as independent CDFs, as well the average relative and absolute deviations in E_{latt} compared with the optimisations with the $CDFs_{all}$.

	CDF_{all}	$CDF_{torsion}$		$CDF_{AUTODOF}$		$CDF_{AUTODOF+}$	
Molecule	**Torsion, bond-angles**	**Torsion bond-angles**	**Mean, mean absolute ΔE_{latt}/kJ·mol^{-1}**	**Torsion, bond-angles**	**Mean, mean absolute ΔE_{latt}/kJ·mol^{-1}**	**Torsion, bond-angles**	**Mean, mean absolute ΔE_{latt}/kJ·mol^{-1}**
XXVI	59, 60	7, 0	6.87, 6.87	9, 2	5.3, 5.3	17, 10	2.96, 2.96
XXIII	40, 41	7, 0	3.62, 3.62	8, 2	3.05, 3.05	13, 10	1.47, 1.43
Mebendazole A	32, 33	6, 0	3.17, 3.17	8, 2	2.14, 2.14	11, 8	1.73, 1.73
Mebendazole C	32, 33	6, 0	1.61, 1.65	8, 2	1.25, 1.30	11, 8	0.73, 0.75
Naproxen	28, 29	6, 0	1.60, 1.60	6, 1	1.34, 1.34	9, 6	0.42, 0.60
		MEAN	3.40, 3.41	**MEAN**	2.65, 2.66	**MEAN**	1.47, 1.51

Results in Figure 6.5, Figure 6.6 and Appendix Tables 6.6-6.10 clearly show that the energies calculated in the benchmark optimisations with the $CDFs_{all}$ were lower than for any alternative, with only a few exceptions that are probably due to the tolerances that are used by CrystalOptimizer to determine the convergence of a minimisation.[13] The lattice energy differences for treating various sets of CDFs as independent varied drastically for different crystal structures of the same molecule. However, this does not seem to be associated with specific structural or packing characteristic, but it can be ascribed to subtle and unpredictable balances between intra- and intermolecular interactions.

Looking at Table 6.1, it is clear that the optimisations with the $CDFs_{torsion}$ produced energies of poor quality compared to the benchmark calculations, with an average increase in E_{latt} of ~3.4 $kJ{\cdot}mol^{-1}$ and with a maximum increase of ~6.9 $kJ{\cdot}mol^{-1}$ for molecule XXVI. Using the $CDFs_{AUTODOF}$ slightly improved the energies, as the average increase in E_{latt} was reduced to ~2.7 $kJ{\cdot}mol^{-1}$ and to ~5.3 $kJ{\cdot}mol^{-1}$ for XXVI. Finally, the addition of some intuitively rigid torsion and bond-angles within the $CDFs_{AUTODOF+}$ decreased the discrepancies to an average of ~1.5 $kJ{\cdot}mol^{-1}$, and to ~3 $kJ{\cdot}mol^{-1}$ for molecule XXVI. However, none of the attempted selection criteria could exactly reproduce the energies obtained with the $CDFs_{all}$.

These results show that it is important to optimise some intuitively rigid CDFs to obtain optimal E_{latt} values. It also appears that even CDFs with a small influence on molecular shape, like bond-angles or cyclic torsion angles, can affect the overall balance of intra- and intermolecular interactions. Furthermore, this limited analysis on five flexible molecules seems to suggest that neglecting intuitively rigid CDFs becomes more detrimental as molecules get larger. As shown in Table 6.1, the number of torsion and bond-angles in the $CDFs_{torsion}$, which are purely selected from intuition, and in the $CDFs_{AUTODOF}$, which are chosen from a combination of intuition and experience derived from modelling small molecules, is similar for the five molecules. However, the difference becomes larger when some additional intuitively rigid torsion and bond-angles are added in the $CDFs_{AUTODOF+}$, which suggests that large and flexible molecules tend to have a higher proportion of CDFs that are ignored when selection criteria based exclusively (or mostly) on chemical intuition are adopted. Furthermore, Table 6.1 shows that the increase of E_{latt} compared to the benchmark calculations with the $CDFs_{all}$ is broadly related to the number of CDFs that are not treated as independent. Hence as molecules get larger obtaining the best possible energies upon optimisations with the Ψ_{mol} model becomes more problematic: the inclusion of many more independent CDFs is required, but even the most comprehensive selection approach still produces results that are of worse quality than for smaller molecules.

Although explicitly optimising more CDFs produces lower E_{latt} values, it also shifts the balance between inter- and intramolecular interactions. Table 6.2 shows the average values of the intramolecular energy penalty (ΔE_{intra}) obtained for each molecule with the various sets of independent CDFs.

Table 6.2: Average ΔE_{intra} for the 20 crystal structures of each molecule when different criteria for selecting the independent CDFs in the CrystalOptimizer optimisations were adopted.

	Mean ΔE_{intra} for the optimisations with different $CDFs_{all}$/ kJ·mol^{-1}			
Molecule	**$CDFs_{all}$**	**$CDFs_{torsion}$**	**$CDFs_{all}$**	**$CDFs_{AUTODOF+}$**
XXVI	14.29	11.02	10.81	12.16
XXIII	9.27	7.72	7.86	8.27
Mebendazole A	3.96	2.78	2.84	3.18
Mebendazole C	4.30	2.69	2.98	3.15
Naproxen	2.50	1.04	1.25	2.05
MEAN	6.86	5.05	5.15	5.76

These data clearly show that crystal structures optimised with more independent CDFs have higher average values for ΔE_{intra}. This implies that the explicit optimisation of more CDFs causes a larger degree of conformational adjustment than when they are considered as dependent.[13] These distortions increase ΔE_{intra}, but this is more than compensated by better intermolecular interactions (with an associated decrease in U_{inter}),[1] like improved electrostatics or better dispersion interactions due to denser packings.[2, 3] Indeed, crystal structures optimised with more independent CDFs were denser, as shown in Appendix Tables 6.11-6.15.

In summary, this analysis shows how choosing the independent CDFs with criteria derived exclusively (*i.e.* the $CDFs_{torsion}$) or mostly (*i.e.* the $CDFs_{AUTODOF}$) from chemical intuition leads to a sub-optimal balance of intra- and intermolecular interactions, and this problem becomes more relevant for larger molecules, as they have a larger proportion of CDFs that are ignored by these approaches. Integrating the sets of independent CDFs derived mostly from chemical intuition with some intuitively rigid CDFs that underwent significant distortions in the benchmark calculations with the $CDFs_{all}$ leads to smaller E_{latt} deviations. Hence, the $CDFs_{AUTODOF+}$ appear to be the most adequate set of independent CDFs among those that were tested, although no approach was accurate enough to exactly reproduce the E_{latt} values obtained in the benchmark calculations. This becomes more problematic as a function of molecular size. This indicates that separating CDFs into dependent and independent is an unideal approximation and that a CSP methodology for refining crystal structures should optimise all molecular degrees of freedom.[3, 6]

6.3.2 Reproduction and ranking of the experimentally known crystal structures

As already mentioned in Chapter 6.2.1, the sample of 100 crystal structures contained six matches to experimentally-characterised forms, which are indicated in Figure 6.6 and Appendix Tables 6.6-6.10. Table 6.3 summarises how the crystal structures matching the experimental forms were ranked for each molecule in terms of E_{latt}, as well as the 15-molecule root mean square deviation ($RMSD_{15}$) of the overlays with the experimental crystal structures and the $RMSD_1$ of the overlays with the experimental conformations.

Table 6.3: For each molecule, the energy ranking in terms of E_{latt} of the crystal structures matching the experimental forms, the quality of the reproduction of the experimental crystal structures ($RMSD_{15}$) and the quality of the reproduction of the experimental conformations ($RMSD_1$) when different criteria for selecting the independent CDFs in the CrystalOptimizer optimisations were adopted.

	Optimisations with the CDF_{all}		Optimisations with the $CDF_{torsion}$		Optimisations with $CDF_{AUTODOF}$		Optimisations with $CDF_{AUTODOF+}$	
Molecule (CSD refcode)	**Rank**	**$RMSD_{15}$, $RMSD_1$/Å**	**Rank**	**$RMSD_{15}$, $RMSD_1$/Å**	**Rank**	**$RMSD_{15}$, $RMSD_1$/Å**	**Rank**	**$RMSD_{15}$, $RMSD_1$/Å**
XXVI (XAQHIF)	1	0.292, 0.115	2	0.259, 0.125	1	0.344, 0.138	1	0.306, 0.108
XXIII form B (XAFPAY01)	2	0.303, 0.133	1	0.364, 0.155	2	0.362, 0.152	2	0.312, 0.133
Mebendazole A (TUXPEJ)	2	0.333, 0.140	1	0.301, 0.133	1	0.301, 0.122	1	0.311, 0.124
Mebendazole C (YULGIW)	3	0.265, 0.041	2	0.287, 0.071	3	0.284, 0.068	2	0.259, 0.042
Naproxen (PAPTUX, racemic)	1	0.711, 0.165	2	0.973, 0.211	2	0.974, 0.206	1	0.780, 0.167
Naproxen (COYRUD11, enantiopure)	9	0.275, 0.048	12	0.458, 0.119	12	0.478, 0.129	10	0.300, 0.057
	MEAN	0.363, 0.107	**MEAN**	0.440, 0.136	**MEAN**	0.457, 0.136	**MEAN**	0.378, 0.105

No method could accurately reproduce all the experimental forms; the worst reproduction was for the racemic form of naproxen, and this is probably due to a poor starting point for the optimisation, as the $RMSD_{15}$ value between PAPTUX and the fully minimised computer-generated crystal structure was also high in the original CSP study.[18] It is impossible to determine what differentiates the experimental forms from the crystal structures optimised with the different sets of independent CDFs, since the structural differences cannot be spotted visually. However, a perfect reproduction of the experimental forms cannot be expected: the E_{latt} optimisations are in fact carried out at 0 K, but all the experimental crystal structures were characterised at room temperature, with the exception of YULGIW that was determined at 150 K. Temperature differences modify crystal structures because of thermal expansion,[28, 29] thus even an extremely accurate model cannot exactly reproduce experimental forms characterised at higher temperatures.

Even accounting for the temperature differences, these results are somewhat surprising, since the reproduction of the experimental crystal structures was not clearly

affected by the number of CDFs treated as independent in the optimisations. This differs from the outcome of the initial test of CrystalOptimizer, where better reproductions were achieved when more CDFs were explicitly optimised.[13] The only examples for which the reproduction of the experimental crystal structures was significantly affected by the set of independent CDFs were the two forms of naproxen. For its enantiopure form (COYRUD11), the best representation of both the molecular conformation and the crystal structure was obtained treating some intuitively rigid torsion angles as independent CDFs because of the unusual bend of the naphthalene ring.[18] Table 6.4 compares the values of the torsion angle between the two aromatic rings in the naphthalene group in the conformations of the enantiopure crystal structure of naproxen, COYRUD11, and of the computer-generated structure matching it, af92, after the optimisations with the different sets of independent CDFs.

Table 6.4: Value of the torsion angle between the two aromatic rings in the naphthalene group in the enantiopure crystal structure of naproxen (CSD refcode COYRUD11) and in structure af92 when different criteria for selecting the independent CDFs in the CrystalOptimizer optimisations were adopted.

Conformation	Torsion angle between the aromatic rings in the naphthalene group/°
Experimental crystal structure (COYRUD11)	174.34
Structure af92 after the optimisation with the $CDFs_{all}$	173.63
Structure af92 after the optimisation with the $CDF_{torsion}$	179.86
Structure af92 after the optimisation with the $CDF_{AUTODOF}$	179.87
Structure af92 after the optimisation with the $CDF_{AUTODOF+}$	173.14

The best reproduction of the torsion angle was achieved in the benchmark optimisation with the $CDFs_{all}$, closely followed by $CDFs_{AUTODOF+}$, where treating one torsion angle between the aromatic rings as independent was sufficient to reproduce a bent naphthalene group. The intuitive rigid treatment of the aromatic rings in the optimisations with the $CDFs_{torsion}$ and the $CDFs_{AUTODOF}$ wrongly produced planar naphthalene.

On the other hand, the poor reproduction of the racemic form with the $CDFs_{torsion}$ and $CDFs_{AUTODOF}$ was not due to this issue, since naphthalene is planar in PAPTUX,[18] but to other subtle differences that affect the balance of intra- and intermolecular interactions. For the other molecules, the differences were small and probably not significant considering the temperature effects described above; however, it is surprising how optimising only the $CDFs_{torsion}$ produced the best reproduction of the experimental forms of molecule XXVI and mebendazole form A. It is also interesting to note that the optimisations with the $CDFs_{AUTODOF+}$ reproduced very well the experimental conformations, as they have the lowest mean $RMSD_1$ values in Table 6.3, confirming that this is a reliable approach for selecting the independent CDFs.

The relative stability of the crystal structures matching the experimental forms relative to the other competitors was less affected by the set of independent CDFs than the absolute energies, suggesting that the energy differences tend to cancel out. The addition of at least some intuitively rigid torsion and bond-angles within the $CDFs_{AUTODOF+}$ was needed to place the structure matching the racemic form of naproxen as the global minimum in E_{latt}. This also improved the ranking of the enantiopure form of naproxen, which however was always found to be the most stable homochiral crystal structure (as shown in Appendix Table 6.10), as well as metastable with respect to the racemic form, consistently with experimental studies.[18] For molecule XXVI, the experience-based addition of some torsion and bond-angles containing polar hydrogen atoms in the $CDFs_{AUTODOF}$ was needed to place the experimentally known form of XXVI as the global minimum in E_{latt}, as it was ranked second in the optimisations with the purely intuition-based $CDFs_{torsion}$ (like in the original CSP study, see Chapter 3.3.2.2). On the other hand, for the A-tautomer of mebendazole one structure becomes slightly more stable than the experimentally known form when the optimisations are performed with the $CDFs_{all}$.

In summary, the differences in terms of reproduction of the experimental forms and their relative stabilities are less affected by the choice of the independent CDFs than the absolute E_{latt} values. Nonetheless, among the three sets of CDFs selected with different approaches, the $CDFs_{AUTODOF+}$ appear once again to be the most adequate, as they produce results of very similar quality to the benchmark optimisations with the $CDFs_{all}$.

6.3.3 Comparison of computational cost

Table 6.5: Average computational cost per crystal structure for performing the optimisations with the different sets of independent CDFs.

	Average cost of the optimisations performed with different CDFs /CPU hours			
Molecule	**$CDFs_{all}$**	**$CDFs_{torsion}$**	**$CDFs_{AUTODOF}$**	**$CDFs_{AUTODOF+}$**
XXVI	131	70	61	65
XXIII	46	10	10	19
Mebendazole A	39	5	6	9
Mebendazole C	60	22	28	28
Naproxen	22	5	5	5

For all molecules, the optimisations with the $CDFs_{all}$ were the most computationally expensive. As shown in Figure 6.7, the CPU cost increases with the number of independent CDFs, consistently with what had been noticed in the initial testing of CrystalOptimizer.[13]

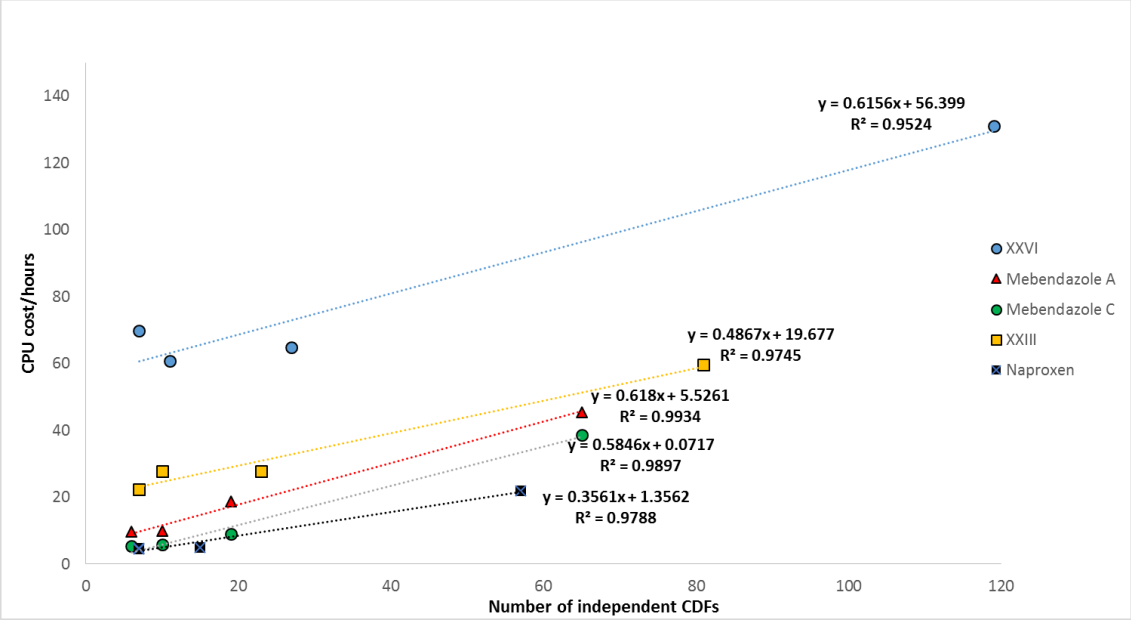


Figure 6.7: For each molecule, average CPU cost of the optimisations as a function of the number of independent CDFs.

For this limited number of data point, Figure 6.7 suggests that each independent CDF increases the computational cost by an average of 0.4-0.6 CPU hours per crystal structure. This cost increase would make the explicit optimisation of the $CDFs_{all}$ unaffordable when thousands of crystal structures need to be minimised, in particular for large and flexible molecules. On the other hand, optimising the crystal structures with the $CDFs_{AUTODOF+}$ does not significantly increase the overall CPU cost compared to the approaches based exclusively or mostly on chemical intuition. Since optimisations with the $CDFs_{AUTODOF+}$ produce better energies and geometries, it can be concluded that this selection approach provides the best balance between accuracy and computational feasibility within the Ψ_{mol} framework among those that were tested in this study, although it still includes some unideal approximations.

6.3.4 CSD validation of the findings of this analysis

The data shown so far have illustrated that the explicit optimisation of only the CDFs that are flexible according to intuition and/or experience from modelling small model molecules is not sufficient. In particular, three categories of intuitively rigid torsion and bond-angles (listed in Chapter 6.2.5.2) seem to vary significantly in the optimisations with the $CDFs_{all}$. The addition of independent CDFs belonging to these three categories in the $CDFs_{AUTODOF+}$ improves the reproduction of the energies compared to the benchmark optimisations with the $CDFs_{all}$. However, this may just be an artefact of the CrystalOptimizer energy model, as Table 6.3 shows that increasing the number of independent CDFs does not improve the reproduction of the experimental forms to the same extent to which it decrease the lattice energies (as shown in Table 6.1 and Appendix Tables 6.6-6.10). Hence, to verify whether the distortion of these three

categories of intuitively rigid CDFs is realistic, a CSD survey was performed. For each category listed in Chapter 6.2.5.2, crystal structures containing the fragments shown in Figure 6.8a-c respectively were extracted with Conquest.[30] In order to increase the confidence in the results, only organic, non-ionic crystal structures without disorder or errors and with an R-factor smaller than 5% were considered. Furthermore the search was limited to the CSD entries within the 'best R-factor'[31] list to limit the double-count of redeterminations of identical crystal structures.

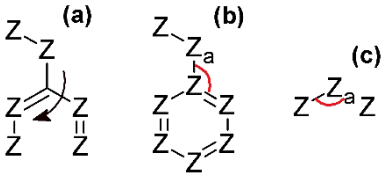


Figure 6.8: Fragments used to perform the CSD surveys with Conquest. The black arrow represents a torsion angle, red arcs bond-angles. Z indicates any non-H atom, while the 'a' subscript indicates that the atom is acyclic. Fragments (a), (b) and (c) represent categories 1-3 of intuitively rigid CDFs respectively (see section 6.2.5.2).

For the second category (Figure 6.8b) only 6-membered aromatic rings were considered, since rare 5-membered rings that are aromatic and/or consist of 3-coordinated carbon atoms and nitrogen in any combination have different idealised values for the bond-angles. The values of all torsion and bond-angles in the CSD structures containing these fragments were extracted.

Conquest searches were also performed with the fragments in Figure 6.9 to extract the values taken by the torsion and bond-angles in rigid rings when none of the central atoms is bonded to a heavy substituent, which in the optimisations with the $CDFs_{all}$ are not significantly distorted by packing forces (see Figure 6.3).

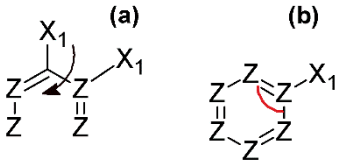


Figure 6.9: Fragments used to perform the CSD surveys with Conquest, representing torsion and bond-angles in rigid rings in the absence of heavy substituents in the central atoms. The black arrow represents a torsion angle, the red arc a bond-angle. Z indicates any non-H atom, X any atom, while the '1' subscript indicates that the atom forms only one bond.

Note that the fragments in Figure 6.9 exclude some relevant configurations, like angles with central 2-coordinated nitrogen atoms, or methyl substituents. Nevertheless, these were the broadest possible criteria using the Conquest search parameters.

Figure 6.10 and Figure 6.11 show the CSD distribution of the values of torsion angle in Figure 6.8a and Figure 6.9a respectively.

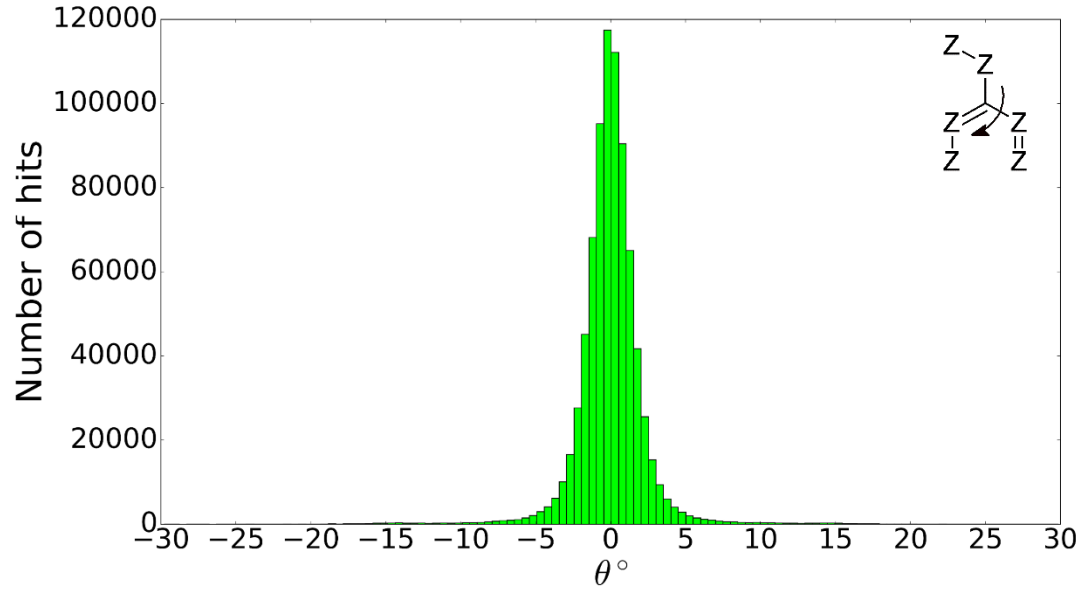


Figure 6.10: Histogram showing the CSD distribution of the values of the torsion angle in the fragment in Figure 6.8a.

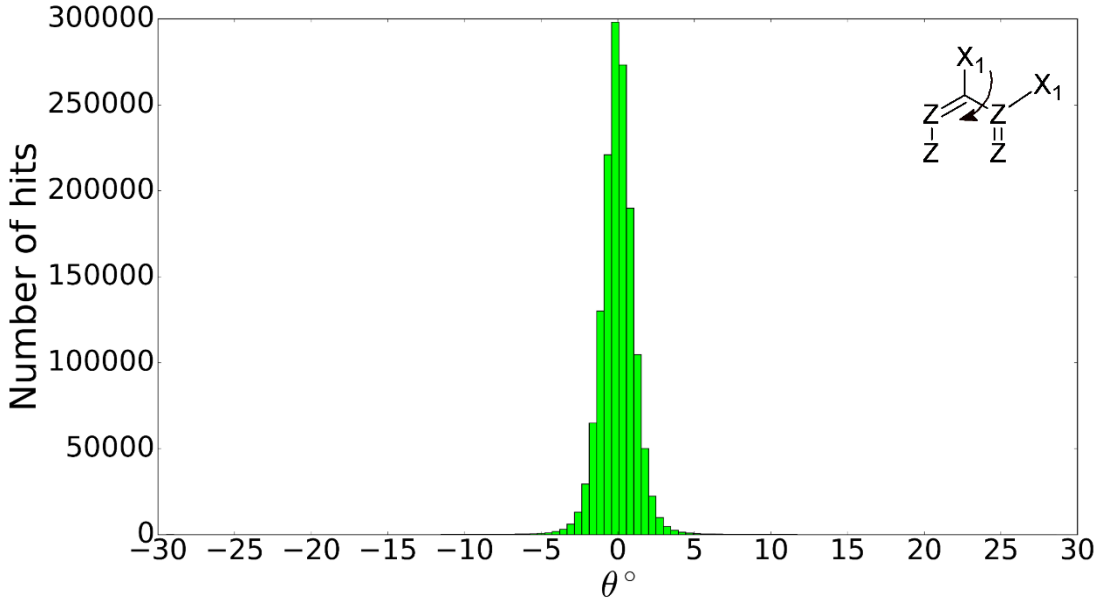


Figure 6.11: Histogram showing the CSD distribution of the values of the torsion angle in the fragment in Figure 6.9a.

Although the torsion angle values in both distributions are tightly clustered around the idealised value of 0°, there is a spread of possible values around both maxima. Although this is not sufficient to justify a flexible treatment of aromatic torsion angles in a CSP search (see Chapter 5),[10] the histograms show that some deviations from the distribution maximum are possible in the solid-state. However, the presence of a substituent with more than one non-H atom in one of the central atoms increases the flexibility range, confirming the results obtained in this chapter. In the presence of a heavy substituent in one of the central atoms, ~3% of the torsion angles are outside the -5° to 5° range, while this occurs in only ~0.4% of the entries in its absence. Figure 6.10 and Figure 6.11 show that the most realistic option is to treat all torsion angles in rigid

rings as independent variables, but they also confirm that the explicitly flexible treatment of torsion angles in rigid rings is more important when one of the central atoms is bonded to a heavy substituent.

Figure 6.12 shows the CSD distribution of the values of the bond-angle in Figure 6.8b, while Figure 6.13 shows the distribution for the bond-angle in Figure 6.9b.

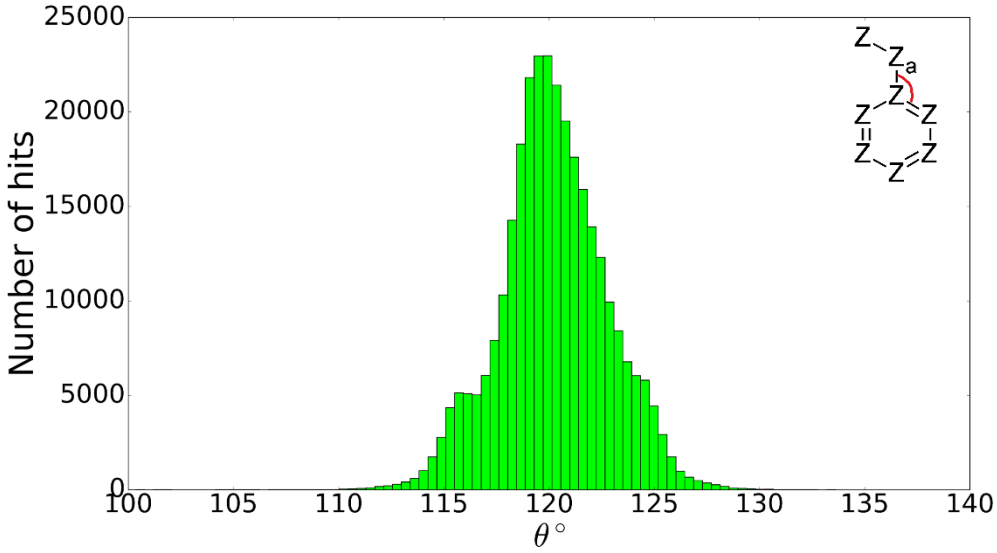


Figure 6.12: Histogram showing the CSD distribution of the values of the bond-angle in the fragment in Figure 6.8b.

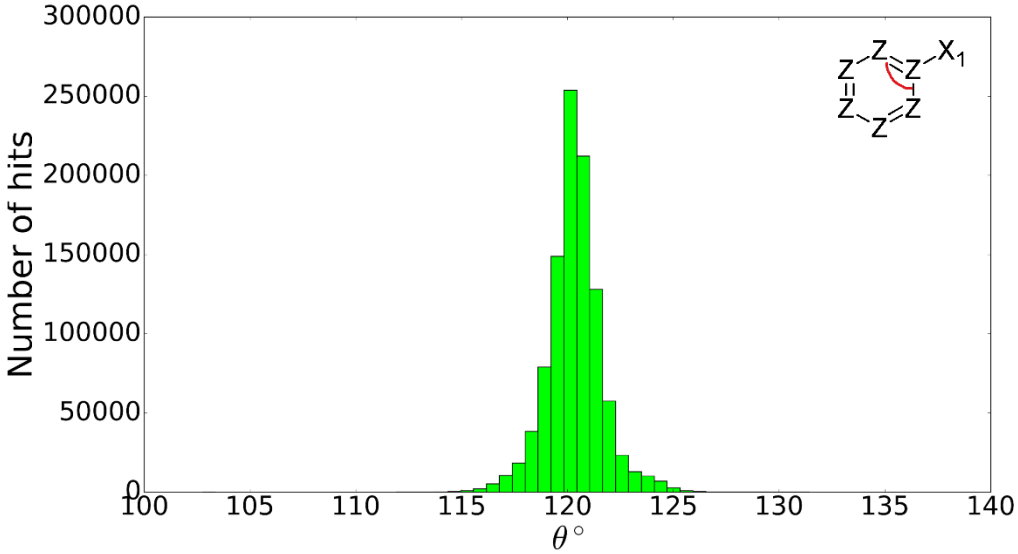


Figure 6.13: Histogram showing the CSD distribution of the values of the bond-angle in the fragment in Figure 6.9b.

Both distributions are clustered around the idealised bond-angle value of 120°, but in both cases there is a spread of possible values around this maximum. However, this spread is larger for the bond-angles between rigid rings and heavy substituents. While ~66% of the values of bond-angles between rigid rings and heavy substituents deviate by more than 1° from the maximum in the distribution, this is true only in ~35% of cases when no heavy substituent is connected to the rigid ring. Figure 6.12 and Figure

6.13 illustrate that treating all bond-angles in rigid rings as explicitly flexible in the optimisations is the most realistic solution, but they also confirm that the presence of a heavy substituent increases the likelihood of a bond-angle deviating from the most common value.

Finally Figure 6.14 shows the CSD distributions of the values of the bond-angle in Figure 6.8c, *i.e.* between atoms in two adjacent acyclic flexible bonds.

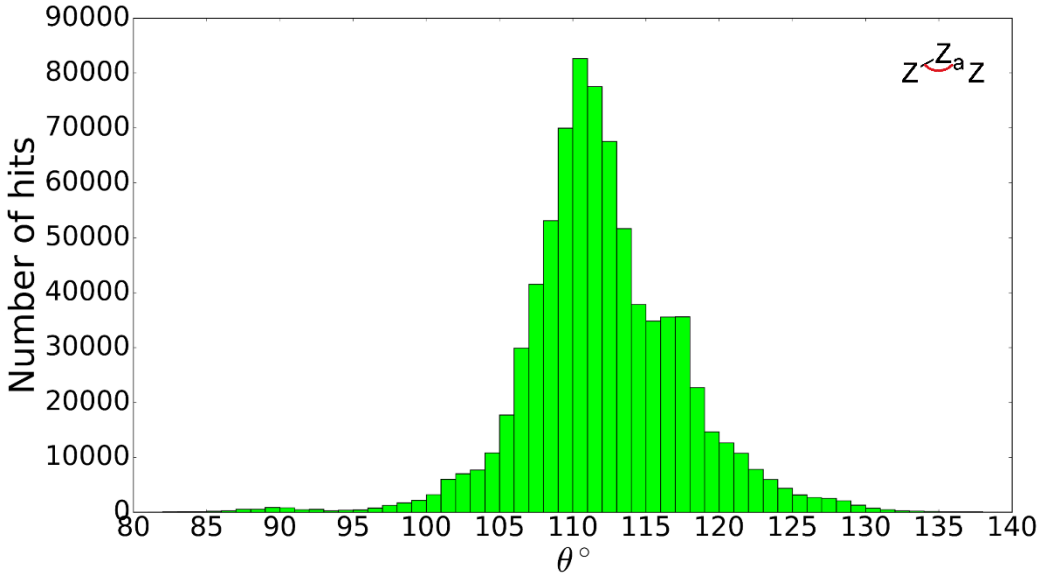


Figure 6.14: Histogram showing the CSD distribution of the values of the bond-angle in the fragment in Figure 6.8c.

Figure 6.14 shows that these bond-angles can take a broad range of values. Hence, this sort of CDFs can show a lot of variation in response to packing forces, and should be treated as independent in the optimisations, confirming the findings of the computational analysis.

In summary, these CSD surveys show that all torsion and bond-angles, even the most rigid ones, can take a range of values in the solid-state to adjust to packing forces and so should be treated as variables when optimising crystal structures. Nonetheless, their degree of flexibility varies. In particular, the intuitively rigid torsion and bond-angles identified in Chapter 6.2.5.2 seem to be more affected by packing than others that did not significantly vary in the benchmark optimisations with the $CDFs_{all}$, confirming that the findings of the calculations performed in this study are not just an artefact of the E_{latt} model.

6.4 Conclusion

Crystal Optimizer optimisations of 100 generated crystal structures of five flexible molecules were performed treating different sets of CDFs as independent to study which torsion and bond-angles can vary significantly to respond to packing. Optimisations with the intuition-based $CDFs_{torsion}$ and the intuition and experience-based $CDFs_{AUTODOF}$

produced results of low quality compared to the ideal but computationally expensive benchmark minimisations where all torsion and bond-angles were treated as independent CDFs, in particular in terms of absolute lattice energies. This showed that some extra CDFs need to be considered to produce better E_{latt} values. Analysing the results of the benchmark optimisations with the $CDFs_{all}$ three categories of intuitively rigidly CDFs were found to vary significantly: torsion angles in rigid rings where one of the central atom is bonded to a heavy substituent, bond-angles between rigid rings and heavy substituents and bond-angle between atoms in two adjacent acyclic flexible bonds. Adding these three categories of torsion and bond-angles to the $CDF_{AUTODOF}$, forming the $CDFs_{AUTODOF+}$, produced lower E_{latt} values at a reasonable computational cost. The importance of treating at least these intuitively rigid CDFs as independent in the final refinement stage of CSP was confirmed by a CSD analysis showing that they take a relatively broad range of values in experimentally-determined crystal structures.

Several insights that are very relevant to the overall purpose of this thesis are provided by this analysis. The explicit optimisation of all torsion and bond-angles in the final refinement stage of CSP is the most accurate solution within the Ψ_{mol} framework, but it is too computationally expensive. Hence an effective selection criterion for independent CDFs is needed, and the approach used to select the $CDFs_{AUTODOF+}$ seems to provide a good balance between accuracy and cost.

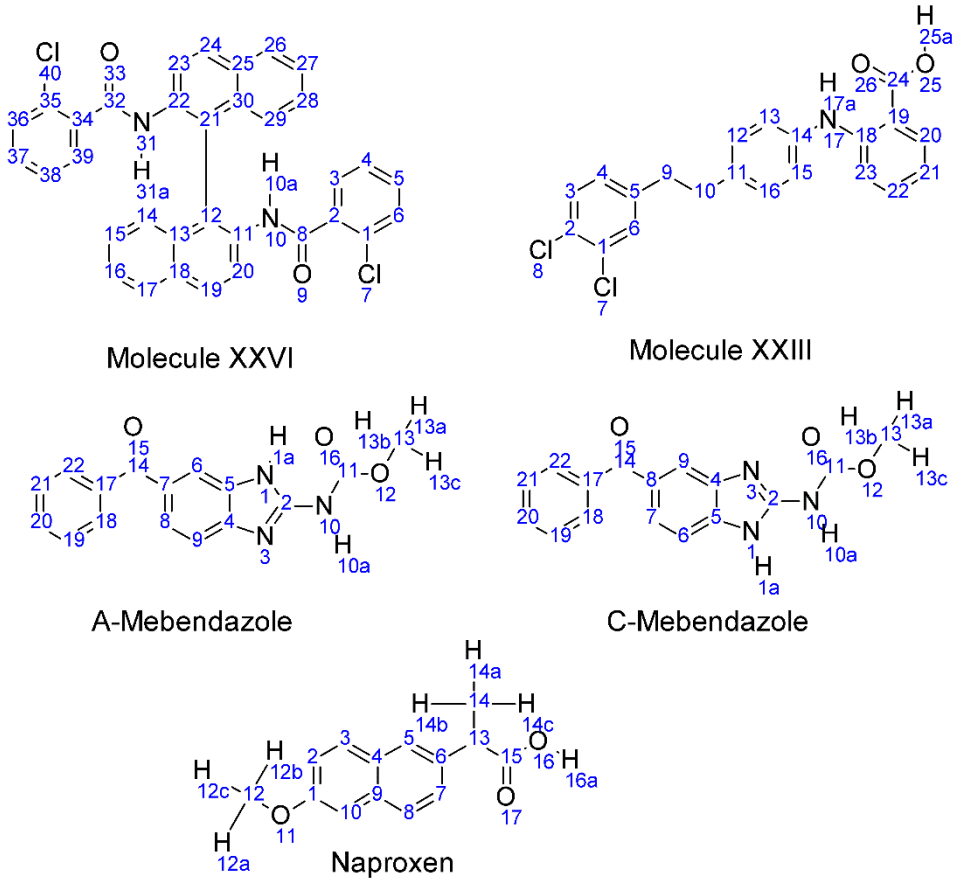
A fundamental weakness of the Ψ_{mol} model is also shown: limiting the explicitly flexible CDFs in crystal structure optimisations to keep the computational expense manageable is a necessary but unideal and not completely realistic approximation.[6] In fact the CSD surveys show that even the most rigid CDFs can take limited but non-insignificant flexibility ranges of values, and no selection approach was capable to reproduce the energies of the benchmark minimisations with the $CDFs_{all}$. Hence, the Ψ_{crys} model is more realistic as it treats all molecular CDFs as optimisation variables; however, similarly to utilising the $CDFs_{all}$ within the Ψ_{mol} model, it is very computationally demanding.[32] As already outlined in Chapter 3, using a very accurate but cost ineffective methodology in the final refinement stage of CSP, when hundreds or thousands of crystal structures need to be optimised, can lead to failures when large and flexible molecules of pharmaceutical interest are targeted. Hence, the use of cheaper Ψ_{crys} models that can explicitly optimise all intermolecular interactions and molecular CDFs at the same time would be the best solution to perform accurate, realistic but computationally affordable optimisations. This possibility is explored in detail in Chapter 7.

6.5 References

1. Cruz-Cabeza, A. J.; Bernstein, J., Conformational Polymorphism. *Chemical Reviews* **2014,** *114* (4), 2170-2191.
2. Thompson, H. P. G.; Day, G. M., Which conformations make stable crystal structures? Mapping crystalline molecular geometries to the conformational energy landscape. *Chemical Science* **2014,** *5* (8), 3173-3182.
3. Price, S. L., Predicting crystal structures of organic compounds. *Chemical Society Reviews* **2014,** *43* (7), 2098-2111.
4. 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.; Vacarro, F.; Vasileiadis, M.; Vazquez-Mayagoitia, A.; Vogt, L.; Wang, Y.; Watson, R. E.; de Wijs, G. A.; Yang, J. Z.; Zhu, Q.; Groom, C. R., Report on the sixth blind test of organic crystal-structure prediction methods. *Acta Crystallographica Section B - Structural Science* **2016**.
5. Iuzzolino, L.; McCabe, P.; Price, S. L.; Brandenburg, J. G., Crystal structure prediction of flexible pharmaceutical-like molecules: Density functional tight-binding as an intermediate optimization method and for free energy estimation. *Faraday Discussions* **2018,** *Advance article*.
6. Pantelides, C. C.; Adjiman, C. S.; Kazantsev, A. V., General Computational Algorithms for Ab Initio Crystal Structure Prediction for Organic Molecules. *Topics in Current Chemistry* **2014,** *345*, 25-58.
7. Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C.; Price, S. L.; Galek, P. T.; Day, G. M.; Cruz-Cabeza, A. J., Successful prediction of a model pharmaceutical in the fifth blind test of crystal structure prediction. *International Journal of Pharmaceutics* **2011,** *418* (2), 168-178.
8. Vasileiadis, M.; Pantelides, C. C.; Adjiman, C. S., Prediction of the crystal structures of axitinib, a polymorphic pharmaceutical molecule. *Chemical Engineering Science* **2015,** *121*, 60-76.
9. Price, S. L.; Brandenburg, J. G., Chapter 11 - Molecular Crystal Structure Prediction. In *Non-Covalent Interactions in Quantum Chemistry and Physics*, Elsevier: 2017; pp 333-363.
10. Iuzzolino, L.; Reilly, A. M.; McCabe, P.; Price, S. L., Use of Crystal Structure Informatics for Defining the Conformational Space Needed for Predicting Crystal Structures of Pharmaceutical Molecules. *Journal of Chemical Theory and Computation* **2017,** *13* (10), 5163-5171.
11. Price, S. L.; Reutzel-Edens, S. M., The potential of computed crystal energy landscapes to aid solid form development. *Drug Discovery Today* **2016**.
12. 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. *International Journal of Pharmaceutics* **2011,** *418* (2), 168-178.
13. Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C., Efficient Handling of Molecular Flexibility in Lattice Energy Minimization of Organic Crystals. *Journal of Chemical Theory and Computation* **2011,** *7* (6), 1998-2016.
14. Nyman, J.; Day, G. M., Static and lattice vibrational energy differences between polymorphs. *Crystengcomm* **2015,** *17* (28), 5154-5165.
15. Price, L. S.; McMahon, J. A.; Lingireddy, S. R.; Lau, S. F.; Diseroad, B. A.; Price, S. L.; Reutzel-Edens, S. M., A molecular picture of the problems in ensuring structural purity of tazofelone. *Journal of Molecular Structure* **2014,** *1078*, 26-42.
16. 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. *Crystal Growth & Design* **2013,** *13* (6), 2396-2406.

17. Braun, D. E.; Orlova, M.; Griesser, U. J., Creatine: Polymorphs Predicted and Found. *Crystal Growth & Design* **2014,** *14* (10), 4895-4900.
18. Braun, D. E.; Ardid-Candel, M.; D'Oria, E.; Karamertzanis, P. G.; Arlin, J. B.; Florence, A. J.; Jones, A. G.; Price, S. L., Racemic Naproxen: A Multidisciplinary Structural and Thermodynamic Comparison with the Enantiopure Form. *Crystal Growth & Design* **2011,** *11* (12), 5659-5669.
19. Nyman, J.; Reutzel-Edens, S., Crystal structure prediction is changing from basic science to applied technology. *Faraday Discussions* **2018,** *Advance article*.
20. Nyman, J.; Day, G. M., Modelling temperature-dependent properties of polymorphic organic molecular crystals. *Physical Chemistry Chemical Physics* **2016,** *18* (45), 31132-31143.
21. Corpinot, M. K.; Iuzzolino, L.; Price, S. L.; Bučar, D.-K., Screening work in progress, 2017.
22. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C., The Cambridge Structural Database. *Acta Crystallographica Section B-Structural Science Crystal Engineering and Materials* **2016,** *72*, 171-179.
23. Ferreira, F. F.; Antoni, S. G.; Rosa, P. C. P.; Paiva-Santos, C. D., Crystal Structure Determination of Mebendazole Form A Using High-Resolution Synchrotron X-Ray Powder Diffraction Data. *Journal of Pharmaceutical Sciences* **2010,** *99* (4), 1734-1744.
24. Martins, F. T.; Neves, P. P.; Ellena, J.; Cami, G. E.; Brusau, E. V.; Narda, G. E., Intermolecular Contacts Influencing the Conformational and Geometric Features of the Pharmaceutically Preferred Mebendazole Polymorph C. *Journal of Pharmaceutical Sciences* **2009,** *98* (7), 2336-2344.
25. Kim, Y. B.; Song, H. J.; Park, I. Y., Refinement of the structure of naproxen, (+)-6-methoxy-α-methyl-2-naphthaleneacetic acid. *Archives of Pharmacal Research* **1987,** *10* (4), 232-238.
26. Price, S. L.; Leslie, M.; Welch, G. W. A.; Habgood, M.; Price, L. S.; Karamertzanis, P. G.; Day, G. M., Modelling Organic Crystal Structures using Distributed Multipole and Polarizability-Based Model Intermolecular Potentials. *Physical Chemistry Chemical Physics* **2010,** *12* (30), 8478-8490.
27. van Rossum, G. *Python: a computer language*, 1.5.1; 1998.
28. van de Streek, J.; Motherwell, S., Searching the Cambridge Structural Database for polymorphs. *Acta Crystallographica Section B - Structural Science* **2005,** *61*, 504-510.
29. Threlfall, T. L.; Gelbrich, T., The crystal structure of methyl paraben at 118 K does not represent a new polymorph. *Crystal Growth & Design* **2007,** *7* (11), 2297-2297.
30. Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kessler, M.; Macrae, C. F.; McCabe, P.; Pearson, J.; Taylor, R., New software for searching the Cambridge Structural Database and visualizing crystal structures. *Acta Crystallographica Section B - Structural Science* **2002,** *58*, 389-397.
31. van de Streek, J., Searching the Cambridge Structural Database for the 'best' representative of each unique polymorph. *Acta Crystallographica Section B - Structural Science* **2006,** *62*, 567-579.
32. Brandenburg, J. G.; Hochheim, M.; Bredow, T.; Grimme, S., Low-Cost Quantum Chemical Methods for Noncovalent Interactions. *Journal of Physical Chemistry Letters* **2014,** *5* (24), 4275-4284.

6.6 Appendix



Appendix Figure 6.1: Chemical diagrams of the five molecules considered in this test, showing the atomic numbering that is used in Appendix Tables 6.1-6.5 to precisely define the torsion and bond-angles.

Appendix Table 6.1: Definition of the torsion and bond-angles of molecule XXVI treated as independent CDFs with when different selection approaches were adopted.

	Torsion angle definition	Bond-angle definition
$CDFs_{torsion}$	31-32-34-35, 22-31-32-34, 21-22-31-32, 11-12-21-22, 8-10-11-12, 2-8-10-11, 1-2-8-10	/
Added to the $CSDs_{AUTODOF}$	31a-31-22-21, 2-8-10-10a	31a-31-22, 8-10-10a
Added to the $CSDs_{AUTODOF+}$	38-39-34-35, 21-12-11-10, 25-30-21-22, 28-29-30-21, 15-14-13-12, 14-13-12-21, 20-11-12-21, 4-3-2-8	32-34-35, 23-22-31, 12-21-22, 10-11-12, 1-2-8, 11-12-21, 34-32-31, 2-8-10

Appendix Table 6.2: Definition of the torsion and bond-angles of molecule XXIII treated as independent CDFs with when different selection approaches were adopted.

	Torsion angle definition	Bond-angle definition
$CDFs_{torsion}$	11-10-9-5, 12-11-10-9, 10-9-5-6, 18-17-14-15, 23-18-17-14, 26-24-19-18, 25a-25-24-19	/
Added to the $CSDs_{AUTODOF}$	17a-17-14-15	17a-17-14, 25a-25-24
Added to the $CSDs_{AUTODOF+}$	1-6-5-9, 13-12-11-10, 17-14-15-16, 22-23-18-17, 24-19-18-17	6-5-9, 16-11-10, 17-14-13, 23-18-17, 24-19-18, 5-9-10, 9-10-11, 25-24-19

Appendix Table 6.3: Definition of the torsion and bond-angles of the A-tautomer of mebendazole treated as independent CDFs with when different selection approaches were adopted.

	Torsion angle definition	Bond-angle definition
$CDFs_{torsion}$	10-11-12-13, 2-10-11-12, 1-2-10-11, 17-14-7-6, 22-17-14-7, 13a-13-12-11	/
Added to the $CSDs_{AUTODOF}$	1a-1-2-10, 10a-10-11-12	1a-1-2, 10a-10-11
Added to the $CSDs_{AUTODOF+}$	7-6-5-1, 19-18-17-14, 9-8-7-6	22-17-14, 14-7-6, 1-2-10, 17-14-7, 10-11-12, 11-12-13

Appendix Table 6.4: Definition of the torsion and bond-angles of the C-tautomer of mebendazole treated as independent CDFs with when different selection approaches were adopted.

	Torsion angle definition	Bond-angle definition
$CDFs_{torsion}$	10-11-12-13, 2-10-11-12, 1-2-10-11, 17-14-8-7, 22-17-14-8, 13a-13-12-11	/
Added to the $CSDs_{AUTODOF}$	1a-1-2-10, 10a-10-11-12	1a-1-2, 10a-10-11
Added to the $CSDs_{AUTODOF+}$	19-18-17-14, 14-8-9-4, 8-9-4-5	17-18-14, 14-8-7, 1-2-10, 17-14-8, 10-11-12, 11-12-13

Appendix Table 6.5: Definition of the torsion and bond-angles of naproxen treated as independent CDFs with when different selection approaches were adopted.

	Torsion angle definition	Bond-angle definition
$CDFs_{torsion}$	15-13-14-14a, 7-6-13-14, 16-15-13-14, 12-11-1-10, 12a-12-11-1, 16a-16-15-13	/
Added to the $CSDs_{AUTODOF}$	/	16a-16-15
Added to the $CSDs_{AUTODOF+}$	11-1-2-3, 8-7-6-13, 2-3-4-5	11-1-10, 7-6-13, 12-11-1, 6-3-14, 16-15-13

Appendix Table 6.6: E_{latt} values of the 20 crystal structures of molecule XXVI when different criteria for selecting the independent CDFs in the CrystalOptimizer optimisations were adopted, relative to the global minimum in the benchmark optimisations with the $CDFs_{all}$. The crystal structure matching the known form (XAFQIH) is highlighted in orange.

	ΔE_{latt} for the optimisations with different CDFs/ kJ·mol^{-1}			
Name of the structure from the original CSP	**$CDFs_{all}$**	**$CDFs_{torsion}$**	**$CDFs_{AUTODOF}$**	**$CDFs_{AUTODOF+}$**
1600	0.00	6.33	3.28	1.91
2231	0.57	12.54	10.11	4.05
675	0.65	7.85	6.71	4.05
3525	1.16	4.98	4.65	2.51
239	2.08	11.42	11.56	4.53
2930	3.64	19.57	7.46	5.11
4946	4.08	22.01	20.42	16.76
615	4.23	15.58	11.24	5.47
38	4.98	9.78	9.21	6.72
2496	5.37	12.57	10.76	7.93
508	5.99	13.43	11.65	8.88
6460	6.62	12.11	10.00	8.81
3104	7.58	10.75	10.77	9.74
6335	7.75	12.54	11.90	9.68

851	8.69	12.06	11.71	11.11
314	8.88	13.37	14.13	10.97
354	8.95	11.88	12.08	10.54
185	9.16	13.20	13.24	11.41
221	9.97	13.05	12.72	11.46
4201	10.54	13.25	13.27	12.37

Appendix Table 6.7: E_{latt} values of the 20 crystal structures of molecule XXIII when different criteria for selecting the independent CDFs in the CrystalOptimizer optimisations were adopted, relative to the global minimum in the benchmark optimisations with the $CDFs_{all}$. The crystal structure matching known form B (XAFPAY01) is highlighted in orange.

	ΔE_{latt} for the optimisations with different CDFs/ kJ·mol^{-1}			
Name of the structure from the original CSP study	**$CDFs_{all}$**	**$CDFs_{torsion}$**	**$CDFs_{AUTODOF}$**	**$CDFs_{AUTODOF+}$**
A2073	0.00	3.35	2.64	1.29
A70	0.01	3.34	2.67	1.34
A51	0.93	6.53	5.67	2.29
A424	1.70	8.77	8.48	8.43
A1361	2.09	4.12	3.78	2.58
A691	2.11	3.69	3.40	2.86
A5191	2.18	5.58	5.27	3.33
A771	2.90	4.53	4.26	3.07
A72	3.00	5.65	5.35	4.10
A4890	3.24	5.41	4.90	4.16
A63	3.57	6.99	6.34	4.46
A118	3.65	6.38	5.84	4.30
A2112	3.72	6.41	5.83	4.41
A894	4.05	9.76	12.81	6.02
A75	4.28	5.33	5.06	4.89
A191	4.34	7.06	6.60	5.33
A272	4.45	10.95	10.52	8.09
A1413	4.62	5.58	5.38	4.68
A1752	5.45	13.46	7.44	6.58
A3457	16.47	22.34	21.55	19.89

Appendix Table 6.8: E_{latt} values of the 20 crystal structures of the A-tautomer of mebendazole when different criteria for selecting the independent CDFs in the CrystalOptimizer optimisations were adopted, relative to the global minimum in the benchmark optimisations with the $CDFs_{all}$. The crystal structure matching known form A (TUXPEJ) is highlighted in orange.

	ΔE_{latt} for the optimisations with different CDFs/ kJ·mol^{-1}			
Name of the structure from the original CSP study	**$CDFs_{all}$**	**$CDFs_{torsion}$**	**$CDFs_{AUTODOF}$**	**$CDFs_{AUTODOF+}$**
19	0.00	4.39	2.86	2.45
788	0.11	3.15	1.38	1.17
37	2.00	6.87	5.15	4.94
50	2.72	5.25	3.92	3.89
291	3.00	8.49	7.90	5.98
143	3.47	8.62	8.05	6.84
109	4.02	11.69	9.66	9.06
173	4.91	8.85	6.73	6.65
90	5.25	8.80	6.99	6.66
72	5.29	9.65	6.60	6.59
49	5.37	6.97	6.82	6.63
89	5.45	9.13	8.61	7.65
53	5.63	5.82	5.86	5.99

54	5.90	9.53	8.77	7.76
306	6.30	8.69	7.89	7.49
78	6.34	6.73	6.72	6.82
202	6.90	8.28	8.12	7.90
75	7.92	9.78	9.63	9.21
71	8.68	11.06	9.71	9.37
604	8.96	9.84	9.70	9.85

Appendix Table 6.9: E_{latt} values of the 20 crystal structures of the C-tautomer of mebendazole when different criteria for selecting the independent CDFs in the CrystalOptimizer optimisations were adopted, relative to the global minimum in the benchmark optimisations with the $CDFs_{all}$. The crystal structure matching known form C (YULGIW) is highlighted in orange.

	ΔE_{latt} for the optimisations with different CDFs/ kJ·mol^{-1}			
Name of the structure from the original CSP study	**$CDFs_{all}$**	**$CDFs_{torsion}$**	**$CDFs_{AUTODOF}$**	**$CDFs_{AUTODOF+}$**
27	0.00	2.20	1.91	1.18
10	1.02	2.83	2.03	1.94
5	1.20	2.80	2.15	1.74
406	2.26	4.63	4.28	3.50
23	2.77	4.47	3.80	3.67
73	3.02	4.39	3.67	3.67
248	3.83	6.38	5.79	4.73
24	3.84	6.41	6.17	5.07
115	3.93	6.31	6.04	4.99
46	4.02	7.15	6.64	5.23
25	4.35	5.49	4.84	4.79
53	4.37	5.52	4.85	4.92
908	4.59	7.05	6.81	5.41
106	4.87	6.14	6.19	5.61
583	5.22	6.41	6.34	5.57
244	5.78	7.61	7.47	6.55
199	7.04	8.26	8.05	8.02
111	7.08	7.74	7.64	7.32
132	7.47	7.21	7.15	7.33
220	8.20	8.09	8.09	8.26

Appendix Table 6.10: : E_{latt} values of the 20 crystal structures of naproxen when different criteria for selecting the independent CDFs in the CrystalOptimizer optimisations were adopted, relative to the global minimum in the benchmark optimisations with the $CDFs_{all}$. The crystal structure matching the known racemic form (PAPTUX) is highlighted in orange, while the crystal structure matching the known enantiopure form (COYRUD11) is highlighted in green.

	ΔE_{latt} for the optimisations with different CDFs/ kJ·mol^{-1}			
Name of the structure from the original CSP study	**$CDFs_{all}$**	**$CDFs_{torsion}$**	**$CDFs_{AUTODOF}$**	**$CDFs_{AUTODOF+}$**
CO_1	0.00	2.37	1.97	0.65
ak57	0.60	1.51	1.07	1.00
fc15	2.55	3.42	3.11	2.96
am85	2.88	4.39	4.06	3.59
fb24	5.30	6.66	6.57	5.44
ak35	5.40	6.33	6.27	5.91
fc100	5.96	7.43	7.29	6.46
ab9	6.00	6.63	6.31	6.23
af92	6.28	9.47	9.10	7.10

fa31	6.54	7.02	6.82	6.85
ak63	6.86	9.75	9.49	7.58
am133	6.89	7.87	7.40	7.50
fc125	7.46	10.40	10.09	8.09
fc119	7.75	8.17	7.98	8.03
cc56	8.71	9.65	9.52	9.34
fa104	9.21	12.95	12.84	9.84
ca102	9.45	9.73	9.67	7.68
aq49	9.64	11.51	11.14	10.38
bh18	9.73	15.61	15.39	11.17
ah12	10.29	11.58	11.42	10.90

Appendix Table 6.11: Packing coefficient of the 20 crystal structures of molecule XXVI when different criteria for selecting the independent CDFs in the CrystalOptimizer optimisations were adopted. The crystal structure matching the known form (XAFQIH) is highlighted in orange.

	Packing coefficient for the optimisations with different CDFs/%			
Name of the structure from the original CSP study	**$CDFs_{all}$**	**$CDFs_{torsion}$**	**$CDFs_{AUTODOF}$**	**$CDFs_{AUTODOF+}$**
1600	67.31	67.47	67.4	67.25
2231	69.17	68.53	68.03	68.73
675	69.44	68.57	68.72	69.07
3525	69.05	68.33	68.41	68.7
239	69.58	68.2	68.08	68.83
2930	69.4	68.25	68.92	69.34
4946	69.81	66.18	66.42	67.23
615	67.95	66.14	66.83	67.45
38	68.25	67.41	67.52	67.86
2496	69.37	67.56	67.95	68.47
508	69.27	68.57	68.7	68.91
6460	66.85	65.72	66	66.19
3104	71.07	70.48	70.54	70.65
6335	68.35	67.61	67.45	67.85
851	67.25	66.7	66.68	66.78
314	67.86	65.84	65.9	67.15
354	67.89	67.01	66.93	67.32
185	71.41	70.7	70.63	71.02
221	67.38	66.57	66.53	66.96
4201	68.49	67.95	67.89	68.09
MEAN	68.76	67.69	67.78	68.19

Appendix Table 6.12: Packing coefficient of the 20 crystal structures of molecule XXIII when different criteria for selecting the independent CDFs in the CrystalOptimizer optimisations were adopted. The crystal structure matching known form B (XAFPAY01) is highlighted in orange.

	Packing coefficient for the optimisations with different CDFs/%			
Name of the structure from the original CSP study	**$CDFs_{all}$**	**$CDFs_{torsion}$**	**$CDFs_{AUTODOF}$**	**$CDFs_{AUTODOF+}$**
A2073	71.63	71.04	71.08	71.26
A70	71.89	71.29	71.34	71.5
A51	70.23	68.03	68.6	69.84
A424	70.32	70.73	70.78	70.79
A1361	71.52	70.87	70.86	71.27
A691	69.83	69.35	69.33	69.55
A5191	70.78	70.32	70.30	70.57
A4890	72.59	72.23	72.27	72.46
A72	71.20	70.51	70.54	70.97
A771	71.57	71.10	71.08	71.43

A63	70.89	70.92	70.91	70.76
A118	70.39	70.03	70.03	70.28
A2112	70.87	70.49	70.53	70.68
A894	70.61	69.5	67.55	70.10
A191	71.21	70.96	71.01	71.14
A1413	69.68	69.49	69.45	69.47
A75	71.73	68.48	68.59	69.15
A272	71.82	71.52	71.55	71.66
A1752	70.16	66.91	69.05	69.56
A3457	68.97	66.57	66.67	67.46
MEAN	70.89	70.02	70.08	70.50

Appendix Table 6.13: Packing coefficient of the 20 crystal structures of the A-tautomer of mebendazole when different criteria for selecting the independent CDFs in the CrystalOptimizer optimisations were adopted. The crystal structure matching known form A (TUXPEJ) is highlighted in orange.

	Packing coefficient for the optimisations with different CDFs/%			
Name of the structure from the original CSP study	**$CDFs_{all}$**	**$CDFs_{torsion}$**	**$CDFs_{AUTODOF}$**	**$CDFs_{AUTODOF+}$**
19	73.93	72.61	72.78	72.83
788	72.38	71.72	71.88	72.01
37	73.26	72.03	72.22	72.24
50	72.57	71.83	72.07	72.13
291	73.65	72.88	73.03	73.34
143	74.17	72.82	72.94	73.72
109	71.82	70.66	70.73	71.07
173	72.40	71.21	71.39	71.43
90	70.69	69.82	70.06	70.35
72	72.28	72.11	72.20	72.09
49	71.65	71.05	71.13	71.14
89	72.25	71.75	71.81	72.18
53	71.63	71.52	71.50	71.77
54	73.02	72.43	72.40	72.36
306	72.77	72.30	72.56	72.74
78	71.52	71.55	71.54	71.42
202	70.72	70.29	70.28	70.56
75	71.70	71.46	71.52	71.59
71	71.56	71.35	71.34	71.27
604	70.60	70.52	70.52	70.41
MEAN	72.23	71.60	71.70	71.83

Appendix Table 6.14: Packing coefficient of the 20 crystal structures of the C-tautomer of mebendazole when different criteria for selecting the independent CDFs in the CrystalOptimizer optimisations were adopted. The crystal structure matching known form C (YULGIW) is highlighted in orange.

	Packing coefficient for the optimisations with different CDFs/%			
Name of the structure from the original CSP study	**$CDFs_{all}$**	**$CDFs_{torsion}$**	**$CDFs_{AUTODOF}$**	**$CDFs_{AUTODOF+}$**
27	72.63	71.75	71.79	72.04
10	72.71	72.03	72.26	72.32
5	72.63	71.9	72.04	72.22
406	72.22	71.23	71.37	71.58
23	71.96	71.16	71.28	71.37
73	71.81	71.25	71.39	71.43
248	72.24	71.53	71.65	71.87
24	71.53	70.46	70.49	70.83
115	72.32	72.22	72.30	72.23

46	72.49	71.28	71.38	71.85
25	70.80	70.17	70.29	70.49
53	69.98	69.60	69.70	69.64
908	71.98	71.44	71.47	71.73
106	73.19	72.75	72.80	72.85
583	73.27	72.63	72.68	72.97
244	72.04	72.07	72.03	71.89
199	71.62	70.84	70.88	70.87
111	74.41	74.12	74.19	74.30
132	72.26	72.11	72.12	72.18
220	72.37	72.17	72.17	72.45
MEAN	72.22	71.64	71.71	71.86

Appendix Table 6.15: Packing coefficient of the 20 crystal structures of naproxen when different criteria for selecting the independent CDFs in the CrystalOptimizer optimisations were adopted. The crystal structure matching the known racemic form (PAPTUX) is highlighted in orange, while the crystal structure matching the known enantiopure form (COYRUD11) is highlighted in green.

	Packing coefficient for the optimisations with different CDFs/%			
Name of the structure from the original CSP study	**$CDFs_{all}$**	**$CDFs_{torsion}$**	**$CDFs_{AUTODOF}$**	**$CDFs_{AUTODOF+}$**
CO_1	71.83	70.93	70.96	71.68
ak57	72.12	71.69	71.65	71.94
fc15	70.56	70.15	70.20	70.49
am85	71.44	70.60	70.60	71.15
fb24	70.53	70.24	70.28	70.64
ak35	69.80	69.40	69.39	69.34
fc100	70.26	69.81	69.80	70.06
ab9	69.89	69.46	69.72	69.73
af92	69.64	68.13	68.25	69.22
fa31	69.77	69.51	69.59	69.59
ak63	68.04	66.93	66.97	67.90
am133	68.45	68.2	68.14	68.32
fc125	68.44	66.91	67.02	68.21
fc119	69.14	69.02	69.01	69.05
cc56	70.07	69.46	69.45	69.64
fa104	69.00	66.41	66.57	68.62
ca102	68.82	68.56	68.57	69.09
aq49	68.33	67.54	67.58	68.01
bh18	69.24	65.63	65.70	68.80
ah12	68.45	67.82	67.65	68.14
MEAN	69.69	68.82	68.86	69.48

Chapter 7: Use of density functional tight-binding (DFTB) as an intermediate optimisation method and for free energy estimation

7.1 Introduction

7.1.1 The importance of intermediate optimisations in the final refinement stage of CSP

In Chapter 5, a workflow for using CSD information to speed up CSP searches was developed.[1] Although most of the significant crystal structures from previous CSP studies (*i.e.* that were thermodynamically competitive after the final optimisations),[2-5] including all those matching the experimentally-characterised *Z'*=1 forms, were successfully generated at a reduced computational cost, they were often poorly ranked among a plethora of alternatives of little practical interest.

This is a well-known problem: the lattice energy (E_{latt}) values are generally computed in the searches with approximate models, and are rarely accurate enough to provide reliable energy rankings (see also Chapter 2.4).[2, 6, 7] This poses significant challenges to the final refinement stage of CSP: because of the poor E_{latt} ranking output by the search, thousands of generated crystal structures must be optimised and re-ranked to produce an accurate crystal energy landscape.[6, 7] Since the final refinement requires the use of high-quality expensive models, such as periodic electronic structure calculations with the Ψ_{crys} methods on all crystal structures,[8-11] or the calculation of a high-quality wave-function for each molecular conformation in the crystals (the Ψ_{mol} method, se chapter 2.3.1),[12-14] the computational cost can easily explode if too many search-generated structures have to be refined.[7, 15]

A possible solution to this problem is performing an intermediate optimisation, which can bridge the gap between the cheap and simple models used in the searches and the more accurate and expensive ones that are required in the final refinement of the generated crystal structures.[2, 7, 16] For flexible molecules, several intermediate optimisation methods are used in CSP, such as minimising some torsion angles with a transferable force-field to improve the molecular conformations,[17, 18] performing single-point energy calculations with a more accurate description of the intra- and intermolecular interactions[19, 20] or partial E_{latt} optimisations (as it was done in Chapters 3 and 4).[2, 7] This chapter attempts to test one possible cheap yet accurate method to perform intermediate optimisations with the goal of speeding up the final refinement stage of a CSP procedure, which could be practicable for expanding computational studies to larger and more flexible molecules.

7.1.2 Advantages of periodic semi-empirical method

In Chapter 6 it has been shown how refining search-generated crystal structures with the Ψ_{mol} method (see Chapter 2.3.1) has the inherent weakness of requiring the selection of the conformational degrees of freedom (CDFs) that are to be explicitly optimised under the influence of packing forces.[21] Ideally all torsion and bond-angles should be treated as explicit variables for each crystal structure (bond-lengths can be safely ignored).[6, 21] However, since hundreds or thousands of search-generated crystal structures often need to be refined, this would be unaffordable for large and flexible molecules as the computational cost increases with the number of explicitly optimised CDFs.[21] Although an effective criterion for selecting the explicitly flexible CDFs was outlined in Chapter 6, no practically affordable method could reproduce the energies obtained upon optimising all torsion and bond-angles. Using Ψ_{crys} methods (see Chapter 2.3.2) is more realistic, as all atomic positions are optimised at once.[22] The most commonly used Ψ_{crys} method in CSP is dispersion-corrected density functional theory, DFT-D (Chapter 2.3.2.1).[2, 10] However, the need of balancing the intra- and intermolecular interactions makes DFT-D with any functional and dispersion correction that can give worthwhile accuracy very demanding.[6, 16]

Within the Ψ_{crys} framework, periodic semi-empirical quantum-mechanical methods (see Chapter 2.3.2.2) can be a significantly less costly alternative.[16, 23, 24] Given their approximate Hamiltonian, they cannot be used to produce an accurate final crystal energy landscape. This chapter tests whether optimisations with density functional tight binding methods (DFTB3-D3, see chapter 2.3.2.2.1 for the theory behind it), which are up to three orders of magnitude cheaper than DFT-D,[16] can be used as an intermediate step in CSP, to be followed by a final calculation of E_{latt} with more accurate models.

Another shortcoming of DFT-D and other Ψ_{crys} methods is that the calculation of all the molecular and lattice phonon modes to determine the vibrational component of free energy is very computationally demanding (note that with the Ψ_{mol} method only the rigid-body lattice modes are calculated, see Chapter 2.3.3.1).[2, 9, 14] Hence another aim of this chapter is to verify the ability of DFTB3-D3 to affordably estimate the vibrational component of free energy within the harmonic approximation (see Chapter 2.3.3).

In the first portion of this chapter, the suitability of DFTB3-D3 as an intermediate optimisation method is tested for its ability to improve the geometries and the energy rankings of the thousands of crystal structures generated in the searches that were carried out in Chapter 5 on five large and flexible molecules.[1] The quality of the intermediate DFTB3-D3 optimisations is also assessed by performing a final calculation of E_{latt} with a higher-quality wave-function starting from their output. Finally, free energy estimates are performed on some of the most competitive crystal structures to test whether DFTB3-D3 can be used to carry out fast and reliable phonon calculations.

7.2 Methods

7.2.1 DFTB3-D3 intermediate optimisation of all the search-generated structures

The starting point for the refinement procedure tested in this chapter was the set of unique crystal structures within 40 $kJ \cdot mol^{-1}$ of the global minimum in CrystalPredictor[25] lattice energy that had been generated with the search workflow described in Chapter 5.2.2.2. This set consisted of 9,215 search-generated crystal structures for molecule XXVI, 16,744 for GSK269984B, 28,249 for molecule XXIII, 26,650 for molecule XX, 4,165 for the A-tautomer of mebendazole and 4,284 for its C-tautomer.

These crystal structures were optimised with DFTB3-D3, using the programme dftb+,[26] which required the symmetry of each crystal structure to be reduced to the P1 space group. The 3OB Slater-Koster files[27, 28] were used for the DFTB3 transferable parameters, and the missing dispersion interactions were corrected with the atom-pairwise D3 scheme.[29] The D3 damping parameters were chosen as those that minimise the errors in centre-of-mass distances of a set of small molecular dimers (S66).[30] The DFTB3-D3 optimisations were performed with a quasi-Newton scheme as implemented in the CRYSTAL14 programme,[31] using thresholds of 0.12 a.u. and 0.003 a.u. for root-mean square (RMS) displacement and RMS gradient respectively.

Once all the minimisations were completed, the symmetry of all the crystal structures that had been successfully optimised was reintroduced with Platon.[32] Successively, they were clustered to remove duplicates using the Crystal Packing Similarity tool[33] (see Chapter 2.5.1) available through the CSD Python API (Chapter 2.6.5).[34] Structures were considered as duplicates if they had an energy difference smaller than 5 $kJ \cdot mol^{-1}$, a density difference smaller than 0.1 $g \cdot cm^{-3}$ and if it was possible to match 15/15 molecules, with 20% distance and 20° angle tolerances, with a root mean square deviation ($RMSD_{15}$) smaller than 0.5 Å. When the structures had different numbers of molecules in the asymmetric unit cell (Z'), the $RMSD_{15}$ threshold was reduced to 0.1 Å, to avoid the removal of isostructural polymorphs[35] with different Z' and in different space groups (see Chapter 8.1 for details).

7.2.2 Final re-ranking using an improved molecular wave-function

The energy ranking calculated with DFTB3-D3 was very different from that computed with the Ψ_{mol} model with atomic point charges used by CrystalPredictor, and the E_{latt} values had a larger spread than after the search. The distributions of the E_{latt} values before and after the DFTB3-D3 optimisations are shown for each molecule in Appendix Figures 7.1-7.6.

The high relative energies of the structures matching the experimental forms showed that, contrary to the original hope, DFTB3-D3 had not produced a more accurate

energy ranking. Since the E_{latt} values after the intermediate optimisations were clearly inadequate, they could not be used to significantly limit the number of crystal structures to be taken to the last step of the refinement procedure. However, the intermediate minimisations were successful at improving the reproductions of the known experimental forms (see Table 7.1). In particular, DFTB3-D3 drastically improved the quality of the reproduction of the crystalline conformations, and this was exploited by reverting to the Ψ_{mol} approach used in the crystal structure generation stage, but with a more accurate evaluation of both ΔE_{intra} and U_{inter}.

For all the DFTB3-D3 optimised crystal structures, the wave-function of each molecular conformation was calculated at the PBE0 6-31G(d,p) level of theory using Gaussian09,[36] and distributed multipoles were derived from the charge density using GDMA 2.2[37] (see Chapter 2.3.1.2.2 for details). Finally the intermolecular interactions were optimised with DMACRYS[38] (see Chapter 2.4.2.1), with the repulsion-dispersion component calculated with the empirically-fitted FIT potential.[39] The crystal structures optimised with this specific method, denoted $\Psi_{mol}^{PBE0+FIT}$,[15] were ranked in terms of E_{latt}. ΔE_{intra} was estimated as the difference between the energies of each molecular conformation contained in the crystal structures and the PBE0 6-31G(d,p) energy, calculated with Gaussian09, of the DFTB3-D3 optimised isolated-molecule global minimum conformer. This was used to avoid the absolute values of E_{latt} being affected by a ΔE_{intra} component calculated relative to a conformer optimised with a different wave-function. Note that this method of calculating ΔE_{intra} does not change the relative energy ranking of the computer-generated crystal structures, which is the most important output of CSP studies.

Given the unreliability of the DFTB3-D3 E_{latt} ranking, all the unique crystal structures within a large window of 50 kJ·mol^{-1} of the global minimum were optimised with the $\Psi_{mol}^{PBE0+FIT}$ method. Despite using a large energy cut-off, its application and the removal of duplicates after the DFTB3-D3 optimisations reduced the number of structures by a molecule-dependent proportion, which is broken down in Appendix Table 7.1. A total of 3,346 crystal structures for molecule XXVI, 5,328 for GSK269984B, 13,490 for molecule XXIII, 19,146 for molecule XX, 3,078 for the A-tautomer of mebendazole and 3,352 for its C-tautomer underwent these final optimisations.

For each molecule, all the crystal structures within 50 kJ·mol^{-1} of the global minimum in $\Psi_{mol}^{PBE0+FIT}$ E_{latt} were clustered to remove duplicates. The same clustering method and parameters as after the DFTB3-D3 optimisations were utilised, with the only difference being that the maximum $RMSD_{15}$ for two structures to be considered as duplicates was raised from 0.5 to 1.0 Å. A stricter threshold had been used when performing clustering after the intermediate optimisations to avoid the removal of

structures that were deemed to be sufficiently different to converge to different E_{latt} minima in the subsequent stage.

7.2.3 DFTB3-D3 phonon calculations to compute free energies

DFTB3-D3 phonon calculations were performed on few diverse CSP-generated crystal structures with low relative E_{latt} values after the $\Psi_{mol}^{PBE0+FIT}$ optimisations, as well as on those matching the experimentally-characterised forms, to verify whether this semi-empirical method can cheaply and reliably compute the vibrational component of free energy (F_{vib}, see Chapter 2.3.3). Free energy calculations were also performed on the two known Z'=2 forms of molecule XXIII,[2] which were outside the scope of the Z'=1 searches performed in Chapter 5; experimental forms C (CSD refcode XAFPAY02) and E (XAFPAY04) were optimised with DFTB3-D3 and then with the $\Psi_{mol}^{PBE0+FIT}$ method, and then their phonon modes were calculated.

The crystal structures output by the $\Psi_{mol}^{PBE0+FIT}$ optimisations had to be re-minimised with DFTB3-D3 before performing phonon calculations using symmetric finite displacements. These re-minimisations required one order of magnitude tighter thresholds than the intermediate DFTB3-D3 optimisations. The Brillouin zone sampling was initially formed by constructing supercells with minimum cell length of 10 Å,[40] but some were expanded to ensure that a similar number of atoms were included in each supercell to enhance error compensation.

7.2.4 Assessment of the results

The results of the crystal structure refinement procedure were assessed by whether the experimental forms and the significant computer-generated crystal structures from the original CSP studies had been reproduced, as well as by the quality of the reproductions and the relative energies of the matches. Two structures were considered to match if the Crystal Packing Similarity tool could overlay 15/15 molecules wit 20% distance and 20° angle tolerances. In a few cases the tolerances had to be slightly increased to find matches with some of the computer-generated significant crystal structures.

Some CrystalOptimizer[21] (see Chapter 2.4.2.2) minimisations were also performed to verify how the geometries and energies obtained after the optimisations with DFTB3-D3 and $\Psi_{mol}^{PBE0+FIT}$ compared with those obtained with a more computationally expensive Ψ_{mol} method. Thus the search-generated crystal structures that upon optimisation ended up matching the experimental forms and the significant computer-generated crystal structures from the original CSP studies were re-minimised with CrystalOptimizer at the PBE0 6-31G(d,p) level of theory for both intramolecular interactions and charge density calculations. The CDFs identified by the AUTODOF

programme (see Chapter 6.2.4),[41] which are shown in Appendix Figure 7.7, were treated as explicit CrystalOptimizer variables. AUTODOF was used because it required a small amount of human time to setup the calculations and for the affordable computational cost of the resulting optimisations, despite its weaknesses (see Chapter 6).

Furthermore, the F_{vib} values obtained with DFTB3-D3 were compared with those obtained using the $\Psi_{mol}^{PBE0+FIT}$ model. Rigid-body harmonic phonon calculations were performed with DMACRYS on the same crystal structure for which DFTB3-D3 F_{vib} values had been computed. Supercells were constructed using the methodology proposed by Nyman and Day,[42] with a target *k*-point distance of 0.12 $Å^{-1}$ (see Chapter 2.3.3.1). This approach assumes that all the molecular modes make an almost identical contribution to F_{vib},[42, 43] in contrast to the DFTB3-D3 calculations that allow the coupling of the molecular and lattice modes.

7.3 Results and discussion

7.3.1 Structures matching the experimentally known forms

7.3.1.1 Quality of the reproductions

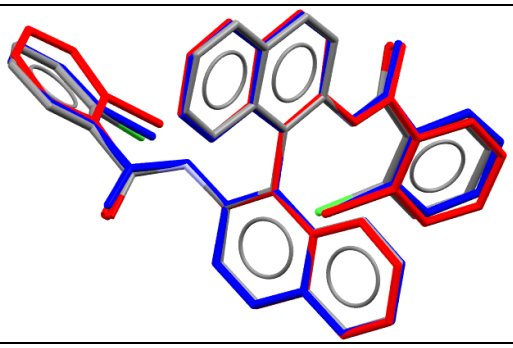
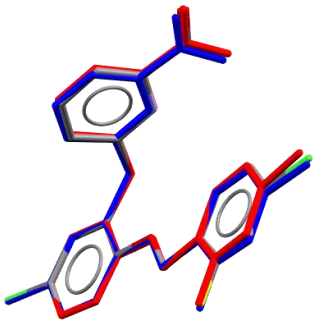
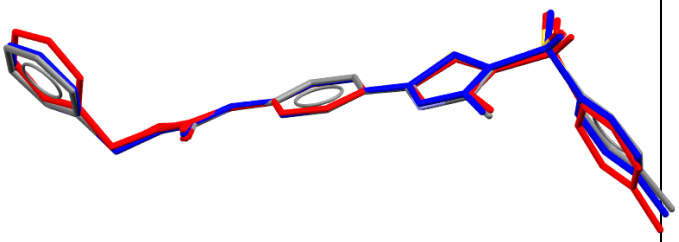
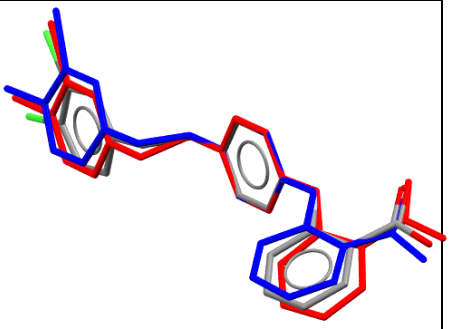
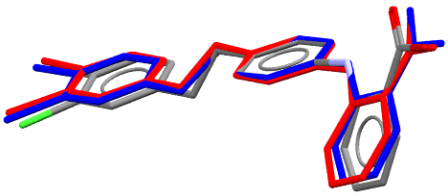
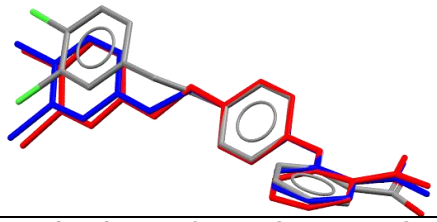
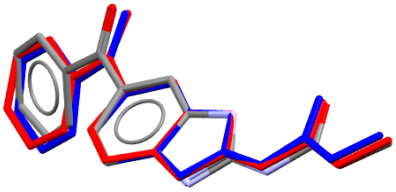
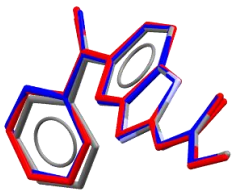
The quality of reproduction of each solved experimental Z'=1 crystal structure of the five molecules and of the molecular conformations within them are shown in Table 7.1. The comparisons were performed with the lowest energy matching crystal structures after the DFTB3-D3 and $\Psi_{mol}^{PBE0+FIT}$ optimisations. These reproductions are contrasted with those achieved upon performing more demanding CrystalOptimizer re-minimisations.

Table 7.1: Accuracy of the reproductions of the experimental crystal structures ($RMSD_{15}$) and experimental conformations ($RMSD_1$, see Table 7.2 for a visual comparison). Note that the molecular conformations were treated as rigid in the final $\Psi_{mol}^{PBE0+FIT}$ optimisations with DMACRYS, and so the $RMSD_1$ values were not affected.

	Reproduction of the experimental conformations ($RMSD_1$)			Reproduction of the experimental crystal lattices ($RMSD_{15}$)			
Molecule (CSD refcode)	**search /Å**	**DFTB3-D3/Å**	**CrystalOptimizer/Å**	**search/ Å**	**DFTB3-D3/Å**	**$\Psi_{mol}^{PBE0+FIT}$/Å**	**CrystalOptimizer/Å**
XXVI (XAFQIH)	0.200	0.133	0.123	0.533	0.425	0.376	0.294
GSK269984B (BIFHOP)	0.093	0.071	0.094	0.219	0.329	0.212	0.210
XXIII (A, XAFPAY)	0.236	0.216	0.212	0.689	0.680	0.663	0.443
XXIII (B, XAFPAY01)	0.278	0.180	0.159	0.626	0.528	0.415	0.373
XXIII (D, XAFPAY03)	0.310	0.278	0.237	0.511	0.544	0.583	0.530
XX (OBEQIX)	0.227	0.137	0.101	0.455	0.469	0.390	0.218
Mebendazole A (TUXPEJ)	0.201	0.176	0.143	0.465	0.408	0.402	0.321
Mebendazole C (YULGIW)	0.096	0.067	0.066	0.825	0.763	0.313	0.280
AVERAGE	**0.205**	**0.157**	**0.142**	**0.540**	**0.518**	**0.419**	**0.334**

Table 7.2 shows visually how well each experimental conformation was reproduced with the various methods.

Table 7.2: For each molecule, overlay of the experimental molecular conformation/s (coloured by element) with the conformation/s contained in the lowest-energy matching crystal structure after the DFTB3-D3 optimisations (in red), and after the CrystalOptimizer re-minimisations (in blue). The $RMSD_1$ for overlaying the experimental and optimised conformations are also indicated. Hydrogen atoms are not shown for clarity.

	
Molecule XXVI (XAFQIH), DFTB3-D3 $RMSD_1$ = 0.133 Å, CrystalOptimizer $RMSD_1$ = 0.123 Å	**GSK269984B (BIFHOP)**, DFTB3-D3 $RMSD_1$ = 0.071 Å, CrystalOptimizer $RMSD_1$ = 0.094 Å
	
Molecule XX (OBEQIX), DFTB3-D3 $RMSD_1$ = 0.137 Å, CrystalOptimizer $RMSD_1$ = 0.101 Å	**Molecule XXIII form A (XAFPAY)**, DFTB3-D3 $RMSD_1$ = 0.216 Å, CrystalOptimizer $RMSD_1$ = 0.212 Å
	
Molecule XXIII form B (XAFPAY01), DFTB3-D3 $RMSD_1$ = 0.180 Å, CrystalOptimizer $RMSD_1$ = 0.159 Å	**Molecule XXIII form D (XAFPAY03)**, DFTB3-D3 $RMSD_1$ = 0.278 Å, CrystalOptimizer $RMSD_1$ = 0.237 Å
	
Mebendazole form A (TUXPEJ), DFTB3-D3 $RMSD_1$ = 0.176 Å, CrystalOptimizer $RMSD_1$ = 0.143 Å	**Mebendazole form C (YULGIW)**, DFTB3-D3 $RMSD_1$ = 0.067 Å, CrystalOptimizer $RMSD_1$ = 0.066 Å

DFTB3-D3 optimisations were effective at improving all the molecular conformations from the search-generated crystal structures, with a 23% reduction in the

average $RMSD_1$. The improvement in the overall crystal packing was more variable, with some becoming slightly worse and some slightly better, with a mere 4% improvement in the average $RMSD_{15}$. Keeping the conformations rigid but optimising the intermolecular interactions with a higher quality wave-function improved the reproductions even further, leading to an overall reduction of the average $RMSD_{15}$ of 22% after the $\Psi_{mol}^{PBE0+FIT}$ minimisations.

Both the molecular and crystalline geometries were of a slightly worse quality than those obtained after the CrystalOptimizer re-minimisations, which improved the average $RMSD_1$ by 31% and an average $RMSD_{15}$ by 38%. However, the intermediate DFTB3-D3 refinements followed by the $\Psi_{mol}^{PBE0+FIT}$ optimisations could be carried out at a much lower computational cost and yet successfully reproduce the experimental crystal structures.

7.3.1.2 Relative E_{latt} rankings

The energy ranking of the crystal structures matching the experimentally known forms and their stability relative to the global minimum in E_{latt} at each stage of the refinement procedure are shown in Table 7.3, and compared to those obtained in the original studies. Visual representations of the crystal energy landscapes obtained after the optimisations with the $\Psi_{mol}^{PBE0+FIT}$ method are shown in Figure 7.1.

Table 7.3: Energy ranking and energy difference with the global minimum at each stage of the refinement of the crystal structures matching the solved experimental forms. These values are compared with their ranking in the original CSP studies. Note that the crystal structures of the two tautomers of mebendazole are ranked together.

Molecule (CSD refcode)	Ranking after search	ΔE_{latt} after search /kJ mol^{-1}	Ranking after DFTB3-D3	ΔE_{latt} after DFTB3-D3/kJ mol^{-1}	Ranking after Ψ_{mol}^{PBE0+F}	ΔE_{latt} after $\Psi_{mol}^{PBE0+FIT}$ /kJ mol^{-1}	Ranking in original CSP study	ΔE_{latt} in original CSP study/kJ mol^{-1}
XXVI (XAFQIH)	262	20.5	208	25.44	2	1.96	2	0.49
GSK269984B (BIFHOP)	54	6.63	1803	37.2	1	0	1	0
XXIII (A, XAFPAY)	4311	20.57	3670	33.8	232	12.31	280	13.52
XXIII (B, XAFPAY01)	55	7.72	805	23.49	3	0.82	1	0
XXIII (D, XAFPAY03)	1318	16.28	3748	33.99	114	10.25	85	9.21
XX (OBEQIX)	10	3.86	10	2.94	1	0	1	0
Mebendazole A (TUXPEJ)	2	1.37	14	4.32	1	0	1	0
Mebendazole C (YULGIW)	92	9.64	101	8.96	15	5.78	4	2.55

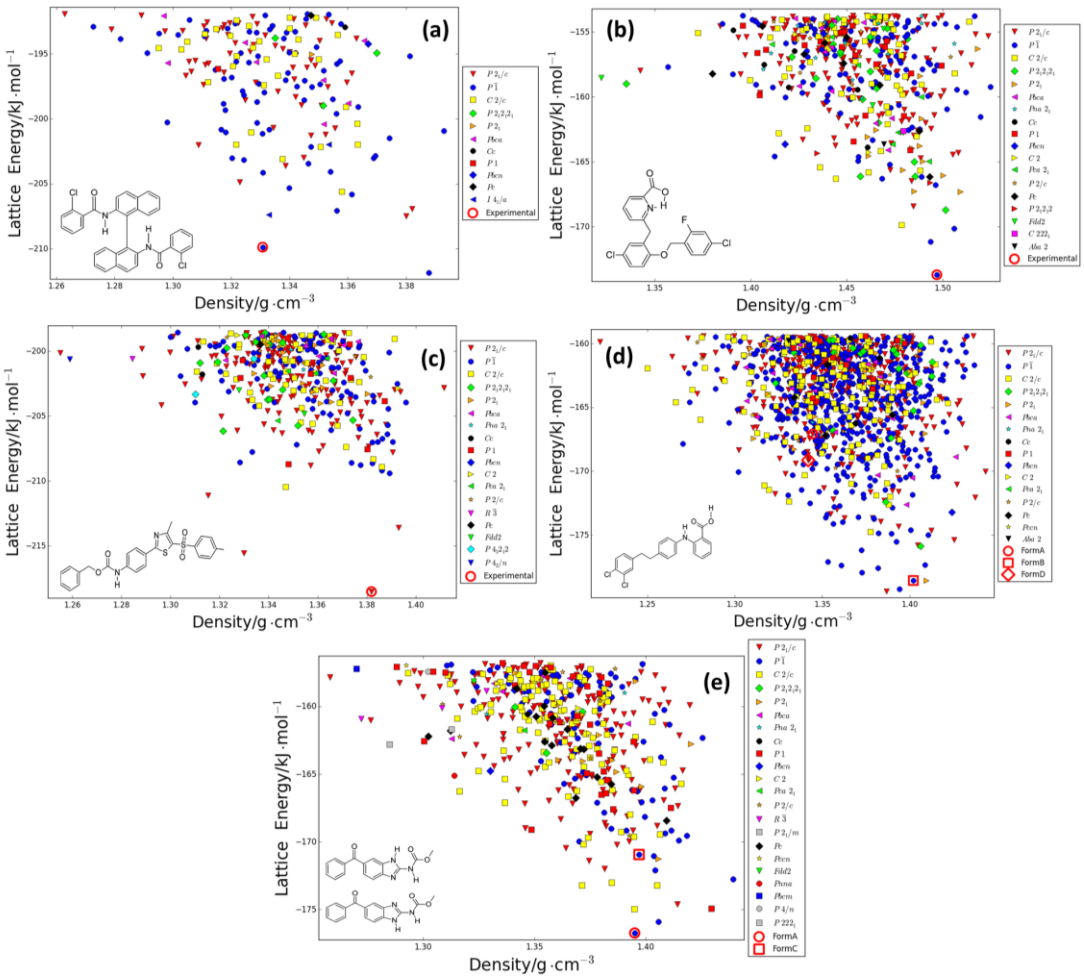


Figure 7.1: Plots of the crystal energy landscapes obtained after the final optimisations with the $\Psi_{mol}^{PBE0+FIT}$ method of (a) molecule XXVI (b) GSK269984B (c) molecule XX (d) molecule XXIII and (e) mebendazole. Each point on the plots corresponds to a separate crystal structure, labelled according to its space group. The structures matching the known experimental forms are indicated. The relative energies of the Z'=2 polymorphs of XXIII (XAFPAY02 for form C and XAFPAY04 for form E),[2] which were outside the scope of the searches performed in Chapter 5, were calculated independently and are only shown in Figure 7.3.

Table 7.3 shows that DFTB3-D3 energies are inadequate for CSP and in most cases the experimental crystal structures are ranked worse than after the searches. On the other hand, the final optimisations of just U_{inter} (*i.e.* keeping the DFTB3-D3 optimised molecular conformations rigid) with the $\Psi_{mol}^{PBE0+FIT}$ model drastically improved the relative energies. The crystal structures matching the experimental forms were ranked similarly to the original CSP studies, which had been performed with more expensive methods. For three experimental crystal structures (molecule XX, GSK269984B, and mebendazole form A) a match was the global minimum. Matches to experimental form B of molecule XXIII, form C of mebendazole and XXVI were all within a few kJ·mol^{-1} of the global minimum, while the structures representing forms A and D of molecule XXIII were higher in energy, although still within a sensible energy window considering the

issues with predicting the stability of the polymorphs of XXIII with any method (see Chapter 7.3.5).[2]

Given the modest improvements they brought to the reproductions of the experimental crystal structures, and the poor E_{latt} ranking, the DFTB3-D3 intermediate optimisations may appear to not have been important to obtain the promising results shown in Table 7.3 and Figure 7.1. To verify whether this was really the case, for the 100 lowest energy crystal structures of each molecule (except for molecule XXIII where 250 were optimised to include all three matches to the known forms), the search-generated structures were optimised directly with the $\Psi_{mol}^{PBE0+FIT}$ model using DMACRYS, without performing the intermediate DFTB3-D3 optimisations. Table 7.4 shows the results.

Table 7.4: Comparison of the relative energies of the structures matching the experimental forms shown in Table 7.3 with those obtained after optimising the same search-generated structures directly with the $\Psi_{mol}^{PBE0+FIT}$ model, *i.e.* without performing the intermediate DFTB3-D3 optimisations.

Experimental crystal structure (CSD refcode)	ΔE_{latt} after $\Psi_{mol}^{PBE0+FIT}$ with DFTB3-D3 intermediate optimisation/kJ mol^{-1}	ΔE_{latt} after $\Psi_{mol}^{PBE0+FIT}$ without DFTB3-D3 intermediate optimisation/kJ mol^{-1}
XXVI (XAFQIH)	1.96	51.67
GSK269984B (BIFHOP)	0	0
XXIII A (XAFPAY)	12.31	17.25
XXIII B (XAFPAY01)	0.82	10.51
XXIII D (XAFPAY03)	10.25	14.58
XX (OBEQIZ)	0	13.13
Mebendazole A (TUXPEJ)*	0	0
Mebendazole C (YULGIW)*	5.78	12.78

With the exceptions of GSK269984B and the A-tautomer of mebendazole, the relative energies of the structures matching the known experimental forms were significantly higher in the absence of the intermediate DFTB3-D3 optimisations. The values in Table 7.4 could have been worse if the same analysis had been carried out on all the thousands of structures that had been optimised with DFTB3-D3 and $\Psi_{mol}^{PBE0+FIT}$ (see Chapter 7.2.2). The importance of the DFTB3-D3 optimisations is particularly evident for XXVI, the largest and most flexible molecule: the match to the experimental form becomes so energetically unfavourable in the absence of the intermediate step that it would be discarded as an irrelevant crystal structure in a CSP study. The most probable explanation is that the geometries of the search-generated structures were poor, because of the approximate models used by CrystalPredictor.[25] However, the DFTB3-D3 intermediate optimisations of all the intra- and intermolecular degrees of freedom produce better starting points for the cheap final optimisations of just the intermolecular interactions with an improved wave-function. In the absence of this intermediate step, expensive optimisations of both intra- and intermolecular degrees of freedom with a high-quality wave-function (*e.g.* with CrystalOptimizer) are required. This confirms that performing an intermediate DFTB3-D3 optimisation of search-generated

crystal structures can limit the overall computational cost without worsening the quality of the results.

7.3.2 The other significant crystal structures

A question that arises is whether the refinement methodology produced not only matches to the known experimental forms but also to the other as yet unfound significant crystal structures found in the original CSP studies (see Chapter 5.2.2.3 for details). As shown in Appendix Tables 7.2-7.6, the vast majority of these structures were reproduced by the refinement procedure: out of the 180 targeted PPMs for the five molecules, only 21 were not found in the set of $\Psi_{mol}^{PBE0+FIT}$ optimised crystal structures. However, in eight of those cases they were already missing after the search, and for a further nine the generated structures were poor matches that had been classified as 'probably found' in Chapter 5. Only four significant structures that had been considered as 'certainly found' after the searches were lost in the refinement, and these are either towards the high energy end of the sample, or similar types of packing are present in the crystal energy landscape.

Nonetheless, some of the matches to the significant structures had high relative energies, above 20 kJ·mol^{-1} of the global minimum, and/or had poor overlays with those found in the original CSP studies. This can reflect a huge sensitivity to the starting points of the optimisations and to subtle changes in conformation, particularly when hydrogen bonds can be intra- or intermolecular as in GSK269984B.[5] This is further highlighted by how the relative energies and the geometries were often different from the original CSP studies in both the DFTB3-D3 and $\Psi_{mol}^{PBE0+FIT}$ optimisations and in the CrystalOptimizer re-minimisations.

Since the great majority of these crystal structures have not yet been experimentally found, it is impossible to determine which of these energies and geometries are more realistic. However, it is a success of the refinement methodology used in this chapter that most of the competitive structures found in the original CSP studies were reproduced as low energy and at a reduced computational cost.

7.3.3 Phonon calculations

The calculation of phonons by DFTB3-D3 methods proved difficult to automate. The recipe of only needing supercells to have a minimum length of 10 Å proved insufficient to converge the Brillouin sampling so that the difference in F_{vib} between structures could be converged below 1 kJ·mol^{-1}. However, expanding some supercells to guarantee that a similar number of atoms was present in each improved the convergence.

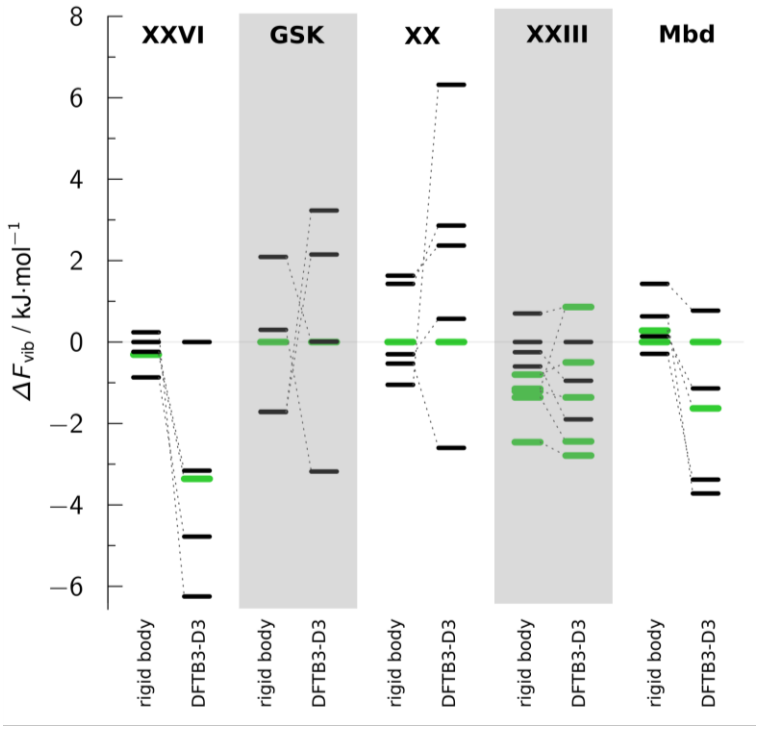


Figure 7.2: For some key CSP-generated crystal structures of each molecule, the relative vibrational contributions ΔF_{vib}, which can be added to E_{latt} to calculate the Helmholtz free energy (A, see Equation 2.37). For each molecule, ΔF_{vib} is calculated relative to the structure which is the E_{latt} global minimum after the $\Psi_{mol}^{PBE0+FIT}$ optimisations, for which ΔF_{vib} = 0. The structures matching the experimentally known forms are indicated in green. The rigid-body modes are pure lattice modes, calculated with the $\Psi_{mol}^{PBE0+FIT}$ model using DMACRYS. Full details about the crystal structures and their energies can be found in Appendix Table 7.7.

Overall, the DFTB3-D3 phonons do not cause significant re-ranking; the most notable exception is for molecule XXVI, where the experimental crystal structure becomes lower in free energy than the E_{latt} global minimum (see Appendix Table 7.7).

It is clear that the inclusion of all the molecular and lattice modes in the DFTB3-D3 calculations leads to a wider range of free energy differences than when only the rigid-body phonons are considered. This is because the packing forces in different polymorphs can affect the lowest frequency (*i.e.* the most flexible) molecular phonon modes, which makes the calculation of the coupling of intra- and intermolecular modes important to compute accurate free energy differences. Hence the rigid-body phonons calculated with the $\Psi_{mol}^{PBE0+FIT}$ method described in Chapter 7.2.4 can be unrealistic for large and flexible molecules as they ignore this coupling.

7.3.4 Computational cost

Table 7.5: **Comparison between the computational costs needed to optimise and re-rank the search-generated structures in the original CSP studies and with the refinement method outlined in this chapter. See Appendix Table 7.8 for a breakdown of the computational cost of the latter.**

Molecule	Saving in CPU hours	% difference
XXVI	~220,000	-90
GSK269984B	*Not recorded*	/
XX	~70,000	-70
XXIII	~40,000	-70
Mebendazole	~2,000	-30

Table 7.5 shows that the savings in computational cost with the refinement procedure used in this study were large. The smallest savings (~30%) were for the two tautomers of mebendazole, which is the least complex molecule in terms of size and conformational flexibility, and the greatest were for molecule XXVI (~90%). Some of this reduction in computational cost would be due to the use of different computer clusters, but the savings from replacing CrystalOptimizer and sometimes intermediate optimisations with DFTB3-D3 followed by $\Psi_{mol}^{PBE0+FIT}$ calculations are drastic and increase with molecular size and flexibility. This is very important in the context of this thesis, as it seems that introducing an intermediate DFTB3-D3 optimisation step could aid the scaling of CSP studies to molecules of pharmaceutical interest.

The DFTB3-D3 phonon calculations were much more expensive than the $\Psi_{mol}^{PBE0+FIT}$ ones, as shown in Appendix Table 7.8. However, DFTB3-D3 computes both the lattice and molecular modes at a reasonable computational cost, while in the Ψ_{mol} method the latter are assumed not to vary between different polymorphs. On the other hand, calculating all the phonon modes with DFT-D for several crystal structures is often unfeasibly expensive.[2, 44]

7.3.5 Discussion

This chapter pioneers the use of DFTB methods for CSP studies of pharmaceuticals. Full DFTB3-D3 optimisations seem to be effective as an intermediate step in the final refinement stage of CSP. Although they provide unreliable energy rankings, they improve the geometries of both the crystal structures and the molecular conformations. In particular, the results shown in Table 7.1 and Table 7.2 suggest that DFTB3-D3 is of similar accuracy to CrystalOptimizer optimisations in reproducing the experimental conformations, as judged by $RMSD_1$ values. The only exception is form D of molecule XXIII, where the optimisations have found a visually different conformation from the experimental one because of a poor search-generated starting point.

This intermediate refinement step provides a good starting point for a final calculation of E_{latt} with an improved wave-function, which in this study was performed via

an optimisation of the intermolecular interactions with the $\Psi_{mol}^{PBE0+FIT}$ model using DMACRYS. This procedure led to slightly worse reproductions of the experimental crystal structures in terms of $RMSD_{15}$ compared to the CrystalOptimizer optimisations, but to energy rankings similar to those obtained in the original CSP studies at a lower computational cost (see Table 7.5).

The cost savings brought by the intermediate DFTB3-D3 optimisations are important for extending CSP to the large and flexible molecules in pharmaceutical development. DFTB3-D3 scales as $N \cdot ln(N)$, in contrast to N^3 for periodic DFT-D methods,[26] where N is the number of atoms in the unit cell. Furthermore, the cost of the Ψ_{mol} optimisations with CrystalOptimizer scales badly with the number of CDFs that are treated as explicit variables (see Chapter 6.3.3).[21]

The disappointment is that the relative energies produced by DFTB3-D3 are too poor (see Appendix Figures 7.1-7.6) to allow a confident, drastic reduction in the number of structures to be investigated with more accurate methods. This is probably due to the theoretical limitations of this methodology.[24, 45] It appears that despite the clear advantages of optimising all intra- and intermolecular degrees of freedom at once, accurate crystal energy landscape cannot be calculated with a simplified wave-function. The long-range electrostatic interactions are described by atomic charges in both DFTB3-D3 and in the crystal generation stage, but by atomic multipoles in the final energy evaluation, and the quality of the electrostatic model is very critical in CSP, particularly in the presence of hydrogen bonds;[46] this may explain why the final energy ranking appears to be much more accurate than the initial and intermediate ones. However, the improvement in the crystalline geometries and in particular in the molecular conformations brought by DFTB3-D3 optimisations was fundamental to perform accurate calculations with the $\Psi_{mol}^{PBE0+FIT}$ model, as shown in Table 7.4.

The phonon calculations with the Ψ_{crys} approach, which can be affordably performed for several crystal structures only using DFTB3-D3, are conceptually different from those performed with the rigid-body Ψ_{mol} method[42] in allowing the coupling of the molecular and lattice modes. This leads to an even larger spread of F_{vib} values than using the rigid-body model (see Figure 7.2), which is already comparable with polymorphic energy differences,[42, 43] raising the question as to whether thermal effects are likely to be more significant in determining the relative thermodynamic stability of pharmaceutical polymorphs than they are for more rigid molecules, particularly in the presence of large conformational and density differences. DFTB3-D3 has the potential of making the modelling of all phonon contributions to free energies more widespread in CSP. The DFTB3-D3 F_{vib} values can be added to E_{latt} values calculated with the Ψ_{mol} method (like in this chapter) or with an accurate Ψ_{crys} method like DFT-D. However, some

benchmarking against experimental data (*e.g.* by Terahertz time-domain spectroscopy)[47] is needed to test the actual accuracy of the DFTB3-D3 phonon calculations.

Furthermore, thermal expansion is likely to be important for both absolute[44] and relative free energies.[14, 43, 48] Although this phenomenon can be modelled with the quasi-harmonic approximation,[43] the validity of this approach is questionable for large molecules where methyl groups may be rotating and phenyl rings may have different large amplitude of motions, depending on the packing. Hence, the potential energy surface should be explored more explicitly, for example by molecular dynamics simulations,[49] as none of the methodologies currently available can fully describe all the possible thermal effects that can affect crystal structures and their relative free energy ranking.

Among the five molecules considered in this chapter, polymorphic molecule XXIII poses the most significant challenges to CSP methodologies. Looking at Figure 7.1d it is clear that although the three Z'=1 polymorphs are all successfully reproduced by the refinement methodology and are ranked among the most competitive crystal structures, nothing differentiates them from a plethora of alternatives. Only form B is close to the bottom of the crystal energy landscape, while forms A and D are above 10 kJ·mol^{-1} of the global minimum together with several unfound competitors. Two further Z'=2 polymorphs have been characterised experimentally. Although they were outside the scope of the searches, their experimental crystal structures were optimised with DFTB3-D3 and $\Psi_{mol}^{PBE0+FIT}$ to calculate the relative energy rankings of the structures matching to the five known forms. A comparison with the energy rankings obtained using different models in the 6th Blind Test[2] is shown in Figure 7.3. It is clear that the relative stabilities of the known forms of molecule XXIII are very sensitive to the energy model. This highlights the challenges of producing an accurate stability ranking between polymorphs, and more CSP studies performed in collaboration with experimentalists in industry and academia are needed to provide benchmarks and improve the models.

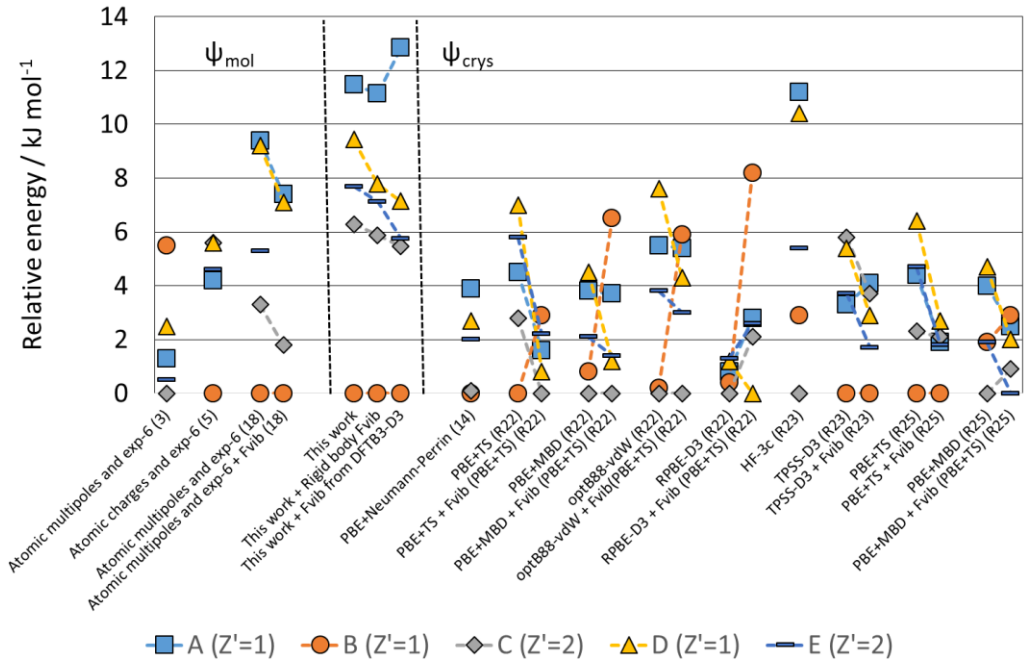


Figure 7.3: Relative energies of the five polymorphs of molecule XXIII calculated in this study compared with those reported by the participants of the 6th Blind Test. Note that values linked by dashed lines denote changes from adding free energy estimates. The number in parentheses corresponds to the group identifier, with R denoting participants who only optimised crystal structures generated by others.[2]

7.4 Conclusion

An intermediate optimisation with the Ψ_{crys} semi-empirical quantum-mechanical method DFTB3-D3 revealed to be suitable for reducing the computational cost of the final refinement stage of CSP of five large and flexible molecules. Although DFTB3-D3 does not produce a sufficiently accurate E_{latt} energy ranking, it improves the geometrical representation of the crystal structures and in particular of the molecular conformations, and it outputs good starting points for better energy calculations to be carried out with a more accurate wave-function. In this chapter, this final evaluation of E_{latt} was performed by optimising the intermolecular interactions with the $\Psi_{mol}^{PBE0+FIT}$ model, which produced results comparable in quality to those obtained in the original CSP studies but at a largely reduced computational cost. On top of producing accurate matches to the experimental forms, and placing them close to the bottom of the crystal energy landscapes, this refinement methodology was also successful at reproducing most of the significant as yet unfound crystal structures found in the original CSP studies. It is likely that accurate crystal energy landscapes could also be produced following the DFTB3-D3 optimisations with other methods to accurately evaluate E_{latt}, *e.g.* single-point DFT-D calculations, although some testing is required. Furthermore DFTB3-D3 can perform affordable calculations of both the molecular and lattice phonon modes, which can be used to calculate the vibrational component of free energy. These calculations suggest that the

thermal effects may cause more significant re-ranking for flexible pharmaceuticals than for smaller, more rigid molecules.

These results are very important in the context of this thesis. First of all, DFTB3-D3 seems to be suitable to bridge the gap between the cheap approximate models used in the searches and the expensive accurate models needed to produce an accurate final crystal energy landscape. A combination of this refinement approach (or suitable alternatives based on the same principles) with the search workflow described in Chapter 5 could produce a complete CSP method cheaper than other competitors and more scalable to larger and more flexible targets, but of similar accuracy.

Furthermore, this study has confirmed that crystal structure vibrations can have important effects on the relative stability of polymorphs, and that the vibrational component of free energy depends not only on the lattice but also on the molecular modes, whose neglecting in Ψ_{mol} methods[42] can be the source of inaccuracies. DFTB3-D3 allows an affordable calculation of all phonon modes, and the resulting F_{vib} can be combined with E_{latt} computed with higher-quality models to produce a relatively cheap estimate of the Helmholtz free energy. However, some benchmarking of the DFTB3-D3 F_{vib} values is needed.

Finally, this study confirms that the difficulty of a CSP study is very molecule-dependent and not easily predictable. Two similarly linear extended molecules like XX and XXIII have very different crystal energy landscapes: while the prediction of the only known crystal structure of the first is straightforward, the latter has a complex polymorphic behaviour that reveals how current models still face significant challenges in determining the set of thermodynamically competitive crystal structures, let alone in predicting how predicted forms could be experimentally produced. CSP is far from a solved problem. Although the work discussed in this chapter is a step towards extending CSP studies to more molecules of pharmaceutical interest, significant developments are needed before computational modelling can predict which crystal structures would form under specific experimental conditions.

7.5 References

1. Iuzzolino, L.; Reilly, A. M.; McCabe, P.; Price, S. L., Use of Crystal Structure Informatics for Defining the Conformational Space Needed for Predicting Crystal Structures of Pharmaceutical Molecules. *Journal of Chemical Theory and Computation* **2017,** *13* (10), 5163-5171.

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

3. Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C.; Price, S. L.; Galek, P. T.; Day, G. M.; Cruz-Cabeza, A. J., Successful prediction of a model pharmaceutical in the fifth blind test of crystal structure prediction. *International Journal of Pharmaceutics* **2011,** *418* (2), 168-178.

4. 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 Crystallographica Section B-Structural Science* **2011,** *67*, 535-551.

5. 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. *Crystal Growth & Design* **2013,** *13* (6), 2396-2406.

6. Pantelides, C. C.; Adjiman, C. S.; Kazantsev, A. V., General Computational Algorithms for Ab Initio Crystal Structure Prediction for Organic Molecules. *Topics in Current Chemistry* **2014,** *345*, 25-58.

7. Vasileiadis, M.; Pantelides, C. C.; Adjiman, C. S., Prediction of the crystal structures of axitinib, a polymorphic pharmaceutical molecule. *Chemical Engineering Science* **2015,** *121*, 60-76.

8. Grimme, S.; Hansen, A.; Brandenburg, J. G.; Bannwarth, C., Dispersion-Corrected Mean-Field Electronic Structure Methods. *Chemical Reviews* **2016,** *116* (9), 5105-5154.

9. Hoja, J.; Tkatchenko, A., First-principles stability ranking of molecular crystal polymorphs with the DFT+MBD approach. *Faraday Discussions* **2018,** *Advance article*.

10. Beran, G. J. O., Modeling Polymorphic Molecular Crystals with Electronic Structure Theory. *Chemical Reviews* **2016,** *116* (9), 5567-5613.

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

12. Price, S. L.; Reutzel-Edens, S. M., The potential of computed crystal energy landscapes to aid solid form development. *Drug Discovery Today* **2016,** *21* (6), 912-923.

13. Price, S. L.; Brandenburg, J. G., Chapter 11 - Molecular Crystal Structure Prediction. In *Non-Covalent Interactions in Quantum Chemistry and Physics*, Elsevier: 2017; pp 333-363.

14. Buchholz, H. K.; Hylton, R. K.; Brandenburg, J. G.; Seidel-Morgenstern, A.; Lorenz, H.; Stein, M.; Price, S. L., Thermochemistry of Racemic and Enantiopure Organic Crystals for Predicting Enantiomer Separation. *Crystal Growth & Design* **2017,** *17* (9), 4676-4686.

15. Iuzzolino, L.; McCabe, P.; Price, S. L.; Brandenburg, J. G., Crystal structure prediction of flexible pharmaceutical-like molecules: Density functional tight-binding as an intermediate optimization method and for free energy estimation. *Faraday Discussions* **2018,** *Advance article*.

16. Brandenburg, J. G.; Grimme, S., Accurate Modeling of Organic Molecular Crystals by Dispersion-Corrected Density Functional Tight Binding (DFTB). *Journal of Physical Chemistry Letters* **2014,** *5* (11), 1785-1789.

17. Day, G. M.; Motherwell, W. D. S.; Jones, W., A strategy for predicting the crystal structures of flexible molecules: the polymorphism of phenobarbital. *Physical Chemistry Chemical Physics* **2007,** *9* (14), 1693-1704.

18. Day, G. M.; Cooper, T. G., Crystal packing predictions of the alpha-amino acids: methods assessment and structural observations. *CrystEngComm* **2010,** *12* (8), 2443-2453.

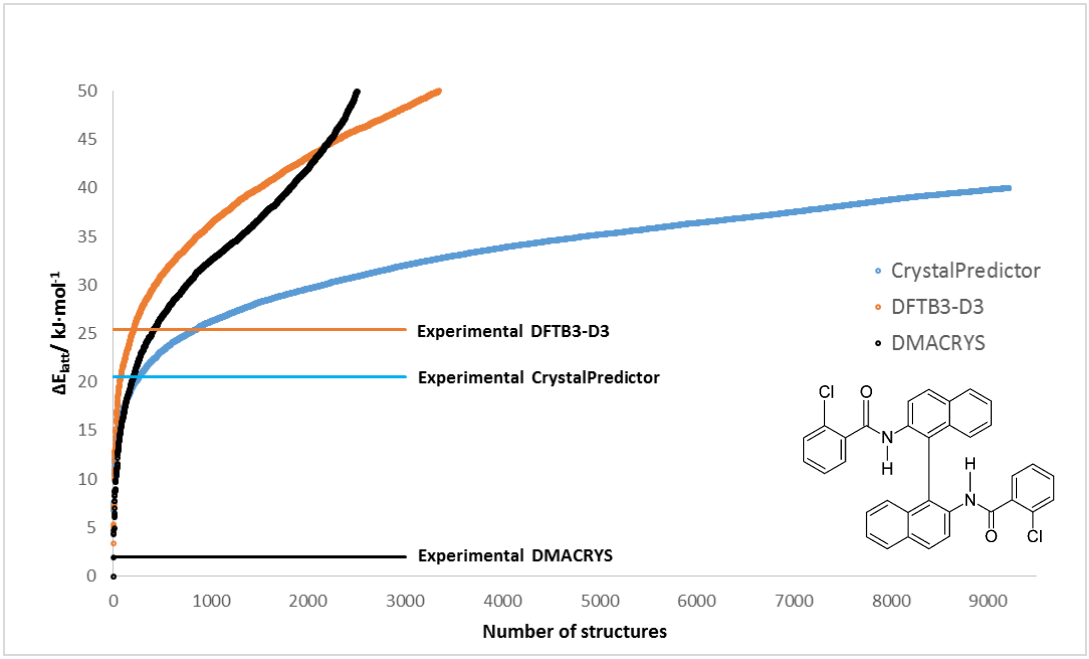
19. Bhardwaj, R. M.; Price, L. S.; Price, S. L.; Reutzel-Edens, S. M.; Miller, G. J.; Oswald, I. D. H.; Johnston, B.; Florence, A. J., Exploring the Experimental and Computed Crystal Energy Landscape of Olanzapine. *Crystal Growth & Design* **2013,** *13* (4), 1602-1617.

20. Price, L. S.; McMahon, J. A.; Lingireddy, S. R.; Lau, S. F.; Diseroad, B. A.; Price, S. L.; Reutzel-Edens, S. M., A molecular picture of the problems in ensuring structural purity of tazofelone. *Journal of Molecular Structure* **2014,** *1078*, 26-42.

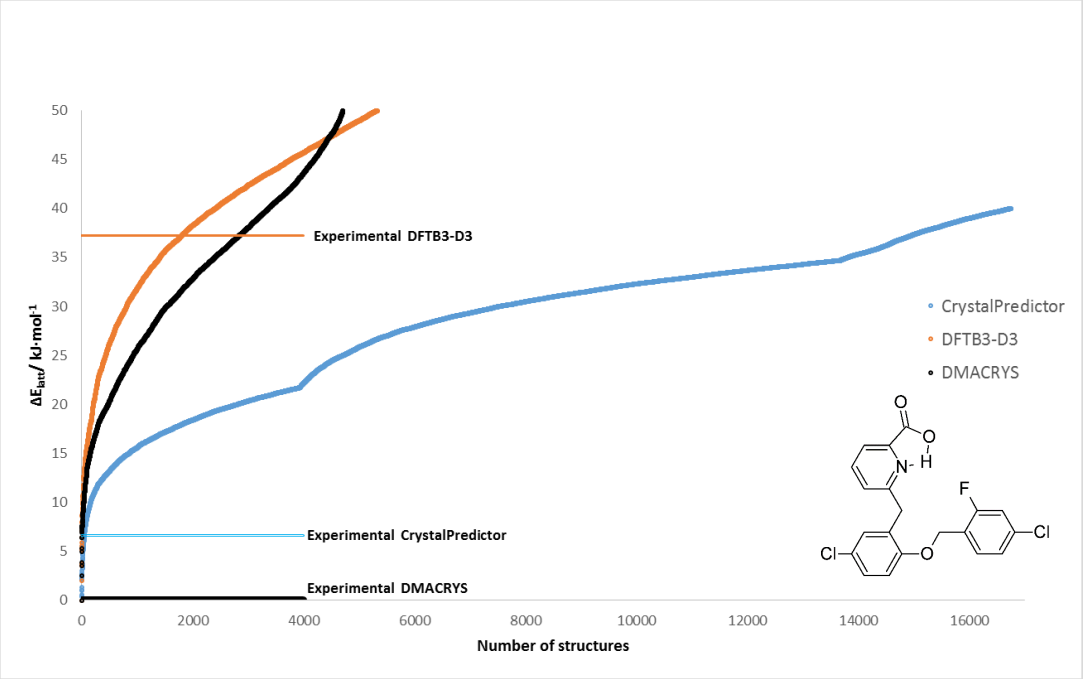
21. Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C., Efficient Handling of Molecular Flexibility in Lattice Energy Minimization of Organic Crystals. *Journal of Chemical Theory and Computation* **2011,** *7* (6), 1998-2016.
22. Price, S. L., Predicting crystal structures of organic compounds. *Chemical Society Reviews* **2014,** *43* (7), 2098-2111.
23. Brandenburg, J. G.; Hochheim, M.; Bredow, T.; Grimme, S., Low-Cost Quantum Chemical Methods for Noncovalent Interactions. *Journal of Physical Chemistry Letters* **2014,** *5* (24), 4275-4284.
24. Christensen, A. S.; Kubař, T.; Cui, Q.; Elstner, M., Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications. *Chemical Reviews* **2016,** *116* (9), 5301-5337.
25. Karamertzanis, P. G.; Pantelides, C. C., Ab initio crystal structure prediction. II. Flexible molecules. *Molecular Physics* **2007,** *105* (2-3), 273-291.
26. Aradi, B.; Hourahine, B.; Frauenheim, T., DFTB+, a Sparse Matrix-Based Implementation of the DFTB Method. *The Journal of Physical Chemistry A* **2007,** *111* (26), 5678-5684.
27. Gaus, M.; Goez, A.; Elstner, M., Parametrization and Benchmark of DFTB3 for Organic Molecules. *Journal of Chemical Theory and Computation* **2013,** *9* (1), 338-354.
28. Kubillus, M.; Kubař, T.; Gaus, M.; Řezáč, J.; Elstner, M., Parameterization of the DFTB3 Method for Br, Ca, Cl, F, I, K, and Na in Organic and Biological Systems. *Journal of Chemical Theory and Computation* **2015,** *11* (1), 332-342.
29. Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H., A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J Chem Phys* **2010,** *132* (15), 154104.
30. Brauer, B.; Kesharwani, M. K.; Kozuch, S.; Martin, J. M. L., The S66x8 benchmark for noncovalent interactions revisited: explicitly correlated ab initio methods and density functional theory. *Physical Chemistry Chemical Physics* **2016,** *18* (31), 20905-20925.
31. Dovesi, R.; Orlando, R.; Erba, A.; Zicovich-Wilson, C. M.; Civalleri, B.; Casassa, S.; Maschio, L.; Ferrabone, M.; De La Pierre, M.; D'Arco, P.; Noël, Y.; Causà, M.; Rérat, M.; Kirtman, B., CRYSTAL14: A program for the ab initio investigation of crystalline solids. *International Journal of Quantum Chemistry* **2014,** *114* (19), 1287-1317.
32. Spek, A. L. *PLATON, A Multipurpose Crystallographic Tool*, Utrecht University: Utrecht, The Netherlands, 2003.
33. Chisholm, J. A.; Motherwell, S., COMPACK: a program for identifying crystal structure similarity using distances. *Journal of Applied Crystallography* **2005,** *38*, 228-231.
34. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C., The Cambridge Structural Database. *Acta Crystallographica Section B-Structural Science Crystal Engineering and Materials* **2016,** *72*, 171-179.
35. Coles, S. J.; Threlfall, T. L.; Tizzard, G. J., The Same but Different: Isostructural Polymorphs and the Case of 3-Chloromandelic Acid. *Crystal Growth & Design* **2014,** *14* (4), 1623-1628.
36. 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. 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, D. J. *Gaussian 09, Revision D.01*, 2009.
37. Stone, A. J., Distributed Multipole Analysis: Stability for Large Basis Sets. *Journal of Chemical Theory and Computation* **2005,** *1* (6), 1128-1132.
38. Price, S. L.; Leslie, M.; Welch, G. W. A.; Habgood, M.; Price, L. S.; Karamertzanis, P. G.; Day, G. M., Modelling Organic Crystal Structures using Distributed Multipole and Polarizability-Based Model Intermolecular Potentials. *Physical Chemistry Chemical Physics* **2010,** *12* (30), 8478-8490.
39. 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. *Journal of Physical Chemistry* **1996,** *100* (18), 7352-7360.

40. Reilly, A. M.; Tkatchenko, A., Understanding the role of vibrations, exact exchange, and many-body van der Waals interactions in the cohesive properties of molecular crystals. *The Journal of Chemical Physics* **2013,** *139* (2), 024705-024705.
41. Nyman, J.; Reutzel-Edens, S. M., Crystal structure prediction is changing from basic science to applied technology. *Faraday Discussions* **2018,** *Advance Article*.
42. Nyman, J.; Day, G. M., Static and lattice vibrational energy differences between polymorphs. *Crystengcomm* **2015,** *17* (28), 5154-5165.
43. Nyman, J.; Day, G. M., Modelling temperature-dependent properties of polymorphic organic molecular crystals. *Physical Chemistry Chemical Physics* **2016,** *18* (45), 31132-31143.
44. Brandenburg, J. G.; Potticary, J.; Sparkes, H. A.; Price, S. L.; Hall, S. R., Thermal Expansion of Carbamazepine: Systematic Crystallographic Measurements Challenge Quantum Chemical Calculations. *Journal of Physical Chemistry Letters* **2017,** *8* (17), 4319-4324.
45. Elstner, M.; Seifert, G., Density functional tight binding. *Philosophical Transactions of the Royal Society A* **2014,** *372* (2011).
46. Day, G. M.; Motherwell, W. D. S.; Jones, W., Beyond the isotropic atom model in crystal structure prediction of rigid molecules: Atomic multipoles versus point charges. *Crystal Growth & Design* **2005,** *5* (3), 1023-1033.
47. Li, R. Y.; Zeitler, J. A.; Tomerini, D.; Parrott, E. P. J.; Gladden, L. F.; Day, G. M., A study into the effect of subtle structural details and disorder on the terahertz spectrum of crystalline benzoic acid. *Physical Chemistry Chemical Physics* **2010,** *12* (20), 5329-5340.
48. Cervinka, C.; Beran, G. J. O., Ab initio prediction of the polymorph phase diagram for crystalline methanol. *Chemical Science* **2018,** *9*, 4622-4629.
49. Rossi, M.; Gasparotto, P.; Ceriotti, M., Anharmonic and Quantum Fluctuations in Molecular Crystals: A First-Principles Study of the Stability of Paracetamol. *Physical Review Letters* **2016,** *117* (11).

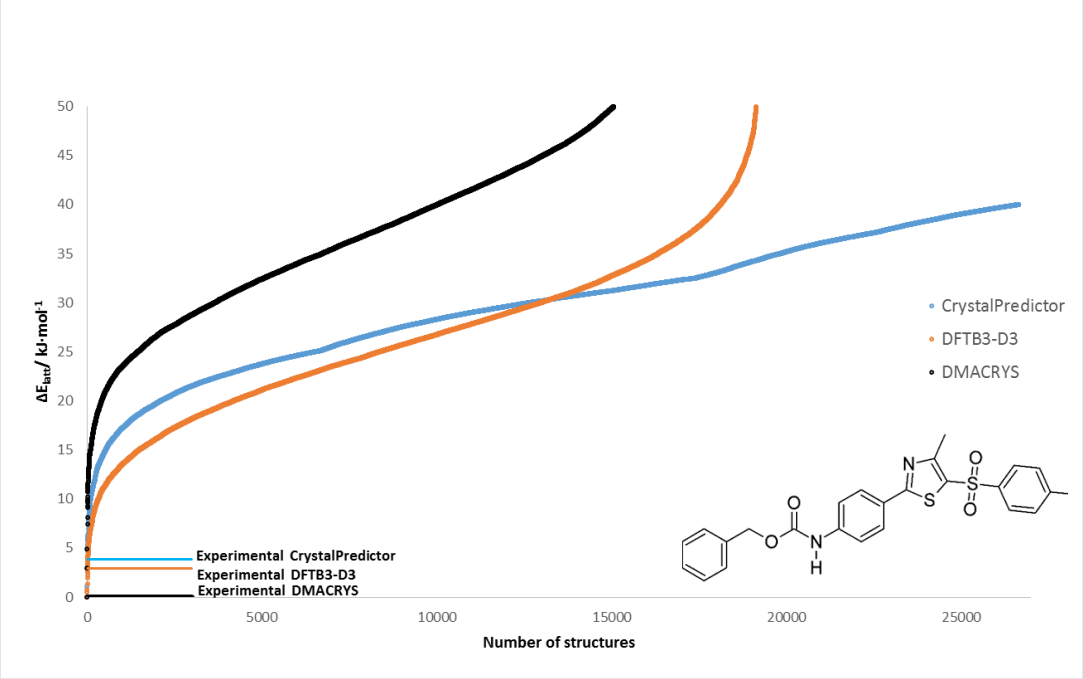
7.6 Appendix



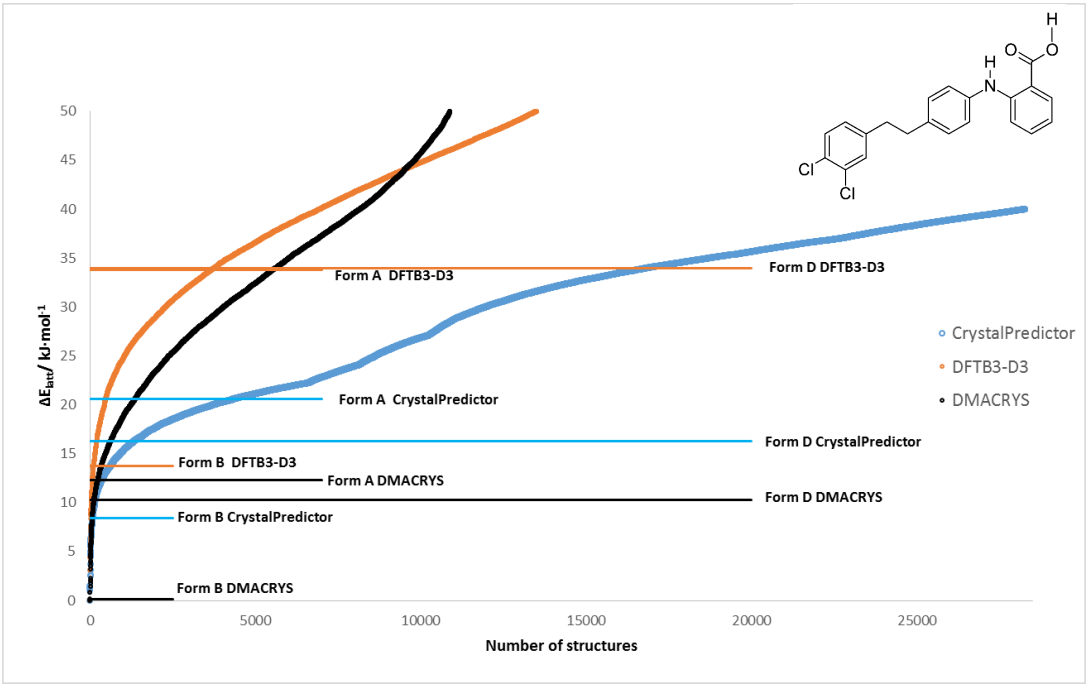
Appendix Figure 7.1: Plot showing the energy distribution of the computer-generated crystal structures of molecule XXVI at the various stages of the CSP refinement procedure outlined in this chapter.



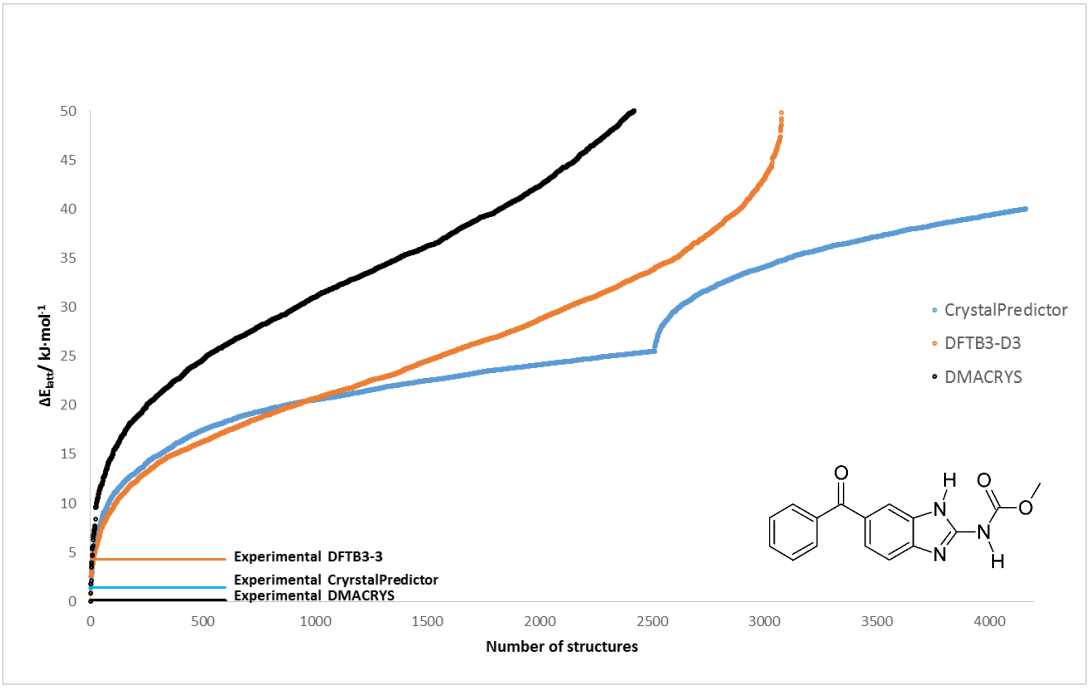
Appendix Figure 7.2: Plot showing the energy distribution of the computer-generated crystal structures of GSK269984B at the various stages of the CSP refinement procedure outlined in this chapter.



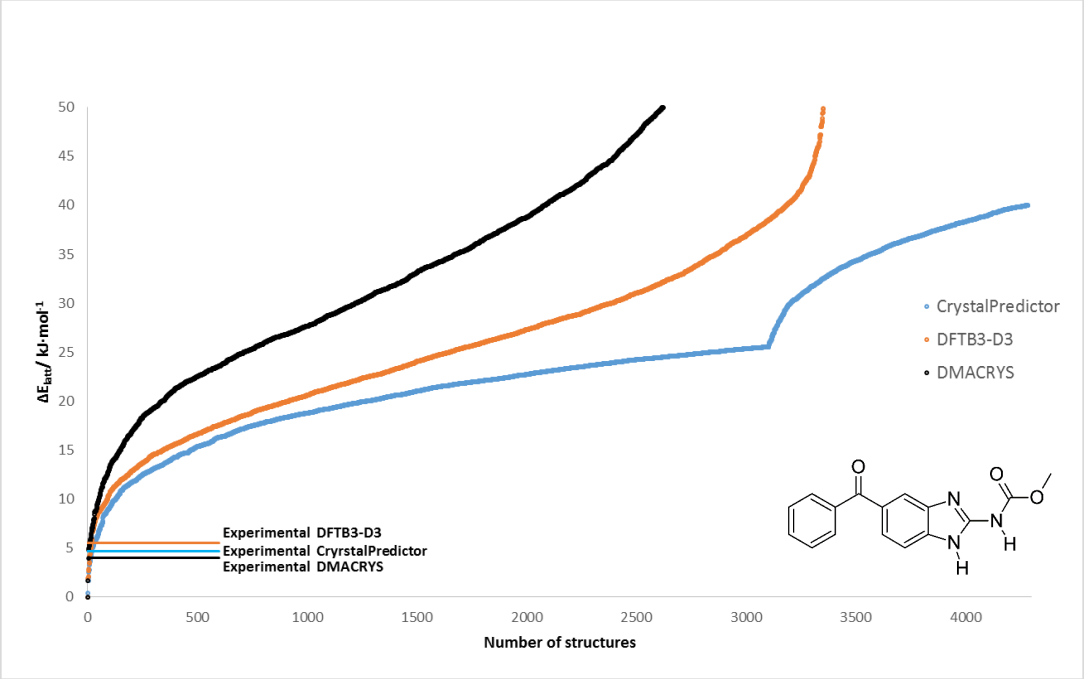
Appendix Figure 7.3: Plot showing the energy distribution of the computer-generated crystal structures of molecule XX at the various stages of the CSP refinement procedure outlined in this chapter.



Appendix Figure 7.4: Plot showing the energy distribution of the computer-generated crystal structures of molecule XXIII at the various stages of the CSP refinement procedure outlined in this chapter.



Appendix Figure 7.5: Plot showing the energy distribution of the computer-generated crystal structures of the A-tautomer of mebendazole at the various stages of the CSP refinement procedure outlined in this chapter.

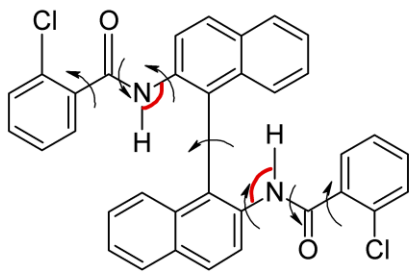


Appendix Figure 7.6: Plot showing the energy distribution of the computer-generated crystal structures of the C-tautomer of mebendazole at the various stages of the CSP refinement procedure outlined in this chapter.

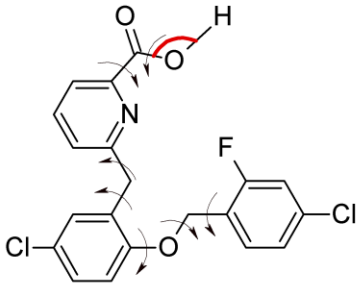
Appendix Table 7.1: For each molecule, origin of the reduction in the number of crystal structures that were taken to the final optimisation with the $\Psi_{mol}^{PBE0+FIT}$ model.

	# structures of mol. XXVI	#structures of GSK269984B	# structures of mol. XXIII	# structures of molecule XX	# structures of mbz A-tautomer	# structures of mbz C-tautomer
Original CrystalPredictor structures	9215	16744	28249	26650	4165	4284
Removing structures becoming duplicates with DFTB3-D3	8821	16569	27866	26278	4147	4267
Removing wrong molecules*	8583	16270	27866	26098	4124	4238
Clustering DFTB3-D3 with looser criteria	7534	13829	23682	19251	3083	3362
Applying the 50 kJ·mol^{-1} cut-off to DFTB3-D3 energies	3346	5238	13490	19146	3078	3352

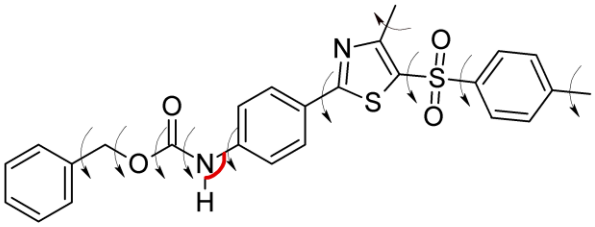
*In a few cases the DFTB3-D3 optimisations wrongly changed the covalent bonding of the molecule.



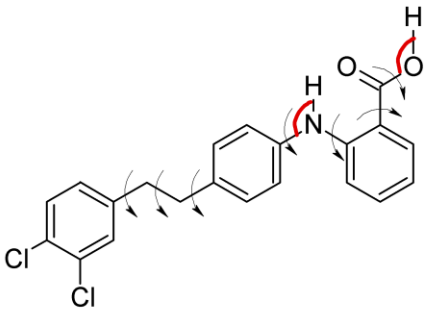
Molecule XXVI



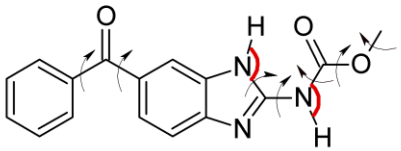
GSK269984B



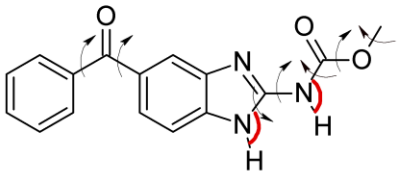
Molecule XX



Molecule XXIII



A-Mebendazole



C-Mebendazole

Appendix Figure 7.7: Chemical diagram of each molecule considered in this study, with indicated torsion angles (black arrows) and bond-angles (red arcs) treated as independent degrees of freedom in the CrystalOptimizer re-minimisations of some key search-generated crystal structures. The optimised CDFs were selected by AUTODOF.[41]

Appendix Table 7.2: Reproduction and ranking of the matches to the significant crystal structures of molecule XXVI that were found in the original CSP study. The structure highlighted in green matches the experimental form. The significant crystal structures for which a good match was not found are classified: in blue where they were not found by the search, in turquoise where they had a poor match in the search, and in red where structures were missed despite having a good match in the search. The structures in orange were found after $\Psi_{mol}^{PBE0+FIT}$ at ΔE_{latt} > 20 kJ·mol^{-1}. In a few cases, the distance and angle tolerances had to be increased to 30% and 30° respectively to obtain a match, and the $RMSD_{15}$ values for these structures are shown in italics.

Structure name	Found ?	Old CSP rank	Ranking after $\Psi_{mol}^{PBE0+FIT}$	Old CSP study ΔE_{latt} /kJ·mol^{-1}	$\Psi_{mol}^{PBE0+FIT}$ ΔE_{latt} /kJ·mol^{-1}	$RMSD_{15}$ after $\Psi_{mol}^{PBE0+FIT}$	CrystalOptimizer ΔE_{latt} /kJ·mol^{-1}
3525	YES	1	7	0.00	6.03	0.48	0.23
1600	YES	2	2	0.49	1.96	0.528	0.00
675	YES	3	10	2.60	6.97	0.284	3.45
38	YES	4	4	4.15	4.45	0.195	5.17
421	YES	5	10	5.43	6.97	0.286	3.43
3104	YES	6	33	5.65	10.90	0.178	6.63
615	YES	7	91	6.28	16.12	0.37	6.79
239	YES	8	29	6.38	10.63	0.389	6.94
2930	YES	9	161	6.56	18.99	0.421	3.70
354	YES	10	147	6.88	18.58	0.398	8.31
851	YES (high energy)	11	272	7.04	22.54	0.359	7.43
6460	YES	12	31	7.11	10.79	*0.931*	8.26
6335	YES	13	100	7.45	16.41	0.452	8.15
221	YES	14	31	7.46	10.79	0.483	8.29
2231	YES	15	5	7.57	4.77	0.33	6.32
2496	NOT IN SEARCH	16	\	7.93	\	\	\
185	YES (high energy)	17	369	8.10	24.61	0.621	9.05
4201	POOR IN SEARCH	18	\	8.21	\	\	\
314	YES	19	23	8.22	9.83	0.322	10.55
508	YES	20	132	8.29	18.02	0.298	7.16
4946	YES	21	165	19.14	16.30	0.475	5.09
6879	YES	22	115	8.51	17.18	0.229	10.33
506	YES	23	59	8.62	13.98	0.481	10.05
4842	YES	24	46	8.83	12.85	0.645	7.03
43	YES	25	3	9.02	4.35	0.439	9.03
1236	YES	26	8	9.15	6.24	0.318	9.48
1537	YES	27	12	9.16	7.76	0.334	9.73
188	YES	28	14	9.41	8.39	0.942	9.47
5126	YES	29	180	10.05	19.77	0.513	10.54
444	YES	30	6	10.12	4.92	0.137	9.64
544	YES (high energy)	31	272	10.28	22.54	0.377	7.44
686	YES	32	25	10.34	9.86	0.211	10.78
89	POOR IN SEARCH	33	\	10.44	\	\	\
20	YES	34	70	10.69	14.78	*0.932*	\
83	YES	35	88	10.82	15.93	0.398	10.74
2591	NO	132	\	17.03	\	\	\

Appendix Table 7.3: Reproduction and ranking of the matches to the significant crystal structures of GSK269984B that were found in the original CSP study. The structure highlighted in green matches the experimental form. The significant crystal structures for which a good match was not found are classified: in blue where they were not found by the search, in turquoise where they had a poor match in the search, and in red where structures were missed despite having a good match in the search. The structures in orange were found after $\Psi_{mol}^{PBE0+FIT}$ at ΔE_{latt} > 20 kJ·mol^{-1}. In a few cases, the distance and angle tolerances had to be increased to 30% and 30° respectively to obtain a match, and the $RMSD_{15}$ values for these structures are shown in italics.

Structure name	Found ?	Old CSP rank	Ranking after $\Psi_{mol}^{PBE0+FIT}$	Old CSP study ΔE_{latt} /kJ·mol^{-1}	$\Psi_{mol}^{PBE0+FIT}$ ΔE_{latt} /kJ·mol^{-1}	$RMSD_{15}$ after $\Psi_{mol}^{PBE0+FIT}$	CrystalOptimizer ΔE_{latt} /kJ·mol^{-1}
180Intra10	YES	1	1	0.00	0.00	0.135	0.00
90InterB36	YES (high energy)	2	684	0.53	22.50	0.525	17.13
180InterA11	YES (high energy)	3	543	2.06	20.84	0.407	16.72
180InterA8	YES	4	17	2.76	7.59	0.581	10.83
180InterB6	YES (high energy)	5	820	3.26	23.91	0.34	19.79
180Intra8	YES	6	179	3.55	15.50	0.375	8.56
180Intra38	POOR IN SEARCH	7	\	3.62	\	\	\
180InterB9	YES	8	278	3.80	17.36	0.394	8.58
180Intra76	YES	9	98	4.24	13.15	0.47	8.51
180InterA22	YES	10	269	4.36	17.24	0.726	17.97
90InterB6	NOT IN SEARCH	11	\	4.61	\	\	\
180Intra19	YES	12	23	4.66	8.28	0.189	10.27
180Intra74	YES	13	34	5.00	9.38	0.366	9.42
180Intra4	YES	14	42	5.04	9.93	0.385	9.26
180InterA60	YES	15	404	5.08	19.19	0.161	16.73
180Intra2	YES	16	6	5.16	5.28	0.204	6.85
180InterA3	NOT IN SEARCH	17	\	5.32	\	\	\
180InterA30	YES	18	329	5.39	18.30	0.094	17.62
180Intra83	NOT IN SEARCH	19	\	5.40	\	\	\
180Intra56	POOR IN SEARCH	20	\	5.44	\	\	\
180InterA7	YES	21	417	5.81	19.34	0.428	17.44
90Intra31	YES	22	464	5.88	19.81	0.368	16.86
180Intra32	YES	23	169	6.15	15.25	0.18	12.68
180InterA18	POOR IN SEARCH	24	\	6.19	\	\	\
180Intra92	YES	25	49	6.35	10.40	*0.932*	8.34
180InterA12	YES	26	483	6.47	19.97	0.355	18.20
180InterA29	YES (high energy)	27	765	6.51	23.34	0.298	19.19
180InterB10	YES (high energy)	28	1418	6.53	29.12	0.471	21.50
90InterA14	NOT IN SEARCH	29	\	6.62	\	\	\
180Intra84	YES	30	61	6.67	11.15	0.258	8.54
180Intra47	YES	31	77	6.72	12.05	0.162	11.28
180Intra65	YES	32	28	6.81	8.78	0.116	12.00
90InterA32	YES (high energy)	33	1412	6.89	29.09	0.26	24.66
180Intra5	YES	34	56	6.94	10.99	0.329	11.61
180Intra28	YES	35	21	7.08	8.06	0.208	11.36
180InterA26	YES	36	483	7.16	19.97	*1.441*	18.28
180InterB87	YES (high energy)	37	1051	7.18	26.08	0.265	20.29
180Intra57	YES	38	24	7.43	8.50	0.764	12.21

Appendix Table 7.4: Reproduction and ranking of the matches to the significant crystal structures of molecule XX that were found in the original CSP study. The structure highlighted in green matches the experimental form. The structures in red were missed despite having a good match in the search. The structures in orange were found after $\Psi_{mol}^{PBE0+FIT}$ at ΔE_{latt} > 20 kJ·mol^{-1}. In a few cases, the distance and angle tolerances had to be increased to 30% and 30° respectively to obtain a match, and the $RMSD_{15}$ values for these structures are shown in italics.

Structure name	Found ?	Old CSP rank	Ranking after $\Psi_{mol}^{PBE0+FIT}$	Old CSP study ΔE_{latt} /kJ·mol^{-1}	$\Psi_{mol}^{PBE0+FIT}$ ΔE_{latt} /kJ·mol^{-1}	$RMSD_{15}$ after $\Psi_{mol}^{PBE0+FIT}$	CrystalOptimizer ΔE_{latt} /kJ·mol^{-1}
dfAa132	YES	1	1	0.00	0.00	0.308	0.00
dfAc102	YES	2	2	0.78	2.98	0.217	2.16
dfAa180	YES	3	5	2.38	8.08	0.357	3.01
dfAc14	YES	5	7	5.59	9.34	0.19	5.37
dfAc48	YES	10	3	6.15	4.91	0.429	8.83
dfAc19	YES	6	6	6.42	9.13	*0.658*	5.87
dfAc7	YES	12	14	7.26	9.97	0.186	8.12
dfAc43	NO	14	\	7.69	\	\	\
dfAc17	YES	15	9	7.86	9.72	0.307	10.97
dfAc172	YES	16	25	7.97	11.53	0.528	7.76
dfAc29	YES	17	129	8.19	15.70	0.49	6.93
dfAb181	YES (high energy)	22	1135	9.12	24.00	0.671	10.14
dfAd152	YES	23	314	9.13	18.96	0.839	8.33
dfAc86	YES	24	13	9.36	9.92	0.329	8.91
dfAc67	YES	25	20	9.48	11.14	0.142	9.39
dfAa277	YES	27	43	9.70	12.85	0.16	9.12
dfAa4	YES	28	54	9.76	13.50	0.579	10.44
dfAa1	YES	29	115	9.78	15.51	0.419	11.51
dfAb161	YES	31	59	9.88	13.74	0.395	9.51
dfAb1	YES	32	34	9.90	12.41	0.245	10.20
dfAd79	YES	33	112	9.93	15.43	0.346	8.20
dfBa28	NO	47	\	11.44	\	\	\

Appendix Table 7.5: Reproduction and ranking of the matches to the significant crystal structures of molecule XXIII that were found in the original CSP study. The structures highlighted in green match the experimental forms. The significant crystal structures for which a good match was not found are classified: in blue where they were not found by the search, in turquoise where they had a poor match in the search, and in red where structures were missed despite having a good match in the search. The structures in orange were found after $\Psi_{mol}^{PBE0+FIT}$ at ΔE_{latt} > 20 kJ·mol^{-1}. In a few cases, the distance and angle tolerances had to be increased to 30% and 30° respectively to obtain a match, and the $RMSD_{15}$ values for these structures are shown in italics.

Structure name	Found ?	Old CSP rank	Ranking after $\Psi_{mol}^{PBE0+FIT}$	Old CSP study ΔE_{latt} /kJ·mol^{-1}	$\Psi_{mol}^{PBE0+FIT}$ ΔE_{latt} /kJ·mol^{-1}	$RMSD_{15}$ after $\Psi_{mol}^{PBE0+FIT}$	CrystalOptimizer ΔE_{latt} /kJ·mol^{-1}
A1361	YES	1	1	0.00	0.00	0.117	1.27
A70	YES	2	3	1.66	0.83	0.168	0.00
A6494	POOR IN SEARCH	3	\	2.13	\	\	\
A691	YES	4	5	3.38	1.46	0.293	0.21
A3457	YES	5	7	3.68	2.04	0.312	1.40
A72	YES	6	24	3.81	5.47	0.355	2.15
A424	YES	7	6	4.41	1.81	0.253	1.64
A771	YES	8	4	4.64	0.85	0.173	1.20
A191	NO	9	\	5.07	\	\	\

A4890	YES	10	66	5.46	8.35	*0.703*	9.04
A5191	NOT IN SEARCH	11	\	5.52	\	\	\
A272	YES	12	35	5.68	6.53	0.664	6.17
A63	POOR IN SEARCH	13	\	6.05	\	\	\
A118	YES	14	2	6.13	0.17	0.281	2.69
A75	YES	15	27	6.29	5.73	0.58	2.17
A1413	YES	16	12	6.33	3.59	0.133	2.28
A2457	YES	17	34	6.66	6.41	*0.594*	11.13
A587	YES	18	59	6.85	7.93	0.358	7.41
A2417	YES	19	40	6.97	6.91	*0.434*	8.25
A138	YES	20	111	7.17	10.02	0.594	5.36
A227	YES	21	52	7.34	7.64	0.3	5.20
A1949	YES	22	279	7.61	13.00	*1.107*	5.34
A3174	NOT IN SEARCH	23	\	7.76	\	\	\
A2054	NOT IN SEARCH	24	\	7.81	\	\	\
A3023	YES	25	153	7.85	10.89	0.27	6.96
A2311	YES	26	216	7.86	12.04	0.233	11.15
A3513	YES	27	82	7.97	9.09	0.52	8.47
A1109	YES	28	231	7.99	12.30	0.424	6.74
A894	POOR IN SEARCH	29	\	8.07	\	\	\
A1422	YES	30	68	8.15	8.44	0.603	9.77
A1127	YES	31	99	8.15	9.74	0.376	7.05
A6634	POOR IN SEARCH	32	\	8.34	\	\	\
A282	YES	33	155	8.81	10.93	0.213	13.68
A323	YES	34	85	8.85	9.16	0.865	8.96
A2715	YES	35	141	8.92	10.68	0.226	10.99
A24995	YES	36	55	8.98	7.79	0.248	7.23
A3746	NO	37	233	8.99	12.34	*0.618*	5.26
A368	YES	38	239	9.06	12.36	0.509	7.76
A6738	YES	39	1080	9.07	19.28	*1.009*	11.00
A4228	YES	40	93	9.08	9.58	0.533	4.59
A1752	YES	41	13	9.16	3.87	0.227	4.26
A113	YES	42	31	9.17	6.02	0.315	4.49
A3750	YES	43	87	9.19	9.21	0.198	12.60
A505	YES	44	217	9.27	12.06	0.314	11.16
A12658	YES	45	61	9.56	7.96	0.224	6.78
A1918	YES	46	37	9.64	6.69	*0.797*	7.99
A1411	YES	47	950	9.72	18.49	*1.035*	20.35
A5145	YES	48	711	9.92	16.92	*0.674*	7.48
A710	YES	49	302	9.98	13.37	*0.677*	12.31
B204	YES	66	401	10.93	14.45	0.672	5.39
B60	YES	83	160	11.65	11.04	0.441	8.80
B184	YES (high energy)	100	1379	12.36	20.69	0.717	10.22
Exptal A	YES	(167	232	13.60	12.31	0.664	10.78

Appendix Table 7.6: Reproduction and ranking of the matches to the significant crystal structures of mebendazole that were found in the original CSP study. The structures highlighted in green match the experimental forms. The structures in orange were found after $\Psi_{mol}^{PBE0+FIT}$ at ΔE_{latt} > 20 kJ·mol^{-1}. In a few cases, the distance and angle tolerances had to be increased to 30% and 30° respectively to obtain a match, and the $RMSD_{15}$ values for these structures are shown in italics.

Structure name	Found ?	Old CSP rank	Ranking after $\Psi_{mol}^{PBE0+FIT}$	Old CSP study ΔE_{latt} /kJ·mol^{-1}	$\Psi_{mol}^{PBE0+FIT}$ ΔE_{latt} /kJ·mol^{-1}	$RMSD_{15}$ after $\Psi_{mol}^{PBE0+FIT}$	CrystalOptimizer ΔE_{latt} /kJ·mol^{-1}
A788	YES	1	1	0.00	0.00	0.164	0.00
A19	YES	2	2	2.15	0.85	0.258	\
C27	YES	3	3	2.54	1.79	0.128	2.66
C5	YES	4	15	2.54	5.78	0.142	2.51
C10	YES	5	6	2.63	3.51	0.103	2.31
A50	YES	6	10	3.15	4.65	*0.795*	4.69

A37	YES	7	5	4.07	2.11	0.234	4.41
C23	YES	8	36	4.33	7.89	0.255	6.12
C73	YES	9	51	4.33	9.63	0.219	7.95
C406	YES	10	93	4.68	11.62	0.172	12.75
A53	YES	11	11	4.81	4.75	0.347	4.67
C53	YES	12	35	5.48	7.74	0.121	5.25
C25	YES	13	18	5.50	6.57	0.097	5.22
A173	YES	14	126	5.66	12.80	0.333	10.69
A72	YES	15	101	5.75	11.90	0.214	5.79
A49	YES	16	17	5.79	6.52	0.88	6.18
A78	YES	17	12	5.93	5.34	0.097	5.73
A90	YES	18	7	5.97	3.53	0.193	5.67
A291	YES	19	23	6.14	7.13	0.273	6.85
C248	YES	20	29	6.17	7.47	0.166	6.03
A306	YES	21	91	6.27	11.53	0.413	6.53
C46	YES	22	25	6.30	7.20	0.106	7.08
C24	YES	23	22	6.36	7.04	0.228	6.69
C115	YES	24	37	6.47	7.91	0.32	6.50
C509	YES	25	72	6.61	10.60	0.374	6.81
C583	YES	26	448	6.62	19.25	0.304	14.60
A202	YES	27	20	6.72	6.78	*0.764*	7.78
C106	YES	28	53	6.72	9.65	0.506	6.62
A143	YES	29	28	7.19	7.36	0.19	6.94
A89	YES	30	40	7.33	8.42	0.17	7.33
C908	YES	31	43	7.42	8.77	0.296	7.51
CCis32	YES (high energy)	67	578	18.00	20.86	0.106	18.82

Appendix Table 7.7: Structures selected for phonon calculations for each molecule. For each structure the identifier is indicated, together with its density and E_{latt} after the optimisations with the $\Psi_{mol}^{PBE0+FIT}$ model, the variations in F_{vib} and in the Helmholtz free energies (ΔA), calculated using the rigid body and the DFTB3-D3 models, relative to the global minimum in E_{latt}, and the supercell used to calculate the DFTB3-D3 phonons. Structures matching the experimentally known forms are indicated in green and as yet unobserved E_{latt} minima are in orange. Note that forms C and E of XXIII were not present in the search, since they are Z'=2, and were optimised independently for comparison purposes. The first letter in the identifier of the mebendazole crystal structures indicates whether they contained the A or C tautomers.

Molecule XXVI							
Structure name	**Density /g·cm^{-3}**	**E_{latt} /kJ·mol^{-1}**	**ΔF_{vib} rigid body/kJ·mol^{-1}**	**ΔA rigid body /kJ·mol^{-1}**	**ΔF_{vib} DFTB3-D3/kJ·mol^{-1}**	**ΔA DFTB3-D3/kJ·mol^{-1}**	**DFTB3-D3 supercell**
C26_863	1.388	-211.84	0.00	0	0.00	0	222
C124_2	1.331	-209.88	-0.31	1.65	-3.36	-1.40	222
C1_134	1.333	-207.39	-0.24	4.21	-3.16	1.29	111
C41_619	1.361	-205.82	-0.87	5.15	-4.78	1.25	222
C805_7	1.393	-200.95	0.24	11.14	2.03	12.93	222
GSK269984B							
Structure name	**Density /g·cm^{-3}**	**E_{latt} /kJ·mol^{-1}**	**ΔF_{vib} rigid body/kJ·mol^{-1}**	**ΔA rigid body /kJ·mol^{-1}**	**ΔF_{vib} DFTB3-D3/kJ·mol^{-1}**	**ΔA DFTB3-D3/kJ·mol^{-1}**	**DFTB3-D3 supercell**
C1_60	1.497	-173.71	0.00	0	0	0	222
C1_240	1.494	-171.17	-1.71	0.83	3.23	5.78	222
C1_1165	1.506	-170.14	-1.72	1.86	2.15	5.72	222
C1_685	1.479	-169.86	0.30	4.15	-3.18	0.67	211
C1_2_1	1.484	-165.01	2.09	10.79	0.01	8.71	141

Molecule XX							
Structure name	**Density /g·cm^{-3}**	**E_{latt} /kJ·mol^{-1}**	**ΔF_{vib} rigid body/kJ·mol^{-1}**	**ΔA rigid body /kJ·mol^{-1}**	**ΔF_{vib} DFTB3-D3/kJ·mol^{-1}**	**ΔA DFTB3-D3/kJ·mol^{-1}**	**DFTB3-D3 supercell**
C1_9	1.382	-218.52	0.00	0	0	0	111
C1_60	1.330	-215.54	-1.05	1.93	6.32	9.30	411
C1_250	1.393	-213.61	1.43	6.34	2.86	7.77	121
C78_1191	1.315	-211.12	1.63	9.03	2.37	9.78	121
C78_28	1.347	-210.45	-0.53	7.54	-2.60	5.48	111
C245_91	1.328	-199.71	-0.30	18.5	0.57	19.39	111
Molecule XXIII							
Structure name	**Density /g·cm^{-3}**	**E_{latt} /kJ·mol^{-1}**	**ΔF_{vib} rigid body/kJ·mol^{-1}**	**ΔA rigid body /kJ·mol^{-1}**	**ΔF_{vib} DFTB3-D3/kJ·mol^{-1}**	**ΔA DFTB3-D3/kJ·mol^{-1}**	**DFTB3-D3 supercell**
C1_13	1.387	-179.40	0.00	0	0	0	212
C1_60	1.394	-179.22	-0.25	-0.08	-0.95	-0.78	222
C1_43	1.402	-178.57	-0.80	0.03	-0.50	0.32	231
C103_31	1.410	-178.55	0.70	1.55	0.86	1.71	311
C1_889	1.342	-169.14	-2.46	7.81	-2.79	7.47	221
C49_1002	1.411	-169.09	-0.60	9.71	-1.90	8.41	221
C103_847	1.345	-167.08	-1.15	11.17	0.86	13.17	221
Form C	1.402	-172.28	-1.21	5.91	-1.33	5.79	221
Form E	1.366	-170.88	-1.36	7.16	-2.44	6.07	221
Mebendazole							
Structure name	**Density /g·cm^{-3}**	**E_{latt} /kJ·mol^{-1}**	**ΔF_{vib} rigid body/kJ·mol^{-1}**	**ΔA rigid body /kJ·mol^{-1}**	**ΔF_{vib} DFTB3-D3/kJ·mol^{-1}**	**ΔA DFTB3-D3/kJ·mol^{-1}**	**DFTB3-D3 supercell**
A_C1_5	1.395	-176.73	0.00	0	0	0	221
A_C1_47	1.406	-175.87	0.14	1.00	-1.14	-0.29	211
A_C1_6	1.430	-174.91	1.43	3.25	0.77	2.58	111
C_C1_39	1.395	-174.94	0.63	2.42	-3.72	-1.93	111
C_C1_28	1.397	-170.94	0.28	6.06	-1.63	4.15	221
C_C1_227	1.370	-169.96	-0.29	6.48	-3.38	3.39	211

Appendix Table 7.8: Breakdown of the computational cost for refining and re-ranking the CrystalPredictor generated crystal structures of each molecule with the method described in this chapter. The cost of the phonon calculations is not included in the total, to have a more meaningful comparison with the original CSP studies where only E_{latt} values had been calculated.[2-5]

	XXVI	**GSK269984B**	**XXIII**	**XX**	**Mebendazole**
DFTB Optimisation cost/ hours	10,927	4,829	8,386	12,723	2,215
DFTB clustering cost/ hours	109	156	618	1,306	37
DMACRYS Optimisation cost/ hours	4,426	2,911	6,903	14,674	2,300
Total cost/ hours	15,462	7,896	15,907	28,703	4,552
DFTB3-D3 phonon calculations/hours	1,844	344	984	512	99
Rigid body phonon calculations/hours	5	3	9	6	7

Chapter 8: The intricacies of discriminating between polymorphs and duplicates.

8.1 Introduction

A problem of CSP studies is that most search algorithms tend to generate the same crystal structure several times.[1-3] This is true especially for the less accurate models used for crystal structure generation, which tend to have a rough energy surface, with more minima than more accurate periodic models.[4] Since CSP studies aim to find all the plausible distinct polymorphs of a molecule, the presence of duplicate crystal structures is detrimental to the assessment of the results of a CSP study, as well as to its overall efficiency given the waste of computational resources in performing several calculations on the same crystal structure.[2, 3, 5, 6] On the other hand, there is a danger in removing as duplicates similar crystal structures that may correspond to different polymorphs.[7] Throwing away an experimental crystal structure as a duplicate would turn a successful CSP study into a failure.

A number of approaches have been developed to remove duplicates, a process known as clustering.[3, 5, 8-10] All these methods share the underlying problem that there is not a clear quantitative degree of dissimilarity that can be used to distinguish polymorphs from variations of the same crystalline form.[7, 11] Clustering approaches are all based on arbitrary thresholds, mostly derived from human experience and intuition.[10] For this reason, finding some quantitative parameters to assess whether two crystal structures are duplicates or polymorphs would be very important in CSP. This could also be used by experimentalists who are in doubt about how to classify the crystal structures they crystallise, as well as for database analyses and surveys. But can thresholds actually be found?

The main problem is that there is some ambiguity regarding what actually constitutes a polymorph.[12-15] The most widely accepted definition of polymorph is still the one given by McCrone in 1965: “A polymorph is a solid crystalline phase of a given compound resulting from the possibility of at least two different arrangements of the molecules of that compound in the solid-state”.[16] Other definitions have been proposed, but they have not changed the general view of the phenomenon, and Bernstein pointed out that they often create more confusion than clarity.[17] McCrone’s definition is important in stating a polymorph is not necessarily a different crystal structure, but a different phase, which is defined by IUPAC as “an entity of a material system which is uniform in chemical composition and physical state.”[18] Hence two polymorphs are essentially two solid-states exhibiting different physical properties, but sharing the same liquid and vapour states.[17, 19] This has the fundamental implication that, although some structural

dissimilarity is needed to have a phase change (and hence a polymorph),[20] it must just be large enough to cause some deviation in the solid-state physical properties. These differences can be rather small in many cases, and the term “isostructural polymorphs” has been coined to describe distinct phases with a strong degree of structural similarity.[12, 21] Although Gavezzotti stated that subtle modulations without any significant change in crystal structures should not be termed as polymorphs,[13] the polymorphic nature of almost identical crystal structures has in several cases been unequivocally established by discovering phase transitions with experimental techniques like differential scanning calorimetry (DSC)[12], magnetic susceptibility measurements,[22] neutron diffraction,[23] Raman Spectroscopy,[24] or by changes in properties of the solid.[25] In the absence of a clearly identified phase transition, the distinction lines get blurry, and the identification of isostructural polymorphs can often be dependent on a personal interpretation by the crystallographer. Isostructural polymorphs are a major challenge in understanding the structural differences that separate polymorphs from duplicates.

To complicate the matter even further, a different crystal structure is not necessarily a different polymorph.[20, 26] The same polymorph produced in a different experiment could have a slightly different molecular arrangement, particularly if it is characterised at different temperatures and pressures. Thermal expansion can change unit cell dimensions, and the differences can be remarkable over large temperature ranges; furthermore, this lattice expansion can often be anisotropic and lead to differences in packing without any phase transition occurring.[20, 27, 28] The amount of structural change caused by variations in pressure can be even larger.[26] Hence, several ambiguous cases exist, in which quantitative structural parameters may not be sufficient to separate polymorphs from redeterminations; it is possible that different determinations of the same phase could be more structurally different by certain measures than some polymorphs.

However, these ambiguities are known to be rare, and in most cases identical phases characterised under different conditions are structurally similar, while polymorphs have very different and easily distinguishable packings, if not conformations.[11, 29] Furthermore, it can be safely assumed that two crystal structures of the same molecule with very different packing arrangements, *i.e.* with their atoms occupying very different positions within the lattice, are polymorphs, since large structural differences are associated with large differences in properties.[30, 31] Hence, it is worth investigating the differences that can be generally found in polymorphs, with the aim of determining criteria that in the great majority of cases can separate polymorphs from duplicates. At a first glance, it may appear that the computation of lattice energies, enthalpies or free energies would be adequate for this purpose: while duplicates should optimise to the same energy minimum, polymorphs should optimise to different minima.

However, the energy surface described even by the most accurate theoretical models is not perfectly realistic,[32] and polymorphs often have very small energy differences, within the range of accuracy of the calculations.[33] This problem becomes even more serious for isostructural polymorphs forming upon phase transitions with temperature and/or pressure, since the two phases could belong to different free energy minima but to the same lattice energy minimum (*i.e.* at 0 K and 0 GPa), or one form could be a higher symmetry phase averaging over several lower symmetry E_{latt} minima.[1] The detection of this sort of differences is beyond the capability of the available modelling methods, which makes the discrimination of any polymorph from duplicates purely on energy grounds very complicated.[10] Several other methods have been proposed to separate polymorphs from duplicates,[15] such as fingerprints based on energies of molecule-molecule pairs[13], configurational distances between crystal structures[34] or analysis of the Hirshfield surfaces.[35, 36] However, they all lack quantitative thresholds, and the user has to decide where to draw the boundary between polymorphs and duplicates.

In this work, the focus is on the differences in 3-dimensional structural coordinates and crystallographic parameters between crystal structures. From the analysis outlined above, it is clear how the only possible approach for finding clustering criteria is a heuristic one, based on analysing what sort of differences are common between polymorphs and what similarities are common between duplicates. A few failures in identifying isostructural polymorphs, or different modulations of the same crystal structures, are expected. While these failures would probably not be problematic in CSP, for experimental studies the clustering criteria can be treated more flexibly: they can give an idea of what sort of differences can suggest the presence of polymorphs or duplicates, while more ambiguous cases can be investigated more in depth with the appropriate experimental techniques.

The Cambridge Structural Database (CSD),[37] which contains more than 900,000 experimentally-determined crystal structures, is an obvious choice for performing this heuristic analysis. In the CSD, every distinct chemical compound is characterised by a six-letter ‘refcode’, and there are thousands of compounds for which multiple crystal structures have been deposited. Enantiopure and racemic compounds of the same molecules are different in the liquid phase and so belong to different refcode families.[38] The different crystal structures within the same refcode family are not necessarily polymorphs, but can also be multiple redeterminations or refinements of the same crystal structures. Polymorphic crystal structures are often flagged in the CSD: different polymorphs are flagged differently, while duplicates share the same flag.[27, 38] However, not all polymorphs are flagged, since this depends on whether the polymorphic nature of the molecule was mentioned in the original publication.[27] Furthermore, polymorph flags are not always specific: some clearly indicate what polymorph each crystal structure

corresponds to (for example flagging polymorphs as polymorph I, II *etc*.), while others are more generic (for example 'monoclinic', '$P2_1/c$'). In this chapter, it is assumed that each identical flag characterises the same polymorphs and vice versa, independently of its specificity, unless a manual check reveals otherwise.

Given the imperfect nature of the polymorph flags, the CCDC has tried to find ways to distinguish polymorphs and redeterminations that is independent of whether the crystallographers mentioned the polymorphic nature of the deposited crystal structures. These efforts have led to the creation with every release of the CSD of the 'best R-factor' list.[38] This subset of the CSD is selected through an algorithm that uses simulated PXRD similarities[39] (see Chapter 2.5.2) to remove all the redeterminations from the CSD, leaving only the unique crystal structures with the best R-factor. Hence, every refcode family with more than one crystal structure in the 'best R-factor' list should be polymorphic. Also, disordered or wrong crystal structures are removed, which should guarantee that each crystal structure in the 'best R-factor' list is ordered and of high quality. This list, or variations built with a similar method, have been utilised in several high-quality studies on crystalline polymorphism.[29, 30, 40]

However, since PXRD similarities alone do not seem to solve the clustering problem in CSP,[7] the 'best R-factor' list was just used to heuristically determine a set of differences that are common in most polymorphic pairs and hence can be used as criteria to discriminate between polymorphs and duplicates. These same criteria were then applied on the entire CSD to test their effectiveness. Finally an efficient clustering algorithm based on the CSD Python API[37] (see Chapter 2.6.5) and implementing all of these criteria was written for use in CSP studies, accounting for the different types of problems that can be encountered in computational studies compared to database analysis. This algorithm was tested on a set of CSP-generated crystal structures of molecule XXVI (see Chapter 3) and mebendazole (see Chapter 4).

8.2 Methods

8.2.1 Analysis of the 'best R-factor list' to determine the differentiation criteria

Conquest 1.19[41] (see Chapter 2.6.1) was used to retrieve the 499,981 crystal structures in the 'best R-factor' list. Successively, all the refcode families with only one entry were removed, leaving 10,042 crystal structures of 4,748 compounds. These were grouped in 6,007 possible polymorphic pairs; two crystal structures of the same compound formed one pair, three crystal structures formed two pairs *etc.* No limitation in the number of molecular components was introduced. Only crystal structures of organic molecules determined at ambient pressures were kept; this was because even small pressure differences can cause large structural variations.[26, 42] This resulted in 3,925 pairs of crystal structures. This sample included polymorphs determined at different

temperatures, since temperature differences tend to have a smaller structural effect than pressure differences.[43, 44]

The next important choice was whether to trust the methodology used to form the list and consider all of those pairs as actual polymorphs, or if further limitations were required. The 'best R-factor list' is known to contain a small but not insignificant number of false polymorphs, *i.e.* pairs of crystal structures that appear to be polymorphs but that are effectively duplicates, and false duplicates, *i.e.* pairs of crystal structures that are polymorphs but look like duplicates.[38] This is because the list is compiled using PXRD similarity, which can sometimes be misleading. To check how prevalent this phenomenon may be, the PXRD similarities were tested for their ability to reveal the degree of structural similarity. For each of the 3,925 pairs, the CSD Python API was used to estimate both the simulated PXRD similarities (with its default settings) and to perform structural overlays. Clusters of 15 molecules were overlaid using the Crystal Packing Similarity tool (see Chapter 2.5.1),[45] with 20% distance and a 20° angle tolerances. The position of hydrogen atoms was ignored since it is often uncertain in experimentally-determined crystal structures.[29, 46] Figure 8.1 shows how the PXRD similarity compares with the number of molecules that were successfully overlaid for the 3,925 pairs of crystal structures.

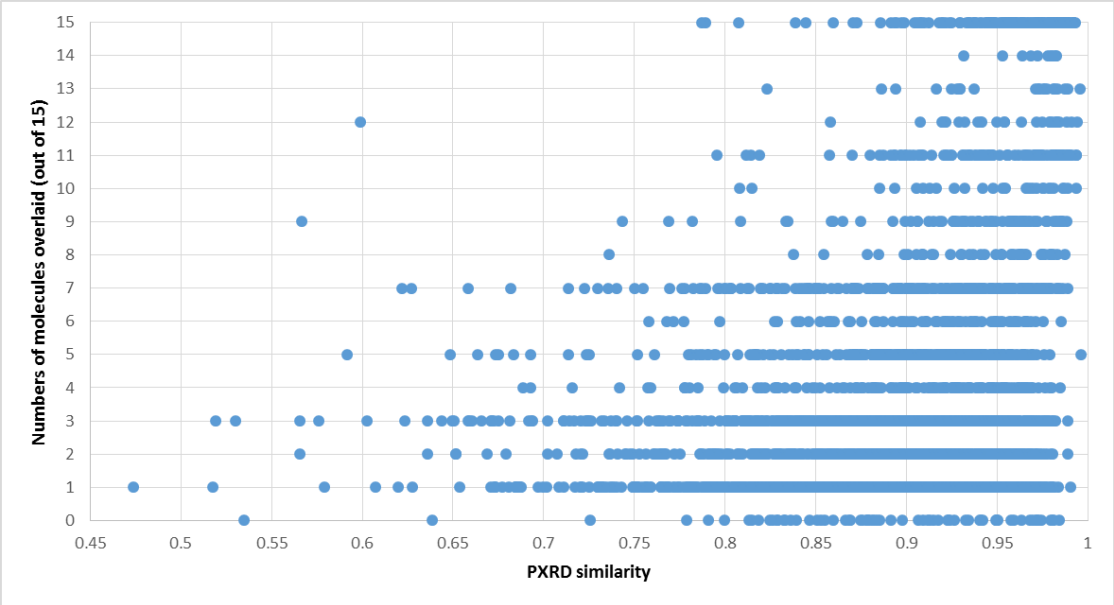


Figure 8.1: Comparison of the PXRD similarity and the number of molecules matched (out of 15) for the 3,925 pairs of crystal structures under consideration.

Crystal structures for which more molecules can be overlaid tend to show higher PXRD similarities. However, the relationship is far from perfect, and structures with high or low PXRD similarities can be found across the whole spectrum of numbers of overlaid molecules. This generalises the worrying findings of an earlier study limited to isostructural hydrates.[47] Some fully matching crystal structures exhibit PXRD similarities smaller than 0.80 and very different ones have similarities above 0.99. This weakness

suggests that not all the 3,925 pairs can be safely considered as polymorphic, and also makes it likely that the 'best R-factor' list does not contain all the possible pairs of polymorphs within the CSD. Hence, since certainty about the nature of the crystalline pairs is fundamental for the effectiveness of the analysis, it was decided to further limit the analysis only to those pairs with both crystal structures flagged as polymorphic. This limited the data set to 3,371 pairs. In 3,331 pairs the flags were different, while 40 pairs they were identical. Those 40 pairs were manually analysed. In 16 of those cases the flag was found to be wrong (*e.g.* AZELAC04-AZELAC15) and those polymorphic pairs were kept, while in 24 cases (see Appendix Table 8.1 for details) they were determined to be duplicates, and were removed from the data set, leaving 3,347 pairs.

The next step was to analyse the remaining polymorphic pairs to heuristically find what sort of differences they generally share. The first check was on the possibility to overlay all the molecules within 15-molecule clusters. For most of those pairs it was not possible, since in 3,288 cases a full 15-molecule overlay was not achieved, leaving 59 pairs of fully overlaid polymorphs. This showed that for most polymorphic pairs 15/15 molecule overlays fail, so structural differences are very common, as expected, and are a very effective criterion to recognise polymorphs. However, the possibility of overlaying 15/15 molecule is not sufficient, and further criteria are needed to recognise a minority of structurally similar polymorphs. An analysis of the 59 remaining polymorphic pairs revealed that in 43 of these cases polymorphs have at least one of these crystallographic differences:

- A different number of molecules in the unit cell (Z)
- The same number of molecules in the unit cell (Z) and the same number of molecules in the asymmetric unit cell (Z'), but a different space group
- The same number of molecules in the unit cell (Z) and the same space group, but a different number of molecules in the asymmetric unit cell (Z')

Only for 16 pairs of crystal structures was it possible to perform both a 15/15 overlay and not find any of the crystallographic differences above.

It was noted that polymorphs with different Z' and space group tended to be more similar (*i.e.* to have smaller $RMSD_{15}$ values) than those with the same Z' and space group. This is probably because phase transitions between structurally similar polymorphs are often associated with a loss of symmetry elements and an associated increase in the number of molecules in the asymmetric unit.[12, 13, 23, 48] In 5 out of the 6 polymorphic pairs of structures with the same space group and Z' $RMSD_{15}$ was larger than 0.5 Å, and in 8 out of 10 pairs with different space group and Z' it was larger than 0.1 Å. Hence, including these $RMSD_{15}$ differences leaves only three pairs of polymorphs (see Appendix Table 8.2 for details), which cannot be separated with any sensible structural or crystallographic criterion.

Since the differences in terms of number of molecules overlaid, crystallographic parameters and $RMSD_{15}$ between crystal structures that were heuristically found in this analysis are effective in recognising the great majority of polymorphic pairs in the 'best R-factor' list, they were considered as effective criteria to separate polymorphs from duplicates. The results of this analysis and the set of developed criteria are summarised in the decision tree in Figure 8.2.

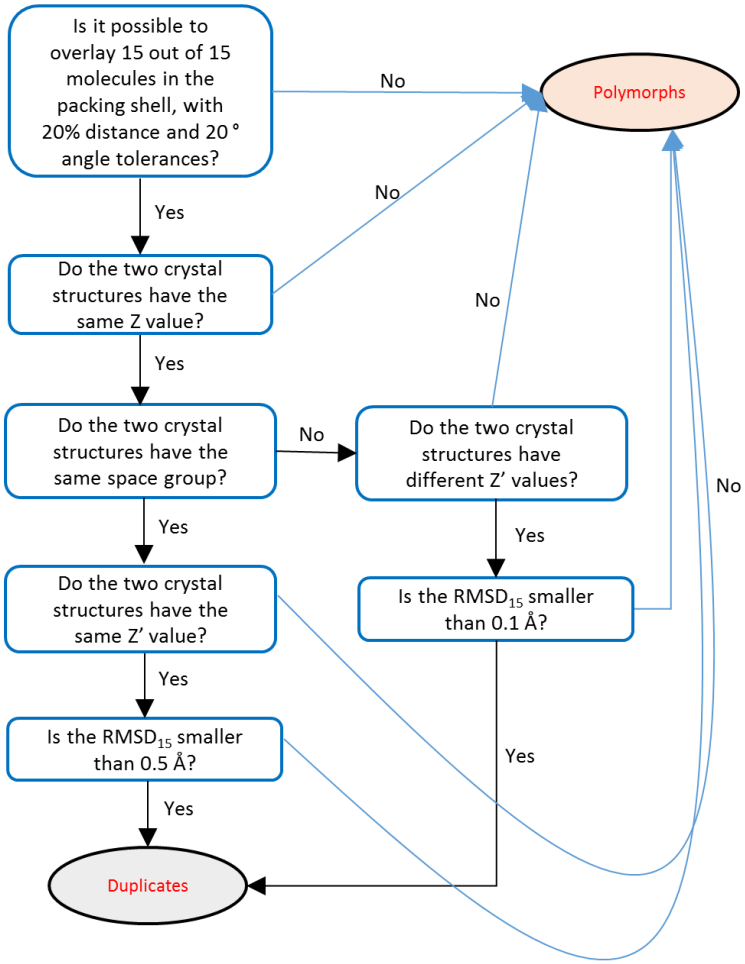


Figure 8.2: Decision tree to discriminate whether two organic crystal structures characterised at similar pressures are duplicates or polymorphs.

8.2.2 Testing the criteria on the whole CSD

The criteria determined from the analysis of the difference between the polymorphic pairs in the 'best R-factor' list were then applied on the entire CSD to test if they were effective at separating polymorphs from duplicates. The full CSD is ideal because, differently from the 'best R-factor' list, it purposely includes both duplicates and polymorphs.[38] Also, this sample of structures is not limited to a selection of high-quality entries like the 'best R-factor' list, and can reveal if the approach is suited to recognise polymorphs even in the presence of poorer determinations.

The 876,747 individual crystal structures in the 2017 release of the CSD were the starting point for this analysis. They were reduced by selecting only organic, not

disordered, with determined 3D coordinates structures determined at ambient pressure. For an objective assessment of the criteria, only polymorph-flagged crystal structures were retained. Since misassignments of refcodes do occasionally occur,[27, 38] crystal structures were grouped according to their compound rather than just by refcode. Hence, structures were first grouped according to their chemical formulae; only those with at least two entries with the same chemical formula were kept. D and H atoms were considered as equivalent. The crystal structures were then grouped according to the isomer they contained. To do that, the Crystal Packing Similarity tool was used to perform 1-molecule overlays between all the crystal structures with the same chemical formulae; when a match was not possible this indicated that the crystal structures contained different compounds. Although effective in most cases, this method carried the risk of grouping together racemates, stereoisomers and multi-component compounds with a common molecule. Hence all the groups containing more than one refcode family were manually checked to remove these occurrences.

This initial grouping resulted in 3,110 families of compounds containing 8,978 polymorph-flagged crystal structures; in the great majority of cases these compound families matched refcode families, with few exceptions due to chirality or misassignments. The crystal structures in each of the 3,110 families were ordered according to their R-factors and clustered using an algorithm, implemented in Python, which followed the decision tree shown in Figure 8.2. The overlays were performed with the Crystal Packing Similarity tool and once again hydrogen atoms were ignored. Crystal structures were clustered in ascending order of R-factor, by comparing each entry with those within the same family already identified as belonging to a separate cluster.

The outcome was this analysis was the identification of 6,550 clusters of crystal structures, and each cluster should contain a separate polymorph. More than one crystal structure was present in 1,130 clusters; multiple crystal structures in the same cluster should be duplicates. This set of results was checked for the presence of false polymorphs and false duplicates. The assumption was made that if the polymorph flag and the clustering results coincided then the answer was correct and no further validation was undertaken.

To identify the false duplicates, the 1,130 clusters containing more than one crystal structure were checked to verify whether they contained entries with different polymorph flags. This was the case for 280 clusters, containing 930 crystal structure, which were checked manually to verify whether the flag was correct, or the discrepancy was caused by errors. The observed errors in the polymorph flags were:

- Typos or different writings of the same flag (*e.g.* ‘1 Polymorph’ and ‘Polymorph 1’). This type of error was the most common.

- The CSD polymorph flags present in the CSD are inconsistent with the publication/s.
- The publication/s do not mention that a structure with a different polymorph flag is actually a different polymorph.
- The publication/s state that one crystal structure is just a redetermination with a more accurate symmetry. This occurred rarely and it was the least common type of error.

In 239 clusters the presence of different polymorph flags was due to one of these errors, leaving 41 containing false duplicates (see Appendix Tables 8.4-8.5 for details), *i.e.* different polymorphs wrongly classified as duplicate using the criteria in Figure 8.2. Chapter 8.3.2.2 discusses the origin of these false duplicates.

On the other hand, to identify the false polymorphs all the crystal structures in the same input compound group and output in different clusters were checked to verify whether they had the same polymorph flag. This was the case for 122 pairs of structures. A further manual check was performed on all those pairs, and in 41 cases the polymorph flag was found to be wrong, meaning the structures were actual polymorphs, leaving 81 pairs of false polymorphs, *i.e.* duplicates wrongly classified as polymorphs. These are listed in Appendix Table 8.3 and their origin is discussed in Chapter 8.3.2.3.

8.2.3 Clustering CSP-generated structures

The final step of this analysis was to test the criteria in Figure 8.2 on CSP-generated crystal structures. First of all a clustering algorithm, based on the CSD Python API tools, was written for use in CSP, and then it was tested on crystal structures generated in the CSP searches of molecule XXVI and mebendazole (see Chapters 3 and 4).

8.2.3.1 Algorithm to perform the clustering

Clustering CSP-generated crystal structures requires an algorithm that is more sophisticated than the one used to cluster CSD structures. First of all, the problem is different, since in the CSD there are only few experimentally-determined crystal structures for each compound, while a CSP study often generates thousands of candidates for a given molecule. Computational cost can be problematic and must be tackled effectively, by parallelising the task between multiple processors and/or limiting the number of comparisons to a feasible number. Furthermore, the position of hydrogen atoms is accurately determined in CSP, and in some cases it can be important in fully understanding the spectrum of possible crystal structures, as exemplified by gallic acid.[49] Finally, CSP studies provide information about the energy of crystal structures, as well as a number of structural characteristic, such as density and unit cell parameters at

constant temperatures and pressures, that the user may want to consider when performing the clustering procedure.

Hence a clustering algorithm for use in CSP has to be effective, flexible and computationally efficient. The algorithm that was written for this purpose analyses structural and crystallographic differences between crystal structures. It takes into account all the criteria in Figure 8.2 but the user can decide the number of molecules to be overlaid (*x*), the distance and angle tolerances for the overlays, whether or not to account for the position of hydrogen atoms, and the maximum $RMSD_x$ differences, in the presence of the same or different space groups, for structures with *x/x* molecules matching to be considered as duplicates. In order to improve the computational efficiency and give the user more flexibility, it is also possible to choose the maximum differences in calculated energy, density and unit cell parameters (*i.e.* unit cell lengths and angles) for two structures to be overlaid, as well as a minimum PXRD similarity. This clustering algorithm can work in parallel to take advantage of multi-core computer clusters; the user decides the number of processors that he wants to utilise. Finally a path to a crystal structure containing the correct molecule can be specified for removing those computer-generated structures with unrealistic close contacts, whose presence is common in CSP searches of flexible molecules due to the failings of the codes, usually caused by the minimisation going past the maximum in the exp-6 repulsion-dispersion potential.[50] The algorithm functions as follows:

1) Two input files are read; the first contains the user-defined criteria listed above and the second one contains a list of the crystal structures to be analysed, their energies and their densities. The structures are ordered according to their energies. The full set is broken down into several smaller sub-sets, each containing the same number of crystal structures. The number of sub-sets corresponds to the number of processors specified in the first input file.
2) Each of these sub-sets is clustered in parallel, using the Python multiprocessing library. In each sub-set, the crystal structures are analysed from the lowest energy one upwards.
    - If the user has chosen the option, a first check is performed to verify whether the molecule in the crystal structure is correct. An overlay with the crystal structure containing the right molecule is attempted, with a 1-molecule shell size. If the overlay fails, the crystal structure contains the wrong molecular geometry, and it is removed.
    - Secondly, the actual clustering step is performed. Each crystal structure in the sub-set is compared, in reversed energy order, with each form already identified as unique. This list of unique forms is one of the outputs of the clustering analysis. If the differences in terms of energy, density or unit-cell

parameters between two crystal structures are above the thresholds specified in the input, or if the crystallographic criteria in Figure 8.2 are not met, they are not compared. If all those criteria are met, then PXRD similarity is identified using the CSD Python API, and if it is above the user-defined threshold, then the Crystal Packing Similarity Tool is used to overlay the crystal structures. If the tool manages to overlay each molecule within the cluster with an $RMSD_x$ value lower than the thresholds specified in the input, then the structure is considered to be a duplicate and thrown out of the sub-set, otherwise it is added to the list of unique crystal structures. The method for deciding whether two crystal structures belong to different clusters or if they are duplicates is summarised in Figure 8.3. The process is then repeated until each crystal structure within the sub-set has been analysed. A list of structures that have already been compared in this first clustering cycle is saved, in order to avoid wasting computational resources performing identical comparisons in the following step.

3) The outputs of the clustering analyses performed on each processor are connected, and the unique crystal structures found in the previous step are combined to form a new group of sub-sets. Two groups of unique crystal structures are combined only if the energy difference between the lowest energy crystal structure in the higher-energy sub-sets and the highest-energy crystal structure in the lower-energy sub-set is smaller than the user-defined energy difference threshold. Once these new sub-sets have been generated, they are assigned to different processors and clustered in parallel in the same manner as for the original sub-sets (see Figure 8.3 for a summary). Two crystal structures are compared only if they had not already been contrasted in step 2. Once a sub-set has been fully analysed, the output is saved and a new sub-set is sent for clustering to the same processor. This process is repeated until all the sub-sets have been fully analysed.
4) Clustering is finally complete and the user is provided with a set of outputs: a list of duplicate and wrong crystal structures, as well as a spreadsheet containing the unique crystal structures, which can be used as an input for further clustering.

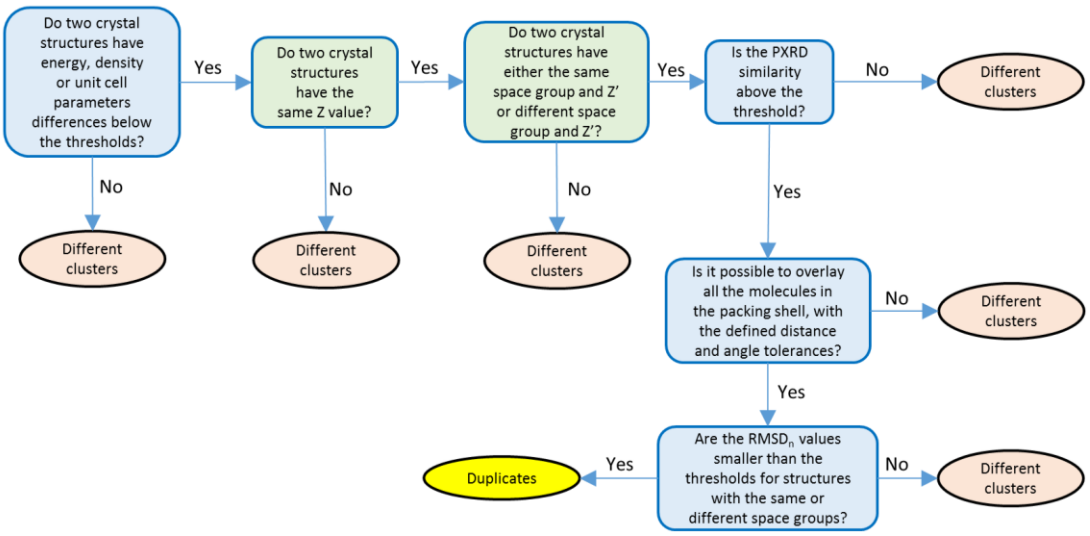


Figure 8.3: Schematics of the criteria used in the algorithm to decide whether two computer-generated crystal structures are duplicates or belong to different clusters. The blue boxes indicate steps where the user can define the thresholds for success or failure, while the green boxes represent hard-coded criteria.

8.2.3.2 Testing the algorithm and the criteria on CSP structures

The algorithm described above was tested the 9,400 search-generated crystal structures of molecule XXVI, the 855 of the A-tautomer of mebendazole, and the 964 of its C-tautomer that were taken forward to the successive refinement stages. Given the large size of these samples, the criteria shown in Figure 8.2 were integrated with some limitations in order to limit the computational cost: structures were compared only if their energies did not differ by more than 8 kJ·mol^{-1}, their densities by not more than 0.1 g·cm^{-3}, and if they had PXRD similarities larger than 0.900.

The results were analysed and compared with those obtained in the original CSP studies, where no clustering of the search-generated crystal structures had been undertaken besides the one automatically performed by CrystalPredictor (see Chapters 3.2.2 and 4.2.2). The goal was to verify whether this clustering method and criteria can effectively reduce the number of generated structures to be taken to the refinement stage, limiting in this way the number of calculations to be performed on what are effectively identical crystal structures but without throwing away matches to the experimental form/s or other signicant PPMs. The potential savings of this clustering step were estimated by calculating how much computational cost would have been saved in the original CSP studies if the crystal structures removed by this clustering analysis had not been refined.

8.3 Results and Discussion

8.3.1 Determination of the clustering criteria from the 'best R-factor' list

An analysis of the polymorphic pairs in the 'best R-factor' list allowed to find the most common structural and crystallographic differences. These differences appear to be present independently of the number of components in the crystal structure, as well as of the size of the molecule/s under consideration. The few failures (*i.e.* false polymorphs and false duplicates) are indeed for single-component crystal structures of relatively small molecules.

The impossibility of performing 15/15 overlays is the most common difference, as shown in in Figure 8.4.

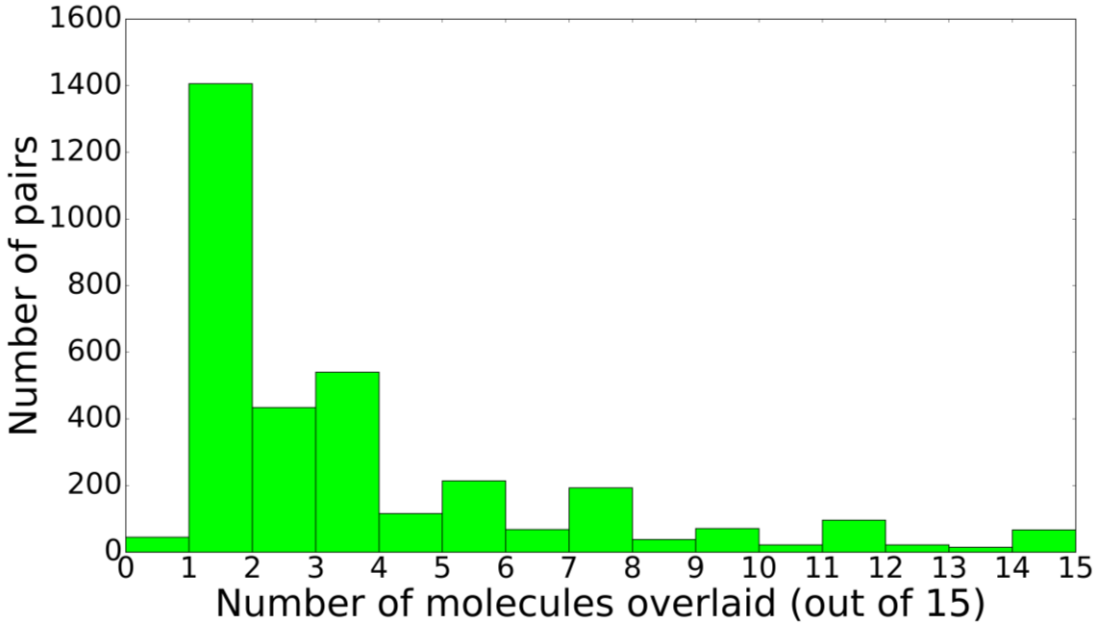


Figure 8.4: Distribution of the number of molecules that could be overlaid for the 3,347 pairs of polymorphs. They correspond to the 3,371 pairs of crystal structures considered, excluding the 24 identified as duplicates.

Only ~1.8% of flagged polymorphic pairs have a 15-molecule match, and in ~82% of cases five or fewer molecules can actually be overlaid, indicating a high degree of structural dissimilarity. The polymorphic pairs without any of the crystallographic differences (*i.e.* different Z, same space group and different Z', or different space group and same Z') are more common, around ~29% of the total. However, only ~0.5% of the polymorphic pairs have both a 15/15 molecule overlay and have none of the crystallographic differences. Putting a limit on $RMSD_{15}$ values reduces this percentage to ~0.1%, with only three pairs of false duplicates (see Appendix Table 8.2) in the 'best R-factor' list passing all the tests in Figure 8.2. Furthermore, out of the 24 pairs of flagged duplicates, 20 meet all the criteria in Figure 8.2, with only four being failures, which can be classified as false polymorphs. These are listed in Appendix Table 8.1.

An analysis of the few pairs of crystal structure whose nature is not identified by the heuristically developed approach reveals interesting information. All the three pairs of false duplicates (refcodes DIMETH01-DIMETH06, LIHXUW01-LIHXUW02 and MOSTIX-MOSTIX01, more details in Appendix Table 8.2) were determined at different temperatures, and they were subject to a phase transition occurring within that temperature range. In Figure 8.5 an overlay of the crystal structures of DIMETH01 and DIMETH06 shows how similar the crystal structures of distinct polymorphs can be when a phase transition with temperature occurs.

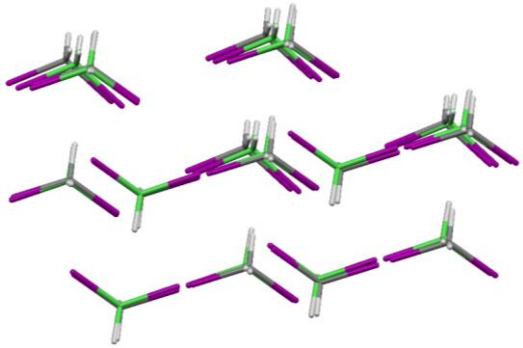


Figure 8.5: 15-molecule overlay between the crystal structures of DIMETH01 (coloured by elements) and DIMETH06 (in green). The $RMSD_{15}$ is 0.226 Å. The publication clearly states that they are different polymorphs.[51]

Hence, phase transitions appear to be problematic, and in a few cases they can cause changes of properties with barely noticeable structural changes. In the view of a CSP study, this is not particularly concerning, since crystal structures are generated and compared under the same conditions, generally at 0 K and 0 GPa.[1, 2, 52] On the other hand, for crystal structures experimentally determined at different temperatures or pressures ambiguous cases can occur, and DSC measurements or other experimental validations may be required.

Out of the four pairs of false polymorphs (see the highlighted entries in Appendix Table 8.1), one was not identified because the structures failed to meet the 15/15 molecule overlay criterion and three because they failed to meet the $RMSD_{15}$ thresholds. The only case where the 15/15 molecule overlay was not achieved was for FORMAC-FORMAC01, where the crystal structures are nearly identical if single and double bonds are not distinguished, while being very different if bond types are accounted for, as illustrated in Figure 8.6. FORMAC does not have the position of hydrogen atoms determined, so in this case the error is probably due to the Crystal Packing Similarity tool incorrectly defining the double bond in the carboxyl group. This is a problem related to the history of crystallography, since FORMAC is a determination from 1953, and the lack of hydrogen position is a common feature of old CSD entries.[38] Changing the settings of the Crystal Packing Similarity to 'Ignore each atom's hydrogen count' and 'Ignore each atom's bond count' produces a 15/15 overlay with an $RMSD_{15}$

of just 0.065 Å. However, a routine use of these settings is not suitable, since comparisons would not be restricted to crystal structures of the same compounds; for example, tautomers would not be distinguished. Furthermore, not accounting for bond types when comparing crystal structures generated in a CSP study would hinder the detection of possible static disorder.[1, 7]

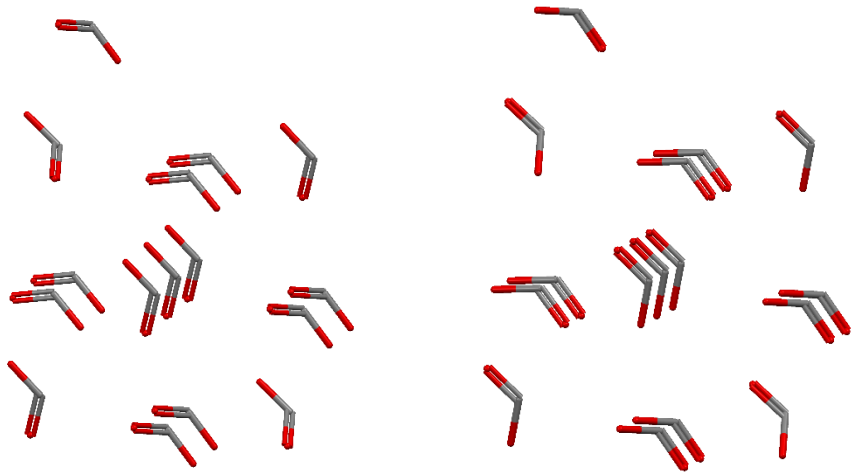


Figure 8.6: Comparison between the crystal structures of FORMAC (left) and FORMAC01 (right).

Out of the three pairs of duplicates not identified because they exceed the $RMSD_{15}$ thresholds, two (GLYCIN16-GLYCIN82 and GLYCIN71-GLYCIN81) are redeterminations where one crystal structure was determined in the P1 space groups, despite its real symmetry being consistent with its counterpart. Since the crystal structures had been incorrectly reported in a different space group and with a different Z', the methodology in Figure 8.2 applies a different similarity threshold (0.1 rather than 0.5 Å, see Chapter 8.2.1 for the motivation of this difference), which is exceeded in both cases. Although this should not be a problem in CSP, where symmetry is generally imposed when a crystal structure is generated, this shows that the $RMSD_{15}$ thresholds can in some cases lead to the presence of false polymorphs. For the other false polymorph (HOJQII-HOJQII01) the two structures have the same space group and Z', but a very large $RMSD_{15}$ value of ~0.77 Å, significantly above the 0.5 Å threshold. However a more careful analysis of the publication reveals that the crystal structure of HOJQII01 was determined at 2.5 GPa,[53] although this is not flagged in the CSD. This shows how careful analysis of database information data is often needed, since the CSD contains some errors, and also illustrates how much structural change pressure can cause without inducing phase transformation. This effect is much larger than it is for temperature, which at least in this test does not cause the presence of false polymorphs.

8.3.2 Test on the whole CSD

8.3.2.1 Overall results on the entire CSD

The algorithm based on the decision tree in Figure 8.2 produced 6,550 clusters of crystal structures, in most cases matching the polymorph flags given by the experimentalists. Overall, the algorithm identified 4,197 pairs of crystal structures of the same compounds belonging to different clusters, which should be polymorphs. It has also identified 8,200 pairs of structure belonging to the same clusters, which should be duplicates/redeterminations. Although not all those pairs of duplicates were directly compared with one another, it can be assumed that since they all match with the same crystal structure, then they all match with one another (*i.e.* if structure 2 and structure 3 both meet all the similarity criteria when compared to structure 1, then it is assumed structures 2 and 3 would meet all the criteria if compared with one another). The overall outcome of the analysis is summarised in Figure 8.7.

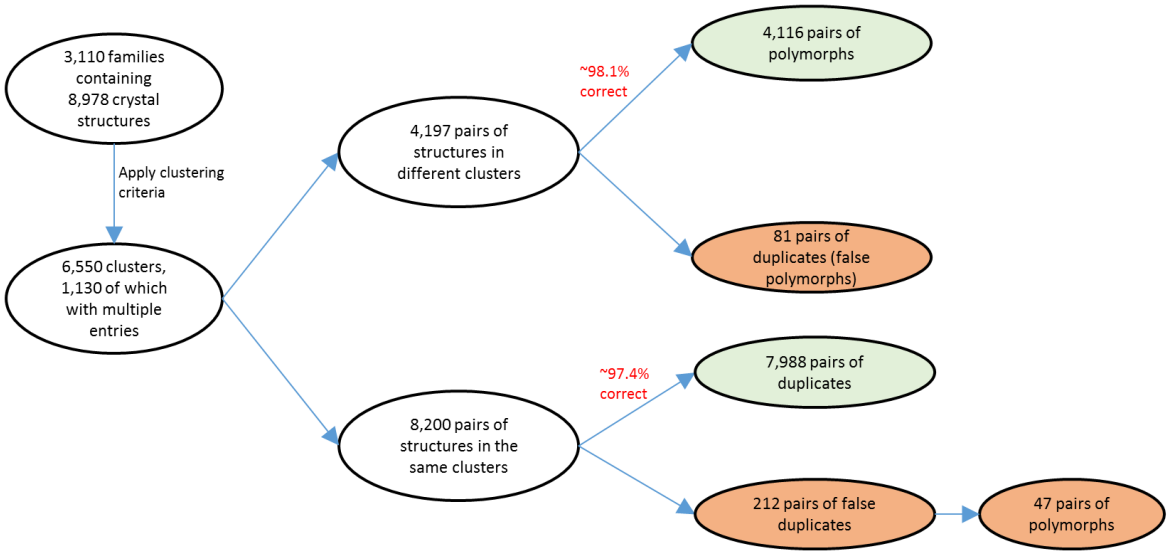


Figure 8.7: Summary of the results of the analysis of the polymorph-flagged CSD crystal structures. The pairs of false polymorphs are listed in Appendix Table 8.3 and the pairs of false duplicates in Appendix Table 8.4.

The methodology is successful in the great majority of cases, with ~98% of the pairs of structures in different clusters being actual polymorphs and ~97% of the pairs of structures in the same clusters being actual duplicates. This analysis confirms that polymorphs are usually structurally different: for 3,911 of the 4,163 polymorph pairs (*i.e.* the 4,116 found pairs of polymorphs and the 47 missed ones listed in Appendix Table 8.5) it was impossible to overlay 15/15 molecules. Of the 252 pairs with a 15/15 overlay, only 128 do not have any of the crystallographic differences; 47 of those 128 do not have the $RMSD_{15}$ differences and are output as false duplicates.

On the other hand, very few of the structures that fail to meet any of these criteria are found to be duplicates: only ~1.4% of the pairs of structures for which a 15/15 overlay

is impossible, although this may be overestimated because of some of the problems highlighted in Chapter 8.3.2.2, and only ~1% of the structures that do not meet the crystallographic criteria are duplicates. Furthermore, out of the structures for which a 15/15 overlay is possible but exceed the $RMSD_{15}$ thresholds, ~9% are duplicates. This means that the heuristically developed $RMSD_{15}$ thresholds are effective in the majority of cases. Figure 8.8 summarises the overall analysis for each step of the decision tree in Figure 8.2.

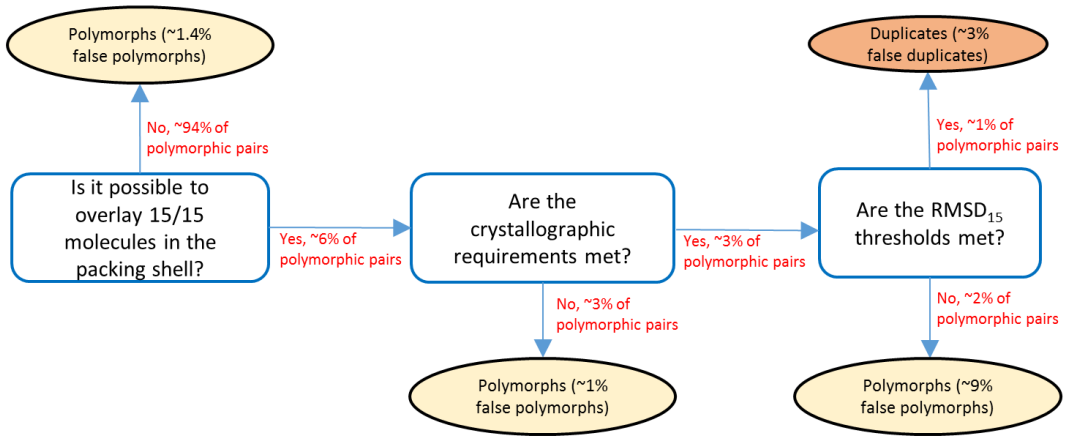


Figure 8.8: Summary of the results of the test on the polymorph-flagged CSD crystal structures for each step of the methodology in Figure 8.2.

The next question is whether the degree of success of the decision tree depends on the characteristic of the compounds present in the crystal structures. Table 8.1 shows a breakdown of the compounds in the crystal structures that were considered in this work, in terms of the number of components, flexibility and size. A crystal structure was considered to contain a 'large' molecule if the molecule or the largest molecular component (for multi-component compounds) contained more than 18 heavy atoms, a criterion used in previous studies on polymorphism,[30] and 'small' otherwise. Furthermore, it was considered as 'flexible' if the molecule or the largest molecular component had at least one rotatable bond, as defined by Mogul,[54] and 'rigid' otherwise.

Table 8.1: Breakdown of the results of the analysis on polymorphism, in terms of the number of molecular components and the molecular properties.

	Number of components		Flexibility		Size	
	Single-component	Multi-component	Flexible	Rigid	Small	Large
Total number of compounds	2,452	658	2,641	469	2,039	1,071
Compounds containing false polymorphs	50 (~2.0%)	16 (~2.4%)	49 (~1.8%)	17 (~3.6%)	34 (~1.7%)	32 (~3.0%)
Compounds containing false duplicates	30 (~1.2%)	11 (~1.7%)	30 (~1.1%)	11 (~2.3%)	27 (~1.3%)	14 (~1.3%)

An analysis of these data reveals that the percentage of compounds containing at least one pair of false duplicates and/or false polymorphs seems to be almost

independent of the number of molecular components (although they are both slightly over-represented among multi-component compounds), indicating that criteria in Figure 8.2 are equally effective in discriminating polymorphs from duplicates in single-component and multi-component crystal structures. On the other hand, compounds containing rigid-molecules are slightly over-represented in both false polymorphs and false duplicates. The over-representation of rigid molecules in false duplicates may be caused by the fewer degrees of freedom, which can make them exhibit similar crystal structures even in the presence of different phases;[12] although slightly tighter $RMSD_{15}$ thresholds may help for rigid molecules, the discrepancy is small and probably not statistically significant. On the other hand, the over-representation of rigid molecules within compounds containing false polymorphs is counterintuitive and probably due to historical reasons: ~26% of the lowest R-factor crystal structures of the rigid compounds were characterised before 1980, compared to the ~14% of the whole set, and for older entries errors tend to be more prevalent. Finally, it is surprising that there is no correlation between molecular size and number of false duplicates, since one may expect that smaller molecules can exhibit smaller differences for similar but distinct polymorphs. On the other hand, larger molecules are over-represented among false polymorphs, which is sensible since the presence of more atoms means that duplicates may have slightly larger structural differences than smaller molecules. Once again, although more relaxed $RMSD_{15}$ thresholds may help for larger molecules, these small differences are probably not significant enough. Hence, the criteria outlined in Figure 8.2 appear to be broadly applicable and are effective in the great majority of cases for any sort of compound.

In summary, although certain properties associate with a higher proportion or false polymorphs and/or false duplicates, the overall number of compounds containing at least one error is always between ~3-6% for any molecular characteristic and number of components. The decision tree in Figure 8.2 is effective in discriminating the majority of ambiguous cases, independently of the characteristic of the compounds. The next sections will analyse the few failures and their causes.

8.3.2.2 Identification of false duplicates and their causes

A total of 41 families of compounds containing false duplicates were identified. Out of those 41 families, in 39 two polymorphs were present within each family, in one case three polymorphs were present and in one case four. This results in 47 unfound polymorphic pairs, forming 212 pairs of false duplicates (this number includes pairs of duplicates of these 47 polymorphs), which are listed in Appendix Tables 8.4 and 8.5. Analysing these false duplicates in details, it was found that in only one case (ZZZHQU01-ZZZHQU02) the polymorphic nature was not identified because of the inadequacy of the 15-molecule overlay itself, since there are longer-range differences,

as shown by the possibility of matching only 23 molecules within a larger 30-molecule shell. This is very promising, and it appears that 15-molecule clusters are almost always adequate and that there is no need to use larger and more computationally demanding packing shells. In the remaining 211 pairs, the structures were nearly identical. In 27 of 47 unfound polymorphic pairs the publication reports that the polymorphic difference was caused by a phase transition happening at a given temperature, and in 6 pairs to a phase transition happening with pressure (although the crystal structures were determined at 0 GPa).[55] This confirms what had been found in the analysis of the 'best R-factor' list, *i.e.* that phase transitions can cause polymorphic changes with very small structural differences, making them the biggest source of false duplicates. The remaining 14 cases are more ambiguous, no phase transition is described despite the publications stating presence of different phases. This shows that the distinction between experimental polymorphs and duplicates can in some cases be difficult and subject to the interpretation of the crystallographer.

8.3.2.3 Identification of false polymorphs and their causes

A total of 81 pairs of false polymorphs were identified; this number might be underestimated, since crystal structures classified as polymorphs but with identical polymorph flags may have not been identified because of some of the possible errors in the flags that are listed in Chapter 8.2.2.

In 35 pairs the crystal structures would meet all three criteria in the decision tree if double and single bonds were not distinguished. In 23 of those pairs one of the crystal structures does not have the hydrogen atoms explicitly indicated, which may cause errors in the automatic determination of the double-bonds, like in the example shown in Figure 8.6. On the other hand, an interesting pair of false polymorphs in which hydrogen atoms of both crystal structures are indicated is DBEZLM01-DBEZLM05, shown in Figure 8.9. The different position of the hydrogen atoms and the double bonds may be due to an uncertainty possibly caused by tautomeric disorder, although this is not mentioned in the publication. Tautomeric disorder is explicitly mentioned for QQQFDJ19 and QQQFDJ20,[56] indicating how this phenomenon can cause some ambiguous cases.

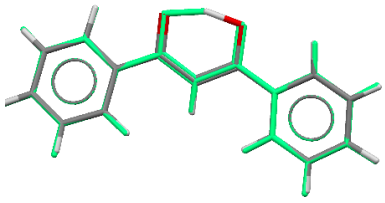


Figure 8.9: 1-molecule overlay of the crystal structures of DBEZLM01 (coloured by elements) and DBEZLM05 (in green). See Appendix Table 8.3 for details.

Out of the remaining 46 pairs of false polymorphs, in 22 cases the algorithm did not match 15/15 molecules. A further 11 pairs were considered polymorphic because

they differed in the crystallographic parameters, and in 13 cases because they met both the overlay and crystallographic criteria but had $RMSD_{15}$ values above the thresholds.

Out of the 22 cases where the Crystal Packing Similarity tool failed to produce a 15/15 match, 13 were due to the 20% distance and a 20° angle tolerances being too tight, and increasing the tolerances to 40% and 40° respectively found a full overlay; generally in those pairs one of the crystal structures is a poor determination, often with some wrong covalent bonds, as exemplified in Figure 8.10 for the intuitively wrong phenyl rings of ANTCYB11, compared to the correct ones of ANTCYB13.

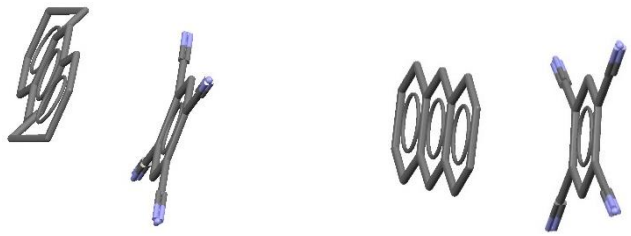


Figure 8.10: Molecular conformations contained in (left) ANTCYB11 and (right) ANTCYB13. See Appendix Table 8.3 for details. The phenyl rings in ANTCYB11 are intuitively wrong.

Hence, in a limited number of cases poor determinations can make discriminating polymorphs from duplicates complicated. However, increasing the tolerances would be counterproductive since this could generate more false duplicates.

A further seven cases were caused by the presence of disorder in the hydrogen atoms. In the presence of disorder, the overlay often fails since the Crystal Packing Similarity tool cannot physically match the suppressed atoms. In the algorithm utilised for this test, crystal structures are compared only with the lowest R-factor counterpart within each cluster. For those seven pairs, the overlay fails with the crystal structure with the lowest R-factor entry within the cluster, but all the criteria are met when a comparison is attempted with other duplicates of that structure. Of the remaining two cases, one was caused by one crystal structure within the pair having a wrong molecular geometry (TFMETH, with a large 15% R-factor, and TFMETH02) and one by the Crystal Packing similarity tool showing a 3/15 molecule overlay despite a visually perfect match for NOETNA01-NOETNA02. Those problems signal the need of being cautious when poor or disordered crystal structures are compared.

Nine of the 11 pairs of duplicates with the crystallographic differences that according to the criteria in Figure 8.2 indicate the presence of polymorphs contained symmetric molecules. More specifically, one crystal structure had a lattice symmetry element correspondent to a molecular symmetry element, while the counterpart did not have it; this leads to different Z values. The pair formed by BENZID04-BENDIZ08, shown in Figure 8.11, is an example of this: BENZID04 has an inversion centre at the symmetric centre of the molecule that is absent in BENZID08. Hence BENZID04 has a Z value of 1

and BENZID08 of 2, although they are both in the $P\bar{1}$ space group and have nearly identical crystal structures.

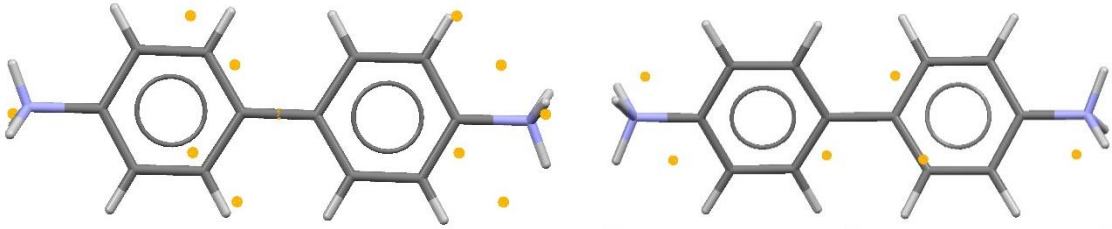


Figure 8.11: Molecule and symmetry elements of (right) BENZID04 and (left) BENZID08. See Appendix Table 8.3 for details. Those crystal structures are both in the $P\bar{1}$ space group and a 15/15 molecule overlay is possible with an $RMSD_{15}$ of 0.181 Å. The yellow dots represent crystalline inversion centres, and in BENZID04 one of the inversion centre is located half the way through the central C-C bond.

This shows that symmetric molecules can sometimes be challenging, and that they may need careful consideration. One possible solution could be not to consider Z values when comparing crystal structures; however, this would lead to several more false duplicates. The other two pairs of false polymorphs caused by crystallographic differences are OCRSOL-OCRSOL01 and GLYCIN16-DOLBIR09: these structures have the same Z and Z' value, but OCRSOL and DOLBIR09 are in the $P3_1$ space group, while OCRSOL01 and GLYCIN16 are in $P3_2$. This highlights how the determination of crystalline symmetry can in some cases be arbitrary, which can cause problems in the separation of polymorphs from duplicates.

Out of the 13 pairs of false polymorphs caused by the $RMSD_{15}$ thresholds, in four cases the Crystal Packing Similarity tool gave unrealistically high $RMSD_{15}$ values, despite visually excellent matches, probably because of errors in the overlays themselves. One example is DXYLEN14-DXYLEN15, shown in Figure 8.12.

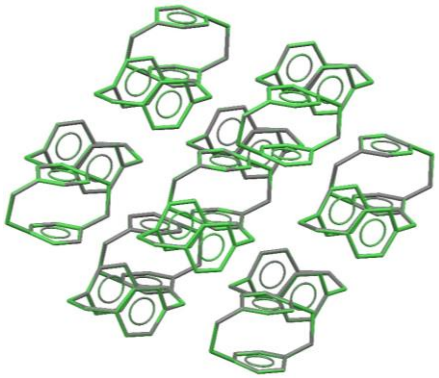


Figure 8.12: 15-molecule overlay of DXYLEN14 (coloured by elements) and DXYLEN15 (in green). See Appendix Table 8.3 for details. Despite a clearly excellent agreement, the tool returns an unrealistic $RMSD_{15}$ of 2.335 Å. This has to be due to an error in the Crystal Packing Similarity tool.

Of the remaining nine pairs of false polymorphs, six pairs had different Z' and space group and an $RMSD_{15}$ value slightly larger than the 0.1 Å threshold, while three

had identical Z' and space group and an $RMSD_{15}$ slightly above the 0.5 Å threshold. These few cases once again fall into that category of ambiguity, where there is a degree of ambivalence on whether crystal structures may be polymorphic. In five of these nine pairs the crystal structures were determined at different temperatures. Loosening the $RMSD_{15}$ thresholds would once again be dangerous since it could produce more false duplicates. However, despite those few failures, the heuristically-determined $RMSD_{15}$ thresholds appear to be effective in most cases, as shown in Figure 8.7 and Figure 8.8.

8.3.2.4 Polymorphic differences and possible alternative methodologies

The results of the analysis on the whole CSD are very similar to those from the 'best R-factor list', from which the methodology in Figure 8.2 was heuristically developed, although the percentage of false duplicates is slightly higher. However, most of those false duplicates are not present in the 'best R-factor list' because of their high level of simulated PXRD similarity (as shown in Figure 8.14), which would have not allowed their identification as separate polymorphs,[38] and because some of them are old entries with poor R-factors. On the other hand, comparing the number of false polymorphs would be misleading, since the 'best R-factor list' contains very few flagged duplicates.

Although in the great majority of cases polymorphs have very different crystal structures and duplicates nearly identical ones, there are several polymorphic pairs with very similar crystal structures, and some duplicates with rather different packings; the most extreme examples found in this study are BIZWAJ01-BIZWAJ02 and SUWMIG-SUWMIG03, which are shown in Figure 8.13.

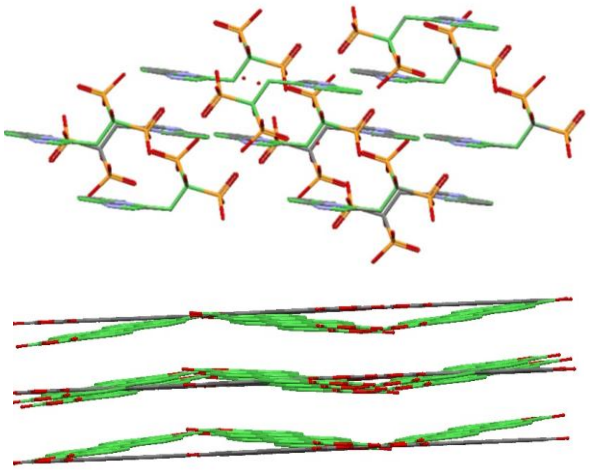


Figure 8.13: 15-molecule overlays of (above) BIZWAI01 and BIZWAI02, with an $RMSD_{15}$ of just 0.007 Å, despite being explicitly indicated as polymorphs in the publication[57] (below) SUWMIG and SUWMIG03, which require looser 40% distance and 40° angle tolerances to obtain a 15/15 overlay with an $RMSD_{15}$ of 1.16 Å, despite being explicitly indicated as redeterminations in the publication.[58]

Polymorphs with similar crystal structures are more frequent than duplicates with dissimilar packings. In fact, out of the 78 pairs of false polymorphs, the majority are not

identifiable as duplicates either because of errors in the experimental data and in the overlaying tools or short-comings of the adopted clustering methodology. There are only a handful of ambiguous cases due to the actual crystalline packing. On the other hand, out of the 212 pairs of polymorphs found in the same clusters, only one was not identified because of short-comings in the clustering methodology, while the other 211 are different phases with very similar crystal structures. The majority of those problematic polymorphic pairs are associated with phase transitions with temperature or pressure. Although phase transition seem to be associated with only a small fraction (~10%) of the total number of polymorphs in the CSD,[59] they represent a much larger portion (~70%) of the false duplicates identified in this study. Overall it is highly unlikely that very structurally similar crystal structures are polymorphic, in particular when determined at the same temperature and pressure; hence, structurally similar CSP-generated structures can be safely clustered with a low chance of removing potential polymorphs.

Using the criteria in Figure 8.2 for performing a clustering analysis seems to be more effective than utilising simulated PXRD similarities. Figure 8.14 shows the distribution of the simulated PXRD similarities for the pairs structures identified as duplicates (including the false polymorphs) and polymorphs (including the false duplicates) in this study

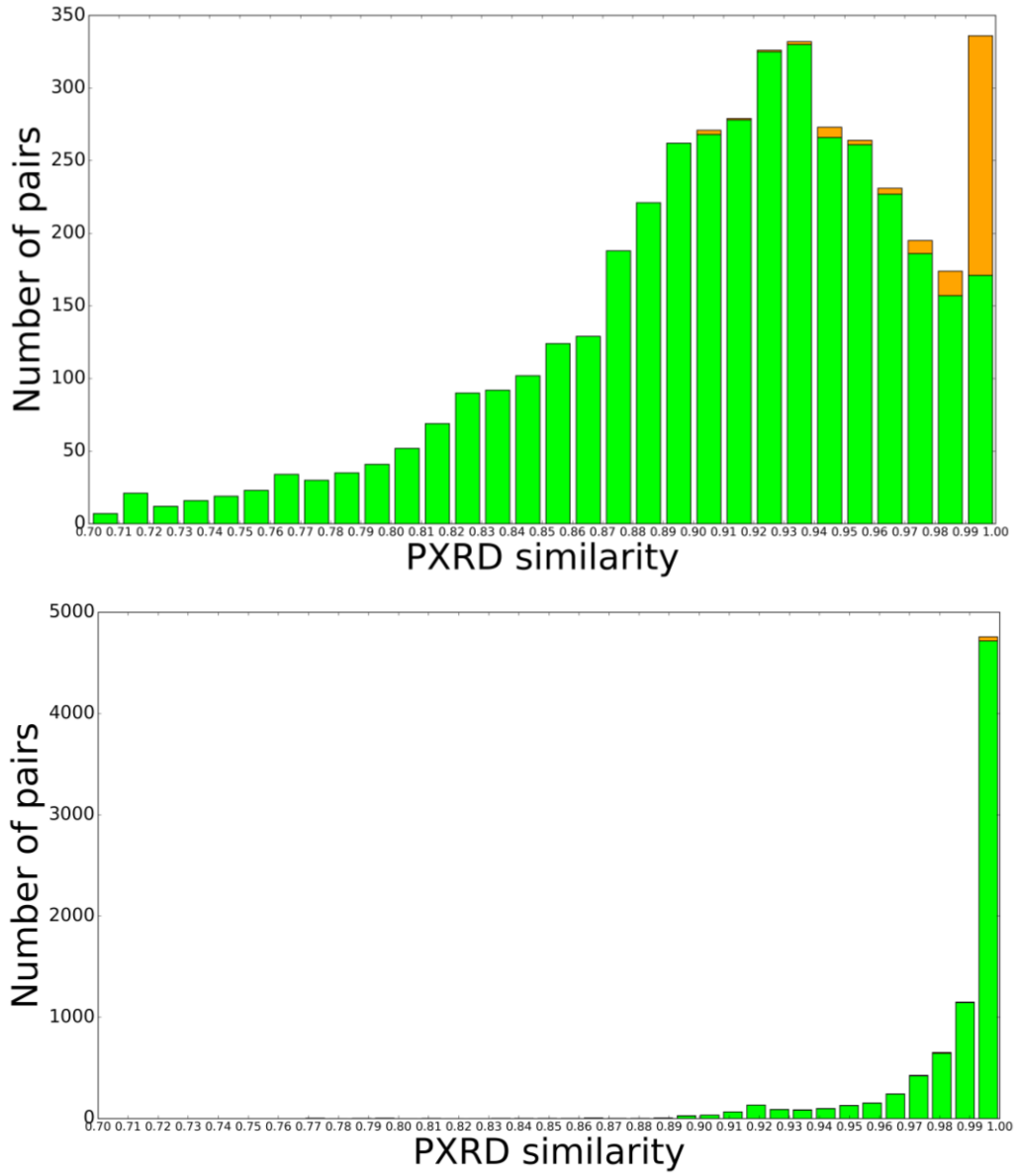


Figure 8.14: Histograms showing the simulated PXRD similarities of (above) the polymorphic pairs identified; the false duplicates are indicated by orange bars (below) the pairs of duplicates identified; false polymorphs are indicated by orange bars.

Duplicates clearly have much higher simulated PXRD similarities than polymorphs, which is expected since duplicates tend to have very similar crystal structures and polymorphs very different ones. However, there are large numbers of polymorphs with high simulated PXRD similarities and vice versa, confirming the results shown in Figure 8.1. The method used to compile the 'best R-factor' considers as definite duplicates structures with a simulated PXRD similarity above 0.990,[38] and 336 polymorphic pairs of crystal structures would be considered as duplicates (*i.e.* they would be false duplicates), compared to the 212 in this study; as shown by the orange bars in Figure 8.14, simulated PXRD similarities would have failed to identify as polymorphs most of the false duplicates produced by the clustering methodology in Figure 8.2. On the other hand, 1,114 pairs of duplicates would be considered as certain polymorphs,

since they have a PXRD similarity smaller than 0.970 (*i.e.* they would be false polymorphs), compared to the 81 in this study. However, this may be an exaggeration since the simulated PXRD similarities calculated here are not exactly consistent with those used to produce the 'best R-factor' list, as they are not compared to those calculated with a reduced unit cell, the volumes are not normalised and the similarities are not adjusted for temperature and/or pressure differences.[38] Also, 372 pairs of polymorphs and 1,717 pairs of duplicates have a simulated PXRD similarity between 0.970 and 0.990 that would qualify them as unknown.[38] However they are classified, this would likely further increase both false polymorphs and false duplicates. On the other hand, simulated PXRD similarities often succeed to recognise some of the duplicates that were mistakenly identified as polymorphs using the method in Figure 8.2 because of differences in double bonds, experimental errors or errors in the overlay tools.

Overall, the heuristic methodology identifies wrongly ~2% of the structures it classifies as polymorphs and ~3% of the structures it classifies as duplicates, while applying the simulated PXRD similarity-based criteria used to compile the 'best R-factor' list these percentages increase to ~4% for duplicates and ~26% of polymorphs. Although the final number is probably exaggerated, both percentage would likely be increased by classifying structures falling into the ambiguous category in terms of PXRD similarity. Hence, the method utilised in this chapter appears to be more reliable as a whole.

8.3.3 Test on CSP-generated structures

The results of clustering the generated crystal structures of molecule XXVI and the two tautomers of mebendazole are shown in Figure 8.15-17 and Table 8.2.

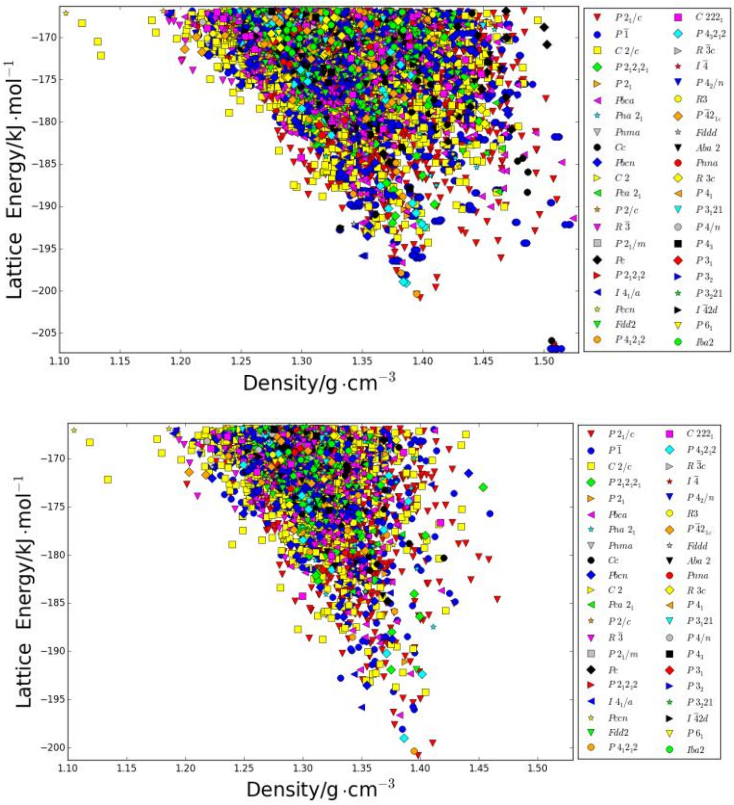


Figure 8.15: Lattice energy vs density plot obtained after the search with CrystalPredictor of molecule XXVI (above) before and (below) after clustering.

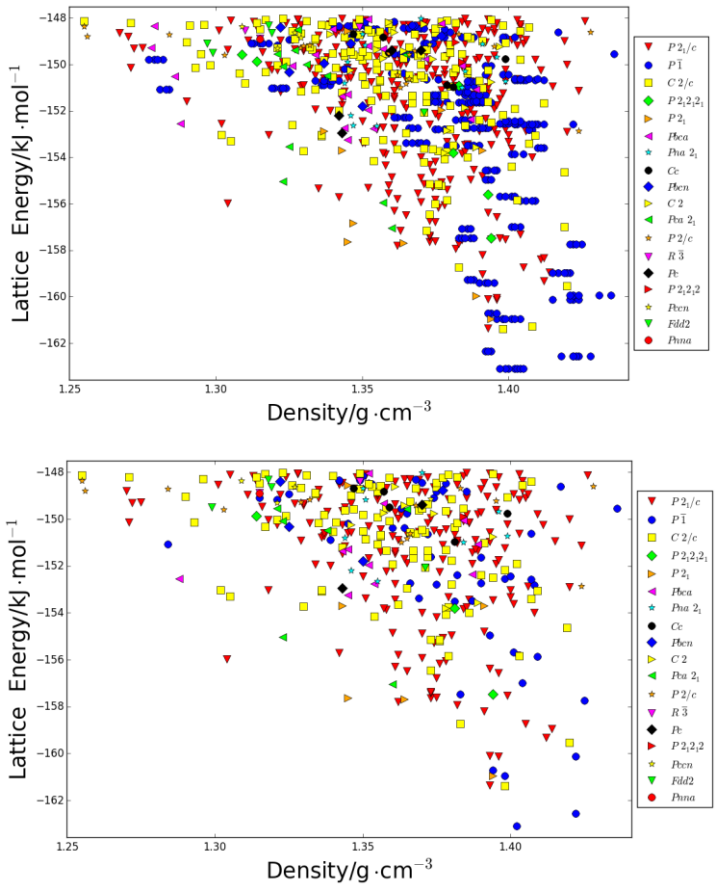


Figure 8.16: Lattice energy vs density plot obtained after the search with CrystalPredictor of the A-tautomer of mebendazole (above) before and (below) after clustering.

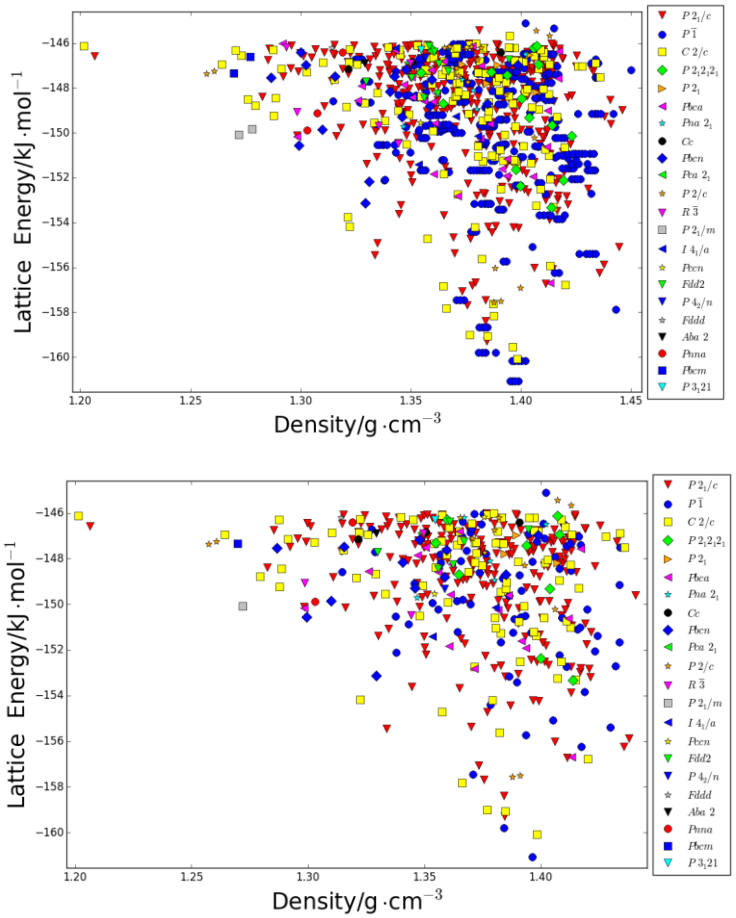


Figure 8.17: Lattice energy vs density plot obtained after the search with CrystalPredictor of the C-tautomer of mebendazole (above) before and (below) after clustering.

Table 8.2: Summary of the results of clustering the search-generated crystal structures of molecule XXVI and the two tautomers of mebendazole.

Molecule	# generated structures	# duplicates	# wrong structures	% of structures eliminated	# lost minima (% of duplicates)	Clustering cost /CPU hours	Estimated saving /CPU hours
XXVI	9,400	2,580	2,623	55	623 (24%)	1,625	88,000
Meb. A	855	437	0	51	29 (7%)	5	1,250
Meb. C	964	496	0	51	43 (9%)	6	1,060

Clustering removed a large portion of CSP-generated structures for all three molecules. If the structures removed by this clustering analysis had not been taken to the final refinement stage of the original CSP studies, this would have led to important cost savings. The estimated savings are much larger than the computational cost required to perform the clustering step, and amount to approximately 30-40% of the computational expense for the final refinement stage, which is shown in Table 3.3 for molecule XXVI and Table 4.2 for mebendazole. Hence this could be an advantageous intermediate stage within a CSP workflow.

However, undertaking this extra step can be dangerous, since a significant proportion of the removed structures (7 to 24% in these three cases) did not optimise to

the same minimum as the structures they were considered duplicates of. This is a known problem associated with removing presumptive duplicates at early stages of CSP studies.[3] In this case, this loss is probably due to the E_{latt} model in CrystalPredictor being different from the one used in CrystalOptimizer,[60] which was used to refine the generated crystal structures. The extent to which changing the energy surface can take similar structures to different end-points is shown in Figure 8.18 for CSP-generated structures 2494 and 2495 of molecule XXVI.

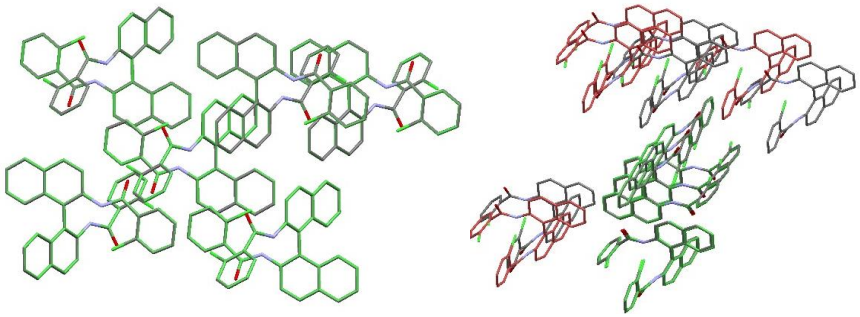


Figure 8.18: Overlay of structures 2494 and 2495 (right) after the search, where it is possible to overlay 15/15 molecules with an $RMSD_{15}$ of 0.006 (left) after the single-iteration of CrystalOptimizer, where it is only possible to overlay 5/15 molecules. Molecules of 2494 are coloured by elements, while for 2495 matched molecules are coloured in green, non-matched molecules in red.

Hence there is no guarantee that two crystal structures that are duplicates at one stage of the CSP workflow will remain duplicates at later stages. Although for these three molecules the structures matching the experimental forms were not thrown away, it appears to be risky to remove duplicates produced exploring a poor energy surface before refinement with a more accurate one.

The decision of whether or not to undertake a clustering step comes down to the priority of the CSP study itself. If what is prioritised is the generation of a very accurate crystal energy landscape, and time or CPU cost constraints are not a problem, it is probably best to optimise all the search-generated structures within a certain energy window. On the other hand, if the resources are limited, then the risk of losing some potentially important minima may be accepted if clustering search-generated structures significantly reduces number of expensive optimisations. The larger and more flexible the molecule is, the more the decrease in CPU cost can become convenient and worth the partial loss in accuracy; however, this limited analysis seems to suggest that the number of lost minima could increase with molecular size.

An alternative solution is the use of tighter constraints in terms of energy or density differences, PXRD similarity or unit-cell parameters, or even the $RMSD_{15}$ thresholds to discriminate between duplicates and polymorphs, which could be adapted depending on the size and flexibility of the molecule. Although this would reduce the cost of the clustering step and the risk of removing crystal structures that would optimise to

separate minima, the number of identified duplicates would also decrease, limiting the savings in computational cost and the value of the intermediate clustering step.

8.4 Conclusion

An analysis of the CSD 'best R-factor list' allowed to define a set of criteria that can be used to separate the great majority of duplicates from polymorphs. A more complete test on the polymorph-flagged crystal structures in the CSD revealed that these criteria work in ~98% of cases. The ~2% of few failures are either due to errors in the experimental entries or in the overlaying tools, or to the presence of ambiguous pairs of crystal structures; the latter are generally associated with phase-transitions with temperature or pressure. In many cases, the ambiguity requires a certain level of human judgment to decide whether two structures are polymorphic or not,[11] or more accurate experimental measurements. However, most polymorphs and duplicates are easy to distinguish, and these ambiguous cases are exceptions rather than the rule.

The development of a new clustering script and the application of these criteria to a set of the search-generated crystal structures of molecule XXVI and the two tautomers of mebendazole shows that removing duplicates at early stages of a CSP procedure can be both cost-effective and risky: the large number of removed crystal structures vastly reduces the computational cost of later refinement stages, but several structures that optimise to distinct minima are eliminated, limiting completeness.

This chapter provides several insights that are useful within the overall context of this thesis. First of all, removing duplicates produced after the generation stage can be important in reducing the computational cost of CSP studies on large molecules; however, a compromise must be found, since there is a risk of removing structures that are effectively distinct minima. Comparing and removing only structures that are similar not only in crystal structure but also in energy, density, unit cell parameters or PXRD similarity could reduce such risks, but would also reduce the worth of the clustering procedure itself. Secondly, the CSD analysis has produced some widely applicable criteria for separating polymorphs from duplicates that are not drastically affected by molecular size or flexibility, as well as the number of components in the crystal structures. Finally, the risk of similar crystal structures forming separate polymorphic phases is extremely small, in particular if they are obtained or predicted under the same conditions. Although this risk cannot be ruled out, this study increases the confidence that a careful analysis of CSP-generated crystal structures can identify a set of distinct PPMs.

8.5 References

1. Price, S. L., Predicting crystal structures of organic compounds. *Chemical Society Reviews* **2014,** *43* (7), 2098-2111.

2. 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.; Vacarro, F.; Vasileiadis, M.; Vazquez-Mayagoitia, A.; Vogt, L.; Wang, Y.; Watson, R. E.; de Wijs, G. A.; Yang, J. Z.; Zhu, Q.; Groom, C. R., Report on the sixth blind test of organic crystal-structure prediction methods. *Acta Crystallographica Section B - Structural Science* **2016**.
3. van Eijck, B. P.; Kroon, J., Fast clustering of equivalent structures in crystal structure prediction. *Journal of Computational Chemistry* **1997,** *18* (8), 1036-1042.
4. Ghasemi, S. A.; Amsler, M.; Hennig, R. G.; Roy, S.; Goedecker, S.; Lenosky, T. J.; Umrigar, C. J.; Genovese, L.; Morishita, T.; Nishio, K., Energy landscape of silicon systems and its description by force fields, tight binding schemes, density functional methods, and quantum Monte Carlo methods. *Physical Review B* **2010,** *81* (21), 214107.
5. Verwer, P.; Leusen, F. J. J., Computer Simulation to Predict Possible Polymorphs. In *Reviews in Computational Chemistry Volume 12*, Lipkowitz, K. B.; Boyd, D. B., Eds. John Wiley and Sons Inc.: New York, 1998; pp 327-365.
6. Day, G. M., Current approaches to predicting molecular organic crystal structures. *Crystallography Reviews* **2011,** *17* (1), 3-52.
7. Copley, R. C. B.; Barnett, S. A.; Karamertzanis, P. G.; Harris, K. D. M.; Kariuki, B. M.; Xu, M. C.; Nickels, E. A.; Lancaster, R. W.; Price, S. L., Predictable disorder versus polymorphism in the rationalization of structural diversity: A multidisciplinary study of eniluracil. *Crystal Growth & Design* **2008,** *8* (9), 3474-3481.
8. Case, D.; Campbell, J.; Bygrave, P.; Day, G., Convergence Properties of Crystal Structure Prediction by Quasi-Random Sampling. *Journal of Chemical Theory and Computation* **2016,** *12* (2), 910-924.
9. Hofmann, D. W. M.; Kuleshova, L., New similarity index for crystal structure determination from X-ray powder diagrams. *Journal of Applied Crystallography* **2005,** *38*, 861-866.
10. Valle, M.; Oganov, A. R. In *Crystal structures classifier for an evolutionary algorithm structure predictor*, 2008 IEEE Symposium on Visual Analytics Science and Technology, 19-24 Oct. 2008; 2008; pp 11-18.
11. Desiraju, G. R., Polymorphism: The same and not quite the same. *Crystal Growth & Design* **2008,** *8* (1), 3-5.
12. Coles, S. J.; Threlfall, T. L.; Tizzard, G. J., The Same but Different: Isostructural Polymorphs and the Case of 3-Chloromandelic Acid. *Crystal Growth & Design* **2014,** *14* (4), 1623-1628.
13. Gavezzotti, A., A solid-state chemist's view of the crystal polymorphism of organic compounds. *Journal of Pharmaceutical Sciences* **2007,** *96* (9), 2232-2241.
14. Bernstein, J.; Dunitz, J. D.; Gavezzotti, A., Polymorphic Perversity: Crystal Structures with Many Symmetry-Independent Molecules in the Unit Cell. *Crystal Growth & Design* **2008,** *8* (6), 2011-2018.
15. Bernstein, J., Polymorphism - A Perspective. *Crystal Growth & Design* **2011,** *11* (3), 632-650.
16. 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.
17. Bernstein, J., *Polymorphism in Molecular Crystals*. Clarendon Press: Oxford, 2002.
18. McNaught, A. D.; Wilkinson, A., *IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book")*. WileyBlackwell; 2nd Revised edition edition.
19. Stahly, G. P., Diversity in single- and multiple-component crystals. The search for and prevalence of polymorphs and cocrystals. *Crystal Growth & Design* **2007,** *7* (6), 1007-1026.
20. Threlfall, T. L.; Gelbrich, T., The crystal structure of methyl paraben at 118 K does not represent a new polymorph. *Crystal Growth & Design* **2007,** *7* (11), 2297-2297.
21. Fabian, L.; Kalman, A., Isostructurality in one and two dimensions: isostructurality of polymorphs. *Acta Crystallographica Section B - Structural Science* **2004,** *60*, 547-558.

22. Beer, L.; Brusso, J. L.; Cordes, A. W.; Haddon, R. C.; Itkis, M. E.; Kirschbaum, K.; MacGregor, D. S.; Oakley, R. T.; Pinkerton, A. A.; Reed, R. W., Resonance-Stabilized 1,2,3-Dithiazolo-1,2,3-dithiazolyls as Neutral π-Radical Conductors. *Journal of the American Chemical Society* **2002,** *124* (32), 9498-9509.
23. Barnett, S. A.; Broder, C. K.; Shankland, K.; David, W. I. F.; Ibberson, R. M.; Tocher, D. A., Single-crystal X-ray and neutron powder diffraction investigation of the phase transition in tetrachlorobenzene. *Acta Crystallographica Section B - Structural Science* **2006,** *62*, 287-295.
24. Torrie, B. H.; Anderson, A.; Andrews, B.; Laurin, D. G.; White, J. K.; Zung, W. W. E., Raman and far-infrared spectra of crystalline methylene iodide. *Journal of Raman Spectroscopy* **2005,** *18* (3), 215-220.
25. Silvestru, A.; Haiduc, I.; Ebert, K. H.; Breunig, H. J., Novel Coordination Pattern of Dithiophosphorus Ligands. Crystal and Molecular Structure of (Diphenylphosphinodithioato)phenyltellurium(II), PhTeS2PPh2. Supramolecular Association through Monodentate Biconnective Dithiophosphorus Ligands. *Inorganic Chemistry* **1994,** *33* (7), 1253-1254.
26. Abbas, N.; Oswald, I.; Pulham, C., Accessing Mefenamic Acid Form II through High-Pressure Recrystallisation. *Pharmaceutics* **2017,** *9* (2).
27. van de Streek, J.; Motherwell, S., Searching the Cambridge Structural Database for polymorphs. *Acta Crystallographica Section B - Structural Science* **2005,** *61*, 504-510.
28. Stephenson, G., Anisotropic lattice contraction in pharmaceuticals: The influence of cryo-crystallography on calculated powder diffraction patterns. *Journal of Pharmaceutical Sciences* **2006,** *95* (4), 821-827.
29. Cruz-Cabeza, A. J.; Bernstein, J., Conformational Polymorphism. *Chemical Reviews* **2014,** *114* (4), 2170-2191.
30. Cruz-Cabeza, A. J.; Reutzel-Edens, S. M.; Bernstein, J., Facts and fictions about polymorphism. *Chemical Society Reviews* **2015,** *44*, 8619-8635.
31. Storey, R. A.; Ym,n, I., *Solid-state characterization of pharmaceuticals*. Wiley: Chichester, 2012.
32. Yang, J.; Hu, W.; Usvyat, D.; Matthews, D.; Schutz, M.; Chan, H., Ab initio determination of the lattice energy in crystalline benzene to sub-kilojoule per mole accuracy. *Science* **2014,** *345* (6197), 640-643.
33. Nyman, J.; Day, G. M., Static and lattice vibrational energy differences between polymorphs. *Crystengcomm* **2015,** *17* (28), 5154-5165.
34. Zhu, L.; Amsler, M.; Fuhrer, T.; Schaefer, B.; Faraji, S.; Rostami, S.; Ghasemi, S. A.; Sadeghi, A.; Grauzinyte, M.; Wolverton, C.; Goedecker, S., A fingerprint based metric for measuring similarities of crystalline structures. *Journal of Chemical Physics* **2016,** *144* (3).
35. Spackman, M. A.; Jayatilaka, D., Hirshfeld surface analysis. *CrystEngComm* **2009,** *11* (1), 19-32.
36. Carter, D. J.; Raiteri, P.; Barnard, K. R.; Gielink, R.; Mocerino, M.; Skelton, B. W.; Vaughan, J. G.; Ogden, M. I.; Rohl, A. L., Difference Hirshfeld fingerprint plots: a tool for studying polymorphs. *CrystEngComm* **2017,** *19* (16), 2207-2215.
37. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C., The Cambridge Structural Database. *Acta Crystallographica Section B-Structural Science Crystal Engineering and Materials* **2016,** *72*, 171-179.
38. van de Streek, J., Searching the Cambridge Structural Database for the 'best' representative of each unique polymorph. *Acta Crystallographica Section B - Structural Science* **2006,** *62*, 567-579.
39. de Gelder, R.; Wehrens, R.; Hageman, J. A., A generalized expression for the similarity of spectra: Application to powder diffraction pattern classification. *Journal of Computational Chemistry* **2001,** *22* (3), 273-289.
40. Nyman, J.; Day, G., Static and lattice vibrational energy differences between polymorphs. *Crystengcomm* **2015,** *17* (28), 5154-5165.
41. Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kessler, M.; Macrae, C. F.; McCabe, P.; Pearson, J.; Taylor, R., New software for searching the Cambridge Structural Database and visualizing crystal structures. *Acta Crystallographica Section B - Structural Science* **2002,** *58*, 389-397.
42. Moggach, S.; Parsons, S.; Wood, P., *High-pressure polymorphism in amino acids*. 2008; Vol. 14, p 143-184.
43. Budd, L.; Ibberson, R.; Marshall, W.; Parsons, S., The effect of temperature and pressure on the crystal structure of piperidine. *Chemistry Central Journal* **2015,** *9*.
44. Boldyreva, E. V.; Shakhtshneider, T. P.; Vasilchenko, M. A.; Ahsbahs, H.; Uchtmann, H., Anisotropic crystal structure distortion of the monoclinic polymorph of acetaminophen at high

hydrostatic pressures. *Acta Crystallographica Section B - Structural Science* **2000,** *56* (2), 299-309.
45. Chisholm, J. A.; Motherwell, S., COMPACK: a program for identifying crystal structure similarity using distances. *Journal of Applied Crystallography* **2005,** *38*, 228-231.
46. Taylor, R.; Cole, J.; Korb, O.; McCabe, P., Knowledge-Based Libraries for Predicting the Geometric Preferences of Druglike Molecules. *Journal of Chemical Information and Modeling* **2014,** *54* (9), 2500-2514.
47. Wood, P. A.; Oliveira, M. A.; Zink, A.; Hickey, M. B., Isostructurality in pharmaceutical salts: How often and how similar? *CrystEngComm* **2012,** *14* (7), 2413-2421.
48. Leitch, A. A.; Reed, R. W.; Robertson, C. M.; Britten, J. F.; Yu, X.; Secco, R. A.; Oakley, R. T., An Alternating π-Stacked Bisdithiazolyl Radical Conductor. *Journal of the American Chemical Society* **2007,** *129* (25), 7903-7914.
49. Braun, D. E.; Bhardwaj, R. M.; Florence, A. J.; Tocher, D. A.; Price, S. L., Complex Polymorphic System of Gallic Acid-Five Monohydrates, Three Anhydrates, and over 20 Solvates. *Crystal Growth & Design* **2013,** *13* (1), 19-23.
50. Iuzzolino, L.; Reilly, A. M.; McCabe, P.; Price, S. L., Use of Crystal Structure Informatics for Defining the Conformational Space Needed for Predicting Crystal Structures of Pharmaceutical Molecules. *Journal of Chemical Theory and Computation* **2017,** *13* (10), 5163-5171.
51. Prystupa, D. A.; Torrie, B. H.; Powell, B. M.; Gerlach, P. N., Crystal structures of methylene bromide and methylene iodide. *Molecular Physics* **1989,** *68* (4), 835-851.
52. Price, S. L.; Reutzel-Edens, S. M., The potential of computed crystal energy landscapes to aid solid form development. *Drug Discovery Today* **2016**.
53. Lekin, K.; Phan, H.; Winter, S. M.; Wong, J. W. L.; Leitch, A. A.; Laniel, D.; Yong, W.; Secco, R. A.; Tse, J. S.; Desgreniers, S.; Dube, P. A.; Shatruk, M.; Oakley, R. T., Heat, Pressure and Light-Induced Interconversion of Bisdithiazolyl Radicals and Dimers. *Journal of the American Chemical Society* **2014,** *136* (22), 8050-8062.
54. 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. *Journal of Chemical Information and Computer Sciences* **2004,** *44* (6), 2133-2144.
55. Takamizawa, S.; Takasaki, Y., Superelastic Shape Recovery of Mechanically Twinned 3,5-Difluorobenzoic Acid Crystals. *Angewandte Chemie-International Edition* **2015,** *54* (16), 4815-4817.
56. Moré, R.; Busse, G.; Hallmann, J.; Paulmann, C.; Scholz, M.; Techert, S., Photodimerization of Crystalline 9-Anthracenecarboxylic Acid: A Nontopotactic Autocatalytic Transformation. *The Journal of Physical Chemistry C* **2010,** *114* (9), 4142-4148.
57. Airoldi, A.; Bettoni, P.; Donnola, M.; Calestani, G.; Rizzoli, C., Crystal structure of zwitterionic 3-(2-hydroxy-2-phosphonato-2-phosphonoethyl)imidazo 1,2-a -pyridin-1-ium monohydrate (minodronic acid monohydrate): a redetermination. *Acta Crystallographica Section E-Crystallographic Communications* **2015,** *71*, 51-+.
58. Tojo, K.; Mizuguchi, J., Refinement of the crystal structure of beta-3,4 : 9,10-perylenetetracarboxylic dianhydride, C24H8O6, at 223 K. *Zeitschrift Fur Kristallographie-New Crystal Structures* **2002,** *217* (2), 255-256.
59. Kersten, K.; Kaur, R.; Matzger, A., Survey and analysis of crystal polymorphism in organic structures. *Iucrj* **2018,** *5*, 124-129.
60. Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C., CrystalOptimizer. An efficient Algorithm for Lattice Energy Minimisation of Organic Crystal using Isolated-Molecule Quantum Mechanical Calculations. In *Molecular System Engineering*, Adjiman, C. S.; Galindo, A., Eds. WILEY-VCH Verlag GmbH & Co.: Weinheim, 2010; Vol. 6, pp 1-42.

8.6 Appendix

Appendix Table 8.1: The 24 pairs of duplicate crystal structures in the 'best R-factor' list. The four highlighted in yellow are false polymorphs (*i.e.* duplicates wrongly identified as polymorphs) using the decision tree in Figure 8.2.

Structure 1	Structure 2	Overlaid molecules/ out of 15	$RMSD_{15}$	Comment
PYRAZI01	PYRAZI	15	0.070	/
NTBZAM10	NTBZAM01	15	0.106	/
NIJHUJ03	NIJHUJ01	15	0.108	/
TIPWIY07	TIPWIY06	15	0.110	/
VEGNAX02	VEGNAX	15	0.115	/
FACRIK06	FACRIK05	15	0.128	/
TMPPIO13	TMPPIO01	15	0.141	/
ACEMID03	ACEMID01	15	0.144	/
MENSEE01	MENSEE	15	0.145	/
GLYCIN82	GLYCIN16	15	0.147	above $RMSD_{15}$ threshold
TIPWIY07	TIPWIY	15	0.155	/
FUYVUQ03	FUYVUQ01	15	0.157	/
DCLBEN06	DCLBEN02	15	0.162	/
GLYCIN81	GLYCIN71	15	0.173	above $RMSD_{15}$ threshold
CHAPEP10	CHAPEP02	15	0.177	/
HACTPH15	HACTPH11	15	0.230	/
HIXHIF03	HIXHIF02	15	0.231	/
TIPWIY06	TIPWIY	15	0.263	/
NMBYAN22	NMBYAN01	15	0.274	/
MUROXA02	MUROXA01	15	0.341	/
DIFVET02	DIFVET	15	0.471	/
HOJQII01	HOJQII	15	0.767	different pressure
DCLBEN07	DCLBEN01	15	0.098	/
FORMAC01	FORMAC	1	0.036	different double bond

Appendix Table 8.2: The three pairs of false duplicates (*i.e.* polymorphs wrongly identified as duplicates) left after applying the heuristically developed criteria in Figure 8.2 to the 'best R-factor' list.

Structure 1	Structure 2	Overlaid molecules/ out of 15	$RMSD_{15}$	Comment
DIMETH06	DIMETH01	15	0.226	Phase transition with temperature
LIHXUW02	LIHXUW01	15	0.095	Phase transition with temperature
MOSTIX01	MOSTIX	15	0.498	Phase transition with temperature

Appendix Table 8.3: The 78 pairs of false polymorphs (*i.e.* duplicates wrongly identified as polymorphs) found after applying the criteria in Figure 8.2 to the polymorph-flagged crystal structures in the CSD. The pairs highlighted in green are examples mentioned in the chapter.

Structure 1	Structure 2	Overlaid molecules/ out of 15	$RMSD_{15}$	Comment
PDABZA	PDABZA01	7	/	different double bond
ACRLAC02	ACRLAC01	1	/	different double bond
VEKSOT	VEKSOT01	1	/	different double bond
ZZZZCB04	ZZZZCB03	5	/	different double bond
AMBNAC12	AMBNAC	3	/	different double bond
DXYLEN15	DXYLEN14	15	2.335	Wrong $RMSD_{15}$
ANTCYB11	ANTCYB13	9	/	same with higher tolerances
HEVXUB	HEVXUB03	2	/	different double bond
MBPHOL15	MBPHOL14	failed	/	failed because overlaid only with lowest R-factor
MBPHOL02	MBPHOL15	failed	/	failed because overlaid only with lowest R-factor
MBPHOL02	MBPHOL14	failed	/	failed because overlaid only with lowest R-factor

ZZZNQS08	ZZZNQS09	7	/	different double bond
BENZID04	BENZID08	15	0.181	different Z because of molecular symmetry
MONTIS02	MONTIS01	4	/	same with higher tolerances
BZDMAZ01	BZDMAZ02	1	/	different double bond
BETANC	BETANC01	2	/	different double bond
MBZPNA12	MBZPNA13	failed	/	failed because overlaid only with lowest R-factor
NMBYAN25	NMBYAN26	failed	/	failed because overlaid only with lowest R-factor
BIPHEN06	BIPHEN08	15	0.091	different Z because of molecular symmetry
PYRDNO10	PYRDNO11	15	0.017	different Z because of molecular symmetry
CASHOT	CASHOT02	7	/	same with higher tolerances
CBENPH	CBENPH01	15	0.519	above $RMSD_{15}$ threshold
CILHIO02	CILHIO10	15	1.503	Wrong $RMSD_{15}$
COMXAD03	COMXAD05	7	/	different double bond
OCRSOL	OCRSOL01	15	0.100	same Z and Z', different space group
QQQFDJ04	QQQFDJ19	3	/	different double bond
QQQFDJ20	QQQFDJ04	3	/	different double bond
HAVBIQ06	HAVBIQ02	1	/	different double bond
DAWFUE02	DAWFUE01	15	0.680	above $RMSD_{15}$ threshold
DAWGAL	DAWGAL02	15	0.756	above $RMSD_{15}$ threshold
DAZABZ10	DAZABZ02	7	/	different double bond
DBEZLM05	DBEZLM01	1	/	different double bond
DHNAPH17	DHNAPH	7	/	different double bond
DIVHOF01	DIVHOF	10	/	different double bond
DIXNIH02	DIXNIH03	15	0.105	different Z because of molecular symmetry
SUCACB10	SUCACB03	3	/	different double bond
PIMELA13	PIMELA02	3	/	different double bond
PIMELA	PIMELA13	1	/	different double bond
PIMELA	PIMELA02	1	/	different double bond
GLYCIN71	GLYCIN81	15	0.173	above $RMSD_{15}$ threshold
GLYCIN82	DOLBIR09	15	0.144	above $RMSD_{15}$ threshold
GLYCIN16	DOLBIR09	15	0.017	same Z and Z', different space group
GLYCIN82	GLYCIN16	15	0.147	above $RMSD_{15}$ threshold
EDTAXX02	EDTAXX01	1	/	different double bond
LGLUAC	LGLUAC11	1	/	different double bond
FORMAC	FORMAC01	1	/	different double bond
FOXNEL01	FOXNEL10	15	2.615	Wrong $RMSD_{15}$
FUCTIG03	FUCTIG02	15	0.068	different Z because of molecular symmetry
INDMET01	INDMET03	2	/	different double bond
PRONAC01	PRONAC	1	/	different double bond
KASXUY02	KASXUY01	6	/	different double bond
KTCYQM	KTCYQM01	15	0.127	different Z because of molecular symmetry
SIKLIH06	SIKLIH07	1	/	different double bond
LUTDUR	LUTDUR02	9	/	same with higher tolerances
MCHTEP17	MCHTEP06	7	/	same with higher tolerances
XENXUL01	XENXUL02	3	/	different double bond
NIJHUJ	NIJHUJ01	15	0.008	different Z because of molecular symmetry
NOETNA02	NOETNA01	3	/	same but error in the overlay
OXACDH03	OXACDH	7	/	different double bond
OXACDH13	OXACDH03	7	/	different double bond
OXACDH13	OXACDH	7	/	different double bond
OXIBZN10	OXIBZN02	15	0.116	above $RMSD_{15}$ threshold
PEYMIQ04	PEYMIQ02	failed	/	failed because overlaid only with lowest R-factor
PHTHCY07	PHTHCY12	failed	/	different double bond
PHENOL01	PHENOL03	15	0.287	above $RMSD_{15}$ threshold
SABCII	SABCII02	15	0.213	different Z because of molecular symmetry
SRFORA06	SRFORA07	9	/	same with higher tolerances
SUWMIG02	SUWMIG01	7	/	same with higher tolerances
SUWMIG03	SUWMIG	4	/	same with higher tolerances, most structurally dissimilar duplicate
TASCUM02	TASCUM03	12	/	same with higher tolerances
TEDAPC01	TEDAPC29	7	/	different double bond
TEDAPC06	TEDAPC22	7	/	different double bond
TETDAM01	TETDAM	15	0.127	above $RMSD_{15}$ threshold

TFMETH	TFMETH02	3	/	same but very poor reproduction
TGLYSU23	TGLYSU22	failed	/	failed because overlaid only with lowest R-factor
TGLYSU20	TGLYSU23	7	/	same with higher tolerances
TGLYSU20	TGLYSU22	7	/	same with higher tolerances
TGLYSU	TGLYSU16	7	/	different double bond
TIYQAU	TIYQAU02	11	/	same with higher tolerances
TIZWAA01	TIZWAA02	15	0.005	different Z because of molecular symmetry
XINRUH01	XINRUH	15	3.306	Wrong $RMSD_{15}$

Appendix Table 8.4: The 212 pairs of false duplicates (*i.e.* polymorphs wrongly identified as duplicates) found after applying the criteria in Figure 8.2 to the polymorph-flagged crystal structures in the CSD. These include several pairs of duplicates of the 47 missed polymorph pairs listed in Appendix Table 8.5. The pairs highlighted in green are examples mentioned in the chapter.

Structure 1	Structure 2	Comment	Structure 1	Structure 2	Comment
QQQAUJ03	QQQAUJ05	/	DOWVOC	DOWVOC32	Phase transition with temperature
QQQAUJ03	QQQAUJ06	/	DOWVOC	DOWVOC35	Phase transition with temperature
AHOXLH02	AHOXLH	Phase transition with temperature	DOWVOC	DOWVOC30	Phase transition with temperature
AKIJIP	AKIJIP01	/	DOWVOC	DOWVOC34	Phase transition with temperature
MEDLUE12	MEDLUE15	Phase transition with temperature	DOWVOC	DOWVOC33	Phase transition with temperature
MEDLUE12	MEDLUE19	Phase transition with temperature	DOWVOC	DOWVOC28	Phase transition with temperature
MEDLUE12	MEDLUE16	Phase transition with temperature	DOWVOC	DOWVOC38	Phase transition with temperature
MEDLUE12	MEDLUE	Phase transition with temperature	DOWVOC	DOWVOC37	Phase transition with temperature
MEDLUE12	MEDLUE18	Phase transition with temperature	DOWVOC	DOWVOC36	Phase transition with temperature
MEDLUE12	MEDLUE17	Phase transition with temperature	DOWVOC	DOWVOC31	Phase transition with temperature
MEDLUE11	MEDLUE15	Phase transition with temperature	GAKNAH	GAKNAH02	Phase transition with temperature
MEDLUE11	MEDLUE19	Phase transition with temperature	GAKNAH01	GAKNAH02	Phase transition with temperature
MEDLUE11	MEDLUE16	Phase transition with temperature	EZIZAP	EZIZAP01	Phase transition with temperature
MEDLUE11	MEDLUE	Phase transition with temperature	ZZZHQU02	ZZZHQU01	Different for 30-molecule shell
MEDLUE11	MEDLUE18	Phase transition with temperature	FIBYIY	FIBYIY01	/
MEDLUE11	MEDLUE17	Phase transition with temperature	GERZAG02	GERZAG03	Phase transition with temperature
MEDLUE10	MEDLUE15	Phase transition with temperature	PIDGOZ03	PIDGOZ01	Phase transition with pressure
MEDLUE10	MEDLUE19	Phase transition with temperature	PIDGOZ05	PIDGOZ03	Phase transition with pressure
MEDLUE10	MEDLUE16	Phase transition with temperature	PIDGOZ05	PIDGOZ01	Phase transition with pressure
MEDLUE10	MEDLUE	Phase transition with temperature	PIDGOZ02	PIDGOZ05	Phase transition with pressure
MEDLUE10	MEDLUE18	Phase transition with temperature	PIDGOZ02	PIDGOZ03	Phase transition with pressure
MEDLUE10	MEDLUE17	Phase transition with temperature	PIDGOZ04	PIDGOZ02	Phase transition with pressure

MEDLUE09	MEDLUE15	Phase transition with temperature	PIDGOZ04	PIDGOZ05	Phase transition with pressure
MEDLUE09	MEDLUE19	Phase transition with temperature	PIDGOZ04	PIDGOZ03	Phase transition with pressure
MEDLUE09	MEDLUE16	Phase transition with temperature	PIDGOZ04	PIDGOZ01	Phase transition with pressure
MEDLUE09	MEDLUE	Phase transition with temperature	SINZIY01	SINZIY	Phase transition with temperature
MEDLUE09	MEDLUE18	Phase transition with temperature	HECXAO01	HECXAO	Phase transition with temperature
MEDLUE09	MEDLUE17	Phase transition with temperature	IJUXEQ01	IJUXEQ	Phase transition with temperature
DXYLEN22	DXYLEN21	Phase transition with temperature	TAYTUI	TAYTUI01	/
DXYLEN22	DXYLEN14	Phase transition with temperature	TAYTUI	TAYTUI02	/
DXYLEN19	DXYLEN22	Phase transition with temperature	KIBBUS01	KIBBUS	/
DXYLEN20	DXYLEN19	Phase transition with temperature	LIHXUW01	LIHXUW02	Phase transition with temperature
DXYLEN20	DXYLEN21	Phase transition with temperature	MAMPUM01	MAMPUM02	Phase transition with temperature
DXYLEN20	DXYLEN14	Phase transition with temperature	MOBBAH02	MOBBAH	/
DXYLEN24	DXYLEN15	Phase transition with temperature	MOSTIX	MOSTIX01	Phase transition with temperature
DXYLEN23	DXYLEN15	Phase transition with temperature	OPUPOG	OPUNIY	/
DXYLEN13	DXYLEN15	Phase transition with temperature	UFOCAU01	UFOCAU	/
DXYLEN17	DXYLEN13	Phase transition with temperature	QUBPIN01	QUBPIN	Phase transition with temperature
DXYLEN17	DXYLEN23	Phase transition with temperature	QUBPIN03	QUBPIN01	Phase transition with temperature
DXYLEN17	DXYLEN24	Phase transition with temperature	QUBPIN03	QUBPIN	Phase transition with temperature
DXYLEN18	DXYLEN17	Phase transition with temperature	RBHTCA01	RBHTCA02	/
DXYLEN18	DXYLEN15	Phase transition with temperature	RBHTCA	RBHTCA02	/
DXYLEN16	DXYLEN17	Phase transition with temperature	ROKQUF01	ROKQUF02	Phase transition with temperature
DXYLEN16	DXYLEN15	Phase transition with temperature	ROKQUF	ROKQUF01	Phase transition with temperature
PRMDIN01	PRMDIN02	Phase transition with temperature	RUCJUV04	RUCJUV01	/
PRMDIN07	PRMDIN01	Phase transition with temperature	SARCAC	SARCAC01	Phase transition with temperature
PRMDIN04	PRMDIN01	Phase transition with temperature	TETBBZ	TETBBZ04	Phase transition with temperature
PRMDIN06	PRMDIN01	Phase transition with temperature	TETBBZ	TETBBZ03	Phase transition with temperature
PRMDIN05	PRMDIN01	Phase transition with temperature	TETBBZ	TETBBZ02	Phase transition with temperature
PRMDIN03	PRMDIN01	Phase transition with temperature	TETBBZ01	TETBBZ	Phase transition with temperature
PRMDIN	PRMDIN01	Phase transition with temperature	UBUQIR04	UBUQIR03	Phase transition with temperature
BARBAD07	BARBAD10	Phase transition with temperature	UBUQIR05	UBUQIR03	Phase transition with temperature
BARBAD07	BARBAD11	Phase transition with temperature	UBUQIR06	UBUQIR03	Phase transition with temperature
BARBAD07	BARBAD13	Phase transition with temperature	UBUQIR08	UBUQIR03	Phase transition with temperature

BARBAD07	BARBAD14	Phase transition with temperature	UBUQIR07	UBUQIR03	Phase transition with temperature
BARBAD07	BARBAD15	Phase transition with temperature	UBUQIR13	UBUQIR07	Phase transition with temperature
BARBAD12	BARBAD07	Phase transition with temperature	UBUQIR13	UBUQIR08	Phase transition with temperature
BARBAD09	BARBAD12	Phase transition with temperature	UBUQIR13	UBUQIR06	Phase transition with temperature
BARBAD09	BARBAD10	Phase transition with temperature	UBUQIR13	UBUQIR05	Phase transition with temperature
BARBAD09	BARBAD11	Phase transition with temperature	UBUQIR13	UBUQIR04	Phase transition with temperature
BARBAD09	BARBAD13	Phase transition with temperature	UBUQIR14	UBUQIR13	Phase transition with temperature
BARBAD09	BARBAD14	Phase transition with temperature	UBUQIR14	UBUQIR03	Phase transition with temperature
BARBAD09	BARBAD15	Phase transition with temperature	UBUQIR01	UBUQIR13	Phase transition with temperature
BARBAD08	BARBAD12	Phase transition with temperature	UBUQIR01	UBUQIR03	Phase transition with temperature
BARBAD08	BARBAD10	Phase transition with temperature	UBUQIR09	UBUQIR01	Phase transition with temperature
BARBAD08	BARBAD11	Phase transition with temperature	UBUQIR09	UBUQIR14	Phase transition with temperature
BARBAD08	BARBAD13	Phase transition with temperature	UBUQIR09	UBUQIR07	Phase transition with temperature
BARBAD08	BARBAD14	Phase transition with temperature	UBUQIR09	UBUQIR08	Phase transition with temperature
BARBAD08	BARBAD15	Phase transition with temperature	UBUQIR09	UBUQIR06	Phase transition with temperature
BARBAD06	BARBAD12	Phase transition with temperature	UBUQIR09	UBUQIR05	Phase transition with temperature
BARBAD06	BARBAD10	Phase transition with temperature	UBUQIR09	UBUQIR04	Phase transition with temperature
BARBAD06	BARBAD11	Phase transition with temperature	UBUQIR10	UBUQIR01	Phase transition with temperature
BARBAD06	BARBAD13	Phase transition with temperature	UBUQIR10	UBUQIR14	Phase transition with temperature
BARBAD06	BARBAD14	Phase transition with temperature	UBUQIR10	UBUQIR07	Phase transition with temperature
BARBAD06	BARBAD15	Phase transition with temperature	UBUQIR10	UBUQIR08	Phase transition with temperature
BARBAD01	BARBAD06	Phase transition with temperature	UBUQIR10	UBUQIR06	Phase transition with temperature
BARBAD01	BARBAD08	Phase transition with temperature	UBUQIR10	UBUQIR05	Phase transition with temperature
BARBAD01	BARBAD09	Phase transition with temperature	UBUQIR10	UBUQIR04	Phase transition with temperature
BARBAD01	BARBAD07	Phase transition with temperature	UBUQIR12	UBUQIR01	Phase transition with temperature
BARBAD	BARBAD06	Phase transition with temperature	UBUQIR12	UBUQIR14	Phase transition with temperature
BARBAD	BARBAD08	Phase transition with temperature	UBUQIR12	UBUQIR07	Phase transition with temperature
BARBAD	BARBAD09	Phase transition with temperature	UBUQIR12	UBUQIR08	Phase transition with temperature
BARBAD	BARBAD07	Phase transition with temperature	UBUQIR12	UBUQIR06	Phase transition with temperature
BIZWAI01	BIZWAI02	Most structurally similar polymorph	UBUQIR12	UBUQIR05	Phase transition with temperature
SUCROS16	SUCROS17	Phase transition with temperature	UBUQIR12	UBUQIR04	Phase transition with temperature
HOLVAG	HOLVAG01	Phase transition with temperature	UBUQIR11	UBUQIR01	Phase transition with temperature

CAZLAR02	CAZLAR01	Phase transition with temperature	UBUQIR11	UBUQIR14	Phase transition with temperature
CAZLAR	CAZLAR02	Phase transition with temperature	UBUQIR11	UBUQIR07	Phase transition with temperature
CELWIB04	CELWIB03	Phase transition with temperature	UBUQIR11	UBUQIR08	Phase transition with temperature
CHNAPQ	CHNAPQ01	/	UBUQIR11	UBUQIR06	Phase transition with temperature
DCLNAP	DCLNAP02	/	UBUQIR11	UBUQIR05	Phase transition with temperature
DECYIU01	DECYIU	/	UBUQIR11	UBUQIR04	Phase transition with temperature
DIMETH01	DIMETH06	Phase transition with temperature	UBUQIR	UBUQIR11	Phase transition with temperature
DIMETH10	DIMETH06	Phase transition with temperature	UBUQIR	UBUQIR12	Phase transition with temperature
DOWVOC40	DOWVOC38	Phase transition with temperature	UBUQIR	UBUQIR10	Phase transition with temperature
DOWVOC40	DOWVOC37	Phase transition with temperature	UBUQIR	UBUQIR09	Phase transition with temperature
DOWVOC40	DOWVOC36	Phase transition with temperature	UBUQIR	UBUQIR13	Phase transition with temperature
DOWVOC40	DOWVOC31	Phase transition with temperature	UBUQIR	UBUQIR03	Phase transition with temperature
DOWVOC28	DOWVOC40	Phase transition with temperature	UBUQIR02	UBUQIR	Phase transition with temperature
DOWVOC33	DOWVOC40	Phase transition with temperature	UBUQIR02	UBUQIR01	Phase transition with temperature
DOWVOC34	DOWVOC40	Phase transition with temperature	UBUQIR02	UBUQIR14	Phase transition with temperature
DOWVOC30	DOWVOC40	Phase transition with temperature	UBUQIR02	UBUQIR07	Phase transition with temperature
DOWVOC35	DOWVOC40	Phase transition with temperature	UBUQIR02	UBUQIR08	Phase transition with temperature
DOWVOC32	DOWVOC40	Phase transition with temperature	UBUQIR02	UBUQIR06	Phase transition with temperature
DOWVOC39	DOWVOC40	Phase transition with temperature	UBUQIR02	UBUQIR05	Phase transition with temperature
DOWVOC	DOWVOC39	Phase transition with temperature	UBUQIR02	UBUQIR04	Phase transition with temperature

Appendix Table 8.5: The 47 pairs that are the best representative of each polymorph pair among the false duplicates in Appendix Table 4. The pairs highlighted in green are examples mentioned in the chapter.

Structure 1	Structure 2	Overlaid molecules/ out of 15	$RMSD_{15}$	Comment
AHOXLH02	AHOXLH	15	0.082	Phase transition with temperature
AKIJIP	AKIJIP01	15	0.046	/
MEDLUE12	MEDLUE17	15	0.045	Phase transition with temperature
DXYLEN22	DXYLEN14	15	0.082	Phase transition with temperature
PRMDIN01	PRMDIN02	15	0.097	Phase transition with temperature
BARBAD07	BARBAD15	15	0.072	Phase transition with temperature
BIZWAI01	BIZWAI02	15	0.007	Most structurally similar polymorph
SUCROS16	SUCROS17	15	0.043	Phase transition with temperature
HOLVAG	HOLVAG01	15	0.114	Phase transition with temperature
CAZLAR02	CAZLAR01	15	0.062	Phase transition with temperature
CELWIB04	CELWIB03	15	0.045	Phase transition with temperature
CHNAPQ	CHNAPQ01	15	0.054	/
DCLNAP	DCLNAP02	15	0.115	/
DECYIU01	DECYIU	15	0.320	/
DIMETH01	DIMETH06	15	0.226	Phase transition with temperature
DOWVOC40	DOWVOC31	15	0.075	Phase transition with temperature

GAKNAH	GAKNAH02	15	0.138	Phase transition with temperature
EZIZAP	EZIZAP01	15	0.274	Phase transition with temperature
ZZZHQU02	ZZZHQU01	15	0.123	Different but not shown with 15 molecules
FIBYIY	FIBYIY01	15	0.013	/
GERZAG02	GERZAG03	15	0.070	Phase transition with temperature
PIDGOZ03	PIDGOZ01	15	0.133	Phase transition with pressure
PIDGOZ05	PIDGOZ03	15	0.013	Phase transition with pressure
PIDGOZ05	PIDGOZ01	15	0.129	Phase transition with pressure
PIDGOZ04	PIDGOZ05	15	0.015	Phase transition with pressure
PIDGOZ04	PIDGOZ03	15	0.017	Phase transition with pressure
PIDGOZ04	PIDGOZ01	15	0.130	Phase transition with pressure
SINZIY01	SINZIY	15	0.058	Phase transition with temperature
HECXAO01	HECXAO	15	0.120	Phase transition with temperature
IJUXEQ01	IJUXEQ	15	0.130	Phase transition with temperature
TAYTUI	TAYTUI02	15	0.295	/
KIBBUS01	KIBBUS	15	0.059	/
LIHXUW01	LIHXUW02	15	0.095	Phase transition with temperature
MAMPUM0	MAMPUM0	15	0.058	Phase transition with temperature
MOBBAH02	MOBBAH	15	0.180	/
MOSTIX	MOSTIX01	15	0.498	Phase transition with temperature
OPUNIY	OPUPOG	15	0.087	/
UFOCAU01	UFOCAU	15	0.039	/
QUBPIN01	QUBPIN	15	0.068	Phase transition with temperature
QUBPIN03	QUBPIN01	15	0.066	Phase transition with temperature
QUBPIN03	QUBPIN	15	0.021	Phase transition with temperature
RBHTCA02	RBHTCA01	15	0.016	/
ROKQUF01	ROKQUF02	15	0.238	Phase transition with temperature
RUCJUV04	RUCJUV01	15	0.039	/
SARCAC	SARCAC01	15	0.078	Phase transition with temperature
TETBBZ	TETBBZ02	15	0.483	Phase transition with temperature
UBUQIR04	UBUQIR03	15	0.165	Phase transition with temperature

Chapter 9: Overall conclusion and future work

9.1 Can CSP be routinely used to complement polymorphs screens?

Chapters 3 and 4 have described the state-of-the-art of CSP studies on large and flexible molecules. In Chapter 3 the only solved crystal structure of molecule XXVI was successfully predicted in the context of the 6th Blind Test of organic CSP.[1] In Chapter 4 a CSP study successfully found the only two solved forms of the antihelminthic drug mebendazole,[2, 3] in the right stability order, but it produced no match to the powder patterns of other six single-component crystal structures that were produced in a parallel academic experimental solid form screening effort. However, these new forms have not been solved yet, and it is impossible to assess whether they were not found because they were outside the scope of the computational study or because of flaws in the CSP procedure. These two studies confirm that CSP can be successful under blind conditions and a useful complement to an experimental effort. The routine industrial use of CSP would be highly desirable as it can establish the completeness of experimental polymorph screens, avoiding the cost of performing unnecessary experiments but also warning about potential risks.[4]

However, the results shown in these two chapters also outline some weaknesses that make the applicability of CSP to the very large and flexible molecules the pharmaceutical industry mostly works with problematic. First of all these two studies reveal that CPU cost scales badly with size and level of flexibility:[5, 6] generating a crystal energy landscape of molecule XXVI was more than twenty times more expensive than it was for mebendazole. As pharmaceutical molecules are often larger and more flexible than XXVI, this hinders a routine use of CSP in pharmaceutical polymorph screening. Current methodologies cannot be scaled to a molecule like ritonavir.[6] Moreover the high computational expense required to perform CSP limits the coverage of the search even for relatively easy targets like mebendazole, as shown in Chapter 4. For example, CSP studies can only affordably predict crystal structures with Z'=1 and, in a few cases, with Z'=2, though Z'=2 searches generally include severe space group limitations and rigid cut-offs in terms of number of structures generated to limit the cost.[1] They are also generally limited to single-component crystal structures, while thorough polymorph screens cover other solid forms like hydrates or co-crystals.[5, 6] Finally, even with these limitations, CSP studies often produce very congested crystal energy landscapes,[7] with no clear indication of which crystal structure/s are actually obtainable and of the experiments needed to produce them. The interpretation of CSP results is complicated by several low-energy crystal structures being very similar, without a clear way to

determine whether they are potential polymorphs, duplicates that are artefacts of the energy model or different components of a disordered phase.

9.2 How can CSP be extended to large and flexible molecules of pharmaceutical interest?

From what has been said so far, it is clear that methodological improvements are needed to make the application of CSP to larger and more flexible molecules, as well as a more complete coverage of the search space, affordable. In Chapter 5 a workflow that uses Cambridge Structural Database (CSD)[8] conformational information to define the molecular flexibility range and to perform crystal structure searches was developed from and tested on five large and flexible molecules.[9] This method successfully reproduced the great majority of the most significant crystal structures from the original CSP studies, including those matching all the experimentally known forms, at a reduced computational cost. A newly discovered polymorph of succinic acid with an unusual conformation[10] was also successfully generated, indicating that the workflow can also be applied to molecules not included in its development. These results illustrate that CSD-derived knowledge-based conformational information can be used in CSP. The workflow parameters in Chapter 5 will probably need to be adapted to apply CSD-retrieved data to other molecules and/or CSP algorithms, and more efficient ways to use this information may also be discovered.

Chapter 6 and 7 deal with the complex issue of refining crystal structures generated in a CSP search. Chapter 6 shows how limiting the conformational degrees of freedom (CDFs) treated as explicit variables in crystal structure optimisations is not a completely realistic approach for pharmaceutical-like molecules. However, as thousands of generated crystal structures often need to be taken to the final refinement stage of CSP because of the approximate search models, for large and flexible molecules expensive full optimisations with the Ψ_{mol} method treating all torsion and bond-angles as explicit variables (bond-lengths can be ignored)[11] would be unaffordable. The alternative solution of performing all the optimisations with Ψ_{crys} methods like DFT-D,[12] which optimise all intra- and intermolecular degrees of freedom without requiring an explicit selection, is once again too expensive for minimising hundreds or thousands of crystal structures. Hence, in Chapter 7 the affordable Ψ_{crys} semi-empirical dispersion corrected density functional tight-binding (DFTB3-D3) method[13] was used to perform full intermediate optimisations of all intra- and intermolecular degrees of freedom of the structures generated in Chapter 5, followed by an optimisation of only the intermolecular interactions with a molecular wave-function calculated at the PBE0 6-31G(d,p) level of theory. This produced crystal energy landscapes of similar quality to the original CSP studies, with all experimentally known forms found within a sensible energy window for

polymorphism, but at a reduced cost.[14] These results show that a cheap intermediate optimisation followed by a more accurate final evaluation of E_{latt} is an effective approach to perform crystal structure refinement. The good scaling of cost with molecular size and level of flexibility suggests that these methods could be scalable to molecules of pharmaceutical interest. Moreover, DFTB3-D3 allows to affordably calculate all molecular and lattice phonon modes, which cause a larger spread of free energy values than the rigid-body lattice modes alone.

Finally, in Chapter 8 a CSD survey was performed to identify some structural and crystallographic similarity thresholds that can separate redeterminations from polymorphs. These thresholds are successful in ~98% of cases. Many of the failures are due to structural ambiguities generally associated with phase transitions with temperature and/or pressure. A clustering algorithm that can apply these criteria to CSP-generated crystal structures was written. This is important to expand CSP studies to larger and more flexible molecules: understanding which structures could be actual polymorphs as opposed to artefacts of the models is important to limit the crystal energy landscape to plausible polymorphs or components of disordered phases.

9.3 Possible future developments

The work performed in this thesis has outlined some of the strengths and weaknesses of current CSP methodologies. Ways to solve some of the limitations of CSP, in particular the high computational cost and its poor scaling with molecular size and flexibility, have been proposed. However, more work is needed in several directions.

One of the main issues of CSP is the trade-off between computational cost and accuracy. This is particularly evident in CSP searches, where speed is fundamental to cover the search space within a feasible time-scale.[15] However, the use of approximate models (*e.g.* the point charges in CrystalPredictor)[16] for limiting cost has important draw-backs: the energy ranking output by the search is often inaccurate, as shown in Chapters 3-5, and thousands of expensive refinements of the generated crystal structures must often be carried out to select all potentially important crystal structures. This can severely limit the ability to use the most accurate models in the final refinement stage and to fully cover all packing possibilities. Furthermore, the use of simplified search models can result in important crystal structures being missed altogether if they are not minima in the approximate potential energy surface.[15] Hence search methods should be improved, and the potential energy surface explored in the searches should resemble more that of the final refinement methods. This would increase the confidence that the searches have not missed important crystal structures, and it would limit the number of candidates to be taken to the final refinement stage since there would be smaller discrepancies between the relative energy rankings. This general idea is already

implemented in GRACE,[17] where a tailor-made force field is fitted that mimics DFT-D results. However, fitting to *ab initio* data is itself a very expensive process, in particular for large molecules,[1] and more efficient methods need to be found. In the recent Faraday Discussion on CSP several groups working on inorganic molecules have applied machine learning algorithms to crystal structure searches with very promising results.[18, 19] The organic community should follow on these ideas, finding good training sets and comparing results with those obtained with the most accurate models. Chapter 5 has shown that CSD information on geometric preferences can be effectively utilised to define the conformational search space that needs to be covered by CSP searches. Thus knowledge-based data may be used as a training set to produce machine learning hypotheses that accurately describe molecular flexibility and that could be applied in the crystal structure generation stage of CSP. Different CSP approaches are ultimately distinguished by the search methods: although some in-house refinement algorithms have been developed (*e.g.* CrystalOptimizer,[11] used in this thesis), several accurate quantum-chemistry codes exist[20-23] that can perform the same task to various degrees of efficiency. In my opinion, the biggest breakthrough in CSP will come from the development of better search methods that affordably cover a sufficient portion of a fairly accurate potential energy surface.

Until improved search algorithms are available, it will be necessary to use some approximations in the refinement stage, in particular for large and flexible molecules where fully optimising hundreds or thousands of generated crystal structures with the most accurate models is not feasible.[24] Chapter 7 shows that a possible solution is to perform an intermediate optimisation with DFTB3-D3. However, the energy ranking output by DFTB3-D3 is not reliable. Developments in DFTB algorithms that could correct for some of the weaknesses outlined in Chapter 2.3.2.2.1, the use of alternative semi-empirical approaches[13] or the development of more accurate force fields[25] could all help to speed up refinement without compromising accuracy. Machine learning could once again be very useful for this purpose, and it has shown some potential in computing accurate energies within a crystal structure refinement procedure.[26]

Another aspect that requires improvements is the calculation of free energies. Chapter 7 has shown that calculating both lattice and molecular phonon modes with a Ψ_{crys} method that allows their coupling can cause more re-ranking of E_{latt} values than the rigid-body lattice modes alone. Accurate free energy calculations should thus be performed with high-quality Ψ_{crys} methods, which are currently too expensive to be applied to a large portion of the crystal energy landscape. Furthermore, using the harmonic approximation does not account for anharmonic effects that may occur in the lattice,[27] even when the quasi-harmonic correction is applied.[28] Molecular dynamics simulations could be a step towards solving this problem,[10] as they explore the potential

energy surface more explicitly, but this would require the development of very accurate potentials that can describe all intra- and intermolecular interactions. Machine learning could once again be employed for developing these potentials.

Finally, the issue of analysing the crystal energy landscape and extracting useful information remains fundamental. In Chapter 8 CSD mining has provided information that can be used for separating polymorphs from redeterminations. More developments however will be required to understand which E_{latt} minima would actually survive at finite temperatures, either as unique free energy minima or as components of disordered phases, to predict the properties of computer-generated crystal structures and to direct experimentalists towards the crystallisation of new forms.[4] This will require a more thorough understanding of temperature effects and of crystallisation kinetics. Molecular dynamics could be used to improve our understanding of both phenomena,[4] but affordable and accurate potentials are needed.

In summary, CSP is far from a solved problem. Although very accurate and widely available algorithms to refine crystal structures exist, they cannot be routinely used in CSP because of their high computational expense and their poor scaling with molecular size and flexibility. Developing more accurate yet affordable search methodologies could be a fundamental breakthrough for CSP, and it could allow a more complete coverage of the search space. Intermediate refinement methods that are more efficient at bridging the gap between the searches and final optimisations could be identified, and more accurate ways to analyse the crystal energy landscape, to calculate free energies and to predict crystallisation kinetics are needed. Any improvement of models and procedures will require a solid collaboration with experimentalists to obtain accurate benchmarks, such as energy differences between polymorphs or phonon modes, which can be used to drive developments and validate results.

This thesis has brought some advancements in these directions, and has provided some hints about what may be done in the near future. The scientific and financial incentives are strong enough to proceed in these directions.

9.4 References

1. 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.; 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 Crystallographica Section B* **2016,** *72* (4), 439-459.
2. Ferreira, F. F.; Antoni, S. G.; Rosa, P. C. P.; Paiva-Santos, C. D., Crystal Structure Determination of Mebendazole Form A Using High-Resolution Synchrotron X-Ray Powder Diffraction Data. *Journal of Pharmaceutical Sciences* **2010,** *99* (4), 1734-1744.
3. Martins, F. T.; Neves, P. P.; Ellena, J.; Cami, G. E.; Brusau, E. V.; Narda, G. E., Intermolecular Contacts Influencing the Conformational and Geometric Features of the Pharmaceutically Preferred Mebendazole Polymorph C. *Journal of Pharmaceutical Sciences* **2009,** *98* (7), 2336-2344.
4. Nyman, J.; Reutzel-Edens, S. M., Crystal structure prediction is changing from basic science to applied technology. *Faraday Discussions* **2018,** *Advance Article*.
5. Price, S. L.; Reutzel-Edens, S. M., The potential of computed crystal energy landscapes to aid solid form development. *Drug Discovery Today* **2016,** *21* (6), 912-923.
6. Price, S. L.; Braun, D. E.; Reutzel-Edens, S. M., Can computed crystal energy landscapes help understand pharmaceutical solids? *Chemical Communications* **2016,** *52*, 7065-7077.
7. Price, S. L., Why don't we find more polymorphs? *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **2013,** *69*, 313-328.
8. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C., The Cambridge Structural Database. *Acta Crystallographica Section B-Structural Science Crystal Engineering and Materials* **2016,** *72*, 171-179.
9. Iuzzolino, L.; Reilly, A. M.; McCabe, P.; Price, S. L., Use of Crystal Structure Informatics for Defining the Conformational Space Needed for Predicting Crystal Structures of Pharmaceutical Molecules. *Journal of Chemical Theory and Computation* **2017,** *13* (10), 5163-5171.
10. Lucaioli, P.; Nauha, E.; Gimondi, I.; Price, L. S.; Guo, R.; Iuzzolino, L.; Singh, I.; Salvalaglio, M.; Price, S. L.; Blagden, N., Serendipitous isolation of a disappearing conformational polymorph of succinic acid challenges computational polymorph prediction. *CrystEngComm* **2018,** *20* (28), 3971-3977.
11. Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C., Efficient Handling of Molecular Flexibility in Lattice Energy Minimization of Organic Crystals. *Journal of Chemical Theory and Computation* **2011,** *7* (6), 1998-2016.
12. Grimme, S.; Hansen, A.; Brandenburg, J. G.; Bannwarth, C., Dispersion-Corrected Mean-Field Electronic Structure Methods. *Chemical Reviews* **2016,** *116* (9), 5105-5154.
13. Christensen, A. S.; Kubař, T.; Cui, Q.; Elstner, M., Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications. *Chemical Reviews* **2016,** *116* (9), 5301-5337.
14. Iuzzolino, L.; McCabe, P.; Price, S. L.; Brandenburg, J. G., Crystal structure prediction of flexible pharmaceutical-like molecules: Density functional tight-binding as an intermediate optimization method and for free energy estimation. *Faraday Discussions* **2018**.
15. Pantelides, C. C.; Adjiman, C. S.; Kazantsev, A. V., General Computational Algorithms for Ab Initio Crystal Structure Prediction for Organic Molecules. *Topics in Current Chemistry* **2014,** *345*, 25-58.
16. Karamertzanis, P. G.; Pantelides, C. C., Ab initio crystal structure prediction. II. Flexible molecules. *Molecular Physics* **2007,** *105* (2-3), 273-291.
17. Neumann, M. A. *GRACE (the Generation, Ranking and Characterisation Engine)*, 1.0; Avant-garde Materials Simulation Deutschland GmbH: 2007.
18. Tong, Q.; Xue, L.; Lv, J.; Wang, Y.; Ma, Y., Accelerating CALYPSO structure prediction by data-driven learning of a potential energy surface. *Faraday Discussions* **2018**.
19. Deringer, V. L.; Proserpio, D. M.; Csányi, G.; Pickard, C. J., Data-driven learning and prediction of inorganic crystal structures. *Faraday Discussions* **2018**.
20. Kresse, G.; Furthmuller, J., Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B* **1996,** *54* (16), 11169-11186.
21. Dovesi, R.; Orlando, R.; Erba, A.; Zicovich-Wilson, C. M.; Civalleri, B.; Casassa, S.; Maschio, L.; Ferrabone, M.; De La Pierre, M.; D'Arco, P.; Noël, Y.; Causà, M.; Rérat, M.; Kirtman, B., CRYSTAL14: A program for the ab initio investigation of crystalline solids. *International Journal of Quantum Chemistry* **2014,** *114* (19), 1287-1317.
22. Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L.; Cococcioni, M.; Dabo, I.; Dal Corso, A.; de Gironcoli, S.; Fabris, S.; Fratesi, G.; Gebauer, R.; Gerstmann, U.; Gougoussis, C.; Kokalj, A.; Lazzeri, M.; Martin-Samos, L.; Marzari, N.; Mauri, F.; Mazzarello, R.; Paolini, S.; Pasquarello, A.; Paulatto, L.; Sbraccia, C.; Scandolo, S.; Sclauzero, G.; Seitsonen, A. P.; Smogunov, A.; Umari, P.; Wentzcovitch, R. M.,

QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *Journal of Physics-Condensed Matter* **2009,** *21* (39), 395502.
23. Clark, S. J.; Segall, M. D.; Pickard, C. J.; Hasnip, P. J.; Probert, M. J.; Refson, K.; Payne, M. C., First principles methods using CASTEP. *Zeitschrift fur Kristallographie* **2005,** *220* (5-6), 567-570.
24. Brandenburg, J. G.; Grimme, S., Accurate Modeling of Organic Molecular Crystals by Dispersion-Corrected Density Functional Tight Binding (DFTB). *Journal of Physical Chemistry Letters* **2014,** *5* (11), 1785-1789.
25. Uzoh, O. G.; Galek, P. T. A.; Price, S. L., Analysis of the conformational profiles of fenamates shows route towards novel, higher accuracy, force-fields for pharmaceuticals. *Physical Chemistry Chemical Physics* **2015,** *17* (12), 7936-7948.
26. Musil, F.; De, S.; Yang, J.; Campbell, J. E.; Day, G. M.; Ceriotti, M., Machine learning for the structure-energy-property landscapes of molecular crystals. *Chemical Science* **2018,** *9* (5), 1289-1300.
27. Dove, M. T., *Introduction to Lattice Dynamics*. Cambridge University Press: Cambridge, 1993.
28. Nyman, J.; Day, G. M., Modelling temperature-dependent properties of polymorphic organic molecular crystals. *Physical Chemistry Chemical Physics* **2016,** *18* (45), 31132-31143