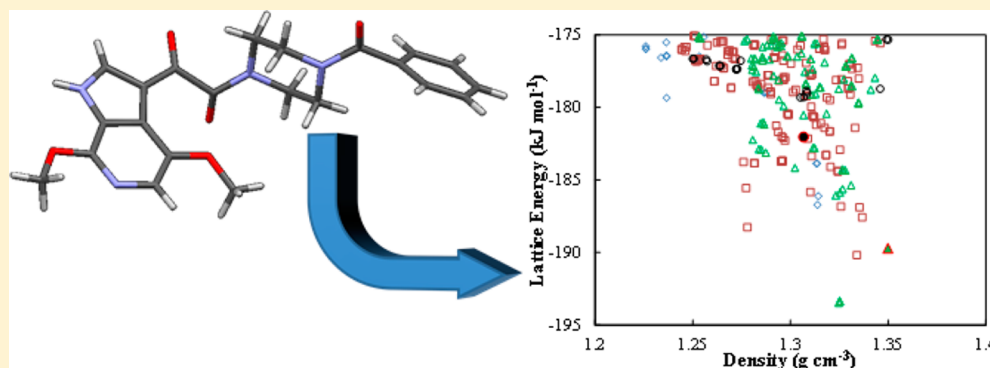


Efficient Handling of Molecular Flexibility in Ab Initio Generation of Crystal Structures

Matthew Habgood,[†] Isaac J. Sugden, Andrei V. Kazantsev, Claire S. Adjiman, and Constantinos C. Pantelides*

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

S Supporting Information



ABSTRACT: A key step in many approaches to crystal structure prediction (CSP) is the initial generation of large numbers of candidate crystal structures via the exploration of the lattice energy surface. By using a relatively simple lattice energy approximation, this global search step aims to identify, in a computationally tractable manner, a limited number of likely candidate structures for further refinement using more detailed models. This paper presents an effective and efficient approach to modeling the effects of molecular flexibility during this initial global search. Local approximate models (LAMs), constructed via quantum mechanical (QM) calculations, are used to model the conformational energy, molecular geometry, and atomic charge distributions as functions of a subset of the conformational degrees of freedom (e.g., flexible torsion angles). The effectiveness of the new algorithm is demonstrated via its application to the recently studied 5-methyl-2-[(2-nitrophenyl)amino]-3-thiophenecarbonitrile (ROY) molecule and to two molecules, β -D-glucose and 1-(4-benzoylpiperazin-1-yl)-2-(4,7-dimethoxy-1H-pyrrolo[2,3-c]pyridin-3-yl)ethane-1,2-dione, a Bristol Myers Squibb molecule referenced as BMS-488043. All three molecules present significant challenges due to their high degree of flexibility.

1. INTRODUCTION

The crystal structure adopted by a molecule influences many solid state properties that significantly affect its utility, including solubility, dissolution rate, bioavailability, hygroscopicity, compactibility, tabletability, shock-sensitivity, and color.¹ This represents both a challenge and an opportunity: the emergence of unanticipated crystal forms, such as polymorphs or solvates, is a significant risk in the fine chemical and pharmaceutical industries; on the other hand, the ability to devise new crystal forms, such as cocrystals or salts, with improved properties (e.g., bioavailability), is highly desirable.

Computational crystal structure prediction (CSP) aims to predict the crystal structures that a given molecule or set of molecules can form, starting from the corresponding molecular diagram(s) only. CSP has made considerable advances in the past decade,^{2–4} as evidenced, for example, by the Cambridge Crystallographic Data Centre's "blind tests". In the two most recent (fourth and fifth) blind tests^{5,6} correct predictions were made for every molecular system, with the exception of a

hydrate that was predicted correctly, with the exception of hydrogen atomic positions. These and other studies have demonstrated that predictions for small molecules are now becoming almost routine. Furthermore, improvements in prediction capabilities have also been demonstrated for larger molecules of pharmaceutical size,^{7–10} and for hydrates,¹¹ cocrystals,^{12–15} and salts.^{16,17}

Based on thermodynamic considerations, the crystal forms that are most likely to appear in nature are those corresponding to low-energy minima in the free-energy surface with respect to all variables determining the structure of an infinite periodic crystal lattice. These comprise the unit cell geometry and the positions, orientations, and conformations of all chemical entities (e.g., molecules or ions) within the unit cell. The stable crystal form is the one corresponding to the global free-energy minimum, whereas metastable forms typically correspond to

Received: July 15, 2014

Published: February 18, 2015

local minima within a few kJ mol^{-1} of the global minimum.¹⁸ However, in view of the theoretical and computational difficulties of determining free energies, most current CSP methods employ lattice energy or enthalpy as surrogate measures for the minimization.

1.1. Two-Stage CSP Methodologies. The complexity of CSP mainly arises from the fact that the lattice energy surface exhibits a very large number of local minima, the correct identification of which tends to be very sensitive to the accuracy of the lattice energy computations. Accordingly, practical CSP methods involve two distinct steps: (1) a global search stage which aims to identify the low-energy minima by performing an extensive search of the decision space using moderately accurate and relatively inexpensive lattice energy computations; and (2) a structure refinement stage during which the structures of lower energy identified by the global search are further refined using a more accurate, but also more computationally expensive, lattice energy model.

The success of the above two-stage approach is predicated on ensuring that no physically relevant structure is missed either because it is not identified by the global search, or because it is determined by the latter to have such a high lattice energy that it is not selected for further refinement. Thus, the more accurate the lattice energy model being used during stage 1, the fewer the structures that need to be refined during stage 2. On the other hand, using too expensive a lattice energy model at stage 1 may render the whole algorithm impracticable given that, typically, hundreds of thousands or millions of structures need to be generated and evaluated. Accordingly, this paper is specifically concerned with techniques for improving the effectiveness and accuracy of the global search stage in identifying all low-energy structures, while staying within acceptable computational limits.

It is worth noting that the two-stage approach was applied by all four groups who achieved successes in the most recent (fifth) blind test of crystal structure prediction,⁶ using one of three different codes for initial structure generation. The program GRACE¹⁹ uses a force field that is custom-parametrized for each molecule, taking into account a full range of variation in bonds, angles and torsions, as well as charges and repulsion/dispersion parameters to model intermolecular interactions; the starting points for local minimization are generated via a parallel-tempering approach that seeks out low-energy regions of the decision variable space. The program UPACK^{20–23} allows for several types of potential energy function and generates starting points across the whole decision variable space, using either random numbers or a grid of points. Finally, two of the four successful groups used the program CrystalPredictor^{24,25} which is described in more detail in section 1.2.

There are other approaches in addition to those above that can be used for the generation of crystal structures. Some workers have applied a genetic algorithm, combined with local minimizations, to home in on low-energy structures.²⁶ Other researchers have dispensed with a potential energy function altogether (cf. the use of close packing/volume minimization to generate structures in MOLPAK²⁷) or have supplemented energy calculations with knowledge of common interactions or packing motifs (cf. the program Promet²⁸ or the work²⁹ of Thakur et al.). Metadynamics has also been used³⁰ to explore the potential energy surface of a molecule in the crystal phase without a variable-space spanning generation of points. However, none of these alternate approaches has yet demonstrated

successful predictions for a wide range of molecules, especially under blind test conditions.

1.2. CrystalPredictor Algorithm. This paper aims to present a significantly improved version of the CrystalPredictor algorithm^{24,25} for the global search stage of CSP methodologies (cf. section 1.1). This section presents a brief outline of the original algorithm to the extent that it is necessary to understand the current paper. This is followed by an analysis of some aspects that are the focus of the proposed improvements.

CrystalPredictor performs an extensive exploration of the lattice energy surface by using low-discrepancy Sobol sequences.³¹ These are sequences of points chosen deterministically to ensure the best coverage of the space of decision variables that is achievable using any given number of points. In particular, they result in more uniform coverage than pseudorandom number generation, and they are better than a uniform grid in that each generated point corresponds to a unique value for each and every decision variable. Furthermore, a Sobol sequence can be restarted from any point, taking into account all previously generated structures; this makes it possible, if necessary, to extend the scope of an already performed global search without having to repeat any calculations. The structures generated are used as initial points for lattice energy minimization using a sequential quadratic programming algorithm. The minimized structures of lowest energy are selected for the refinement stage of the CSP algorithm.

In common with several other CSP algorithms, CrystalPredictor expresses lattice energy, E^{latt} , as the sum of intramolecular and intermolecular (repulsion/dispersion and electrostatic) energy terms. The repulsion/dispersion interaction contribution is described as an exponential- r^{-6} term, typically parametrized by atom type using standard values obtained by fitting to experimental data, for example, the “FIT” potential^{32–34} or the “Williams 01” potential.³⁵ This shorter-range potential is calculated for each atom pair out to a user-defined cutoff (e.g., 15 Å). Electrostatic interactions are represented by distributed point charges situated on the atoms and, optionally, other “satellite” locations.³⁶

The computation of intramolecular lattice energy contributions in the original CrystalPredictor is closely related to the latter’s model of molecular flexibility.^{24,25} In particular, a flexible molecule is considered to comprise two or more rigid fragments connected via flexible torsion angles (FTAs). Thus, the intramolecular energy is a function of the values of the FTAs and can be computed via an isolated-molecule QM calculation which minimizes the conformational energy for chosen values of the FTAs. In practice, these QM calculations are performed a priori to determine the values of the intramolecular energy and its partial derivatives with respect to the FTAs on a regular grid in the space of the FTAs. This allows the construction of restricted multidimensional cubic Hermite interpolants that are used for the efficient computation of intramolecular energy contributions during the global search. The rigid fragments are assumed to retain their in vacuo conformation throughout the search.

The CrystalPredictor algorithm outlined above has been implemented for, and deployed on, machines comprising multiple processors within a distributed computing architecture. Once the grid of intramolecular energy has been derived, the code allows the minimization of a single structure in a few seconds, making it feasible to minimize hundreds of thousands

of structures within reasonable CPU times. In recent years, it has been applied successfully by several research groups to a relatively wide variety of systems including single-compound crystals,^{9,37–45} cocrystals,^{47–49} including chiral cocrystals,^{49–51} hydrates and solvates.^{11,52} It has also provided the basis of several successful predictions in the third,⁵³ fourth,⁵ and fifth⁶ blind tests, including the only two successful predictions for “molecule XX”,⁷ one of the largest and most flexible molecules considered in a blind test to date.

1.3. Need for an Improved Global Search Algorithm.

The application of CrystalPredictor to a wide range of systems has helped identify several areas that would need to be addressed in order to extend the algorithm's applicability to even more challenging systems.

One important concern is the size of the interpolation grid needed to provide a sufficiently accurate representation of the intramolecular energy contributions. In particular, the use of a regular grid spanning the FTA ranges of interest may pose a severe computational barrier given the cost of the QM calculation at each grid point and the fact that the number of grid points increases exponentially with the number of FTAs. Moreover, the number of points in each dimension of the grid depends on the accuracy of the interpolating function being employed, namely, an n -dimensional Hermite interpolant where n is the number of FTAs. In principle, such an interpolant provides a cubic approximation of the function being interpolated on the grid and can be constructed using the values of the function and its partial derivatives up to order n at the grid points. However, in the case of CrystalPredictor, in order to avoid the need for any partial derivatives of the configurational energy above first order, the n -dimensional Hermite interpolant is “restricted” (i.e., some of its coefficients are set to zero).²⁵ As a result, a cubic quality of approximation is actually obtained only for the case $n = 1$; as n increases, the proportion of interpolating coefficients that are set to zero increases, and consequently, the quality of approximation deteriorates. The practical implication of this is that an increasingly finer grid needs to be used to achieve the required accuracy of configurational energy representation as more FTAs are taken into account.

A second difficulty arises from treating the molecular fragments connected via the FTAs as completely rigid during the global search, which is sometimes physically unrealistic. In practice, the primary deformation of the FTAs caused by intermolecular interactions within the crystalline environment may result in non-negligible changes in the other molecular conformational parameters (e.g., torsion angles that are not treated as “flexible” and potentially some of the bond angles). These secondary deformations will, in turn, affect interatomic distances and therefore intermolecular energies.

Moreover, the effects of molecular deformation on the atomic charges used for computing electrostatic contributions to the lattice energy may also need to be considered. In principle, these charges may be described as functions of the FTAs using interpolants similar to those used for the intramolecular energy. However, in practice, this approach is limited by the difficulty in extracting the values of the partial derivatives of the atomic charges with respect to the FTAs from QM calculations. Therefore, most of the applications of CrystalPredictor listed above have made use of conformation-invariant atomic charges.

Finally, the task of dividing the molecule(s) of interest into rigid fragments introduces a degree of subjective judgment that

is undesirable in a general CSP methodology. Although the choice of such fragments is often obvious (especially given the limited number of FTAs that can be handled, as previously mentioned), this is not always the case. For example, in a recent study⁴⁵ on 5-methyl-2-[(2-nitrophenyl)amino]-3-thiophenecarbonitrile (ROY, so-called because of its Red, Orange, and Yellow polymorphs), the existence of an intramolecular hydrogen bond was a key factor that had to be taken into account both in the definition of an appropriate set of rigid fragments and in the choice of FTAs.

1.4. Structure of This Paper. The work reported in this paper aims to extend the applicability of CrystalPredictor by addressing the above issues. Section 2 describes how this can be achieved via an alternative description of molecular flexibility and the use of Local Approximate Models (LAMs) for intramolecular energy and atomic charges.

Section 3 of the paper presents CSP studies for the three molecules shown in Figure 1. The first example revisits a

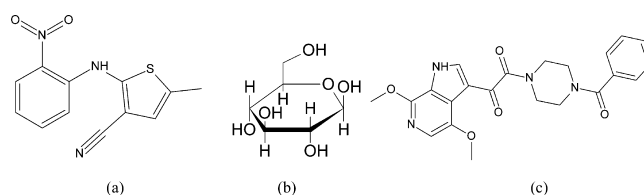


Figure 1. Molecules for which crystal energy landscapes are generated in this study: (a) 5-methyl-2-[(2-nitrophenyl)amino]-3-thiophenecarbonitrile (ROY), (b) β -D-glucose, (c) BMS-488043.

recently reported CSP study⁴⁵ on ROY and demonstrates the ability of the new algorithm to perform the global search at comparable computational cost and with improved accuracy in the ranking of the generated candidate structures. In particular, the rankings of the seven polymorphs with known crystal structures at the end of the global search stage are much improved. The second example is β -D-glucose, a relatively small molecule that nonetheless has a very high degree of flexibility arising from the closely coupled torsions of its five hydroxylic hydrogen atoms. CSP has previously been carried out on this molecule^{20,54,55} by van Eijck et al., but the method was conducted only by effectively ignoring the hydrogen atom positions and then reintroducing them at a later stage in a manner that is restrictive and that may result in a failure to identify low-energy structures for molecules whose crystal-phase behavior is not already understood. In the present study, all five torsions are treated as variables without restriction, and the experimental crystal structure is successfully generated. The final example is the pharmaceutical molecule BMS-488043 (molecular weight 422 g mol⁻¹), which also has five flexible torsions that must be varied during the global search. Although the ranges of torsion angle variations that need to be considered are narrower than in the case of glucose, the size of the molecule makes this another benchmark study for CSP.

Finally, Section 4 summarizes the key points of the paper and draws some general conclusions.

2. CRYSTALPREDICTOR II ALGORITHM

The new CrystalPredictor algorithm seeks to address the issues identified in section 1.3 by adopting a different model of molecular flexibility and also a more efficient approach for the accurate computation of intramolecular energy as a function of

the molecular conformation during the global search. It also makes use of conformation-dependent atomic charges.

In the rest of this paper, we shall refer to the original and new algorithms as “CrystalPredictor I” and “CrystalPredictor II” respectively. We will consider crystal structures that, in principle, contain any number of chemical entities in the asymmetric unit and belong to any space group. Although, in the interest of brevity, we shall refer to “molecules”, the proposed techniques also handle ionic species in a straightforward manner.

2.1. Description of Molecular Flexibility. CrystalPredictor I essentially categorized the conformational degrees of freedom (CDFs) of the molecule(s) under consideration as either “flexible” or “rigid”, the latter being fixed at their values in the in vacuo conformation. Furthermore, it was only possible to treat torsional angles as flexible. In contrast, CrystalPredictor II allows all CDFs, including torsional angles, bond angles, and bond lengths, to vary during the global search. CDFs are further distinguished into *independent* and *dependent* CDFs. The former, denoted as θ , are affected directly by intermolecular interactions in the crystalline environment; they are usually CDFs that have a relatively small effect on the magnitude of conformational energy and can therefore undergo substantial changes from their in vacuo values. On the other hand, the dependent CDFs, denoted as $\bar{\theta}$, have a strong influence on the conformational energy and are therefore not expected to undergo large deviations within the crystal; however, they do assume values that minimize the conformational energy for given values of θ and are, therefore, a well-defined function of θ .

The partitioning between θ and $\bar{\theta}$ provides a useful mechanism for taking account of the trade-offs between physical realism and computational cost. The latter is generally an increasing function of the size of set θ as this affects both the size of the decision space being searched (and therefore the number of candidate structures that need to be generated to ensure sufficient coverage of this space) and the cost of the local minimization of each candidate structure. At the two extremes, an empty set θ corresponds to a rigid molecule search, while an empty set $\bar{\theta}$ represents a fully atomistic computation.

In general, the decision variables that determine a crystal structure comprise the unit cell lattice lengths and angles, collectively denoted by X , the positions of the centers of mass and the orientation of the molecules within the unit cell, collectively denoted by β , and the independent CDFs, θ , of these molecules. If the search is carried out within a given space group (cf. section 2.3), the decision variables other than θ may be constrained by space group symmetry, for example, for structures in the monoclinic space group $P2_1/c$, only one lattice angle is an independent decision variable, the other two being fixed at 90° .

In any case, given the values of θ , we can compute the remaining (i.e., dependent) CDFs $\bar{\theta}$ (see section 2.2 below), at which point we have the complete molecular conformation(s). The latter, together with the values of the decision variables β , allow us to compute the positions of all atoms in the unit cell. These, together with the lattice parameters X , now fully determine all interatomic distances in the crystal, which allows us to compute the intermolecular contributions to the lattice energy, provided we have values of the atomic charges that are appropriate for the current molecular conformation(s). Finally, we can also compute the intramolecular energy contribution as

a function of θ , thereby determining the lattice energy corresponding to the given values of the decision variables.

2.2. Local Approximate Models for QM Calculations.

In principle, the functional dependence of the dependent CDFs $\bar{\theta}$ on the independent CDFs θ can be computed by performing an isolated-molecule constrained-geometry QM optimization. More specifically, the intramolecular energy contribution U^{intra} to the lattice energy is given by

$$U^{\text{intra}}(\theta) = \min_{\bar{\theta}} E^{\text{intra}}(\bar{\theta}, \theta) - E^{\text{vac}} \quad (1)$$

where $E^{\text{intra}}(\bar{\theta}, \theta)$ is the molecular energy for given values $\bar{\theta}, \theta$ of the CDFs, and E^{vac} is a constant representing the energy of the in vacuo conformation of the molecule, usually the globally minimal value of E^{intra} . The solution of the above optimization problem also provides the values of $\bar{\theta}$ and the electrostatic field from which atomic charges may be computed.

CrystalPredictor achieves the efficiency required in the context of global search while retaining the accuracy afforded by the above computations by making use of QM results computed prior to the global search. However, instead of the multidimensional cubic Hermite interpolants used by CrystalPredictor I, CrystalPredictor II makes use of Local Approximate Models (LAMs).⁵⁶ These are quadratic approximations of $U^{\text{intra}}(\theta)$ of the form:

$$U^{\text{intra}}(\theta) = U^{\text{intra}}(\theta_{\text{ref}}) + \mathbf{b}(\theta_{\text{ref}})^T(\theta - \theta_{\text{ref}}) + \frac{1}{2}(\theta - \theta_{\text{ref}})^T \mathbf{C}(\theta_{\text{ref}})(\theta - \theta_{\text{ref}}) \quad (2)$$

and linear approximations of $\bar{\theta}$ of the form:

$$\bar{\theta}(\theta) = \bar{\theta}_{\text{ref}} + \mathbf{A}(\theta_{\text{ref}})(\theta - \theta_{\text{ref}}) \quad (3)$$

where θ_{ref} is a set of reference values of θ , $\mathbf{b}(\theta_{\text{ref}})$ is a vector, and $\mathbf{C}(\theta_{\text{ref}})$ and $\mathbf{A}(\theta_{\text{ref}})$ are matrices computed from the second partial derivatives of $E^{\text{intra}}(\bar{\theta}, \theta)$ with respect to $\bar{\theta}$ and θ .

Equations 2 and 3 are derived by combining a second-order Taylor expansion of $E^{\text{intra}}(\bar{\theta}, \theta)$ at the solution of the minimization problem 1 at a point $\theta = \theta_{\text{ref}}$ with the first-order optimality conditions with respect to $\bar{\theta}$. The complete derivations as well as the expressions for $\mathbf{b}(\theta_{\text{ref}})$, $\mathbf{C}(\theta_{\text{ref}})$, and $\mathbf{A}(\theta_{\text{ref}})$ are given elsewhere.⁵⁶

In the limit of the dependent CDF set $\bar{\theta}$ being empty (i.e., a completely atomistic description of the molecule), the above reduces to the second-order Taylor series expansions used by van Eijck and co-workers.²³ However, the approach proposed here is generally applicable to any partitioning of the CDFs, including those in which only a relatively small subset θ is treated as independent, a crucial consideration in the context of efficient global search.

Expressions 2 and 3 match the results of the energy minimization 1 exactly at the reference values θ_{ref} and provide reasonably accurate approximations to $U^{\text{intra}}\mathbf{b}(\theta)$ and $\bar{\theta}(\theta)$ within a vicinity of θ_{ref} , typically defined as a hyper-rectangle of the form:

$$\theta_{\text{ref}} - \Delta\theta \leq \theta \leq \theta_{\text{ref}} + \Delta\theta \quad (4)$$

where $\Delta\theta$ is a given vector defining the range of validity of the LAM.

The intramolecular energy LAMs given by eq 2 retain their quadratic accuracy irrespective of the number of independent CDFs θ being considered. The restricted Hermite interpolants employed by CrystalPredictor I exhibit higher (cubic) accuracy

for the case of a single independent variable, but their accuracy deteriorates as the grid dimensionality increases. Thus, LAMs have a distinct advantage for systems with higher degrees of configurational flexibility, such as those of interest to this paper.

A sufficient number of LAMs is constructed to cover the entire conformational space of practical interest, usually defined as the set of values of θ for which $U^{\text{intra}}(\theta)$ (as computed by eq 1) is below a given threshold. In our work, the latter is typically taken to be +20–30 kJ mol⁻¹ relative to the global minimum in vacuo energy as it is generally unlikely for the reduction in intermolecular energy caused by the deformation of the molecules in the tightly packed crystalline environment to be able to compensate for intramolecular energy increases above this range.⁵⁷ In some cases, larger variations of the intramolecular energy have been reported⁵⁸ within the crystalline environment and the algorithm is implemented in such a way that the user may choose to explore a wider range of flexibility. The approximate size and shape of this space can be established by a conformational scan (i.e., varying each independent CDF θ around its value in the in vacuo conformation of the molecule while performing an isolated-molecule energy minimization with respect to all other CDFs at each θ -point considered). One may also be guided by information on the values of torsions in analogous molecules, such as those stored in the Cambridge Structural Database (CSD).⁵⁹ Ultimately, this exercise determines one or more domains of the form $[\theta_1^{\text{min}}, \theta_1^{\text{max}}] \times [\theta_2^{\text{min}}, \theta_2^{\text{max}}] \times [\theta_3^{\text{min}}, \theta_3^{\text{max}}] \times \dots$ that need to be considered during the global search. The approximate value of the LAM validity range $\Delta\theta$ may also be established by constructing LAMs at some of the points considered during the scan and comparing their predictions of intramolecular energy at neighboring points against the corresponding values computed by the QM calculations.

Once the relevant θ space(s) and the range of LAM validity are obtained in the manner described above, a sufficient number of LAMs can be constructed to cover the entire region of interest within the required accuracy. In the interest of simplicity, in the current implementation of CrystalPredictor II, we place the LAM reference points θ_{ref} on a regular grid, although this is not strictly required. The results of the QM calculations at each point θ_{ref} are also used for computing a corresponding set of atomic charges. Our current implementation employs a zeroth-order approximate model for atomic charges (i.e., the latter are considered to be constant within the range of validity of each LAM) but differ from one LAM to another. All QM computations are done using Gaussian09,^{60,60} with atomic charges being fitted using the CHELPG algorithm.⁶¹

We note that the construction of LAMs takes place as a preprocessing step prior to performing the global search. In contrast, the CrystalOptimizer algorithm⁵⁶ for structure refinement constructs LAMs “on-the-fly” as and when required, an efficient strategy which in most cases leads to relatively few LAMs actually being generated. However, by design, the global search in CrystalPredictor thoroughly explores the entire space of decision variables, which makes it necessary to compute the intramolecular energy over the entire space of conformational flexibility; therefore, it is better to generate the LAMs a priori in a systematic fashion.

Finally, an important benefit of using LAMs in the global search stage is that the LAM database generated can be reused in the refinement stage, provided that the same level of theory and basis set are used. It is not necessary to use the same set of

independent CDFs in both stages to permit reuse of the database.^{4,56}

2.3. Structure Generation Procedure and Local Lattice Energy Minimization. As mentioned in section 1.2, CrystalPredictor II employs a search procedure that systematically explores the space of decision variables within user-specified bounds. The search is carried out within a set of crystallographic space groups selected by the user, taking account of the symmetry constraints of each space group in order to improve efficiency. The algorithm can in principle account for all 230 space groups. In the current implementation, 63 space groups are handled; they are listed in Appendix A and correspond to those that appear with the highest frequency in the CSD. The user typically specifies the number of candidate structures to be generated and the way these should be allocated among the selected space groups (e.g., in equal numbers, or in proportion to the space group’s relative frequency in the CSD).

Any structure generated is subjected to a set of preliminary checks to determine whether it is physically realistic. This leads to the elimination of structures in which two or more molecules overlap in space, or where the lattice unit cell is unrealistic in terms of shape and density (as specified by user), or where intermolecular and/or intramolecular energies are excessively large.

Structures that pass the above tests are used as starting points of local minimizations of the lattice energy with respect to the decision variables subject to constraints on unit cell shape and volume. The efficient and accurate computation of the various contributions to the lattice energy and their partial derivatives with respect to the decision variables has been described elsewhere.^{24,25} The minimization is carried out using a sequential quadratic programming algorithm, implemented as either E04UFF⁶² or NLPQLP^{63,64} with default optimality/termination tolerances of 10⁻⁶.

Local minimizations starting from different candidate structures are independent of each other and can be run in parallel, a fact that is already exploited by the CrystalPredictor software architecture (cf. section 1.2). The achievable speedup factor is a near-linear function of the number of processors, and therefore, there is no need to make use of the parallelised computation capabilities of solvers such as NLPQLP.

The overall structure of the CrystalPredictor II code is summarized in Figure 2. As indicated, the structures obtained are subjected to a final clustering step in which duplicate structures that may have been generated from different starting points are removed, based on a comparison of energy, density, and root-mean-square deviation of the atomic distances in the 15-molecule coordination sphere.

3. CASE STUDIES

We now consider the application of CrystalPredictor II to three molecules (cf. Figure 1) with significant degrees of flexibility. The independent CDFs for each molecule are marked in Figure 3, with the corresponding ranges searched being reported in Table 1. For ROY and BMS-488043, the ranges shown are consistent with those used in previous studies.^{45,65}

In order to demonstrate the contribution of CrystalPredictor to the overall CSP methodology, we also report results for the subsequent structure refinement stage performed using the CrystalOptimizer⁵⁶ code for two of the three molecules considered, namely, β -D-glucose and BMS-488043. Structure refinement calculations have already been reported elsewhere⁴⁵

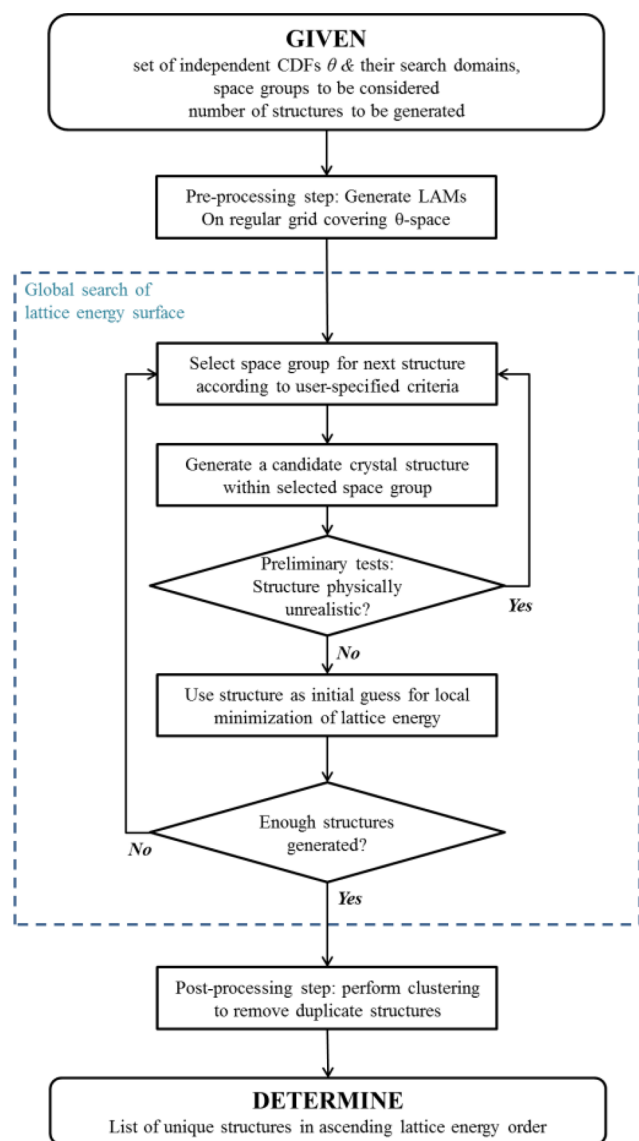
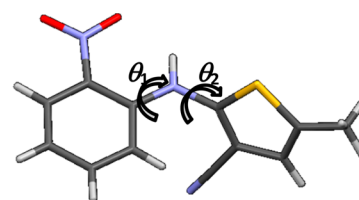


Figure 2. Flowchart of crystal structure generation in CrystalPredictor II.

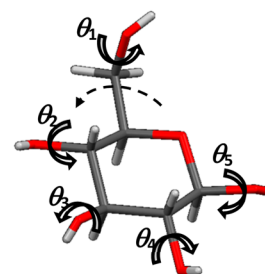
for ROY, albeit for a different set of starting structures, and are not repeated here.

3.1. ROY. The recent study⁴⁵ on ROY using CrystalPredictor I modeled the flexibility of torsions θ_1 and θ_2 , indicated by the solid arrows in Figure 3a, using an elaborate system of rigid fragments aiming to maintain geometrical relationships between nonconnected parts of the molecule. In particular, the hydrogen atom in the secondary amine group had to be included within a single rigid fragment together with the (2-nitrophenyl)amino group (cf. left part of Figure 1a). This was necessary in order to take account of the intramolecular hydrogen bond between the amine hydrogen and one of the oxygen atoms of the nitro group. Albeit effective for this specific example, the need for such system-dependent judgment is not a desirable attribute of what aims to be a widely applicable CSP procedure. As we shall demonstrate in this section, the use of a more natural representation of molecular conformation in CrystalPredictor II allows us to address this issue.

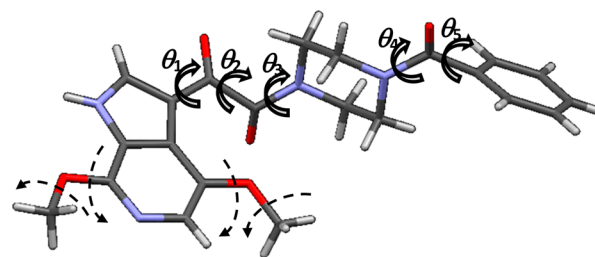
The original study made use of a uniform interpolation grid of 90 points over the independent torsion angle space of



(a) 5-methyl-2-[(2-nitrophenyl)amino]-3-thiophenecarbonitrile (ROY)



(b) β -D-glucose



(c) BMS-488043

Figure 3. Independent CDFs (torsional angles) considered by CrystalPredictor II during global search (block arrows). For β -D-glucose and BMS-488043, broken thin arrows indicate additional torsions treated as independent CDFs during the subsequent structure refinement stage. As reported elsewhere, structure refinement for ROY considered 22 independent CDFs, including 10 bond angles and 12 torsion angles.

$[100^\circ, 260^\circ] \times [0^\circ, 170^\circ]$. Each grid point required a QM calculation performed using Gaussian09^{60,60} at the B3LYP level of theory with a 6-31G(d,p) basis set. In the present study, the same set of flexible torsions was investigated to make it easier to compare the performance of the two CrystalPredictor algorithms. Fifty six LAMs were generated at intervals of 20° (i.e., $\Delta\theta = 10^\circ$), using the same level of theory and basis set to maintain consistency.

Both studies made use of the FIT exp-6 repulsion/dispersion potential,^{33,34,66} expanded to include sulfur,⁶⁷ and generated 400 000 crystal structures. The numbers of unique structures identified within 15 kJ mol^{-1} of the global minimum were 1584 for CrystalPredictor I and 3375 for CrystalPredictor II. The 1097 structures identified by the present study within 10 kJ mol^{-1} of the global minimum are shown in Figure 4.

In both cases, the generated structures include all seven polymorphs of ROY for which crystal structures have been characterized experimentally. However, there were significant differences in the ranking of these experimental structures, as illustrated in Table 2, which lists the seven structures in order of decreasing experimentally determined stability.^{68–70} Overall, the rankings of the computed structures providing a good match of the experimental polymorphs determined by CrystalPredictor II are much improved. In particular, the

Table 1. Domains of Independent CDFs (See Figure 3) Used for Global Search and Numbers of LAMs Used To Cover Them^a

molecule	$\theta_1 / ^\circ$	$\theta_2 / ^\circ$	$\theta_3 / ^\circ$	$\theta_4 / ^\circ$	$\theta_5 / ^\circ$	LAM validity range, $\Delta\theta$	total LAMs generated
ROY	[110, 250]	[0, 170]	-	-	-	$\pm 10^\circ$	56
β -D-glucose	[0, 360]	[0, 360]	[0, 360]	[0, 360]	[0, 360]	$\pm 30^\circ$	7776
BMS-488043						$\pm 10^\circ$	180
region 1	[0, 40]	[-140, -100]	[160, 200]	[150, 210]	[0, 180]		27
region 2	[0, 40]	[-140, -100]	[160, 200]	[-70, 70]	[0, 180]		63
region 3	[0, 40]	[-140, -100]	[-20, 20]	[150, 210]	[0, 180]		27
region 4	[0, 40]	[-140, -100]	[-20, 20]	[-80, 80]	[0, 180]		63

^aAs discussed in Section 3.3, BMS-488043 required four searches in four separate conformational regions. An explanation of the choices of independent CDFs and their ranges can be found in ref 45 for ROY and in the Supplementary Information for BMS-488043.

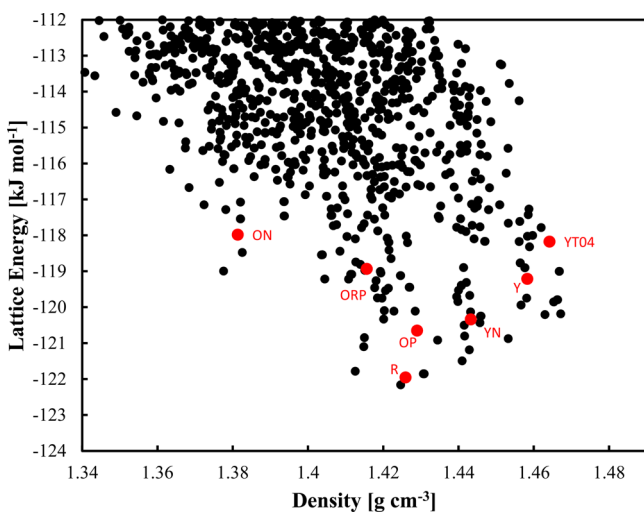


Figure 4. Crystal structures (1097) generated for ROY within 10 kJ mol⁻¹ of the global minimum by CrystalPredictor II. Each lattice energy/density point represents a unique structure. Structures corresponding to all seven experimental polymorphs have been identified and are marked in red.

worst-ranked experimental polymorph is at place 930 (R) with CrystalPredictor I but at place 120 (ON) using CrystalPredictor II. As can be seen in Table 2, the relative rankings obtained with CrystalPredictor II for the experimentally determined polymorphs are in much better agreement with those of CrystalOptimizer than the relative rankings derived from CrystalPredictor I. This indicates that the LAM-based lattice energies are more accurate than those computed with the restricted Hermite interpolants. Moreover, the similarity of the generated structures to the corresponding experimental

structures, measured through the root-mean-square difference in atomic positions between a cluster of 15 molecules from each structure (RMSD₁₅) is similar to CrystalPredictor I results or much superior. This improvement is achieved with the same number of grid points in CrystalPredictor II as in CrystalPredictor I. This indicates that a significant improvement in the accuracy of the energy function has been achieved by introducing LAMs and incorporating a description of the effect of conformational flexibility on dependent CDFs and charges.

Irrespective of which version of CrystalPredictor is used for the global search, the low-ranking structures identified by the latter need to be refined further using CrystalOptimizer,⁵⁶ which employs a more accurate description of electrostatic interactions in terms of distributed multipole expansions and also takes account of a much wider range of independent CDFs; a total of 22 independent CDFs, including 10 bond angles and 12 torsion angles, were considered in the previous study of this molecule. It is worth noting that, in view of the rankings reported in Table 2, using CrystalPredictor II for the global search would lead to much fewer structures having to be refined in order to ensure identification of all experimentally known polymorphs. Given the high computational cost of the refinement stage, this has a significant impact on the reliability and the overall cost of CSP. The improved reliability of the crystal structure landscape is a major benefit of the revision of CrystalPredictor. A further benefit arises from the fact that the LAMs constructed as part of a CrystalPredictor run can be reused at the refinement stage, making the overall process more efficient.

3.2. β -D-Glucose. Because of their relatively low effect on conformational energy, all five of the hydroxylic hydrogens in β -D-glucose could potentially take on any orientation within the crystalline environment. Therefore, the corresponding torsion angles have to be treated as independent CDFs with search

Table 2. Lattice Energies, Rankings, and Deviations (RMSD₁₅) from the Experimental Structures for the Polymorphs of ROY in the Lists of Crystal Structures Generated by CrystalPredictor I in a Previous Study,⁴⁵ Using CrystalPredictor II in This Study^a

polymorph	CrystalPredictor I			CrystalPredictor II			CrystalOptimizer	
	rank	E^{latt} (kJ mol ⁻¹)	RMSD ₁₅ (Å)	rank	E^{latt} (kJ mol ⁻¹)	RMSD ₁₅ (Å)	rank	E^{latt} (kJ mol ⁻¹)
Y	20	-126.7	0.19	57	-119.2	0.208	17	-119.7
YT04	5	-129.4	0.24	103	-118.2	0.184	88	-117.3
YN	6	-129.1	0.36	16	-120.3	0.241	9	-120.3
R	930	-116.3	[7/15]	2	-122.0	0.387	7	-120.6
OP	373	-119.8	[6/15]	13	-120.7	0.271	30	-119.1
ORP	263	-120.7	0.58	75	-118.9	0.210	66	-117.7
ON	31	-126.0	0.73	120	-118.0	0.369	73	-117.6

^aLattice energies and rankings from generated by CrystalOptimizer in a previous study⁴⁵ are also shown. Polymorphs are listed in decreasing order of experimentally determined thermodynamic stability.^{68–70} Where 15 molecules could not be overlaid between the generated structure and the corresponding experimental structure, the number of molecules that could be overlaid is listed in square brackets instead of an RMSD₁₅.

ranges set at $[0, 360^\circ]$. In contrast, conformational scans indicate that the torsion angle shown by the broken thin arrow in Figure 3b is much less flexible, with relatively small deviations from its in vacuo value of 60° leading to significant increases in conformational energy. It is therefore treated as a dependent CDF for the purposes of the global search. There exists a second low-energy minimum at 285° (see S4.1 in the Supporting Information) that would need to be sampled for a comprehensive search, but in this work we treat only the global minimum region in order to illustrate the method.

Handling the flexibility of the five independent CDFs using CrystalPredictor I would have been problematic due to the need for relatively small intervals between grid points arising from the relatively large number of CDFs (e.g., with 30° intervals, nearly 250 000 QM calculations would be required). In CrystalPredictor II, LAMs were generated at intervals of 60° at the HF-SCF level of theory with a 6-31G (d,p) basis set in Gaussian09. The result was a large but computationally manageable set of $7776 = (360/60)^5$ LAMs.

The global search generated 500 000 points, yielding only 14 unique structures within 20 kJ mol^{-1} of the global minimum (see Figure 5a). This is due to the presence of an isolated global minimum, at a comparatively low energy value. Given this unusual landscape, structures within 40 kJ mol^{-1} of the global minimum were considered in order to include over 1000 structures in the refinement step. The FIT exp-6 repulsion/dispersion potential was used.^{33,34} A structure closely matching the experimentally known crystal structure (CSD refcode:

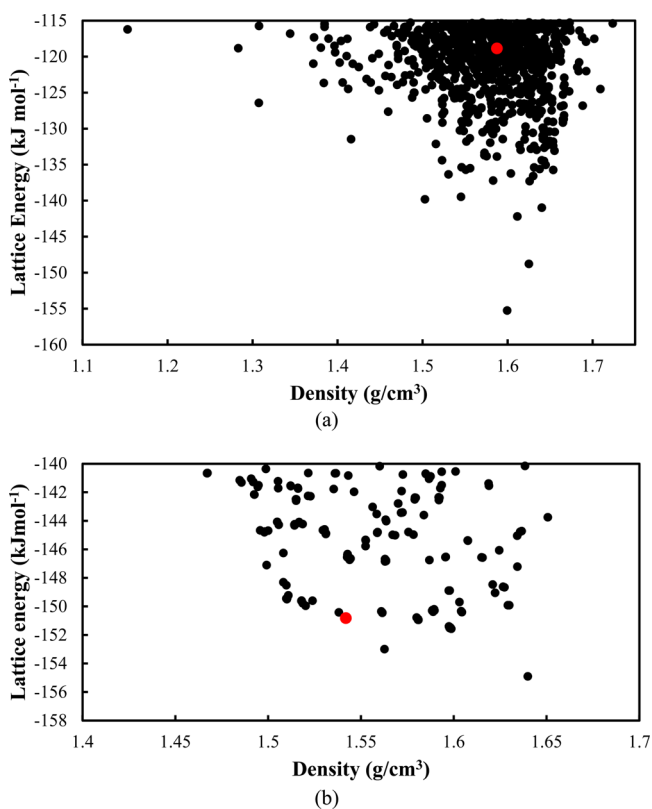


Figure 5. (a) Crystal structures (1160) generated for β -D-glucose by CrystalPredictor II within 40 kJ mol^{-1} and (b) the most stable structures after refinement by CrystalOptimizer and inclusion of PCM. Each lattice energy/density point represents a structure. The structure corresponding to the experimental polymorph is marked in red.

GLUCSE02) was found ranked 489th at 36.4 kJ mol^{-1} above the global minimum with $\text{RMSD}_{15} = 0.206 \text{ \AA}$. Subsequent refinement and reranking was carried out using CrystalOptimizer. The torsion angle shown by the broken thin arrow in Figure 3b was now also treated as an independent CDF. The intramolecular energy was calculated at the HF-SCF level of theory using a 6-31G (d,p) basis set, while the intermolecular energy was calculated using electrostatic multipoles derived using GDMA2⁷¹ from wave functions calculated at the PBE0 level of theory using a 6-31G (d,p) basis set and the FIT exp-6 repulsion/dispersion potential. After this refinement, the experimentally known structure was ranked 137th.

The adjacent alcohol groups in β -D-glucose lead to strong intramolecular OH...O interactions, although the molecular geometry does not permit the formation of full intramolecular hydrogen bonds. On the other hand, intermolecular hydrogen bonding does take place and this, in turn, may result in significant electrostatic induction effects between neighboring molecules. The consequent changes in electron density are not captured by isolated-molecule QM calculations and this may have important consequences on the relative ranking of putative crystal structures.⁷² To compensate for this polarizability effect, the lattice energies of the final crystal structures were recalculated using electrostatic multipoles and intramolecular energies calculated in a polarizable continuum model (PCM)⁷³ with dielectric constant of $\epsilon = 3$, which has been shown⁷⁴ to approximate the polarizing effect of a crystalline environment.

The computational cost of the entire CSP exercise is shown in the corresponding columns of Table 3. The structure refinement cost includes the final reranking calculations mentioned above.

The final, refined set of crystal structures is shown in Figure 5b and contains a match to the experimental crystal structure with $\text{RMSD}_{15} = 0.159 \text{ \AA}$ in sixth place, 4.1 kJ mol^{-1} above the global minimum. The details of the predicted structure and a comparison with experimental measurements are shown in Table 4 below.

3.3. BMS-488043. There are two experimentally known polymorphs of BMS-488043.⁶⁵ Form I is recorded in the CSD as SUTTOR.⁷⁵

Five torsion angles (Figure 3c) were determined to be sufficiently important in terms of their effect on the value of the intramolecular energy to warrant variation during the global search. The conformational energy scan on angles θ_3 and θ_4 found that each one of them exhibits two disjoint low-energy ranges of potential interest to CSP. In such cases, CrystalPredictor can be used to generate structures across the full set of values of both angles, simply allowing structures in the intermediate, high-energy ranges to be discarded because of violations of the preliminary screening rules described in section 2.3. However, in this case, we chose to perform four separate searches, each corresponding to a different combination of the low energy ranges of θ_3 and θ_4 , as shown in the last four rows of Table 1.

Overall, the relatively small range of rotation in these five torsions allows us to model the flexibility in all four conformational regions using a total of only 180 LAMs. These were calculated at the HF-SCF level of theory with a 6-31G (d,p) basis set in Gaussian09. The global search generated 500 000 points in each conformational region, giving a total of two million initial crystal structures. The FIT exp-6 repulsion/dispersion potential was used.

Overall, 871 unique structures were identified within 20 kJ mol^{-1} of the global minimum and are shown in Figure 6a. A structure corresponding to form I was found ranked sixth, and a structure corresponding to form II was found ranked 767th.

Table 3. Approximate CPU Hours (Intel Xeon Westmere Processor, Running at 2.67 GHz) Required for Different Stages of the CSP Process for Each of the Molecules Studied^a

time (CPU hours)	ROY CrystalPredictor I	ROY CrystalPredictor II	β -D-glucose	BMS-488043
QM calculations	2000	2170	6900	4000
global search	800	1220	400	6800
structure refinement (CrystalOptimizer)	43 000 ^b	–	2450	64 000
total for CSP	2800 (excluding structure refinement)	3390 (excluding structure refinement)	9750	74 800

^aStructure refinement for ROY was carried out elsewhere⁴⁵ and was not repeated here. ^bIncludes single-point MP2 calculations for final structures.

Table 4. Predicted Crystal Structure for β -D-Glucose Corresponding to Experimentally Known Structure (CSD Refcode: GLUCSE02)

energetic information		structural information									
E^{latt} (kJ mol ⁻¹)	rank	rms ₁ (Å)	rms ₁₅ (Å)		ρ (g cm ⁻³)	<i>a</i> (Å)	<i>b</i> (Å)	<i>c</i> (Å)	α (°)	β (°)	γ (°)
-150.81	6	0.07	0.154	exptl	1.58	9.01	12.72	6.60	90	90	90
				pred.	1.54	9.33	12.59	6.61	90	90	90

The two generated structures match the experimental structures with RMSD₁₅ = 0.230 and 0.539 Å, respectively.

Subsequent refinement and reranking of all 871 structures was carried out using CrystalOptimizer. The torsion angles indicated by the broken thin arrows in Figure 3c were also treated as independent CDFs at this stage. The intramolecular energy was

calculated at the PBE0 level of theory using a 6-31G (d,p) basis set, while the intermolecular energy was calculated using electrostatic multipoles derived using GDMA2⁷¹ from wave functions calculated at the PBE0 level of theory using a 6-31G (d,p) basis set, and the FIT exp-6 repulsion/dispersion potential.^{33,34} In the final set of crystal structures (see Figure 6b), the structure corresponding to form I (with an RMSD₁₅ of 0.249 Å) is ranked fifth, while that corresponding to form II (with an RMSD₁₅ of 0.278 Å) is ranked 51st. A comparison of predicted and experimental structures for the two forms is shown in Table 5.

By the standards of contemporary crystal structure prediction, BMS-488043 is a relatively large molecule. Successful generation and refinement of structures for a molecule of this size within a fairly standardized workflow and at relatively modest computational cost (see relevant columns in Table 3) represents a notable step forward for crystal structure prediction as a field.

4. CONCLUDING REMARKS

The CrystalPredictor II algorithm and code described in this paper allow a significant degree of molecular flexibility to be taken into account during the global search stage of CSP methodologies. Like its predecessor, the CrystalPredictor I algorithm,^{23,25} it is generally applicable to systems with one or more chemical entities in the asymmetric unit, forming crystal structures belonging to any space group. It derives its effectiveness in identifying low-energy structures from a systematic and extensive search of the lattice energy surface implemented on distributed computing architectures.

By allowing the user to categorize CDFs into “independent” and “dependent” ones, the new algorithm introduces a natural description of molecular flexibility that can easily be tailored to achieve the best trade-off between physical realism and computational tractability without the need for artificial constructs such as “rigid fragments” which has sometimes proven problematic in the past. Moreover, the use of LAMs provides an efficient and accurate description of the variation of intramolecular energy, dependent CDFs and atomic charges during the global search.

As illustrated by the ROY example (cf. section 3.1), the improved accuracy of CrystalPredictor II may lead to an improved ranking of the candidate structures over that achieved by its predecessor. In practical terms, this can significantly reduce the number of structures that need to undergo further refinement in order to ensure identification of all polymorphs

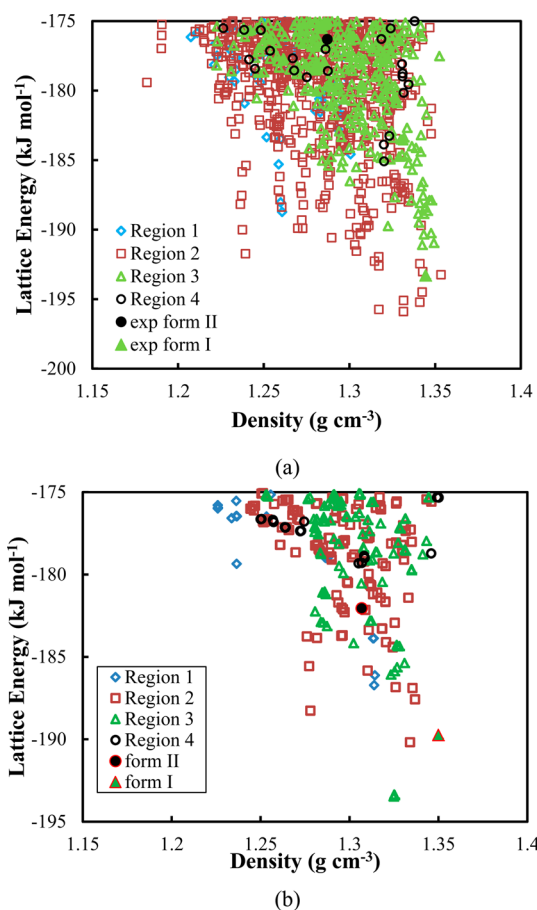


Figure 6. (a) Crystal structures (871) generated for BMS-488043 by CrystalPredictor II within 20 kJ mol⁻¹ of the global minimum and (b) the most stable structures after refinement by CrystalOptimizer. Each lattice energy/density point represents a structure. Those corresponding to experimental polymorphs are marked with solid points. See Table 1 for the definition of the conformational regions.

Table 5. Comparisons of Predicted and Experimental Crystal Structures for the Two Known Forms of Molecule BMS-488043

form	energetic information			structural information								
	E^{latt} (kJ mol ⁻¹)	rank	rms1 (Å)	rms15 (Å)		ρ (g cm ⁻³)	a (Å)	b (Å)	c (Å)	α (°)	β (°)	γ (°)
I	-189.7	5	0.10	0.249	exptl	1.39	20.08	7.27	14.52	90	107.1	90
					pred.	1.35	21.30	7.17	14.76	90	111.0	90
II	-182.1	51	0.09	0.278	exptl	1.38	13.78	11.30	13.22	90	97.9	90
					pred.	1.31	13.28	11.34	14.42	90	81.6	90

of practical interest. Conversely, it can increase the probability of these polymorphs being identified following the refinement of a given number of structures.

The developments reported in this paper render CrystalPredictor II more consistent with its companion CrystalOptimizer⁵⁶ algorithm for structure refinement, as both algorithms now employ the independent/dependent CDF categorization and the LAMs based on it. The higher accuracy required at the structure refinement stage in the computation of lattice energy is achieved by CrystalOptimizer in three different ways, namely by

- using distributed multipole expansions instead of atomic charges for the description of the intermolecular electrostatic interactions;
- treating more CDFs (e.g., more torsion angles and some bond angles) as independent;
- potentially employing a more accurate level of theory/basis set for the QM calculations.

The theoretical consistency between the two algorithms has an important practical implication: if the same level of theory/basis set is used for both CrystalPredictor II and CrystalOptimizer, then any LAMs generated for the former may be stored and reused by the latter, thereby reducing the number of QM calculations that need to be performed during the structure refinement. This is a mathematical property of LAMs that holds^{56,65} even if some of the CDFs treated as dependent by CrystalPredictor II are treated as independent ones by CrystalOptimizer.

The introduction of LAMs in the global search stage of Crystal Structure Prediction is an important step forward in the application of CSP techniques to larger and/or highly flexible molecules because it allows the generation of an improved crystal energy landscape within reasonable computational cost. This increases the reliability of the entire process since a more realistic set of low-energy structures is identified at the end of the global search. It can also reduce significantly the overall computational cost of CSP studies as fewer putative structures need to be refined and LAM databases from the global search can be reused in the refinement step.

APPENDIX A: SPACE GROUPS CURRENTLY HANDLED BY CRYSTALPREDICTOR II

$P1$	$\bar{P}1$	$P2_1$	$P2_1/c$	$P2_12_12$	$P2_12_12_1$	$Pna2_1$	$Pca2_1$	$Pbca$	$Pbcn$
$C2/c$	Cc	$C2$	Pc	Cm	$P2_1/m$	$C2/m$	$P2/c$	$C222_1$	$Pmn2_1$
$Cmc2_1$	$Aba2$	$Fdd2$	$Iba2$	$Pnna$	$Pccn$	$Pbcm$	$Pnmm$	$Pmmm$	$Pnma$
$Cmcm$	$Cmca$	$Fddd$	$Ibam$	$P4_1$	$P4_3$	$I\bar{4}$	$P4/n$	$P4_2/n$	$I4/m$
$I4_1/a$	$P4_12_12$	$P4_32_12$	$P\bar{4}2_1c$	$I\bar{4}2d$	$P3_1$	$P3_2$	$R3$	$P\bar{3}$	$R\bar{3}$
$P3_12_1$	$P3_22_1$	$R3c$	$R\bar{3}c$	$P6_1$	$P6_3$	$P6_3/m$	$P2_13$	$Pa\bar{3}$	$P222_1$
$Pba2$	$P2_1/a$	$P2_1/n$							

ASSOCIATED CONTENT

Supporting Information

Section S1 provides a discussion of the choice of the conformational search space for BMS-488043. Section S2 provides information on the reuse of LAM databases in refinement. Section S3 provides a breakdown of the space groups sampled for each molecule. This material is available free of charge via the Internet at <http://pubs.acs.org/>

AUTHOR INFORMATION

Corresponding Author

*E-mail: c.pantelides@imperial.ac.uk

Present Address

†Visiting Researcher, Department of Chemistry, University College London, 20 Gordon Street, London, WC1H 0AJ

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors gratefully acknowledge funding from the United Kingdom's Engineering and Physical Sciences Research Council (EPSRC) (EP/E016340, EP/J003840/1, EP/J014958/1) and access to computational resources and support from the High Performance Computing Cluster at Imperial College London. We are grateful to Professor S.L. Price for supplying the DMACRYS code for use within CrystalOptimizer. We are also grateful to Dr. San Kiang and Dr. Alexander Marchut of Bristol Myers Squibb for provision of data on BMS-488043 and to Dr. Panagiotis Karamertzanis for useful discussions on the application of CSP to this molecule.

REFERENCES

- Datta, S.; Grant, D. J. W. Crystal Structures of Drugs: Advances in Determination, Prediction and Engineering. *Nat. Rev. Drug Discovery* **2004**, *3*, 42–57.
- Price, S. L. Computed Crystal Energy Landscapes for Understanding and Predicting Organic Crystal Structures and Polymorphism. *Acc. Chem. Res.* **2009**, *42*, 117–126.
- Day, G. M. Current Approaches to Predicting Molecular Organic Crystal Structures. *Crystallogr. Rev.* **2011**, *17*, 3–52.

- (4) Pantelides, C. C.; Adjiman, C. S.; Kazantsev, A. V. General computational algorithms for ab initio crystal structure prediction for organic molecules. In *Prediction and Calculation of Crystal Structures*, 1st ed.; Atahan-Evrenk, S., Aspuru-Guzik, A., Eds.; Springer International Publishing: Switzerland, 2014; Vol. 345, pp 25–58.
- (5) Day, G. M.; Cooper, T. G.; Cruz-Cabeza, A. J.; Heyczyk, K. E.; Ammon, H. L.; Boerrigter, S. X. M.; Tan, J. S.; 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. A.; 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. K.; Schweizer, B. Significant Progress in Predicting the Crystal Structures of Small Organic Molecules - a Report on the Fourth Blind Test. *Acta Crystallogr., Sect. B: Struct. Sci.* **2009**, *65*, 107–125.
- (6) 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.; 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. A. Towards Crystal Structure Prediction of Complex Organic Compounds - a Report on the Fifth Blind Test. *Acta Crystallogr., Sect. B: Struct. Sci.* **2011**, *67*, 535–551.
- (7) Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C.; Price, S. L.; Galek, P. T. A.; Day, G. M.; Cruz-Cabeza, A. J. Successful Prediction of a Model Pharmaceutical in the Fifth Blind Test of Crystal Structure Prediction. *Int. J. Pharm.* **2011**, *418*, 168–178.
- (8) Ismail, S. Z.; Anderton, C. L.; Copley, R. C. B.; Price, L. S.; Price, S. L. Evaluating a Crystal Energy Landscape in the Context of Industrial Polymorph Screening. *Cryst. Growth Des.* **2013**, *13*, 2396–2406.
- (9) Bhardwaj, R. M.; Price, L. S.; Price, S. L.; Reutzel-Edens, S. M.; Miller, G. J.; Oswald, I. D. H.; Johnston, B. F.; Florence, A. J. Exploring the Experimental and Computed Crystal Energy Landscape of Olanzapine. *Cryst. Growth Des.* **2013**, *13*, 1602–1617.
- (10) Price, L. S.; McMahon, J. A.; Lingireddy, S. R.; Lau, S.-F.; Diserod, B. A.; Price, S. L.; Reutzel-Edens, S. M. A Molecular Picture of the Problems in Ensuring Structural Purity of Tazofelone. *J. Mol. Struct.* **2014**, *1078*, 26–42.
- (11) Braun, D. E.; Karamertzanis, P. G.; Price, S. L. Which, If Any, Hydrates Will Crystallize? Predicting Hydrate Formation of Two Dihydroxybenzoic Acids. *Chem. Commun.* **2011**, *47*, 5443–5445.
- (12) Cruz-Cabeza, A. J.; Day, G. M.; Jones, W. Towards Prediction of Stoichiometry in Crystalline Multicomponent Complexes. *Chem.—Eur. J.* **2008**, *14*, 8830–8836.
- (13) Karamertzanis, P. G.; Kazantsev, A. V.; Issa, N.; Welch, G. W. A.; Adjiman, C. S.; Pantelides, C. C.; Price, S. L. Can the Formation of Pharmaceutical Cocrystals Be Computationally Predicted? 2. Crystal Structure Prediction. *J. Chem. Theory Comput.* **2009**, *5*, 1432–1448.
- (14) Habgood, M.; Deij, M. A.; Mazurek, J.; Price, S. L.; ter Horst, J. H. Carbamazepine Co-Crystallization With Pyridine Carboxamides: Rationalization by Complementary Phase Diagrams and Crystal Energy Landscapes. *Cryst. Growth Des.* **2010**, *10*, 903–912.
- (15) Habgood, M.; Price, S. L. Isomers, Conformers and Cocrystal Stoichiometry: Insights From the Crystal Energy Landscapes of Caffeine With the Hydroxybenzoic Acids. *Cryst. Growth Des.* **2010**, *10*, 3263–3272.
- (16) Mohamed, S.; Tocher, D. A.; Price, S. L. Computational Prediction of Salt and Cocrystal Structures - Does a Proton Position Matter? *Int. J. Pharm.* **2011**, *418*, 187–198.
- (17) van de Streek, J.; Neumann, M. A.; Perrin, M.-A. Validation of Dispersion-Corrected Density Functional Theory Calculations for the Crystal Structure Prediction of Molecular Salts: a Crystal Structure Prediction Study of Pyridinium Chloride. *CrystEngComm* **2010**, *12*, 3827–3833.
- (18) Bernstein, J. *Polymorphism in Molecular Crystals*; Oxford University Press: United Kingdom, 2007.
- (19) Neumann, M. A. Tailor-Made Force Fields for Crystal Structure Prediction. *J. Phys. Chem. B* **2008**, *112*, 9810–9829.
- (20) van Eijck, B. P.; Kroon, J. UPACK Program Package for Crystal Structure Prediction: Force Fields and Crystal Structure Generation for Small Carbohydrate Molecules. *J. Comput. Chem.* **1999**, *20*, 799–812.
- (21) Mooij, W. T. M.; van Eijck, B. P.; Kroon, J. Ab Initio Crystal Structure Predictions for Flexible Hydrogen Bonded Molecules. *J. Am. Chem. Soc.* **2000**, *122*, 3500–3505.
- (22) van Eijck, B. P.; Kroon, J. Structure Predictions Allowing More Than One Molecule in the Asymmetric Unit. *Acta Crystallogr., Sect. B: Struct. Sci.* **2000**, *56*, 535–542.
- (23) van Eijck, B. P.; Mooij, W. T. M.; Kroon, J. Ab Initio Crystal Structure Predictions for Flexible Hydrogen-Bonded Molecule. Part II. Accurate Energy Minimization. *J. Comput. Chem.* **2001**, *22*, 805–815.
- (24) Karamertzanis, P. G.; Pantelides, C. C. Ab Initio Crystal Structure Prediction - I. Rigid Molecules. *J. Comput. Chem.* **2005**, *26*, 304–324.
- (25) Karamertzanis, P. G.; Pantelides, C. C. Ab Initio Crystal Structure Prediction. II. Flexible Molecules. *Mol. Phys.* **2007**, *105*, 273–291.
- (26) Zhu, Q.; Oganov, A. R.; Glass, C. W.; Stokes, H. T. Constrained Evolutionary Algorithm for Structure Prediction of Molecular Crystals: Methodology and Applications. *Acta Crystallogr., Sect. B: Struct. Sci.* **2012**, *68*, 215–226.
- (27) Holden, J. R.; Du, Z. Y.; Ammon, H. L. Prediction of Possible Crystal Structures for C-Containing, H-Containing, N-Containing, O-Containing and F-Containing Organic Compounds. *J. Comput. Chem.* **1993**, *14*, 422–437.
- (28) Gavezzotti, A. Generation of Possible Crystal Structures From the Molecular Structure for Low-Polarity Organic Compounds. *J. Am. Chem. Soc.* **1991**, *113*, 4622–4629.
- (29) Thakur, T. S.; Desiraju, G. R. Crystal Structure Prediction of a Co-Crystal Using a Supramolecular Synthron Approach: 2-Methylbenzoic Acid-2-Aminomethylpyrimidine. *Cryst. Growth Des.* **2008**, *8*, 4031–4044.
- (30) Raiteri, P.; Martonak, R.; Parrinello, M. Exploring Polymorphism: the Case of Benzene. *Angew. Chem., Int. Ed.* **2005**, *44*, 3769–3773.
- (31) Sobol, I. M. The Distribution of Points in a Cube and the Approximate Evaluation of Integrals. *USSR Comput. Math. Math. Phys.* **1967**, *7*, 86–112.
- (32) Williams, D. E.; Starr, T. L. Calculation of Crystal Structures of Hydrocarbons by Molecular Packing Analysis. *Comput. Chem.* **1977**, *1*, 173–177.
- (33) Cox, S. R.; Hsu, L. Y.; Williams, D. E. Nonbonded Potential Function Models for Crystalline Oxohydrocarbons. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1981**, *37*, 293–301.
- (34) Williams, D. E.; Cox, S. R. Nonbonded Potentials for Azahydrocarbons: the Importance of the Coulombic Interaction. *Acta Crystallogr., Sect. B: Struct. Sci.* **1984**, *40*, 404–417.
- (35) Williams, D. E. Improved Intermolecular Force Field for Molecules Containing H, C, N, and O Atoms, With Application to Nucleoside and Peptide Crystals. *J. Comput. Chem.* **2001**, *22*, 1154–1166.
- (36) Karamertzanis, P. G.; Pantelides, C. C. Optimal Site Charge Models for Molecular Electrostatic Potentials. *Mol. Simul.* **2004**, *30*, 413–436.
- (37) 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. *Phys. Chem. Chem. Phys.* **2013**, *15*, 8069–8080.

- (38) 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. *Chem.—Eur. J.* **2013**, *19*, 7883–7888.
- (39) Eddleston, M. D.; Hejczyk, K. E.; Bithell, E. G.; Day, G. M.; Jones, W. Polymorph Identification and Crystal Structure Determination by a Combined Crystal Structure Prediction and Transmission Electron Microscopy Approach. *Chem.—Eur. J.* **2013**, *19*, 7874–7882.
- (40) Habgood, M. Solution and Nanoscale Structure Selection: Implications for the Crystal Energy Landscape of Tetrolic Acid. *Phys. Chem. Chem. Phys.* **2012**, *14*, 9195–9203.
- (41) Habgood, M.; Lancaster, R. W.; Gateshki, M.; Kenwright, A. M. The Amorphous Form of Salicylsalicylic Acid: Experimental Characterization and Computational Predictability. *Cryst. Growth Des.* **2013**, *13*, 1771–1779.
- (42) 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*, 6441–6446.
- (43) van Mourik, T.; Karamertzanis, P. G.; Price, S. L. Molecular Conformations and Relative Stabilities Can Be As Demanding of the Electronic Structure Method As Intermolecular Calculations. *J. Phys. Chem. A* **2006**, *110*, 8–12.
- (44) Jones, J. T. A.; Hasell, T.; Wu, X.; Basca, J.; Jelfs, K. E.; Schmidtmann, M.; Chong, S. Y.; Adams, D. J.; Trewin, A.; Schiffman, F.; Cora, F.; Slater, B.; Steiner, A.; Day, G. M.; Cooper, A. I. Modular and Predictable Assembly of Porous Organic Molecular Crystals. *Nature* **2011**, *474*, 367–371.
- (45) 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 Crystallogr., Sect. B: Struct. Sci.* **2012**, *68*, 677–685.
- (46) Frišćić, T.; Lancaster, R. W.; Fábíán, L.; Karamertzanis, P. G. Tunable Recognition of the Steroid α -Face by Adjacent π -Electron Density. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 13216–13221.
- (47) 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*, 2454–2464.
- (48) 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*, 1853–1859.
- (49) 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. *Cryst. Growth Des.* **2011**, *11*, 5659–5669.
- (50) Habgood, M. Analysis of Enantiospecific and Diastereomeric Cocrystal Systems by Crystal Structure Prediction. *Cryst. Growth Des.* **2013**, *13*, 4549–4558.
- (51) Gorbitz, C. H.; Dalhus, B.; Day, G. M. Pseudoracemic Amino Acid Complexes: Blind Predictions for Flexible Two Component Crystals. *Phys. Chem. Chem. Phys.* **2010**, *12*, 8466–8477.
- (52) Braun, D. E.; Bhardwaj, R. M.; Florence, A. J.; Tocher, D. A.; Price, S. L. Complex Polymorphic System of Gallic Acid - Five Monohydrates, Three Anhydrides, and Over 20 Solvates. *Cryst. Growth Des.* **2013**, *13*, 4071–4083.
- (53) 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 Crystallogr., Sect. B: Struct. Sci.* **2005**, *61*, 511–527.
- (54) van Eijck, B. P.; Mooij, W. T. M.; Kroon, J. Attempted Prediction of the Crystal Structures of Six Monosaccharides. *Acta Crystallogr., Sect. B: Struct. Sci.* **1995**, *51*, 99–103.
- (55) van Eijck, B. P.; Mooij, W. T. M.; Kroon, J. Crystal Structure Prediction for Six Monosaccharides Revisited. *J. Phys. Chem. B* **2001**, *105*, 10573–10578.
- (56) Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C. Efficient Handling of Molecular Flexibility in Lattice Energy Minimization of Organic Crystals. *J. Chem. Theory Comput.* **2011**, *7*, 1998–2016.
- (57) Thompson, H. P. G.; Day, G. M. Which Conformations Make Stable Crystal Structures? Mapping Crystalline Molecular Geometries to the Conformational Energy Landscape. *Chem. Sci.* **2014**, *5*, 3173–3182.
- (58) Cruz-Cabeza, A. J.; Bernstein, J. Conformational Polymorphism. *Chem. Rev.* **2013**, *114*, 2170–2191.
- (59) Allen, F. H. The Cambridge Structural Database: a Quarter of a Million Crystal Structures and Rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 380–388.
- (60) 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.; Caicato, 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, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*; Gaussian, Inc.: Wallingford, CT, 2009.
- (61) Breneman, C. M.; Wiberg, K. B. Determining Atom-Centered Monopoles from Molecular Electrostatic Potentials. The Need for High Sampling Density in Formamide Conformational Analysis. *J. Comput. Chem.* **1990**, *11*, 361–373.
- (62) *The NAG Library*, version 24; The Numerical Algorithms Group (NAG): Oxford, United Kingdom, 2013.
- (63) Schittkowski, K. *NLPQLP: a fortran implementation of a sequential quadratic programming algorithm with distributed and non-monotone linesearch - User's Guide*, version 3.0; Department of computer science, University of Bayreuth: Bayreuth, 2009.
- (64) Schittkowski, K. A Robust Implementation of a Sequential Quadratic Programming Algorithm With Successive Error Restoration. *Optimization Letters* **2011**, *5*, 283–296.
- (65) Kazantsev, A. V. *Molecular flexibility in crystal structure prediction*. Ph.D. Thesis, Imperial College College, London, 2011.
- (66) Coombes, D. S.; Price, S. L.; Willock, D. J.; Leslie, M. Role of Electrostatic Interactions in Determining the Crystal Structures of Polar Organic Molecules. A Distributed Multipole Study. *J. Phys. Chem.* **1996**, *100*, 7352–7360.
- (67) 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 Crystallogr., Sect. B: Struct. Sci.* **2000**, *56*, 697–714.
- (68) Yu, L. Color Changes Caused by Conformational Polymorphism: Optical-Crystallography, Single-Crystal Spectroscopy, and Computational Chemistry. *J. Phys. Chem. A* **2001**, *106*, 544–550.
- (69) Chen, S.; Guzei, I. A.; Yu, L. New Polymorphs of ROY and New Record for Coexisting Polymorphs of Solved Structures. *J. Am. Chem. Soc.* **2005**, *127*, 9881–9885.
- (70) Chen, S.; Xi, H.; Yu, L. Cross-Nucleation Between ROY Polymorphs. *J. Am. Chem. Soc.* **2005**, *127*, 17439–17444.
- (71) Stone, A. J. Distributed Multipole Analysis: Stability for Large Basis Sets. *J. Chem. Theory Comput.* **2005**, *1*, 1128–1132.
- (72) 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. *J. Chem. Phys.* **2008**, *128*, 244708.

(73) Cossi, M.; Scalmani, G.; Rega, N.; Barone, V. New Developments in the Polarizable Continuum Model for Quantum Mechanical and Classical Calculations on Molecules in Solution. *J. Chem. Phys.* **2002**, *117*, 43–54.

(74) Cooper, T. G.; Heyczyk, K. E.; Jones, W.; Day, G. M. Molecular Polarization Effects on the Relative Energies of the Real and Putative Crystal Structures of Valine. *J. Chem. Theory Comput.* **2008**, *4*, 1795–1805.

(75) Tobbyn, M.; Brown, J.; Dennis, A. B.; Fakes, M.; Gao, Q.; Gamble, J.; Khimyak, Y. Z.; McGeorge, G.; Patel, C.; Sinclair, W.; Timmins, P.; Yin, S. Amorphous Drug-PVP Dispersions: Applications of Theoretical, Thermal and Spectroscopic Analytical Techniques to the Study of a Molecule With Intermolecular Bonds in Both the Crystalline and the Pure Amorphous State. *J. Pharm. Sci.* **2009**, *98*, 3456–3468.