Filtering & analysing the results of crystal structure prediction

Aim

To filter the results of crystal structure prediction calculations and assess whether any results match experimentally determined solid forms.

Introduction

The occurrence of polymorphism in marketed pharmaceutical compounds is highly important as different polymorphs can exhibit varying bulk physicochemical properties. There are a number of well-documented examples, such as ritonavir[1] and rotigotine,[2] where a new solid form was discovered after the compound went to the market which caused significant processing problems for the associated companies. These cases illustrate the potentially huge financial impact of failing to fully characterise the solid state landscape for a pharmaceutical compound before it is launched.

The main aim of polymorph screening is therefore to avoid the situation described above where a new form is discovered late on in the drug pipeline. There is, however, only a finite amount of resources that can be applied to experimental polymorph screening with regards to the time and drug material that can be used. For this reason crystal structure prediction (CSP) could be a valuable, complementary technique to evaluate the solid form landscape computationally.

CSP is a complex process that generates a huge amount of data, generally many thousands or even millions of predicted structures. Even if only the structures within 5 $kJmol^{-1}$ of the predicted lowest energy structure are analysed, there still can be a large number of structures to look at. This means that it is highly valuable to have automatic structure analysis tools to cluster predicted structures and to match up any known experimental structures to the predictions. This example will focus on the prediction of single component and co-crystal structures of 4-aminobenzoic acid (I), 2,2'-bipyridine (II) and 4-nitrophenylacetic acid (III, Figure 1) as performed by Prof. Sally Price's group.[3]

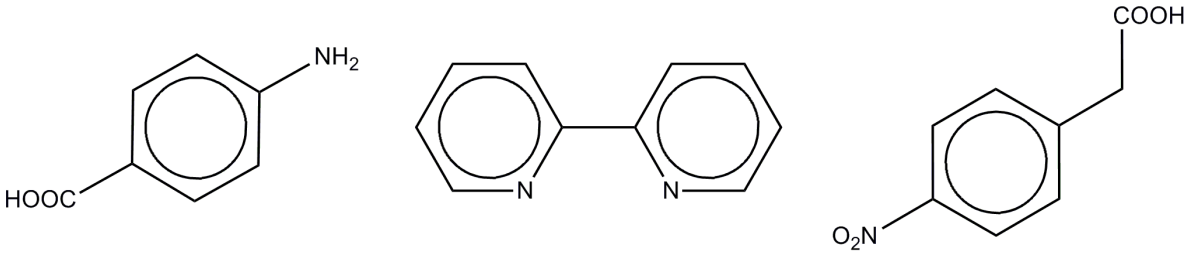

Figure 1 - Chemical structures of compounds I (left), II (middle) & III (right)

Method

The prediction of single and multiple component structures containing compounds I, II and III presents a number of problems to CSP methodology as there is not only conformational flexibility to consider, but also the added number of degrees of freedom associated with multiple components. To tackle these problems a multi-stage methodology was developed (Figure 2) including multiple clustering steps using Crystal Packing Similarity analysis[4] as available in the *Materials* module of Mercury.[5]

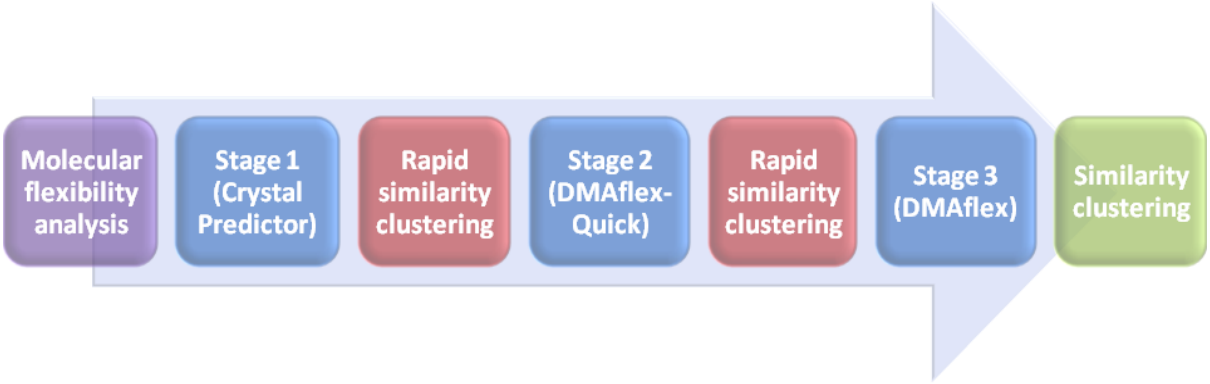

Figure 2 - Overview summary of the multi-stage lattice minimisation method.

The first step of the methodology was to determine the extent of molecular flexibility for each compound. This was done using gas-phase quantum mechanical scans to find the range of geometries that might be energetically accessible (within 20-30 kJ mol^{-1} of the most stable conformations). It was confirmed that all the geometries observed within the Cambridge Structural Database (CSD[6]) lay well within this region. Next, hypothetical crystal structures were generated using CrystalPredictor[7,8] for the 13 most densely populated space groups in the CSD using flexible torsion ranges as determined in the previous step.

The lattice energy minimisation was then performed in stages, with a more realistic and more computationally expensive model for the lattice energy being used in each stage. In the first stage, minimisation by CrystalPredictor, only an atomic charge model for the electrostatic forces was used in quick energy minimisations to optimise the lattice dimensions and flexible torsion angles, taking only a few CPU seconds per structure. These were rapidly clustered (in batch mode) to remove duplicates. This clustering, plus application of an energy cutoff, reduced the 50,000-565,000 crystal structures considered in the search to the 700-3000 most plausible structures. Next, an improved description of electrostatic interactions was introduced (DMAflex-Quick) by using distributed atomic multipoles. The minima determined from DMAflex-Quick were clustered again in batch mode by performing crystal packing similarity calculations using 15-molecule clusters (20-molecule clusters for multi-component structures) and considering clusters to be equivalent based on a root mean square deviation (RMSD) threshold of 0.4 Å.

The lattice energy of each of these unique structures was then evaluated more accurately, by using a single point quantum mechanical evaluation of the intramolecular energy and distributed multipoles

and re-minimising the crystal structure with these. Finally, the 25 most stable, unique resulting structures were *fully* minimised using “on-the-fly" quantum mechanical calculations to achieve conformation-dependent multipoles and intermolecular energies for the minimisations (DMAflex[9]). After this third stage, Crystal Packing Similarity calculations were performed against all predicted structures, to eliminate any duplicates and see the similarities between the predicted thermodynamically feasible structures and experimental structures.

Motif searches, using the *Materials* module of Mercury, can also be performed for the sets of predicted structures in order to categorise the specific hydrogen-bonding patterns observed in each of the structures - this allows better evaluation of the solid state landscape as a whole.

Results

All the searches for single and multiple component structures were found to successfully locate the experimentally determined crystal structure and in 4 out of 5 cases this was found to be at, or very close to, the global lattice energy minimum. Figure 3 shows a structural overlay, produced using Mercury, of the coordination shells of the experimentally observed and computationally predicted structures for the co-crystal of molecules I & II. You can see that the match between the structures is extremely good - in this case the RMSD for the full coordination shell of 20 molecules is just 0.305 Å.

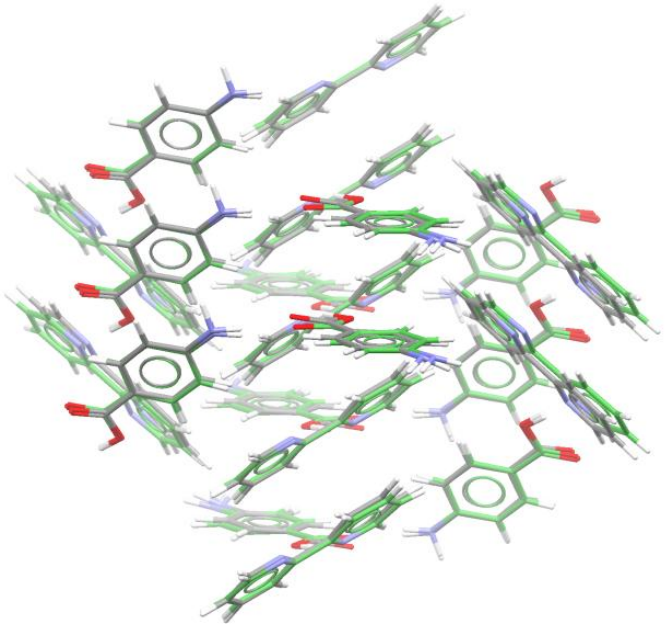

Figure 3 - Structural overlay of observed and predicted structures in the co-crystal of I & II

Figure 4 shows the CSP results for 4-aminobenzoic acid (compound I). In this case there are two known polymorphs of the compound and the packing similarity analysis has shown that the lowest energy calculated structure corresponds to the β polymorph. The metastable α polymorph is Z’=2 and hence could not be found in the search, but quite similar structures were found close in energy. The hydrogen-bonding patterns for all the predicted structures of I were analysed using the Motif tool in the *Materials* module of Mercury. The majority (22 out of 25) of the low energy predicted structures for compound I were found to exhibit carboxylic acid dimers (as found in polymorph α).

The carboxylic acid to amino hydrogen bond observed in polymorph β, the global minimum, was only observed in 2 other predicted structures (3 out of 25 in total) and is a relatively uncommon interaction with relation to the CSD. Motif searches for the carboxylic acid dimer and an interaction from the amino group to the carbonyl of the carboxylic acid group (also present in polymorph β) indicate that whilst the carboxylic acid dimer is more common, the amino to carboxylic acid interaction has the higher frequency of occurrence (ratio of number of structures in which motif is observed to number of structures in which motif is possible).

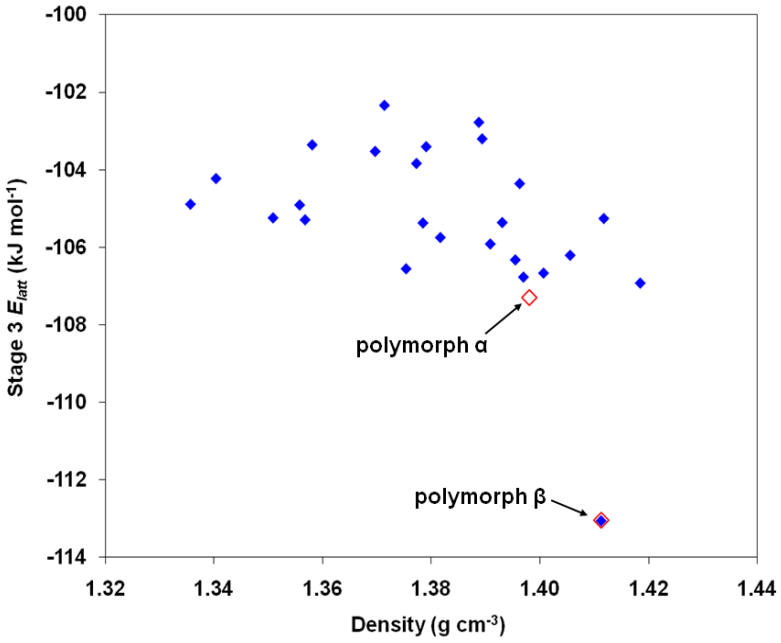

Figure 4 - CSP results for 4-aminobenzoic acid (I) – only the 25 fully minimised structures are shown. The two experimentally determined structures are labelled as polymorphs α & β.

In the predicted 4-aminobenzoic acid structures, the amino group is observed to interact as a donor *via* three different hydrogen-bonding motifs. These motifs are as follows; (i) interactions with the carboxylic acid C=O group (10/25, including polymorphs α and β), (ii) interactions with the carboxylic acid O-H group (15/25), and (iii) hydrogen-bonds to an amino group acceptor (6/25). Motif searches of these possible hydrogen bond motifs indicate that whilst amino to C=O (of a COOH group) interactions have a relatively high frequency of occurrence, amino interactions to the hydroxyl of COOH or an NH_2 group are uncommon, with frequencies of occurrence found to be around 5%. Therefore, if further polymorphs of 4-aminobenzoic acid were found they too could be notable for containing unusual hydrogen bonding patterns.

In the case of the cocrystal of 4-aminobenzoic acid (4-ABA) with 4-nitrophenylacetic acid (4-NPA), the most stable predicted structures were based on 4-ABA carboxylic acid homodimers and 4-NPA carboxylic acid (OH) to 4-ABA amino (NH_2) contacts, whereas the observed form contains only carboxylic acid hetero-dimers between the two co-formers. The donor (OH) to acceptor (NH_2) interactions found in these hypothetical low energy structures, are seen to have a very low frequency of occurrence in the CSD: below 5%. Hence, a 4-ABA homodimer-based co-crystal may yet be found, as predicted, but if so it may be characterised by uncommon hydrogen bonds.

Each of the experimental crystal structures was also compared with the closest predicted structure to evaluate how well the CSP methodology describes the conformational geometries. The heavy atom torsion angles in the predicted structures were not seen to deviate by more than 10° from the experimental values. The biggest conformational discrepancy when visualised using a single molecule overlay in Mercury (Figure 5) was in the rotation and pyramidalisation of the amine group in compound I. This discrepancy may be due to experimental uncertainty in locating the proton positions or the approximation of modelling them as static in a position determined by the balance of separate approximate models for the inter and intramolecular forces.

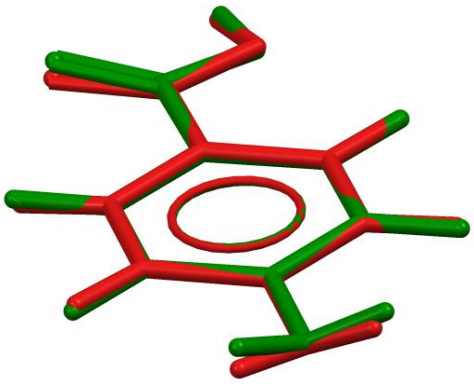

Figure 5 - Capped sticks overlay of the 3D molecular structure of I in the observed (red) & predicted (green) structures

In general this multistage lattice minimisation strategy, including multiple steps of clustering through packing similarity calculations, is seen to efficiently survey the solid state landscape without missing any experimentally observed types of crystal packing. The methodology could also be extended to include more accurate and computationally expensive lattice or free energy minimisation techniques in the future.

Conclusions

This study has illustrated the importance of crystal packing similarity and hydrogen-bond motif tools in the filtering and analysis of predicted crystal structures during CSP. Due to the sheer volume of data produced, it is crucial to be able to filter and categorise the predicted structures as well as pick out matches to experimentally observed structures in an automated fashion. Analysing predicted structures with respect to hydrogen bond motifs found in the CSD may aid the evaluation of whether the computed structures are likely to be found experimentally.

Acknowledgements

Some further structural analyses were performed for this case study beyond the scope of the original published paper. We would therefore like to thank Dr. Louise Price for providing the computed crystal structures from the database developed by the Control and Prediction of the Organic Solid State project (http://www.cposs.org.uk) funded by the Basic Technology program of Research Councils UK.

References

1. J. Bauer, S. Spanton, R. Henry, J. Quick, W. Dziki, W. Porter and J. Morris, *Pharm. Res.*, **2001**, 18, 859-866.
2. H.-M. Wolff, L. Quere and J. Riedner, *WO Pat.*, 068520, **2009**.
3. P. G. Karamertzanis, A. V. Kazantsev, N. Issa, G. W. A. Welch, C. S. Adjiman, C. C. Pantelides and S. L. Price, *J. Chem. Theory Comput.*, **2009**, 5, 1432-1448.
4. J. A. Chisholm and W. D. S. Motherwell, *J. Appl. Crystallogr.*, **2005**, 38, 228-231.
5. C. F. Macrae, I. J. Bruno, J. A. Chisholm, P. R. Edgington, P. McCabe, E. Pidcock, L. Rodriguez-Monge, R. Taylor, J. van de Streek and P. A. Wood, *J. Appl. Crystallogr.*, **2008**, 41, 466-470.
6. F. H. Allen, *Acta Crystallogr., Sect. B*, **2002**, 58, 380-388.
7. P. G. Karamertzanis and C. C. Pantelides, *J. Comput. Chem.*, **2005**, 26, 304-324.
8. P. G. Karamertzanis and C. C. Pantelides, *Mol. Phys.*, **2007**, 105, 273-291.
9. P. G. Karamertzanis and S. L. Price, *J. Chem. Theory Comput.*, **2006**, 2, 1184-1199.

Products

CSD – the world's only comprehensive, fully curated database of crystal structures, containing over 500,000 entries

Mercury – a versatile and feature-rich visualisation tool for molecular structures

Materials module of Mercury – a powerful exploration and comparison tool for solid state structures

For further information please contact Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK. Tel: +44 1223 336408, Fax: +44 1223 336033, Email: admin@ccdc.cam.ac.uk