Prediction of the crystal structures of axitinib, a polymorphic pharmaceutical molecule

Manolis Vasileiadis, Constantinos C. Pantelides, Claire S. Adjiman

Centre for Process Systems Engineering, Department of Chemical Engineering, Imperial College London, London SW7 2AZ, United Kingdom

HIGHLIGHTS

• First ab initio prediction of the polymorphs of axitinib, a challenging pharmaceutical molecule.
• Successful application of a systematic procedure to identify polymorphs with one molecule in asymmetric unit.
• Analysis of strengths and weaknesses of state of the art for crystal structure prediction.
• Identification of future research needs and priorities in this area.

GRAPHICAL ABSTRACT

ABSTRACT

Organic molecules can crystallize in multiple structures or polymorphs, yielding crystals with very different physical and mechanical properties. The prediction of the polymorphs that may appear in nature is a challenge with great potential benefits for the development of new products and processes. A multistage crystal structure prediction (CSP) methodology is applied to axitinib, a pharmaceutical molecule with significant polymorphism arising from molecular flexibility. The CSP study is focused on those polymorphs with one molecule in the asymmetric unit. The approach successfully identifies all four known polymorphs within this class, as well as a large number of other low-energy structures. The important role of conformational flexibility is highlighted. The performance of the approach is discussed in terms of both the quality of the results and various algorithmic and computational aspects, and some key priorities for further work in this area are identified.

1. Introduction

Organic crystals are a key component of the formulated products that are manufactured in many industrial sectors including pharmaceuticals (Storey and Ymén, 2011), agrochemicals, foods, paints, and explosives. The efficacy, stability and other end-use properties of such products are largely influenced by the precise structure of the organic crystals because the molecular packing arrangement affects numerous physical properties such as color, mechanical strength, flowability and solubility, to name but a few (Hilfinger, 2006).

In view of the importance of crystal structure, the propensity of many organic molecules to crystallize readily in multiple metastable structures (“polymorphs”) (Brog et al., 2013; Cruz-Cabeza and Bernstein, 2014) creates significant challenges in many aspects of product development and manufacturing. A well-known example of the problems that can arise as a result of polymorphism is that of ritonavir (Bauer et al., 2001), an active pharmaceutical ingredient (API) marketed as a HIV drug by Abbott Laboratories from 1996. In 1998, a previously unknown form, Form II, appeared, and it became impossible to revert to the production of Form I. Form II was found to be more stable than Form I, with a
significantly lower solubility. The product had to be recalled from the market, leading to an interruption of supply, and was eventually re-developed as a liquid formulation. Further investigation revealed three further forms of ritonavir (Morissette et al., 2003). The existence of polymorphs also poses intellectual property challenges, as patent protection relates to the form of the product. For example, the crystal structures of cefdinir, a drug molecule with at least five polymorphs, have been the subject of multiple patents and of prolonged legal battles (Cabrè et al., 2007).

The magnitude of the risks arising from insufficient knowledge of polymorphism has motivated increasing investment in poly- morph screening, the experimental investigation of the so-called polymorphic landscape of organic molecules (Aaltosen et al., 2009; Newman, 2013). This has been complemented by computational crystal structure prediction (CSP) methodologies aiming to identify possible crystal structures with little or no experimental input. While it has long been clear that achieving this goal would require a very significant research effort (Gavezott, 1994), CSP is increasingly used in combination with experimental screening (Price, in press). The blind tests organized by the Cambridge Crystallographic Data Centre since 1999 (Lommerse et al., 2000) provide a useful series of snapshots of the state-of-the-art and of the progress made in the field. In each blind test, participants are asked to predict the most stable crystal structure for a handful of molecules, salts or co-crystals of varying complexity. The degree of difficulty of each system depends on the number of molecules it contains, the types and number of atoms, the presence of charged species and the flexibility of the molecules. Of particular note in previous blind tests are two milestones: the consistent success achieved with the GRACE approach by Neumann, Leusen, Kendrick, in the fourth blind test (Day et al., 2009) and by Neumann, Leusen, Kendrick, and van de Streek in the fifth blind test (Bardwell et al., 2011), in predicting the polymorphs of small molecules (Kendrick et al., 2011); and the successful prediction in the fifth blind test, by two groups (Kazantsev et al., 2011b), of the most stable structure of “Molecule XX” (benzyl-(4-(4-methyl-5-(p-tolylsulfonyl)-1,3-thiazol-2-yl)phenyl)carbamate), a molecule whose structure, size and flexibility (see Fig. 1g) are representative of those of pharmaceutical compounds.

As a result of the increasing reliability of CSP (Day, 2011; Price, in press), several promising applications to industrially-relevant compounds have been reported in the literature, focusing on the identification of known and potential polymorphs. In the area of pharmaceuticals, these have included studies of some of the compounds shown in Fig. 1, namely (a) naproxen (Braun et al., 2011), (b) GlaxoSmithKline’s molecule GS269984B (Ismail et al., 2013), (c) Pfizer’s crizotinib (Abramov, 2013), (d) a melanotin agonist (Kendrick et al., 2013), (e) Eli Lilly’s olanzapine (Bhardwaj et al., 2013) and (f) Eli Lilly’s tazofoprenol (Price et al., in press). In these different cases, crystal energy landscapes, in which every putative crystal structure is characterized in terms of its energy and density, were generated. The computed crystal structures were ranked in terms of their thermodynamic stability, usually based on the predicted lattice energy, rather than the more difficult to compute Gibbs free energy.

The studies of pharmaceutical compounds reported in the literature to date have focused on “small molecule pharmaceuticals”, typically with up to 10 rotatable bonds. They have generally resulted in the correct identification of all known polymorphic structures as low-energy minima on the energy landscape. The relative stability of the computed polymorphs, however, often differs from the experimental relative stability, as extrapolated to 0 K. Furthermore, many structures that have not been identified experimentally are often found as low-energy minima. This can arise for a number of reasons, including the fact that some computed structures may be found to be unstable when entropic effects are taken into account (Mooij et al., 1998; Zykova-Timan et al., 2008) and the fact that some structures may be difficult to crystallize experimentally (Price, 2013).

Despite these limitations, CSP has found several applications of practical relevance beyond the scientific goal of achieving the blind prediction of all likely polymorphs. Thus, it can be used (i) to provide reassurance that all likely polymorphs have been identified; (ii) to guide the search for further polymorphs by suggesting the specific crystal structures that might be observed, thereby helping to identify appropriate crystallization conditions (Arlin et al., 2011); (iii) to support crystal engineering by providing an understanding of the link between the motifs observed and molecular structure (Uzoh et al., 2012) or crystal composition (Habgood, 2013; Karamertzian et al., 2009); (iv) to help crystallographers interpret data gathered on specific compounds (Baias et al., 2013; Fričič et al., 2010; Wu et al., 2013). This broad array of uses provides impetus for methodological improvements aimed at increasing the accuracy of the predictions and at broadening the range of molecules, co-crystals, salts and solvates that can be tackled in terms of size and complexity.

Several reviews have recently been published on the current state-of-the-art in CSP, covering one or more methodologies (Abramov, 2013; Atahan-Evenk and Aspuru-Guzik, 2014; Day, 2010, 2011, 2012; Kendrick et al., 2011; Pantelides et al., 2014; Price, 2013, 2008a, 2008b, in press). Together with the papers summarizing the results of the five blind tests to date (Bardwell et al., 2011; Day et al., 2009, 2005; Lommerse et al., 2000; Motherwell et al., 2002), these provide an excellent survey of the field. In the present paper, we focus on a specific systematic approach that has been developed in our group (Pantelides et al., 2014). The algorithms on which this approach is based have been successfully used in several of the examples discussed so far (Bardwell et al., 2011; Bhardwaj et al., 2013; Ismail et al., 2013; Kazantsev et al., 2011b; Price et al., in press; Vasileiadis et al., 2012). We aim to provide an introduction to the approach and a perspective on future developments via its application to axitinib (Fig. 2), a Pfizer anti-cancer API that has been noted for its numerous crystal forms, including 5 neat ones and 66 solvates. This provides a great challenge for CSP and a fertile learning ground allowing us to assess the current status of the methodology and to identify directions for further research. In Section 2, we review previous work on the crystal structures of axitinib. In Section 3, we provide an overview of our CSP methodology, and in Section 4 we discuss its application to the prediction of polymorphism in axitinib. Section 5 discusses various key aspects of the performance of the CSP approach when applied to axitinib, aiming to draw some lessons from the results obtained. Section 6 concludes with some general remarks on the current status of CSP methodologies and their limitations, and identifies some relevant research priorities in this area.

2. Earlier work on the crystal structures of axitinib

This section provides a review of the available information on the polymorphism of axitinib. Published experimental data on the five neat polymorphs are summarized, and previous computational work is discussed.

2.1. Experimental investigations

From the point of view of crystallography, axitinib is notable because of its large number of neat polymorphs, solvates and hydrates. The 71 forms that have been reported in the literature to
Fig. 1. Some of the organic compounds that have been the subject of CSP studies: (a) naproxen, \((+)-(S)-2-(6\text{-methoxynaphthalen-2-yl})\)propanoic acid; (b) GSK269984B, 6-\([\text{5-chloro-2-}[(4\text{-chloro-2-fluorophenyl})\text{methyl}](\text{oxy})\text{ phenyl}]\) methyl\[2\text{-pyridinecarboxylic acid}; \) (c) crizotinib, \(3-\{[(1R)-1-(2,6\text{-dichloro-3-fluorophenyl})\text{ethoxy}]\)5-\((1\text{-piperidin-4-yl})\)pyrazol-4-yl\)pyridin-2-amine; (d) a melatonin agonist, \(N\{[(2R)-(6\text{-chloro-5-methoxy-1H-indol-3-yl})\text{propyl}]\text{acetamide}; \) (e) olanzapine, 2-Methyl-4-(4-\text{methyl-1-piperazinyl})\text{10H-thieno}[2,3-\text{b}]\)1,3\text{-benzodiazepine; \}) (f) tazofelone, 2,6-di-\text{tert-butyl-4-}((1,3\text{-thiazolidin-5-yl})\text{methyl})\text{phenol; \}) \text{(g) molecule XX of the fifth blind test, benzyl-(4-(4-methyl-5-(p-tolylsulfonyl))1,3\text{-thiazol-2-yl})phenyl carbamate.}

Fig. 2. Molecular diagram of axitinib \((\text{n-methyl-2-}[[3-\{[(E)-2\text{-pyridin-2-yliethenyl}]1H-indazol-6-yl\} sulfanyl]benzamide). Seven major torsions are highlighted.\)
date are labeled using Roman numerals, i.e. Form I to LXXII, with no form being assigned to number V (Campeta et al., 2010). As a consequence of this rich polymorphic landscape, this molecule has proved a challenge for the pharmaceutical industry, as has been well documented in the context of crystallization process development (Chekal et al., 2009).

There are five known non-solvated forms of axitinib (Campeta et al., 2010; Chekal et al., 2009). Four polymorphs have one molecule in the asymmetric unit (Form I, Form VI, Form XXV, and Form XLI), while Form IV has two independent molecules. The main crystallographic information for these five polymorphs of axitinib, based on spectroscopic data, is summarized in Table 1. All structures have been resolved to a high degree of confidence, as indicated by the low values of the relative R-factor.

Form IV was thought to be the most suitable for development (Chekal et al., 2009) until the discovery, following further experimental screenings, of Form XLI, which is currently considered to be the most stable form of axitinib (Campeta et al., 2010). Campeta et al. further reported that Form XLI is monotypically related to all other forms, while forms IV, VI and XXV are enantiotropically related. Furthermore, Chekal et al. (2009) found that Form IV is more stable than Form XXV at temperatures above 75 °C, based on solubility experiments in a 80:20 water/methanol solution. The transition temperature between Forms XXV and VI has not been determined: the two structures are so close in energy that calorimetric experiments result in conflicting evidence on their relative stability (Campeta et al., 2010). Finally, Form I is the least stable among the five polymorphs, and is known to be unstable in a humid environment and to transform to the monohydrate Form IX.

Overall, differential scanning calorimetry (DSC), solubility measurements and Burger’s rule (Burger and Ramberger, 1979) suggest the order of stability (Campeta et al., 2010):

Form XLI > Form XXV ≈ Form VI > Form IV > Form I,
while measured enthalpies of fusion (ΔHf) indicate the order:
Form XLI > Form VI ≥ Form XXV > Form IV > Form I.

It is interesting to note that these orders do not correlate with the densities and melting points. This is a consequence of the conformational nature of polymorphism in axitinib, arising from the large number of flexible torsion and bond angles (Campeta et al., 2010). Strong intramolecular hydrogen bonds between the benzamide group and the pyrazole and/or the pyridine groups are observed in all polymorphs. In Forms XXV, VI, IV, and I, hydrogen bonding results in the formation of dimers, while in the case of Form XLI, an extended network of hydrogen bonds is created.

The greater stability of Form XLI has been attributed to this extended hydrogen bond network (Campeta et al., 2010).

### 2.2. Computational work

Most of the computational work published to date has focused on the determination of the relative stability of axitinib’s polymorphs and the elucidation of possible relationships between the stability and the structure of the various crystals. Such an assessment was carried out by Campeta et al. (2010) based on hydrogen bond propensity analysis, using atomic cHB surface charges (Klamt, 2005), and the calculation of the relative conformational energies of the various polymorphs. It was found that the strongest intramolecular hydrogen bond is that between the oxygen and the pyrazole amine, and is present in all polymorphs. Relative conformational energies, calculated using the COSMOTHERM (COSMOlogic) software with a PBE functional (Perdew et al., 1996), indicated that the lowest-energy conformation is that of Form VI, while the highest-energy conformer is that of Form XLI. The high stability of Form XLI was then attributed to the presence of the hydrogen bond network, and the fact that it exhibits the shortest hydrogen bond distance between the oxygen and the pyrazole amine; overall, it was asserted that its higher conformational energy is counterbalanced by stronger intermolecular interactions (Campeta et al., 2010).

Abramov (2011) estimated the relative stability of Forms XLI, XXV, VI and IV by applying six computational models, including molecular mechanics (McQuaid et al., 2004), density functional theory (DFT), two versions of dispersion-corrected DFT (DFT+d) (Grimme, 2006), and two versions of the quantum theory of atoms in molecules (QTAIM). QTAIM was used to model molecular clusters derived from experimental crystallography data, with DFT-optimized hydrogen positions. Based on QTAIM, the charge density and the electronic potential energy density, at a point defined as the bond critical point (BCP) along the bond path of one or more hydrogen bonds, were calculated based on a B3LYP/6-31G (d,p)-derived wave function. The relative order of stability computed by DFT+d was found to be in agreement with the experimental order based on ΔHf, while the densities calculated by QTAIM were found to correlate well with the experimental relative stability (Abramov, 2011). The other methodologies tested did not provide good agreement. Interestingly, of the 10 pharmaceutical compounds studied with these different approaches, axitinib was found to present the greatest challenge. At least four of the six computational approaches were found to yield good agreement with experimental order for each of the other molecules.

Finally a limited CSP study was carried out by Lupyan et al. (2012). The authors initially developed an updated parameter set for the intramolecular S…O interaction for the OPLS_2005 force field (Shivakumar et al., 2010). They then used this force field to perform a conformational search using the low-mode search (LMOD) method (Kolossváry and Guida, 1996), in order to find low energy conformations of axitinib. Finally, using the conformations identified by this search that are closest to the experimental conformations, they performed a CSP study for each conformation, restricting the search to the corresponding experimental space groups. The Polymorph Predictor CSP module in Materials Studio 5.5 (Accelrys Software Inc.) with the COMPASS force field was used for this purpose. No prediction was attempted for Form IV because it has two molecules in the asymmetric unit. Forms I, VI and XLI were correctly identified as lattice energy minima with RMSD root mean squared deviation of a cluster of 10 molecules) values equal to 0.66 Å, 0.54 Å and 0.47 Å respectively. Form XXV was not found to correspond to a lattice energy minimum.

### Table 1

Crystallographic information for the five polymorphs of axitinib, listed in decreasing order of stability as determined by heats of fusion. “RT” refers to room temperature. Z denotes the number of molecules in the asymmetric unit, a, b, c the lattice lengths, α, β, γ the lattice angles, ρ the density and R the R-factor.

<table>
<thead>
<tr>
<th>Polymorph</th>
<th>Temperature (K)</th>
<th>Space group</th>
<th>Z</th>
<th>a (Å)</th>
<th>b (Å)</th>
<th>c (Å)</th>
<th>α (deg)</th>
<th>β (deg)</th>
<th>γ (deg)</th>
<th>ρ (g cm⁻³)</th>
<th>R (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>XLI</td>
<td>RT</td>
<td>P2₁/c</td>
<td>1</td>
<td>16.08</td>
<td>8.10</td>
<td>15.58</td>
<td>90.00</td>
<td>112.2</td>
<td>90.00</td>
<td>1.369</td>
<td>5.42</td>
</tr>
<tr>
<td>VI</td>
<td>RT</td>
<td>P1</td>
<td>1</td>
<td>8.15</td>
<td>10.73</td>
<td>12.68</td>
<td>68.1</td>
<td>88.3</td>
<td>70.6</td>
<td>1.330</td>
<td>4.68</td>
</tr>
<tr>
<td>XXV</td>
<td>RT</td>
<td>P2₁/c</td>
<td>1</td>
<td>4.54</td>
<td>11.75</td>
<td>34.83</td>
<td>90.0</td>
<td>92.13</td>
<td>90.0</td>
<td>1.382</td>
<td>6.25</td>
</tr>
<tr>
<td>IV</td>
<td>RT</td>
<td>P1</td>
<td>2</td>
<td>11.86</td>
<td>12.40</td>
<td>15.00</td>
<td>81.7</td>
<td>81.1</td>
<td>65.9</td>
<td>1.293</td>
<td>6.57</td>
</tr>
<tr>
<td>I</td>
<td>213</td>
<td>P1</td>
<td>1</td>
<td>7.74</td>
<td>11.88</td>
<td>12.15</td>
<td>65.67</td>
<td>72.64</td>
<td>76.19</td>
<td>1.333</td>
<td>8.43</td>
</tr>
</tbody>
</table>

M. Vasileiadis et al. / Chemical Engineering Science 121 (2015) 60–76
The computational studies of polymorphism in axitinib carried out to date have allowed the investigation of the effects of flexibility and hydrogen bonding on the stability of the known polymorphs. There have been limited attempts to explore the crystal energy landscape for this challenging molecule. In the remainder of this paper, we investigate the applicability of the approach developed in our group to the \( Z = 1 \) polymorphs of axitinib.

3. Crystal structure prediction: problem formulation and solution

This section provides a brief overview of the CSP problem and its mathematical formulation, and of the CSP methodology that will be used for the ab initio prediction of the polymorphic landscape of axitinib.

3.1. The CSP problem

The crystal structure prediction challenge can be stated as (Pantelides et al., 2014):

“Given the molecular diagrams for all chemical species (neutral molecule(s) or ions) in the crystal, identify the thermodynamically most stable crystal structure at a given temperature and pressure, and, in correct order of decreasing stability, other (metastable) crystal structures that are also likely to occur in nature.”

To be relevant to the pharmaceutical and related industries, a CSP approach must be applicable to organic molecules involving multiple rotatable bonds and having a molecular weight of at least a few hundred daltons. In addition, it is necessary to be able to predict the crystal structures of salts, co-crystals and solvates as such systems are frequently used to enhance product effectiveness or to facilitate manufacturing. In a recent publication (Pantelides et al., 2014), we set out the design requirements that systematic CSP methodologies should meet in order to find wide applicability in practice, such as (i) a high degree of automation, with limited dependence on user insight; (ii) a consistent and general physical basis; (iii) a high degree of reliability; (iv) a high degree of accuracy, but with reasonable computational cost. This last requirement is particularly challenging: experience in crystal structure prediction has shown that it is essential to carry out an exhaustive search of the energy landscape covering millions of potential structures, and that the relative energies of computed crystal structures are highly dependent on the accuracy of the energy model, with electronic structure calculations providing the most reliable results (Bardwell et al., 2011).

In the absence of any defects, a crystal is an inﬁnite periodic structure wholly deﬁned by its space group, the size and shape of the unit cell, the numbers of molecules of each species in the crystal that are within the asymmetric unit and the positions of their atoms. For a given number of molecules in the asymmetric unit (\( Z \)), a crystal structure is fully deﬁned by the following variables:

- the unit cell lattice lengths and angles, denoted by \( X \);
- the positions of the centers of mass and the orientation of the chemical entities within the asymmetric unit, denoted by \( \beta \);
- the conformational degrees of freedom (CDFs) of every chemical entity in the asymmetric unit, \( \theta \), including all bond lengths, bond angles and torsion angles.

From a thermodynamic standpoint, the stable structure at given temperature and pressure is the one with the lowest Gibbs free energy. In addition to this global energy minimum, there are usually many other local minima, corresponding to metastable structures. There is some evidence that metastable structures observed in nature may be over 10 kJ mol\(^{-1}\) above the global minimum, with 25 kJ mol\(^{-1}\) being a possible upper bound (Bernstein, 2002). Hence, we define the CSP problem as the identiﬁcation of all low-energy minima, i.e. global and near-global solutions of the following minimization problem, solved with respect to the variables deﬁning the crystal structure:

\[
\min G = U + pV - TS, \quad x_i \in \{0, 1\}
\]

where \( G \) denotes the Gibbs free energy of the crystal, \( U \) its internal energy, \( p \) the pressure, \( V \) the volume, \( T \) the temperature and \( S \) the entropy.

In practice, the evaluation of the Gibbs free energy of a known crystal structure is a very challenging problem (Frenkel and Smit, 2002; Kofke and Cummings, 1997, 1998; Noya et al., 2008), and it cannot reasonably be performed within an extensive search for putative crystal structures. It is thus common practice to neglect the entropic contribution and to focus on minimizing the crystal enthalpy at 0 K. This is usually justified on the grounds that entropy makes a relatively small contribution to the overall energy at room temperature, the combined contribution of entropy and zero-point energy being estimated to be of the order of 2–5 kJ mol\(^{-1}\) (Day, 2011; Gavezzotti and Filippini, 1995). As mentioned in Section 1, the assumption that entropy can be ignored may result in some inaccuracies in the relative ordering of the structures identiﬁed, as well as in an overestimation of the number of structures. Furthermore, by neglecting the entropic contribution, it is not possible to investigate enantiotropically-related polymorphs (Mukhopadhyay et al., 1985; Murghich and Pissaretzky, 1975). Thus, there are emerging attempts (Day et al., 2003; Karamertzanis et al., 2008; van Eijck, 2001; Vasileiadis, 2013) to take some account of entropic contributions. A further approximation is usually made to neglect the \( pV \) term, as this is very small (of the order of \( J \) mol\(^{-1}\)) at low pressures, only becoming practically important at pressures of the order of \( G \)Pa (Admiraal et al., 1982; Allan et al., 2002; Paliwoda et al., 2012).

Given these approximations, the CSP problem considered by most researchers in the field can be framed as a minimization of the internal energy, usually expressed as a lattice energy, \( E_{\text{latt}} \), namely the difference between the crystal and gas-phase energies. For a crystal involving \( N \) distinct chemical species, this is given by

\[
\min E_{\text{latt}} = U - \sum_{i=1}^{N} x_i U_i^{\text{gas}},
\]

where \( U \) is the specific internal energy of the crystal, \( U_i^{\text{gas}} \) that of component \( i \) at its global minimum energy in the gas phase, and \( x_i \) the mole fraction of component \( i \) in the crystal. The energies are evaluated at 0 K and 0 Pa.

As can be expected based on the close proximity of atoms in the crystalline environment, the specific model chosen for the computation of the lattice energy strongly inﬂuences the outcome of the calculations. Among the approaches commonly used are molecular mechanics force ﬁelds (Kim et al., 2011; Lupypan et al., 2012; Pillardy et al., 2000), plane-wave dispersion-corrected electronic structure calculations (DFT+d) (Kendrick et al., 2013; Neumann and Perrin, 2005), and hybrid models that combine electronic structure calculations (typically on an isolated molecule, although this can be done for several molecules (Mooij et al., 1999)) and empirical terms (Cox et al., 1981; Coobes et al., 1996; Williams, 1999, 2001). Within this latter approach, electrostatic interactions can be modeled in different ways, including the use of point charges (Karamertzanis and Pantelides, 2004, 2007) or distributed multipole (Coobes et al., 1996; Mooij and Leusen, 2001; Price et al., 2010; Stone and Alderton, 1985), and may even
incorporate an anisotropic model of repulsion (Misquitta et al., 2008).

The approach we adopt here is based on hybrid models, in which the lattice energy is partitioned into intramolecular and intermolecular terms as follows:

\[ E^{\text{int}} = \Delta U^{\text{intra}} + U^r + U^d, \]

where \( \Delta U^{\text{intra}} \) is the intramolecular energy (relative to the gas-phase energy), which depends solely on the conformation of the molecules within the unit cell; \( U^r \) and \( U^d \) denote intermolecular electrostatic and repulsion–dispersion interactions respectively, which depend both on the conformation of the molecules and on their positions within the unit cell. The approach adopted to compute each term varies at each stage of our multistage CSP methodology, as described in Section 3.2. In all cases, appropriate summation techniques (e.g., Ewald summation (Ewald, 1921)) are used to capture the infinite periodic nature of the crystal.

3.2. Overview of the CSP methodology

Polymorphism is prevalent in many molecules of practical relevance (e.g., Fig. 1) and often arises from the presence of rotatable bonds (Cruz-Cabeza and Bernstein, 2014) with a deformation energy of the same order of magnitude as the energy change on packing (Kazantsev et al., 2011a). It is thus essential to account accurately for molecular flexibility during the course of energy minimization; a discussion of the pitfalls of neglecting or overly restricting flexibility can be found in Pantelides et al. (2014). The need to explore conformational variation greatly increases the complexity of the optimization problem and has been a key driver for recent theoretical and algorithmic developments of our CSP framework.

Multistage CSP methodologies are based on the idea that one can extract key features of the crystal energy landscape with relatively simple models (the ‘global search stage’), and then improve the accuracy of the results based on much more detailed and computationally expensive models (the ‘refinement stage’). Using a simple model in the global search stage is needed to ensure the computational tractability of the extensive search that is necessary to identify all polymorphs of potential practical interest. On the other hand, if this model is not sufficiently accurate, the global search may also miss important polymorphs because they do not happen to correspond to local minima of the (approximate) lattice energy surface; or, even if it does succeed in identifying them, it may rank them so highly in energy that they would not be considered by the subsequent refinement stage unless it is applied to a very large number of structures. Therefore, a fine balance needs to be struck between the accuracy and the cost of the model used for global search, especially in the consideration of molecular flexibility and its effect on the intramolecular energy contributions.

As illustrated in Fig. 3, the CSP methodology involves three main stages; these are reviewed in more detail below.

3.2.1. Stage 0: conformational analysis and choice of computational model

As already explained, a key success factor for CSP is the correct handling of configurational flexibility at the global search and refinement stages. Therefore, as a first step (Stage 0) of the CSP methodology, we attempt to establish the degree of conformational flexibility of each CDF that needs to be considered in the CSP context. In particular, in vacuo molecular conformations correspond to global or local minima in the conformational energy landscape. Intermolecular interactions in the crystalline environment may cause some CDFs to deviate appreciably from their in vacuo values leading to a conformational energy increase by up to about 20–30 kJ mol\(^{-1}\) in most cases. In some instances where intramolecular hydrogen bonds are broken, intramolecular energy increases greater than 50 kJ mol\(^{-1}\) have sometimes been observed (Cruz-Cabeza and Bernstein, 2014).

Potentially flexible CDFs normally include a subset of the torsion angles and some of the bond angles. They can often be identified using basic chemical understanding, complemented where possible by empirical evidence, such as the geometry of these angles in similar molecules appearing in crystal structures stored in the Cambridge Structural Database (CSD) (Allen, 2002). Their flexibility can be confirmed by performing a 1-dimensional conformational scan for each CDF under consideration; during a scan, the corresponding CDF is fixed at a sequence of values, and at each such point an isolated-molecule quantum mechanical (QM) calculation is performed to minimize conformational energy with respect to all other CDFs. The range of values of the CDF for which the configurational energy increase is within 20 kJ mol\(^{-1}\) are considered to be of interest for the purposes of CSP in this work. In some cases, there may be several non-overlapping ranges satisfying this condition for a given CDF.

The computational model to be used for the CSP is also chosen at Stage 0. This involves the determination of (a) the level of theory and basis set for the isolated-molecule QM calculations used to compute \( \Delta U^{\text{intra}} \) and to characterize the intermolecular electrostatic interactions, and (b) the empirical model used for the repulsion–dispersion component of the intermolecular energy, \( U^d \). For molecules for which one or more crystal structures are already available, a good criterion is the extent to which the selected model can reproduce the experimentally observed structures. This can be assessed by performing a local minimization of lattice energy using CrystalOptimizer (cf. Section 3.2.3), using the experimental structure as an initial guess. A minimized structure that is far from the original experimental structure is a good indication that the selected model is unsuitable and needs to be revised. If no experimental polymorphs are available for the molecule of interest, some validation of the model may be carried out using available crystal structures of similar molecule(s) (e.g., in terms of the functional groups involved).

Another useful, and computationally inexpensive, indication of the validity of the chosen level of QM theory and basis set is provided by the conformational scans mentioned above. The variation of conformational energy over the values of a given torsion angle often correlates well with the frequency of occurrence of these values in crystal structures occurring in nature, as reported in the CSD (Bruno et al., 2004). In particular, the most likely values are expected to be those in regions of low conformational energy.

For the repulsion/dispersion term, an exponential-6 functional form is usually adopted, with parameters obtained from the literature (Williams, 1999, 2001; Williams and Cox, 1984; Coombes et al., 1996), or fitted to existing crystallographic data. We shall return to discuss the implications of these decisions in Section 5.3.

3.2.2. Stage 1: global search

The global search is performed using CrystalPredictor (Karamertzis and Pantelides, 2005, 2007), with the aim of generating a comprehensive energy landscape containing all potentially low-energy structures. The search is carried out over the unit cell parameters \( \mathbf{X} \), the molecular positions and orientations \( \rho \) and a subset \( \theta \) of the CDFs. \( \theta' \). The remaining CDFs, denoted by \( \overline{\theta} \), are kept constant at their values in the in vacuo conformation of the isolated molecule.

The partitioning of the CDFs \( \theta' \) into \( \theta \) and \( \overline{\theta} \) makes use of the information established at Stage 0. CDFs with flexibility ranges
that are wider than a few degrees normally need to be included in \( \theta \). One exception concerns torsion angles that are so flexible that the variation of conformational energy over the entire range \([0, 360]\) of rotation is very small, e.g. of the order of 1 kJ mol\(^{-1}\) or less. We note that the main consideration for any putative structure at the global search stage is the value of its lattice energy relative to that of the global minimum as this will determine whether or not the structure will be selected for subsequent refinement at Stage 2; the structure’s precise geometry is less important at this stage. Since the variation of the very flexible angles mentioned above has little effect on intramolecular energy, they are normally included in \( \theta \) unless they are likely to have a strong effect on intermolecular interactions by significantly altering the positioning of important atoms in the molecule. For example, torsion angles relating to the rotation of methyl groups affect only the positions of the 3 hydrogen atoms and usually have little effect on intermolecular energy; they are therefore allocated to \( \theta \) (i.e. they are fixed during the global search). On the other hand, a very flexible torsion angle describing the rotation around a bond connecting two major parts of a large molecule will significantly affect the relative positioning of many atoms within a crystal, and may consequently have to be included in \( \theta \).

During the global search, the intramolecular energy \( \Delta U_{\text{intra}} \) is computed via Hermite interpolation over a regular grid spanning the space of \( \theta \). The value of \( \Delta U_{\text{intra}} \) at each grid point is obtained via isolated-molecule QM minimization of the conformational energy with respect to \( \theta \). A key complication is that the size of the grid increases exponentially with the number of \( \theta \) under consideration. For large molecules exhibiting significant flexibility, this effect can sometimes be mitigated by assuming that \( \theta \) can be divided into several independent subsets \( \theta_s \); the intramolecular energy being an additive function of the form

\[
\Delta U_{\text{intra}} = \sum_{s=1}^{NS} \Delta U_{\text{intra}}^{\theta_s}(\theta_s)
\]

A lower-dimensional grid can then be generated over each subset \( \theta_s \) to represent the contribution \( \Delta U_{\text{intra}}^{\theta_s} \).

The intermolecular electrostatic interactions are modeled via point charges. Within CrystalPredictor, it is possible to use conformationally-dependent charges (including “satellite” off-atom charges), which are fitted to the electrostatic potential obtained at each grid point. For the study reported in this paper, we use conformationally-invariant atomic charges that are fitted to the charge density of the gas phase conformation with the
CHELPG algorithm (Breneman and Wiberg, 1990), as implemented in GAUSSIAN09.

Within the current implementation of CrystalPredictor, a search can be carried out in up to 63 space groups. The exploitation of space group symmetry allows the number of variables in the lattice energy minimization to be reduced further, contributing to computational efficiency. A large number of candidate structures are generated within the space groups selected by the user; a low-discrepancy Sobol’ sequence (Sobol’, 1967) is used to obtain the initial values of the independent CDFs, the lattice parameters and positions and orientations of the molecules in the asymmetric unit. The use of such a quasi-deterministic sequence to generate several hundreds of thousands or millions of structures ensures comprehensive coverage of the search space. The sampling of the different space groups under consideration is performed according to their frequency of occurrence among all organic crystal structures selected to undergo such refinement. Appropriate space group constraints are imposed for each local minimization, but this does not restrict the ability of the overall algorithm to search over a wide range of space groups.

For each candidate structure, a local lattice energy minimization is carried out subject to space group symmetry constraints. A successive quadratic programming (SQP) algorithm for constrained problems (Nocedal and Wright, 2006) is used for this purpose, making use of analytical partial derivative information for the reliable and efficient identification of the optimal solution. Distributed computing hardware is used to carry out the minimization of multiple structures in parallel. Once all calculations are complete, the generated structures are post-processed to identify any duplicates, and ranked in order of increasing lattice energy.

3.2.3. Stage 2: crystal structure refinement

At this stage, the lowest lattice energy structures generated at Stage 1 are re-optimized with a more accurate energy model, using the CrystalOptimizer algorithm (Kazantsev et al., 2011a). The structures selected to undergo such refinement are typically those within +20–30 kJ mol⁻¹ of the lowest energy structure.

CrystalOptimizer employs an efficient, yet accurate, representation of intramolecular energy based on local approximate models (LAMs) that is capable of handling extensive molecular flexibility. All CDFs \( \sigma \), including torsion and bond angles and bond lengths, are assumed to be affected by the intermolecular interactions in the crystalline environment. For computational efficiency, they are divided into two categories: the independent CDFs \( \theta \) are those which are affected directly by the intermolecular interactions, while the dependent CDFs \( \bar{\sigma} \) always take values that minimize configurational energy for given values of \( \theta \). In addition to all the CDFs considered as flexible at Stage 1, the set \( \theta \) will also include any angles that are so flexible as to have little effect on conformational energy (treated as fixed at Stage 1, see discussion in Section 3.2.2), as well as, typically, several of the remaining torsion and bond angles whose ranges of flexibility were considered to be too narrow to justify inclusion in the global search.

CrystalOptimizer also employs a much more accurate description of electrostatic intermolecular interactions based on distributed multipole expansions up to hexadecapole rather than point charges. The conformational dependence of these multipoles on the independent CDFs \( \theta \) is also described via LAMs.

Overall, the LAMs express the intramolecular energy \( \Delta U_{\text{intra}} \), the dependent CDFs \( \bar{\sigma} \) and the multipoles as functions of the independent CDFs \( \theta \). Each LAM is constructed from information derived from isolated-molecule QM calculation at a given point \( \theta \), and is sufficiently accurate within the vicinity of that point.

The lattice energy minimization in the current implementation of CrystalOptimizer is formulated as a bilevel optimization problem in which the independent CDFs \( \theta \) are treated as outer variables, while the remaining variables, \( \bar{\sigma} \) and \( \bar{\theta} \), are considered in an inner-level minimization using the DMACRYS code (Price et al., 2010). Whenever the lattice energy or its gradients must be evaluated at a given point \( \theta \), a test is applied to determine whether this point is within the range of validity of an already existing LAM, or whether a new LAM needs to be constructed via a new QM calculation. This approach ensures that, even with large and flexible molecules involving large numbers of independent CDFs \( \theta \), the number of LAMs generated (and, consequently, the QM calculations that need to be performed) is very small compared to the total number of lattice energy evaluations.

At the end of the refinement stage, the final structures are post-processed via clustering to remove multiple occurrences of the same structure.

4. Predicting the polymorphs of axitinib

The CSP methodology described in Section 3 is now applied to the axitinib molecule, aiming to identify polymorphs with one molecule in the asymmetric unit \( (Z=1) \). The only input of this study relating specifically to axitinib is the molecular diagram shown in Fig. 2; the primary result of the study is a list of possible crystal structures ranked on the basis of the calculated lattice energy at 0 K.

4.1. Stage 0: conformational analysis and choice of computational model

As outlined in Section 3.2.1, we start by identifying the CDFs that are likely to be affected significantly by intermolecular interactions in the crystalline environment. In this case, these include the 7 torsion angles indicated in Fig. 2. We perform a set of 1-dimensional conformational energy scans using isolated-molecule QM calculations, varying one of the 7 torsion angles at a time while minimizing conformational energy with respect to all other CDFs. All calculations are performed in GAUSSIAN09 (Frisch et al., 2009) using DFT with the M06 functional (Zhao and Truhlar, 2008) and a 6–31G(d,p) basis set. This model was selected mainly on the basis that it offers a reasonable balance between predictive accuracy and computational cost given the size of the axitinib molecule and the large number of QM calculations that would need to be performed with it during the subsequent stages of the CSP procedure (see Sections 3.2.2 and 3.2.3).

Fig. 4 shows the results for the conformational scans for torsions d26, d27 and d28. As might be expected, the methyl group rotation (torsion d28) has only a minor effect on conformational energy. Since this rotation has a relatively small effect on atomic positions in the crystal, d28 belongs to the category of very flexible CDFs that can be fixed at their in vacuo values during the global search.

In contrast to d28, torsion d27 has a strong effect on conformational energy, exhibiting minima at 0° and 180° (solid line in Fig. 4). This is corroborated by the evidence from the CSD shown in Fig. 5 which confirms that this angle is near-planar. Torsion d26 also has a significant effect on conformational energy (dotted line in Fig. 4). As both d27 and d26 exhibit non-negligible ranges of variation over which the conformational energy is within +30 kJ mol⁻¹ of the global minimum, their variation will need to be considered explicitly during the global search. Torsion angles d8 and d10 are also found to be in this category.

The repulsion/dispersion interactions are modeled using the semi-empirical Buckingham potential with the transferable 'FIT'
parameters for C, N, O, H developed by Cox et al. (1981), Williams and Cox (1984), and Coombes et al. (1996). To the best of our knowledge, there exists no generic transferable parameter set for the sulfur atom. We therefore model the sulfur intermolecular interactions using the potential parameters that were developed by the group of Price for the S atom of the thiophene group of 5-cyano-3-hydroxythiophene (Motherwell et al., 2002). This is likely to be an appropriate choice as the S atom in that molecule has a similar environment to the sulfur of axitinib.

4.2. Stage 1: global search with CrystalPredictor

The conformational analysis of Stage 0 identified the CDFs that need to be treated as flexible during the global search. These are the 6 torsion angles listed in Table 2, which also shows the domain of variation that needs to be searched for each angle (3rd and 4th columns); these domains are selected so that their Cartesian product includes all points with conformational energy up to +30 kJ mol⁻¹ above the global in vacuo minimum. In the case of torsion d27, the search is restricted to a small range around the lowest minimum at 180° (Fig. 4).

As mentioned in Section 3.2.2, a multi-dimensional Hermite interpolant is used for the computation of intramolecular energy contributions during the global search. The interpolant is constructed over a regular grid with the spacing indicated in the penultimate column of Table 2.

The last column of Table 2 shows the number of points in the corresponding dimension of the grid. From this, it can be seen that a 6-dimensional grid that would be sufficiently accurate over the entire domain of interest would involve 1,249,248 grid points, each requiring an isolated-molecule QM calculation. This would clearly be prohibitively expensive from a computational point of view. Therefore, taking account of axitinib’s molecular structure (cf. Fig. 2) and the results of the conformational scans, we divide the 6 torsion angles into three groups, each assumed to have an independent effect on conformational energy:

- **Group 1**: d8, d10 described by a 2-dimensional grid of 132 points;
- **Group 2**: d19, d20, d26, described by a 3-dimensional grid of 1183 points;
- **Group 3**: d27, described by a 1-dimensional grid of 8 points.

In particular, it is assumed that torsion d27 can be treated as independent because only small deviations around the minimum are considered. This reduces the required number of grid points considerably. Overall, this decomposition allows us to approximate conformational energy with a total of 1323 QM calculations. Fig. 6 shows some aspects of these grids and the variation of intramolecular energy over them. The energy variation over the d8 × d10 space (Fig. 6a) indicates that planar conformations are favored in the part of the axitinib molecule on the right of Fig. 2. In contrast, Fig. 6b shows that a wide range of combinations of d20 and d26 would result in similar energy values, and therefore these angles may deviate considerably from their in vacuo values.

The global search was performed over the 59 space groups that appear with the highest frequency in CSDSymmetry (Yao et al., 2002). A total of 4,800,000 candidate structures, each involving one molecule in the asymmetric unit, were generated and used as initial guesses for lattice energy minimization subject to space group symmetry constraints. Clustering was used to eliminate any duplicates among the final structures.

Fig. 7 shows the resulting lattice energy landscape which has 5960 unique structures within +30 kJ mol⁻¹ of the global minimum, and 1830 unique structures within +25 kJ mol⁻¹. Structures corresponding to all four Z = 1 polymorphs of axitinib are found in this landscape, and their main characteristics are
summarized in Table 3. The global minimum of the landscape corresponds to Form VI, whilst Form I is the least stable among the experimental polymorphs, with an energy +23.49 kJ mol$^{-1}$ above the global minimum. The quality of experimental structure reproduction is quantified via the root mean squared deviation of the 15-molecule coordination sphere (rms15, also reported in Table 3), calculated using COMPACT (Chisholm and Motherwell, 2005) as implemented in Mercury (Macrae et al., 2008). The computed and experimental structures are in reasonable agreement, especially if one considers the relative simplicity of the computational model used during the global search stage.

4.3. Stage 2: structure refinement with CrystalOptimizer

The structure refinement is performed using the CrystalOptimizer algorithm (Kazantsev et al., 2011a). In addition to employing distributed multipole expansions for the description of intermolecular electrostatic interactions, CrystalOptimizer allows all CDFs in the axitinib molecule to deviate from their in vacuo values. The lattice energy minimization manipulates directly 16 independent CDFs $\theta$ comprising the bond angles and torsion angles listed in Table 4; the remaining CDFs are treated as dependent CDFs $\phi$ which always adjust themselves to minimize the isolated-molecule conformational energy for any given values of $\theta$.

Given the relative simplicity of the lattice energy model used in the global search, and in order to ensure that no structures of practical importance are missed, the CrystalOptimizer refinement would normally need to be applied to all structures appearing within 20–30 kJ mol$^{-1}$ of the global minimum in the lattice energy landscape of Fig. 7. In this case, this would involve the refinement of several thousands of structures, which is impractical given the complexity of the axitinib molecule and the high degree of detail incorporated in the lattice energy model used by CrystalOptimizer.

In view of the above, we adopt a two-stage approach that first attempts to establish a more reliable lattice energy estimate (and therefore ranking) of the structures identified in the global search, before applying the full CrystalOptimizer refinement to the most promising structures. These two steps, referred to as Stages 2a and 2b respectively, are described in more detail below.

4.3.1. Stage 2a: structure re-ranking using a single CrystalOptimizer iteration

A significantly improved estimate of the lattice energy for a structure derived by the global search can be obtained by taking a
A single iteration of CrystalOptimizer is not entirely clear as it involves at least one isolated-molecule QM conformational energy minimization. However, the information generated by these QM calculations is used to construct LAMs which are stored in a database for potential use during Stage 2b.

Here we apply the above procedure to the 3765 structures found within 28 kJ mol\(^{-1}\) of the global minimum in Stage 1. The resulting lattice energy landscape is shown in Fig. 8. Extensive re-ranking of the structures obtained at the end of Stage 1 (cf. results presented in Fig. 7 and Table 3), with the ranks of experimental forms VI, XXV, and XLI now being 14, 77, 95, and 258 respectively.

4.3.2. Stage 2b: structure refinement using CrystalOptimizer

We now apply full refinement to the 500 lowest-energy structures determined at Stage 2a, spanning a range of about +20 kJ mol\(^{-1}\) of the global minimum in Fig. 8. Each of these structures is used as an initial point for a full minimization of lattice energy using CrystalOptimizer. Following the identification of duplicate structures using COMPASS (Chisholm and Motherwell, 2005) as implemented in Mercury (Macrae et al., 2008) and with default settings, 139 minimized structures are removed. This illustrates how allowing a wider range of conformational flexibility allows multiple initial structures to relax into the same final structure.

The final lattice energy landscape, shown in Fig. 9, contains 361 distinct structures spanning a range of just over 20 kJ mol\(^{-1}\) of lattice energies. The density of the landscape confirms the propensity of axitinib to pack in many different ways, and suggests that at least some polymorphs that have not yet been observed experimentally may exist in nature.

We note that Form VI is predicted to have the lowest energy among the four experimental polymorphs, being ranked third, while Form XLI is the highest in energy and is ranked 108th. The lattice energy difference between these two forms is just over 10 kJ mol\(^{-1}\). The computed order of relative stability is thus Form VI > Form XXV > Form I > Form XLI.

Assuming that the lattice energy ranking at 0 K can be compared to the experimental heat of fusion ranking, we can see that the stability of Form XLI is underestimated, but the other three forms are ranked in the correct order of relative stability.

Other characteristics of the four predicted structures are summarized in Table 5. The results of a local minimization of the 2\(Z\) polymorph, Form IV, are also included in Table 5 to allow a more complete assessment of the quality of the computational model. Form IV is found to have the second lowest energy, less than 2 kJ mol\(^{-1}\) above that of Form VI. The computed densities for all polymorphs are found to be in good agreement with the measured values, with Form XXV being the densest, followed by Form XLI. Forms I and VI are found to be less dense, with nearly equal densities. All computed densities are within 2% of the corresponding experimental values. Overlays between the predicted structures and the experimental structures are shown in Fig. 10. The reproduction of the geometrical features, as measured by rms1 for the molecular conformational and rms15 for the crystal structure, is of good quality for Forms VI, XXV, and XLI, with rms1 values around 0.1 Å and rms15 values less than 0.35 Å. For Form I, a poorer reproduction is observed in terms of both conformation and crystal structure. In order to analyze the impact of computational choices on the accuracy of reproduction of the experimental structures, CrystalOptimizer computations starting from the Stage 1 structures that best match the experimental structures and using different levels of theory were carried out. The approaches used, in addition to M06/6-31G(d,p), were HF, PBE0, B3LYP with a 6-31G (d,p) basis, and M06/6-31G(d,p) with the polarizable continuum model (PCM) with a dielectric constant of 3 (Cooper et al., 2008). The resulting rms15 values were found to be similar or worse in all cases. Furthermore, none of these calculations resulted in a better match of the experimentally-determined relative stability of the polymorphs: Form XLI remained the least stable structure across all levels of theory, with energy differences from the consistently most stable form, Form VI, ranging from 9.9 to 19.2 kJ mol\(^{-1}\).
5. Discussion

In this section, we focus on some key aspects of the approach and results presented in Section 4. We consider the quality of the predictions and its interactions with the underlying physical models. We also discuss some methodological and algorithmic considerations that are found to be important in light of the observed performance of the CSP approach when applied to axitinib.

5.1. Quality of the results

The general CSP methodology described in Section 3.2 has been successful in that, starting only from axitinib’s molecular structure shown in Fig. 2, it has managed to identify low-energy crystal structures corresponding to all four known experimental polymorphs for the class considered (i.e., one molecule in the asymmetric unit).

The geometry of the predicted structures is in very good agreement with the experimental structures for three of these polymorphs, and in reasonable agreement for the fourth. The predicted energy ranking is also in good agreement for all but one polymorph. Overall, the results highlight the significant impact of conformational flexibility on crystal structure, with the conformational energies of the axitinib molecule within the experimental polymorphs being ±10–20 kJ mol⁻¹ above the in vacuo value.

### Table 5

Main characteristics of the four structures of the final computed lattice energy landscape that correspond to the experimentally known polymorphs and of the structure of Form IV computed by local minimization with the same model. All quantities are computed, except for the lattice parameters and densities in the rows labeled “expt.” “pred.” denotes results of the CSP study; “min.” denotes the result of a local minimization starting from the experimental structure, as Form IV was not considered in the search. Two rms1 values are reported for Form IV, corresponding to the two molecules in the asymmetric unit.

<table>
<thead>
<tr>
<th>Polymorph</th>
<th>Energetic information</th>
<th>Structural information</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E_latt (kJ mol⁻¹)</td>
<td>Rank</td>
</tr>
<tr>
<td>XLI</td>
<td>– 181.86</td>
<td>108</td>
</tr>
<tr>
<td>VI</td>
<td>– 192.28</td>
<td>3</td>
</tr>
<tr>
<td>XXV</td>
<td>– 187.10</td>
<td>26</td>
</tr>
<tr>
<td>IV</td>
<td>– 190.75  N/A</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>– 186.28</td>
<td>32</td>
</tr>
</tbody>
</table>

Fig. 10. Overlay between predicted (green) and experimental (colored by element) structures. (a) Form XLI, (b) Form VI, (c) Form XXV and (d) Form I. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

### Table 5 continued

<table>
<thead>
<tr>
<th>Polymorph</th>
<th>Energetic information</th>
<th>Structural information</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E_latt (kJ mol⁻¹)</td>
<td>Rank</td>
</tr>
<tr>
<td>XLI</td>
<td>– 181.86</td>
<td>108</td>
</tr>
<tr>
<td>VI</td>
<td>– 192.28</td>
<td>3</td>
</tr>
<tr>
<td>XXV</td>
<td>– 187.10</td>
<td>26</td>
</tr>
<tr>
<td>IV</td>
<td>– 190.75  N/A</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>– 186.28</td>
<td>32</td>
</tr>
</tbody>
</table>
consistent with the already observed propensity of axitinib to pack in different crystals. On the other hand, it is also characteristic of current CSP techniques that many more polymorphs are predicted than are measured. This may be partly a result of inaccuracies in the lattice energy model. Also some structures may be found to disappear once entropic effects are considered, while kinetic considerations could make some metastable structures unlikely to ever form in nature (Price, 2013).

5.2. Space group distribution of predicted low-energy structures

As mentioned in Section 4.2, the global search was performed over 59 space groups. In fact, only 15 space groups actually appear in the final energy landscape of Fig. 9, with the relative frequencies shown in Fig. 11. Approximately 80% of the structures belong to three space groups, P1, P2_1/c and C2/c, with nine space groups having a frequency of no more than 2%. Furthermore, only 12 space groups (P1, P2_1, P2_1/c, P2_2_1, P2_1_2_1, Pna2_1, Pca2_1, C2/c, C2, Fdd2, P2_1/c, and Pbcn) appear among the 100 lowest-energy structures in the final landscape, with 79% of the structures in a P1, P2_1/c or C2/c unit cell.

5.3. Effects of modeling of intermolecular repulsion/dispersion interactions

The significant differences in the underlying computational model of lattice energy between Stages 1, 2a and 2b have a relatively limited impact on the quality of reproduction of the experimental structures, as measured for example by the rms_15. However, the accuracy of lattice energy calculations does affect significantly the relative ranking of different structures. This is arguably the weakest aspect of the current approach. Inaccuracies can arise from the neglect of specific contributions to the crystal energy (e.g., polarizability and entropic effects). In fact, such contributions can be captured partly (Pantelides et al., 2014) within the empirical term of the computational model used here (nominally the “repulsion/dispersion” term) whose parameters are fitted to energetic and structural data from a number of crystal structures. In this context, it is interesting to note that, while the importance of the accurate modeling of the electrostatic and intramolecular contribution has gained wide understanding and acceptance in recent years, much less attention has been given to the impact that the empirical term contribution may have on the final outcome. More specifically, a possible source of error is the fact that the FIT parameters used here were derived from experimental data using models of the intramolecular and electrostatic contributions that are very different to those employed by current CSP methodologies. In particular, consideration of conformational flexibility was limited, QM calculations were performed only at the HF (Hartree–Fock) level of theory, and electrostatic interactions were modeled via atomic point charges rather than distributed multipoles. Thus, there is a potentially severe mismatch between the different contributions to the lattice energy.

To explore this issue, the various contributions to the lattice energy are summarized in Table 6 for the predicted structures that correspond to the four experimental forms. The table shows results obtained using two different repulsion/dispersion potentials, namely the FIT one used throughout this paper, and the W01 potential (Williams, 1999, 2001), which includes a foreshortening of the interaction site for hydrogen; the same sulfur parameter set (Motherwell et al., 2002) is used in both cases.

Considering the FIT results first, it is evident that the variability of the repulsion/dispersion contributions across the four structures is significantly larger than the variabilities of the intramolecular and electrostatic contributions.

A very similar effect is observed in the results obtained using the W01 potential. It is also observed that, with this potential, the predicted lattice energy difference between the most and least stable polymorphs is reduced from 10.42 kJ mol\(^{-1}\) to 7.31 kJ mol\(^{-1}\).

The last four rows of Table 6 show a comparison of the predictions obtained using the two repulsion/dispersion potentials. As a result of the change in the repulsion/dispersion potential, the structures obtained are slightly different in terms of both molecular conformation and relative positioning of the atoms in different molecules in the crystal, and this indirectly leads to small differences in the intramolecular contributions and somewhat larger ones in the electrostatic contributions. The effect on the quality of the experimental structure is generally small.

## Table 6

<table>
<thead>
<tr>
<th>Structure</th>
<th>(U_{\text{fit}}) (kJ mol(^{-1}))</th>
<th>(\Delta U_{\text{fit}}) (kJ mol(^{-1}))</th>
<th>(U_{\text{pol}}) (kJ mol(^{-1}))</th>
<th>(U_{\text{rd}}) (kJ mol(^{-1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td>−192.28</td>
<td>10.96</td>
<td>−55.90</td>
<td>−147.33</td>
</tr>
<tr>
<td>XXV</td>
<td>−187.10</td>
<td>20.68</td>
<td>−44.51</td>
<td>−163.27</td>
</tr>
<tr>
<td>I</td>
<td>−186.28</td>
<td>17.67</td>
<td>−59.42</td>
<td>−144.53</td>
</tr>
<tr>
<td>XLI</td>
<td>−181.86</td>
<td>13.25</td>
<td>−59.41</td>
<td>−135.70</td>
</tr>
<tr>
<td>max−min</td>
<td>10.42</td>
<td>9.73</td>
<td>14.91</td>
<td>27.57</td>
</tr>
<tr>
<td>W01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td>−173.95</td>
<td>11.44</td>
<td>−56.62</td>
<td>−128.78</td>
</tr>
<tr>
<td>XXV</td>
<td>−169.22</td>
<td>20.01</td>
<td>−46.86</td>
<td>−142.38</td>
</tr>
<tr>
<td>I</td>
<td>−166.64</td>
<td>17.75</td>
<td>−59.89</td>
<td>−124.49</td>
</tr>
<tr>
<td>XLI</td>
<td>−166.87</td>
<td>13.26</td>
<td>−62.20</td>
<td>−117.93</td>
</tr>
<tr>
<td>max−min</td>
<td>7.31</td>
<td>8.57</td>
<td>15.34</td>
<td>24.45</td>
</tr>
</tbody>
</table>

**Fig. 11.** Percentage frequency of appearance of the 15 space groups found in the 361 low-energy structures identified following full minimization with CrystalOptimizer.
More interestingly, the differences in the repulsion/dispersion component of the energy between the two sets of results are in the range of about 17.8–20.9 kJ mol$^{-1}$. These are much larger than the lattice energy differences across the four different polymorphs as determined using either the FIT or the W01 potential (10.42 and 7.31 kJ mol$^{-1}$ respectively). The impact of the choice of repulsion/dispersion potential can also be seen by carrying out an extensive re-minimization of all structures found in the final lattice energy landscape with the W01 potential (Fig. 12). The main impact is the stabilization of Form XLI, whose rank is reduced from 108th (cf. Table 5) to 52nd. The least stable experimental polymorph is now Form I, ranked 58th.

The above observations emphasize the importance of accurately characterizing the empirical term contributions in CSP. In the context of empirical potentials, such as FIT and W01, it is not expected that the parameters used are optimal for axitinib and indeed for the level of theory used. This suggests that the corresponding parameters may need to be re-estimated using models that are consistent with the intramolecular and electrostatic descriptions that are employed during the CSP.

5.4. Use of the re-ranking step 2a

The introduction of a re-ranking step 2a (cf. Section 4.3.1) within the refinement stage is based on the assumption that the ranking of structures following a single CrystalOptimizer iteration is a better indicator of final ranking than the ranking obtained following the global search using CrystalPredictor (Stage 1). To investigate this, we consider the final 361 unique polymorphs identified, and trace their rankings through the CSP process. Figs. 13a and 13b plot the rankings of these structures at the ends of Stages 1 and Stage 2a respectively, against the final rankings. The data in Fig. 13b are visibly better correlated than those in Fig. 13a, the corresponding $R^2$ coefficients of correlation being 0.387 and 0.078. Therefore, the lattice energy values after a single CrystalOptimizer iteration provide a better indicator of “promising” structures than the values established by CrystalPredictor during the global search. This may be an important consideration for CSP applied to large, flexible molecules where full structure refinement (Stage 2b) is expensive and needs to be limited to these promising structures only.

6. Concluding remarks

This paper has reviewed a CSP methodology that has been used fairly extensively both by the authors’ research group and by others over the past decade. Its application to axitinib, a relatively large pharmaceutical molecule, has demonstrated the significant progress that has been achieved in recent years in terms of methodological improvements (e.g., in the handling of large extents of conformational flexibility, the systematic and comprehensive search of the lattice energy landscape, and the description of electrostatic interactions), the effective exploitation of high-performance computing architectures for performing this demanding task, and, ultimately, the quality of the results obtained (as demonstrated, for example, by the accurate identification of axitinib’s four experimentally observed polymorphs starting only from the molecular diagram).

The results of the axitinib case study also provide a fairly good illustration of the limitations of current methodologies. Some of these deficiencies are algorithmic, e.g. the large numbers of structures that need to be refined at Stage 2 because of the relatively low accuracy of the model employed during the global search at Stage 1; or the use of a bilevel optimization at Stage 2 (cf. Section 3.2.3) especially if the inner-level optimization does not provide exact partial derivative information needed by the outer-level one. It should be possible to overcome most of these issues via improvements in the underlying algorithms and their implementation (cf. Section 5.5). In addition, the systematic application within our CSP approach of further tests of the viability of the predicted structures, such as mechanical stability, could be used to eliminate some of the putative crystals (Price, in press).

A more serious shortcoming concerns the correct ranking of the polymorphs identified. This is severely affected by the accuracy of the lattice energy calculation. As discussed in Section 5.3, given the significant recent improvements in the descriptions of intramolecular and electrostatic contributions, further progress may be predicated on improving the characterization of repulsive/dispersive interactions.

5.5. Computational efficiency considerations

The computational cost of the different steps is reported in Table 7. The total time spent is approximately 17.4 CPU years, an amount of computation which is rendered practically feasible only via the extensive exploitation of distributed computing architectures. Stage 1 accounts for approximately 40% of the total cost, although the number of structures minimized is 3–4 orders of magnitude larger than that in Stage 2. Once the grid has been generated in Stage 1, the marginal cost of additional structure minimizations is very small (approximately 15 CPU seconds per structure). Given the rapidly increasing cost of individual structure minimizations as the computational model increases in accuracy from Stage 1 to Stages 2a and 2b, it is clear that the overall cost could be reduced if a more accurate energy ranking could be achieved in Stage 1. More specifically, the lattice energy model used for the global search in the current work is based on several approximations that could be removed without an excessive increase in the computational cost. In particular, the electrostatic model (atomic point charges) is assumed to be conformationally invariant and so are all CDFs other than those explicitly considered as “flexible”. These deficiencies may be addressed by adopting the LAM-based approach (Kazantsev, 2011) within CrystalPredictor; this will also allow the consideration of higher degrees of molecular flexibility within the global search. Overall, this can be expected to lead to improved rankings by the end of the global search, thereby reducing the number of structures to be refined during Stage 2.
In this context, it should be recognized that, albeit usually thought of as "expulsion/dispersion potentials", the empirical terms used within current CSP methodologies actually attempt to compensate for a multitude of physical and numerical approximations; and this mismatch between model and reality is likely to persist even if ab initio descriptions are improved (e.g. via the consideration of polarizability effects). We therefore believe that there is a need for more systematic approaches for exploiting all available experimental information within the theoretical framework of CSP.

At a more fundamental level, another likely source of error in the relative ranking of different polymorphs is the use of lattice energy as a proxy measure for free energy, which means that, strictly speaking, any predicted rankings pertain to 0 K rather than to room temperature. The computation of free energies for crystals of organic molecules is now beginning to be feasible (Vasileiadis, 2013), and this should allow an entropic correction to be post-calculated for, and applied to, each of the lowest-energy structures identified by the CSP.

Acknowledgments

We are grateful to Professor S.L. Price for providing the DMACRYS code for use within CrystalOptimizer. Financial support for the work reported here was provided by the United Kingdom’s Engineering and Physical Sciences Research Council (EPSRC) under Grants EP/E016340, EP/J003840/1 and EP/J014958/1. Calculations were performed on the High Performance Computing Cluster at Imperial College London.

Table 7

Approximate CPU time for the different stages of the CSP procedure, and average cost per structure minimization. Calculations were performed using dual intel Xeon dual core 5150 processors running at 2.66 GHz.

<table>
<thead>
<tr>
<th>Type of calculation</th>
<th>Approximate computational cost (CPU hours)</th>
<th>Percentage of total time (%)</th>
<th>Mean CPU time per structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1: grid generation</td>
<td>37,120</td>
<td>24.4</td>
<td>N/A</td>
</tr>
<tr>
<td>Stage 1: CrystalPredictor search</td>
<td>19,500</td>
<td>12.8</td>
<td>15 s</td>
</tr>
<tr>
<td>Stage 2a: CrystalOptimizer single iteration</td>
<td>49,420</td>
<td>32.5</td>
<td>13.1 h</td>
</tr>
<tr>
<td>Stage 2b: CrystalOptimizer full local minimization</td>
<td>46,030</td>
<td>30.3</td>
<td>92.0 h</td>
</tr>
<tr>
<td>Overall</td>
<td>152,070</td>
<td>100</td>
<td>N/A</td>
</tr>
</tbody>
</table>

References


