Ab Initio Crystal Structure Prediction—I. Rigid Molecules

PANAGIOTIS G. KARAMERTZANIS, CONSTANTINOS C. PANTELIDES
Centre for Process Systems Engineering, Department of Chemical Engineering and Chemical Technology, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom

Received 27 May 2004; Accepted 15 October 2004
DOI 10.1002/jcc.20165
Published online in Wiley InterScience (www.interscience.wiley.com).

Abstract: A new methodology for the prediction of molecular crystal structures using only the atomic connectivity of the molecule under consideration is presented. The approach is based on the global minimization of the lattice enthalpy of the crystal. The modeling of the electrostatic interactions is accomplished through a set of distributed charges that are optimally and automatically selected and positioned based on results of quantum mechanical calculations. A four-step global optimization algorithm is used for the identification of the local minima of the lattice enthalpy surface. A parallelized implementation of the algorithm permits a much more extensive search of the solution space than has hitherto been possible, allowing the identification of crystal structures in less frequently occurring space groups and with more than one molecule in the asymmetric unit. The algorithm has been applied successfully to the prediction of the crystal structures of 3-aza-bicyclo(3.3.1)nonane-2,4-dione ($P2_1/a$, $Z' = 1$), allopurinol ($P2_1/c$, $Z' = 1$), 1,3,4,6,7,9-hexa-azacycl(3.3.3)azine (Pbca, $Z' = 2$), and triethylenediamine ($P6_3/m$, $Z' = 1$). In all cases, the experimentally known structure is among the most stable predicted structures, but not necessarily the global minimum.

© 2004 Wiley Periodicals, Inc. J Comput Chem 26: 304–324, 2005

Key words: crystal structure prediction; polymorphism; lattice energy minimization

Introduction

The prediction of the structure of crystals formed by organic molecules is a problem of immense practical significance to many key sectors of the process industries. This is particularly so in view of the propensity of many such compounds to crystallize in multiple stable forms (polymorphs) and the influence of crystal structure on key product characteristics such as the density, color, solubility, rate of dissolution, melting point, chemical stability and optical properties.

The relatively high occurrence of polymorphism in nature makes a theoretical tool for the reliable exploration of the polymorphic landscape highly desirable. This would be useful both for molecules that have not yet been synthesized or crystallized in single crystals or indexable powder, and for molecules for which one or more crystal structures have already been resolved experimentally. Ideally, such a tool would be able to predict all likely crystal structures from a knowledge of simply the molecular structure.

Current Methods for Crystal Structure Prediction

Most current methods for crystal structure prediction are based on the minimization of the crystal lattice energy, with polymorphs being identified as local minima of relatively low energy. The key mathematical difficulty in this context is the fact that the lattice energy surface exhibits very large numbers of local minima. Over the past decade, increasingly general numerical techniques for global optimization have been emerging that can guarantee obtaining the global minimum within finite computational times. However, these techniques are not yet powerful enough to be directly applicable to this particular problem. Moreover, crystal structure prediction aims to identify not just the global energy minimum structure, but also all others of relatively low energy. Consequently, methods for crystal structure prediction tend to rely on carrying out energy minimizations *via* local optimization techniques starting from large numbers of different initial guesses. The latter, commonly referred to as "candidate structures," are generated using some form of deterministic or pseudorandom search procedure.

Correspondence to: C. C. Pantelides; e-mail: c.pantelides@imperial.ac.uk.

Contract/grant sponsor: Mitsubishi Chemical Corporation

Contract/grant sponsor: United Kingdom's Engineering and Physical Sciences Research Council (EPSRC); contract/grant number: GR/N08636

The general framework described above encompasses a number of crystal structure prediction methods that have been proposed over the last decade.[1,2] One such method was MOLPAK,[3] which attempted to build the highest density hypothetical packing arrangements, considering all unique orientations of a central molecule and constructing its surroundings according to the 29 most common coordination geometries encountered in the Cambridge Structural Database.[4,5] The packing calculations are accelerated by the use of only the repulsive contributions of a Lennard–Jones or Buckingham exp-6 potential model. MOLPAK has been used to generate candidate structures serving as initial guesses for local energy minimization with WMIN[6] and DMAREL.[7] The latter performs lattice energy minimizations of rigid molecules, employing an accurate intermolecular potential based on a distributed multipole electrostatic model. The electrostatic energy is calculated from the atomic charges, dipoles, quadrupoles, octapoles, and hexadecapoles derived by a distributed multipole analysis.[8]

Promet[9] uses common symmetry operators to systematically generate dimers or strings of rigid molecules; these are translated in space to form crystals. The energy is minimized under the constraints of space group symmetry using empirical repulsion–dispersion potentials and electrostatic interactions based on the point-charge approximation without forced convergence. The PMC program[10,11] minimizes the potential energy of a crystal composed of several molecular units with fixed internal geometry. The units can be either complete molecules or fragments of a flexible molecule held together by potentials of quadratic type to preserve standard geometry around such a link. CRYSCA[12] is based on energy minimization of randomly generated packings of flexible molecules. The starting set consists of several hundreds of crystal structures with random values for the lattice parameters, orientation, and position of the molecules, as well as selected intramolecular degrees of freedom. MSI-PP (Polymorph Predictor, http://www.accelrys.com/cerius2/index.html), a commercially available program for crystal structure prediction, employs a stochastic generation of initial guesses in each space group considered, using a rigid molecular structure but allowing the relaxation of the molecular structure at a final stage of lattice energy minimization. UPACK[13] generates initial guesses randomly or systematically. At the final lattice energy minimization phase, all-atom potentials and fully flexible molecules are considered. UPACK was further extended to use an *ab initio* derived intermolecular potential,[14,15] including atomic multipoles and polarization effects; for intramolecular interactions, the MM3 force field,[16] as applied in the TINKER program (http://dasher.wustl.edu/tinker/), or *ab initio* computations[17,18] were employed.

Several researchers have advocated the use of a relatively cheaper model for the generation of the initial structures and, possibly, for an initial minimization of these to produce a set of a few hundred structures to be minimized with a more elaborate model. Although this approach can significantly accelerate the search procedure, it can also cause a serious degradation of the ability of the search method to locate candidate structures. Thus, the potential model used in this initial stage needs to be efficient enough for an extensive search to be carried out, yet sufficiently accurate not to exclude good candidate structures from the final local energy minimization phase. These conflicting requirements are not easy to satisfy. For example, the inability of MOLPAK/DMAREL to identify the experimental minimum of the 3-azabicyclo(3.3.1)nonane-2,4-dione test case of the second blind test[19] can be attributed to a failure to meet these requirements (see later).

The number of distinct local minima identified by lattice energy minimization techniques as having energy within the range of polymorphism (typically considered to be 10 kJ mol^{-1} above the global energy minimum) can itself be quite large. Consequently, additional criteria are often applied to determine which of these minima are likely to correspond to stable polymorphs that exist in nature. These examine a crystal structure in terms of its nucleation process, temperature effects, mechanical stability, morphology, and growth rates.[20–23]

Outline of This Article

This article presents a new approach for the determination of crystal structures of organic molecules based on the following four main components:

1. an accurate, yet computationally efficient, distributed charge model that includes electric charges at both atomic and non-atomic (satellite) positions;
2. a systematic global search procedure based on low-discrepancy sequences generating candidate structures that effectively sample the large search space inherent in this problem;
3. an accurate local minimization procedure based on a mathematical formulation and a state-of-the-art optimization algorithm that ensure a high degree of reliability in the optimization of candidate structures;
4. a parallelized implementation of the above that typically allows the examination of hundreds of thousands of promising candidate structures.

The next section presents the mathematical formulation for the lattice energy minimization, while the following section briefly discusses some computational aspects of the lattice energy calculation relating to accuracy, efficiency, and continuity. Then we present our global energy minimization algorithm and discuss some aspects of its parallelized implementation. Finally, we present results obtained by applying our algorithm to four test molecules.

The present, first part of this article focuses on rigid molecules, aiming to establish the basic methodology. The extensions necessary to address flexible molecules are considered in the second part of the article.[24]

Mathematical Formulation

In formulating the lattice energy minimization problem, we will assume that the molecules in the unit cell are rigid and are all characterized by the same molecular conformation or its mirror image.

The Molecular Model

Each molecule comprises a number of sites that can participate in repulsion–dispersion and/or electrostatic interactions. These sites do not necessarily coincide with the atomic positions. A key feature of our overall methodology is the accurate description of a molecule's electrostatic field as computed by a quantum mechanical calculation. This accuracy is achieved by distributed charges, some of which are located at atomic positions and some at "satellite" positions in the vicinity of certain atoms. The magnitudes of all charges and the precise positions of the satellite charges are determined automatically *via* a sophisticated optimization procedure that seeks to minimize the deviation of the electrostatic field induced by these charges from the quantum mechanical one. In addition, our procedure imposes upper bounds on: (1) the allowable deviations of the moments of the molecular charge distribution (e.g., dipoles and quadrupoles) from the corresponding values computed by the quantum mechanical calculations; (2) the accuracy of the estimated charges as assessed *via* the confidence intervals of the parameter estimation procedure.

The above procedure[25] is applicable to any rigid molecule and results in distributed charge models with relative root mean square (RRMS) error well within 5%. This is a measure of the relative difference of the electrostatic fields predicted by, respectively, the distributed charge and the quantum mechanical models. Our experience indicates that this level of accuracy is often necessary for the correct prediction of the crystal structure.

In any case, for the remainder of this article, we will assume that the charge magnitudes and relative positions of all sites in the molecule are known.

Data

Our lattice energy minimization algorithm receives as input the following information:

1. The number of molecules, Z, in the unit cell. When the lattice energy minimization is carried out subject to space group constraints, then Z is determined by the space group and the number of crystallographically independent molecules being considered; see later.
2. The number of sites N^S of each molecule and their body-based coordinates:
$$\mathbf{r}_{ji}^{bf}, \quad j = 1, \ldots, Z, i = 1, \ldots, N^S \tag{1}$$
Note that, in general, different molecules in the unit cell may have different body-based coordinates. This allows the consideration of crystal structures in space groups that involve mirror planes, inversion centers, or glide phases that invert the local axis system.
3. A potential function for the van der Waals interaction between each pair of sites i, i' belonging to different molecules and positioned at a distance r from each other:
$$U_{ii'}^{\text{vdw}}(r) \quad i, i' = 1, \ldots, N^S \tag{2}$$
Intramolecular energy is not important for the rigid molecules considered in this article.
4. The charge associated with each site:
$$q_i, \quad i = 1, \ldots, N^S \tag{3}$$
In this article, we assume that all Z molecules will be characterized by the same set of distributed charges. This is a valid assumption if their molecular conformation is the same or related by means of symmetry relations.
5. The lower and upper bounds for the lattice lengths of the unit cell l^L, l^U. The size of the unit cell is expected to be larger for large molecules or when the number of molecules in the unit cell is large. For small organic molecules, typical values are $l^L = 5$ Å and $l^U = 50$ Å.
6. The lower and upper bounds for the lattice angles of the unit cell ω^L, ω^U. Typical values for these bounds are $\omega^L = 50°$ and $\omega^U = 130°$. This avoids the consideration of cells that are deformed to such an extent that a very large number of periodic cells have to be considered to achieve a certain level of accuracy in the computation of the lattice energy.
7. The pressure P. In some molecules, different polymorphs may form under different pressures. A typical example is benzene.[26–28]

Optimization Decision Variables

The choice of the set of optimization variables is important in designing a robust and efficient energy minimization algorithm. It is normally better to select variables that can be bounded within limits known *a priori* such that the objective function is mathematically well defined for any values of the variables within these limits. With this in mind, the variables chosen for this work are:

1. The lattice lengths $l_a \in [l^L, l^U]$, $a \in \{1, 2, 3\}$.
2. The lattice angles $\omega_a \in [\omega^L, \omega^U]$, $a \in \{1, 2, 3\}$.
3. The *normalized* positions of the molecular centers of mass $\hat{\mathbf{r}}_j$, $j = 2, \ldots, Z$ with respect to the crystallographic frame. Our experience indicates that the use of normalized rather than cartesian coordinates results in better performance in the local minimization phase (see later).

   Normalized positions lie within 0 and 1, and this is taken into account by the global search procedure for the generation of candidate structures (see later). However, it was found to be better not to enforce these bounds during the local minimization phases (see later): if the latter converges to normalized positions outside the domain [0, 1], then these can easily be reduced to values within this domain by exploitation of the crystal structure periodicity.
4. The orientations of the Z molecules in the unit cell with respect to the space fixed frame, defined by three Euler angles ϕ_j, θ_j, ψ_j, $j = 1, \ldots, Z$. No bounds were imposed on these quantities. Thus, the solution of a local lattice energy minimization may lead to a crystal where some of the Euler angles $\{\phi_j, \theta_j, \psi_j\}$ are not in the range $[0, 2\pi] \times [0, \pi] \times [0, 2\pi]$. Again, any such solution can easily be translated to an equivalent one lying within the aforementioned range.
5. The lattice deformation W; the meaning of, and need for this variable are explained below.

Objective Function

The lattice energy minimization problem is formulated as follows:

$$\min_{(\{l_a,\omega_a,\ a=1,2,3\},W,\{\hat{\mathbf{r}}_j,\ j=2,\ldots,Z\},\{\phi_j,\theta_j,\psi_j\ j=1,\ldots,Z\})}$$

$$\frac{1}{2}\sum_{\mathbf{n}\in Z^3}{}' \sum_{j=1}^{Z}\sum_{j'=1}^{Z}\sum_{i=1}^{N^S}\sum_{i'=1}^{N^S} U_{ii'}^{\text{vdw}}(\|\mathbf{r}_{ji}-\mathbf{r}_{j'i'}+\mathbf{R}\cdot\mathbf{n}\|)$$
$$+\frac{2\pi}{V}\sum_{\substack{\mathbf{k}\in Z^3\\ \mathbf{k}\neq\mathbf{0}}}\frac{1}{\|\mathbf{k}\|^2}e^{-\|\mathbf{k}\|^2/4\alpha^2}\vartheta(\mathbf{k})\vartheta^*(\mathbf{k})$$
$$+\frac{1}{2}\sum_{\mathbf{n}\in Z^3}{}' \sum_{j=1}^{Z}\sum_{i=1}^{N^S} q_i \sum_{j'=1}^{Z}\sum_{i'=1}^{N^S} q_{i'}\frac{\operatorname{erfc}(\alpha\|\mathbf{r}_{ji}-\mathbf{r}_{j'i'}+\mathbf{R}\cdot\mathbf{n}\|)}{\|\mathbf{r}_{ji}-\mathbf{r}_{j'i'}+\mathbf{R}\cdot\mathbf{n}\|}+PV$$
$$-\frac{1}{2}\sum_{j=1}^{Z}\sum_{i=1}^{N^S} q_i\left(\frac{2\alpha q_i}{\sqrt{\pi}}+\sum_{\substack{i'=1\\ i'\neq i}}^{N^S} q_{i'}\frac{\operatorname{erf}(\alpha\|\mathbf{r}_{ji}^{bf}-\mathbf{r}_{ji'}^{bf}\|)}{\|\mathbf{r}_{ji}^{bf}-\mathbf{r}_{ji'}^{bf}\|}\right) \qquad (4)$$

where the summation indices $\mathbf{n}$ and $\boldsymbol{\kappa}$ are integer vectors of length 3, $\mathbf{R} = \mathbf{R}\ (l_a,\ \omega_a,\ a = 1,\ 2,\ 3)$ is the cell matrix, $\vartheta(\mathbf{k}) \equiv \sum_{j=1}^{Z} \sum_{i=1}^{N^S} q_i e^{i\mathbf{k}\cdot\mathbf{r}_{ji}}$ and $\mathbf{k}(\boldsymbol{\kappa}) \equiv 2\pi(\mathbf{R}^{-1})^T \cdot \boldsymbol{\kappa}$. The cartesian coordinates $\mathbf{r}_{ji}$ of site i of molecule j within the unit cell are computed from the body-based coordinates *via* the expression $\mathbf{r}_{ji} = \mathbf{A}_j^T\mathbf{r}_{ji}^{bf} + \mathbf{R}\cdot\hat{\mathbf{r}}_j$ where $\mathbf{A}_j = \mathbf{A}_j(\phi_j,\ \theta_j,\ \psi_j)$ is the rotation matrix of molecule j expressed as a function of the Euler angles.

The first term of eq. (4) takes account of repulsion/dispersion interactions, while the next two are, respectively, the reciprocal and real Ewald sums arising from the modeling of electrostatic interactions. The terms account for all interactions between all pairs of sites belonging to different molecules, either within the unit cell or in its periodic neighbors. Because the molecules are assumed to be rigid, interactions between sites within the same molecule are not included; the notation $\sum'$ indicates that the multiple summation excludes elements for which $j = j'$ when $\mathbf{n} = \mathbf{0}$. The electrostatic self correction term [last line of eq. (4)] is an adjustment accounting for the inclusion of intramolecular interactions in the reciprocal summation; this is merely a constant, and can, therefore, be omitted as far as the energy minimization is concerned. However, it is necessary to compute it prior to any comparison of the optimized lattice energy to experimental energies of sublimation.

We note that some formulations[29] in the literature include the dipolar correction term:

$$\mathbf{M}^c = \frac{2\pi}{3V}\left\|\sum_{j=1}^{Z}\sum_{i=1}^{N^S} q_i\mathbf{r}_{ji}\right\|^2 \qquad (5)$$

which attempts to correct for the omission of $\kappa = \mathbf{0}$ in the reciprocal sum of eq. (4). Although this term can be significant for crystal structures in polar space groups, the electrostatic contributions to the lattice energy depend on the external shape of the crystal,[30] which is not known *a priori* in crystal structure predictions. In this article, we neglect this term assuming that the crystal surroundings will annul the surface charge (as may be the case for crystallization in a solvent with a high dielectric constant) or that the crystal will find a shape (e.g., needles or platelets, with the macroscopic dipole moment aligned along the needle direction or lying on the plane of the platelet, respectively) such that this term vanishes and the energy is at the lowest.

The Lattice Deformation Constraint

The cell matrix $\mathbf{R}$ is a function of the square root of the quantity $1 - \sum_{a=1}^{3} \cos^2\omega_a + 2\prod_{a=1}^{3} \cos\omega_a$. The latter is a positive quantity for all physically feasible unit cells. However, in the course of a numerical minimization computation, it may well take negative values as the optimization algorithm examines different values of the lattice angles ω_a. Such an event will, inevitably, cause a failure in the evaluation of the lattice energy. This is highly undesirable in the context of global lattice energy minimization, which may involve the solution of hundreds of thousands of local minimization problems, each starting from an automatically generated candidate structure.

To address this issue, we introduce in our formulation: (a) the lattice deformation W as an additional optimization decision variable; this is bounded to lie in the domain $[\varepsilon_w,\ 1 + \varepsilon_w]$ where ε_w is a small positive number (typically 0.001); (b) the nonlinear equality constraint:

$$W - 1 + \sum_{a=1}^{3}\cos^2\omega_a - 2\prod_{a=1}^{3}\cos\omega_a = 0 \qquad (6)$$

We then express the cell matrix $\mathbf{R}$ in terms of $\sqrt{W}$ rather than $\sqrt{1 - \sum_{a=1}^{3} \cos^2\omega_a + 2\prod_{a=1}^{3} \cos\omega_a}$. Of course, at the solution of the local optimization problems, the equality constraint 6 will be satisfied, and the two formulations for $\mathbf{R}$ will yield the same result. However, *during* the optimization iterations, the optimizer will always maintain W to a positive value even when the quality $1 - \sum_{a=1}^{3} \cos^2\omega_a + 2\prod_{a=1}^{3} \cos\omega_a$ is negative.

Minimization under Space Group Constraints

The mathematical formulation presented so far does not impose any space group symmetry constraints. Its solution, nevertheless, automatically results in crystals that belong to one of the 230 space groups.

On the other hand, the explicit inclusion of symmetry constraints leads to reduced numbers of optimization variables and, consequently, faster solution times for the local lattice energy minimization. For example, in the case of monoclinic and higher symmetry crystal systems, two or more of the lattice angles are fixed; for tetragonal or higher symmetry crystal systems, two or more of the lattice lengths are equal to each other. Also, if the crystal structure involves G crystallographically independent molecules,[31] then for an unconstrained lattice energy minimization to identify a local minimum that happens to be in space group s, it would have to consider explicitly up to $\sigma_s G$ molecules in the unit cell, where σ_s is the total number of symmetry positions of the space group s. On the other hand, a minimization carried out within space group s would need to consider explicitly only G distinct molecules.

As we shall see later, our global optimization algorithm carries out local minimizations both with and without space groups constraints. In the former case, the positions of all sites of all molecules in the unit cell can be determined from the normalized positions of the center of mass and the Euler angles of the G crystallographically independent molecules.

For polar space groups, the crystal structure is invariant under translations of the asymmetric unit along one or more of the crystallographic directions. For every such direction $a \in (\hat{x}, \hat{y}, \hat{z})$, the corresponding normalized position of the center of mass of the first molecule in the asymmetric unit $\hat{r}_1^a$ is excluded from the set of decision variables and all molecules are translated accordingly.

Finally, the lattice deformation variable W and the corresponding equality constraint 6 will be introduced only for those space groups that involve at least one lattice angle to be determined by the optimization problem.

Accurate Computation of Lattice Energy

The solution of the optimization problem 4 to global optimality requires a large number of local minimizations that produce local minima, some of which differ very little in energy. The correct ranking of these minima can be sensitive to the accuracy of the computed lattice energy and, in particular, to the truncation of the infinite summations over real $\mathbf{n}$ and reciprocal space $\boldsymbol{\kappa}$. Moreover, the use of a gradient-based local optimization algorithm imposes further requirements on the continuity and differentiability properties of the objective function.

If r^c is the distance at which the slowest decaying interaction $U^{\text{vdw}}_{ii'}(r)$ is acceptably close to zero, then no repulsion–dispersion interaction between any pair of sites separated by distances greater than r^c needs to be considered. To avoid discontinuities, the truncation is supplemented by the use of a quintic spline interpolation[32] in a region $[r^s, r^c]$, where $r^s < r^c$. For any general, twice-differentiable potential function $U^{\text{vdw}}_{ii'}(r)$, it can be shown that this leads to a modified potential $\hat{U}^{\text{vdw}}_{ii'}$ defined as:

$$\hat{U}^{\text{vdw}}_{ii'} \equiv \begin{cases} U^{\text{vdw}}_{ii'}(r), & r \leq r^s \\ (r - r^c)^3(A_{ii'} + B_{ii'}r + C_{ii'}r^2), & r^s < r \leq r^c \\ 0, & r > r_c \end{cases} \tag{7}$$

where the coefficients $A_{ii'}$, $B_{ii'}$, and $C_{ii'}$ are computed to ensure continuity of $\hat{U}^{\text{vdw}}_{ii'}$, $d\hat{U}^{\text{vdw}}_{ii'}/dr$, and $d^2\hat{U}^{\text{vdw}}_{ii'}/dr^2$ at $r = r^s$. In our computations, we have chosen $r^c = 20$ Å and $r^s = 17$ Å, which are sufficiently long to ensure high accuracy.

If we assume that the center of mass of every molecule is inside the unit cell, that is, $\hat{\mathbf{r}}_j \in [0, 1)^3$, then the infinite summation over real space for the repulsion–dispersion interactions can be truncated to all periodic shells that satisfy the relation:

$$\|\mathbf{r}_j - \mathbf{r}_{j'} + \mathbf{Rn}\| \leq 2\gamma^{\text{mol}} + r^c \tag{8}$$

where γ^{mol} is the radius of the molecules under consideration.[33]

For the real Ewald sum, we employ a cutoff distance r^c of 17 Å, and set the parameter α to a value of 0.376 Å^{-1}. This makes the quantity $\text{erfc}(\alpha r^c)/r^c$ equal to a very small value (10^{-20} Å^{-1}), thereby ensuring that the discontinuity arising from the truncation of the real sum is not detrimental to the performance of the optimization algorithm.[29,34] Similarly, in the reciprocal Ewald sum, we omit all terms for which $(1/\|\mathbf{k}\|^2)e^{-\|\mathbf{k}\|^2/4\alpha^2}$ is smaller than 10^{-20} Å^2. This defines an upper limit on the inverse distances $\|\mathbf{k}\|$ that need to be included in the summations.

Solution Algorithm

Our global minimization algorithm comprises four distinct steps, namely:

1. Global search for the generation of candidate structures.
2. Local minimization with space group constraints, starting from some of the candidate structures generated at step 1.
3. Postprocessing for the identification of unique local minima among those generated at step 2.
4. Confirmation that the low-energy points identified at step 3 are true local minima even without space group constraints.

In general, the above steps are applied repeatedly for different values of the number G of crystallographically independent molecules in the asymmetric unit. The precise nature of each step is explained in more detail below.

Step 1: Global Search

The global search step of our algorithm generates candidate structures that belong to one of the 59 space groups which appear in the CSD with a frequency of more than 0.05%. The most prevalent space groups in the CSD[4] are triclinic (22.3%), monoclinic (53.4%), and orthorhombic (20.4%). Tetragonal, trigonal/rhombohedral, hexagonal, and cubic are relatively rare, and thus usually not explicitly considered by crystal structure prediction algorithms; also, two space groups are not represented at all in the CSD. The rarity of many space groups has been attributed[5] to the presence of mirror planes or rotation axes, which gives rise to inefficient packing and conflicts with Kitaigorodsky's principle[35] of closest packing as a determinant of crystal structure.

The first step in creating a candidate structure is to determine the space group to which it will belong. The probability of generating a candidate in space group s is set to be proportional to $(F_s)^{0.75}\sigma_s$, where F_s is the number of organic/organometallic compounds reported as forming a crystal in space group s in the CSD. The form of this expression is designed to account for the increasing complexity of the potential surface with the number of molecules in the unit cell (cf. Table 1).

Once the space group for a particular candidate structure is selected, the next step is to determine values for the decision variables determining the geometry of the unit cell and the positioning of the molecules within it. The precise number and nature of these variables depend on the space group, as shown in the last two columns of Table 1. (In the interests of brevity, the table shown here is restricted to the 13 most frequent space groups. However, as has already been stated, our global search algorithm considers 59 such space groups.) The values for the lattice lengths l and angles ω are chosen so as to lie within the bounds discussed earlier. The lower and upper bounds for

Table 1. Thirteen Most Frequent Space Groups Investigated during Global Search It Is Assumed that the Asymmetric Unit Consists of One Molecule, that is, $Z' = 1$, that Belongs to the C1 Point Group. Enantiomorphic space groups are shown in bold.

Space group	Z	Percentage of structures generated in global phase	$N^{\text{var a}}$	Optimization decision variables and global search domain
P1	1	0.315	9	$l_1, l_2, l_3, \omega_1, \omega_2, \omega_3$ $\phi_1 \in [0, 2\pi], \theta_1 \in [0, \pi], \psi_1 \in [0, 2\pi]$
P$\bar{1}$	2	5.939	12	$l_1, l_2, l_3, \omega_1, \omega_2, \omega_3$ $\phi_1 \in [0, 2\pi], \theta_1 \in [0, \pi], \psi_1 \in [0, 2\pi]$ $\hat{r}_1^x, \hat{r}_1^y, \hat{r}_1^z \in [0, 0.5]$
P2$_1$	2	2.364	9	l_1, l_2, l_3, ω_2 $\phi_1 \in [0, 2\pi], \theta_1 \in [0, \pi/2], \psi_1 \in [0, 2\pi]$ $\hat{r}_1^x, \hat{r}_1^z \in [0, 0.5]$
P2$_1$/c	4	18.620	10	l_1, l_2, l_3, ω_2 $\phi_1 \in [0, 2\pi], \theta_1 \in [0, \pi/2], \psi_1 \in [0, 2\pi]$ $\hat{r}_1^x, \hat{r}_1^y, \hat{r}_1^z \in [0, 0.5]$
Cc	4	1.288	8	l_1, l_2, l_3, ω_2 $\phi_1 \in [0, 2\pi], \theta_1 \in [0, \pi/2], \psi_1 \in [0, 2\pi]$ $\hat{r}_1^y \in [0, 0.5]$
C2	4	1.104	9	l_1, l_2, l_3, ω_2 $\phi_1 \in [0, 2\pi], \theta_1 \in [0, \pi/2], \psi_1 \in [0, 2\pi]$ $\hat{r}_1^x, \hat{r}_1^z \in [0, 0.5]$
P2$_1$2$_1$2, **P2$_1$2$_1$2$_1$**	4	0.732, 6.706	9	l_1, l_2, l_3 $\phi_1 \in [0, \pi], \theta_1 \in [0, \pi/2], \psi_1 \in [0, 2\pi]$ $\hat{r}_1^x, \hat{r}_1^y, \hat{r}_1^z \in [0, 0.5]$
Pna2$_1$, Pca2$_1$	4	1.771, 1.013	8	l_1, l_2, l_3 $\phi_1 \in [0, \pi], \theta_1 \in [0, \pi/2], \psi_1 \in [0, 2\pi]$ $\hat{r}_1^x, \hat{r}_1^y \in [0, 0.5]$
Pbca, Pbcn	8	6.984, 2.376	9	l_1, l_2, l_3 $\phi_1 \in [0, \pi], \theta_1 \in [0, \pi/2], \psi_1 \in [0, 2\pi]$ $\hat{r}_1^x, \hat{r}_1^y, \hat{r}_1^z \in [0, 0.5]$
C2/c	8	11.284	10	l_1, l_2, l_3, ω_2 $\phi_1 \in [0, 2\pi], \theta_1 \in [0, \pi/2], \psi_1 \in [0, 2\pi]$ $\hat{r}_1^x, \hat{r}_1^y, \hat{r}_1^z \in [0, 0.5]$

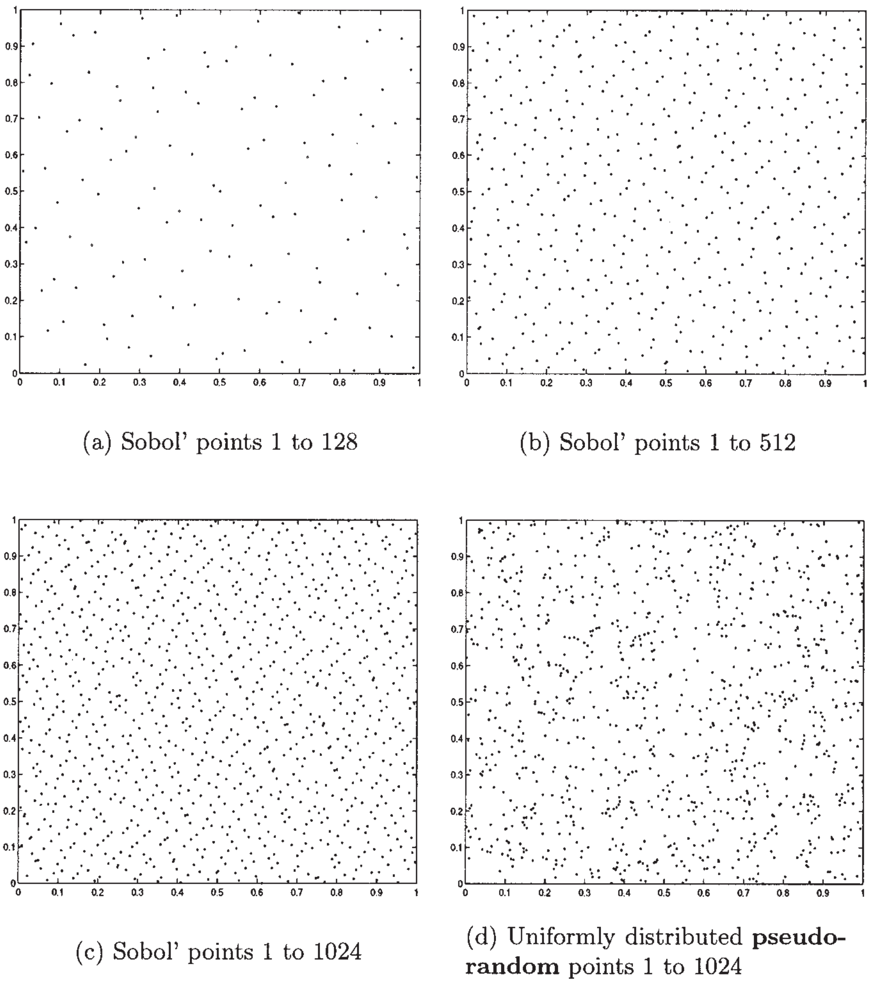
[a] Number of optimization decision variables.

the normalized positions $\hat{\mathbf{r}}$ of the molecular centers of mass and the molecular Euler angles depend on the space group[36] and are also shown in the last column of Table 1. We note that, for crystal systems with symmetry no higher than orthogonal, the crystal is invariant to translations of the asymmetric unit over half a cell length and by rotation over π around any cell axis that is vertical to the plane defined by the other two. Therefore, the range of the Euler angles and the translation vector can be reduced accordingly. The consequent reduction of the size of the search domain ranges from a factor of 8 (for triclinic cells) to 32 (for orthorhombic cells). Further reductions can be made in specific space groups for example, in some orthogonal space groups the three axes are equivalent and thus the condition $l_1 \leq l_2 \leq l_3$ can be applied. Moreover, if the molecule exhibits internal symmetry, the bounds in the Euler angles can be further reduced.

There are a number of well-established ways of searching a multivariable domain of several variables lying within given bounds. One approach is to rely on uniformly distributed pseudorandom numbers,[37] a technique commonly known as Monte Carlo sampling. Another approach is to use a uniform grid of points in multidimensional space. Albeit very widely used, both of these techniques suffer from severe deficiencies. The Monte Carlo sampling offers no guarantee of uniform coverage of the domain of interest for any finite number of points. On the other hand, the uniform grid technique is useful only if one decides *a priori* the number of points to be examined, and then actually proceeds to do so; this may be impractical in cases where the global search may involve a potentially varying number of computer processors operating over several days.

A class of techniques specifically designed to address the above issues is that of low-discrepancy sequences. These are entirely deterministic sequences that aim to achieve the best possible coverage of a domain at each iteration of the sampling. One of the most successful such techniques is that due to Sobol'.[38]

In the Sobol' sequence in a N-dimensional domain, N numbers are generated *simultaneously* as binary fractions of length w bits from a set of w special binary fractions, V_i, $i = 1, \ldots, N$, called *direction numbers.* The reader can refer to the original literature for the details of the method and, in particular, the generation of the direction numbers. Figure 1 provides a graphical impression of how the Sobol' sequence performs in a two-dimensional case in comparison to a random sequence.[39] The more uniform coverage



(a) Sobol' points 1 to 128

(b) Sobol' points 1 to 512

(c) Sobol' points 1 to 1024

(d) Uniformly distributed **pseudo-random** points 1 to 1024

Figure 1. Example of a two-dimensional Sobol' sequence and comparison with a pseudorandom sequence.

achieved by the Sobol' sequence is quite evident. Another characteristic is that the projection of each Sobol' point onto each of the axes corresponds to a distinct point; a practical implication of this fact is that if, for example, we generate a sequence of 59,049 ($=9^5$) crystal structures in the P1 space group (which involves nine decision variables), then these will actually involve 59,049 distinct values of each lattice length, each lattice angle, and so on; this is quite unlike a uniform grid with the same number of points that would only manage to consider five distinct values for each of these nine decision variables.

Once values for all optimization decision variables are created, the resulting cell is checked to determine if it can immediately be excluded from further consideration. This is the case if at least one of the following holds:

1. the crystal structure involves atoms belonging to different molecules being separated by distances smaller than 0.5 Å;
2. the quantity $1 - \Sigma_{a=1}^{3} \cos^2\omega_a + 2 \Pi_{a=1}^{3} \cos \omega_a$ is nonpositive;
3. the density is below a given threshold ρ^{thres};
4. the lattice energy exceeds a given threshold U^{thres}.

If none of the above holds, then the lattice deformation W is set equal to $1 - \Sigma_{a=1}^{3} \cos^2\omega_a + 2 \Pi_{a=1}^{3} \cos \omega_a$, and the resulting point is used as a starting point for local minimization of lattice energy, as described below.

Step 2: Local Minimization Subject to Space Group Constraints

This phase of the algorithm uses the candidate structures generated during the global search phase (see previous section) as starting points for performing local minimizations of the lattice enthalpy defined by eq. (4) with respect to the decision variables listed in the last column of Table 1. If the candidate structure is in a space group that involves the determination of at least one lattice angle, the lattice deformation variable W is added to the set of decision variables, and the optimization is carried out subject to the equality constraint 6.

In all cases, the lattice lengths and angles are restricted to lie within given lower and upper bounds. It has to be recognized that, no matter how loosely the latter are set, some minimizations will converge to structures for which these bounds are active, simply because the global search will have generated initial guesses in the vicinity of these bounds. From the mathematical point of view, such points are valid local minima of the bounds-constrained minimization. However, they do not correspond to physically valid structures as the gradients of the lattice energy with respect to the optimization variables are not zero; consequently, they are immediately discarded from any further consideration. It must be emphasized that this does not lead to the algorithm missing any true low-energy crystal structures: the properties of the Sobol' sequence guarantee that, at some stage during the global search, a point will be generated that is sufficiently close to each low-energy minimum so as to be within the local optimizer's region of attraction. In the local minimizations in this article, we have constrained the lattice lengths to the range 3–100 Å and the lattice angles to 10–170°. These are obtained by somewhat relaxing the bounds used in the global search aiming to achieve a trade-off between having to evaluate the lattice energy for highly deformed cells (something that is computationally expensive and prone to numerical problems due to exceedingly high gradients) and obtaining many local minima lying on these bounds.

Our experience has shown that it is computationally beneficial to force the local minimization algorithm to arrive at reasonably close packed structures during its first few iterations. Because such structures are, in general, energetically favored, they will, eventually, arise simply by the minimization of the objective function 4. However, a more direct and effective way of accelerating their emergence is by the addition of the constraint:

$$l_1 l_2 l_3 \sqrt{W}/Z \leq V^U \tag{9}$$

which imposes an upper bound V^U on the volume per molecule in the unit cell; an appropriate value for V^U can be chosen based on empirical rules.[40] Although this constraint is not active at the solution of the local minimization, its violation during the early stages of the local search forces the algorithm to quickly contract the unit cell, thereby achieving the desired effect.

Routine E04UFF of the NAg Library (Mark 20b, see http://www.nag.co.uk) is used for the local minimization. This is an implementation of a successive quadratic programming algorithm (SQP), which can readily handle the equality constraint[6] and also guarantees that any optimal point produced will be a true local minimum rather than a saddle point. Moreover, it exhibits fast (superlinear) convergence in the region of the local minimum.

In our code, the partial derivatives of the objective function and the lattice deformation constraint with respect to the decision variables are computed exactly[29,33,41] using analytical expressions. More specifically, the derivatives with respect to the positions of the molecular centers of mass and orientations are computed from the gradients of the lattice energy with respect to the cartesian coordinates of the sites *via* a chain-ruling operation. The accuracy of the partial derivatives is important for the reliable identification of local minima and the prevention of premature termination of the optimization algorithm at suboptimal points. Moreover, it generally results in faster convergence of the local minimization, thereby reducing the overall computational cost. In our implementation, the number of iterations required by the SQP algorithm, and consequently, the computational cost, are also reduced by appropriate scaling of the decision variables.

Step 3: Postprocessing of Local Minima

Global optimization techniques applied in crystal structure prediction typically produce large numbers of local minima. Some of them have almost identical lattice energies and densities but differ in the cell parameters, and may well correspond to identical crystals. This can often be verified by the transformation of the cell parameters to the reduced cells.[42,43] However, this approach is not always successful, as two crystals that differ very slightly can have completely different reduced cells.

A number of other techniques have, therefore, been devised to examine the equivalence of two crystals. One such technique[44] uses the matrix of shortest intermolecular distances between pairs of atoms to examine the various hydrogen bonding motifs obtained in the prediction of the crystal packing of 6-azauracil, uracil, and allopurinol. Another method[45] is based on the comparison of the molecular coordination shell and the derivation of the root-mean-square deviation of the non-H atoms for all atoms in the reference molecule and its 12 neighbors. Based on earlier work,[46] another method that is based on the comparison of two structures by computing the rotation that is necessary to transform one orientation to the other has been proposed.[47] Provided this rotation corresponds to a transformation that is compatible with the space group symmetry, the values of the transformed cell lengths, cell angles, and molecular centers of mass are compared. The method was implemented for triclinic, monoclinic, and orthorhombic systems with a single molecule in the asymmetric unit.

In our algorithm, locally optimal crystal structures are compared on the basis of the following quantities: (a) their lattice energies; (b) their densities; and (c) the minimum intermolecular distances between each distinct pair of atom types in the molecule under consideration.

The last quantity above requires some further explanation. For a given molecule, an atom type $\mathcal{A}_t$ is a subset of the molecule's atoms, all of which are related in pairs by at least one molecular symmetry operation. For example, all carbon atoms in the benzene molecule are related by a sixfold axis of rotation and, consequently, belong to the same atom type.

More formally, consider a molecule involving a set of atoms $\mathcal{A}$ and belonging to a point group characterized by a set of symmetry relations $\mathcal{S}$. Each symmetry relation is a triplet of the form $(i, i', \mathcal{G})$ where $i, i' \in \mathcal{A}$ are two distinct atoms and $\mathcal{G}$ is a molecular symmetry operation such as a symmetry plane, a proper axis of rotation, a rotation–reflection axis, or an inversion center. We can now partition set $\mathcal{A}$ into the *minimum* number N^T of disjoint subsets $\mathcal{A}_t$ (i.e., $\mathcal{A} = \cup_{t=1}^{N^T} \mathcal{A}_t$, $\mathcal{A}_t \cap \mathcal{A}_{t'} = \emptyset$, $\forall t \neq t'$) such that two atoms i, i' belong to the same subset $\mathcal{A}_t$ for some $t \in [1, \ldots, N^T]$ if and only if there exists at least one $\mathcal{G}$ such that $(i, i', \mathcal{G}) \in \mathcal{S}$. As the molecules in the unit cell are rigid, they will be either of the same molecular conformation or mirror images of each other, which means that the same partitioning of atoms into types $\mathcal{A}_t$ will hold for all of them.

Given a crystal structure, the minimum intermolecular distance $d_{tt'}$ between two different atom types t and t' is defined as the minimum cartesian distance between any two atoms $i \in \mathcal{A}_t$ and $i' \in \mathcal{A}_{t'}$ residing in different molecules j and j' in the central unit cell or its periodic neighbors. This can be expressed mathematically as:

$$d_{tt'} = \min_{\mathbf{n}, i \in \mathcal{A}_t, i' \in \mathcal{A}_{t'}, j, j' = 1, \ldots, Z} \|\mathbf{r}_{ji} - \mathbf{r}_{j'i'} + \mathbf{Rn}\| \tag{10}$$

The above minimization is understood to exclude elements for which $j = j'$ when $\mathbf{n} = \mathbf{0}$, that is, no *intra*molecular distances are considered because, in the case of rigid molecules, these will be equal for all crystal structures.

Step 3 of our algorithm starts by identifying the locally optimal crystal structure of minimum lattice energy. This is assumed to be the globally optimal structure. All minima whose lattice energy exceed the globally minimal value by more than 10 kJ mol^{-1} are immediately discarded from further consideration as they are very unlikely to occur in nature. For each of the remaining N^L locally optimal crystal structures, we compute the distances $d_{tt'}$ for all pairs of atom types (t, t').

Once this information is available, we classify the N^L crystal structures into a smaller number M of distinct clusters. Two structures are considered similar if the differences in the lattice energies U, the lattice densities ρ and the distances $d_{tt'}$ for all atom type pairs (t, t') are smaller than specified tolerances ε^U, ε^ρ, and ε^d, respectively. To avoid wrong allocation of different crystal structures to the same cluster, we employ rather tight tolerances, that is, $\varepsilon^U = 0.05$ kJ mol^{-1}, $\varepsilon^\rho = 0.50$ kg m^3, and $\varepsilon^d = 0.1$ Å.

Like all methods relying on static similarity indices, our algorithm cannot identify the equivalence of structures that could be converted to each other due to the librational motion at room temperature. Different hydrogen bonding definitely denotes two different structures due to the associated high energy barrier. However, structures with similar strong interactions but with different long-range symmetry are often separated by a low energy barrier.[44]

Step 4: Local Minimization without Space Group Constraints

Step 3 of our algorithm classifies the most promising locally optimal structures determined at step 2 into a smaller number of distinct clusters. However, the local minimizations at step 2 were carried out subject to space symmetry constraints. It is, therefore, possible that some of these structures are local minima only with respect to neighboring structures within the same space group. In fact, such points could be saddle points (rather than true local minima) of the lattice energy surface.

This situation can be tested by considering the second-order optimality conditions. These relate to the positive definiteness of the Hessian matrix of the lattice energy 4 with respect to the lattice lengths and angles, l_a, ω_a, $a = 1, \ldots, 3$, and the molecular positions $\hat{\mathbf{r}}_j$, $j = 2, \ldots, Z$, and orientations ϕ_j, θ_j, ψ_j, $j = 1, \ldots, Z$. The Hessian matrix is obtained by centered finite differences applied to the first-order partial derivatives of the lattice enthalpy, all of which are computed analytically in our code. For the purposes of these computations, the lattice deformation W in the objective function is replaced by its equivalent expression $1 - \sum_{a=1}^{3} \cos^2\omega_a + 2 \prod_{a=1}^{3} \cos \omega_a$.

Once the Hessian matrix is obtained, its eigenvalues are computed. If all of these are positive, then the second-order optimality condition is satisfied and the point is indeed a true local minimum of the lattice energy surface. If not, then it is actually a saddle point. In such cases, we apply to it a small perturbation and use the resulting point as an initial guess for a local energy minimization without space group constraints. This involves the objective function 4 being minimized with respect to all $4 + 6Z$ optimization decision variables listed earlier, subject to the nonlinear equality constraint 6 and appropriate lower and upper bounds on all optimization decision variables. This minimization usually (but not always) leads to a local minimum that belongs to a space group of lower symmetry than that of the crystal structure used as the starting point. It is worth noting that this final space group is not necessarily one of the 59 considered at step 1 of the algorithm.

In the computations reported in this article, these tests have been applied to one representative crystal structure from each of the 20 clusters of lowest energy.

Parallelized Implementation

The success of our algorithm in locating all low energy crystal structures relies on its ability to examine a large number of candidate structures (step 1), performing a local energy minimization from each one of them (step 2). A coarse-grained parallel implementation of the algorithm has been developed to accelerate this process. This involves generating and locally minimizing several candidate structures simultaneously using a network of computer processors.

Our implementation comprises one master process, one generator process, and several slave processes, all executing in parallel. The generator process is responsible for deciding the space group of the next candidate structure to be considered, and for creating vectors of the appropriate length using low-discrepancy sequences. A separate sequence of vectors is maintained for each of the 59 space groups under consideration.

The primary task of the slave processes is to carry out a local minimization within space group constraints. This involves obtaining from the generator the space group of the structure to be examined as well as a low-discrepancy vector of the correct dimensionality for this space group. The slave then uses this vector to construct a crystal lattice, performs on it the preliminary feasibility tests and, if these tests are satisfied, proceeds to perform a local minimization as described earlier. It then computes the energy, density, and minimum intermolecular distances for this locally optimal structure, and sends them, together with the optimal unit cell information, to the master. It then repeats the above sequence until it receives a termination signal from the master.

The master process initiates the generator and the slaves. Its main task is to receive locally optimized structures from the slaves and to use them to maintain an up-to-date list of clusters using the procedure described earlier. It also employs a criterion to decide when to terminate steps 1, 2, and 3 of the algorithm; in our current

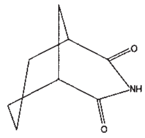
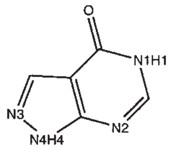
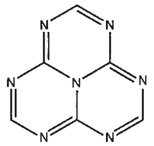
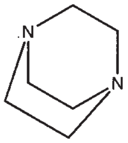
Molecular structure	CSD REFCODE	Experimental crystal characteristics
 3-aza-bicyclo(3.3.1)nonane-2,4-dione	BOQQUT[49] ($T = 297$ K)	P2$_1$/a, $Z = 4.0$, $Z' = 1.0$
 allopurinol	ALOPUR[50] ($T = 283 - 303$ K)	P2$_1$/c, $Z = 4.0$, $Z' = 1.0$
 1,3,4,6,7,9-hexa-azacycl(3.3.3)azine (tri-s-triazine)	BOFBIH[51] ($T = 283 - 303$ K)	Pbca, $Z = 16.0$, $Z' = 2.0$
 Triethylenediamine	TETDAM06[52] ($T = 110$ K)	P6$_3$/m, $Z = 12.0$, $Z' = 0.167$

Figure 2. Test molecules studied.

implementation, this criterion is simply based on an upper bound on elapsed time. Once this criterion is satisfied, the master sends a termination signal to the slaves and generator, and then proceeds to implement step 4 of the algorithm.

The algorithm has been implemented in standard FORTRAN 90, with the interprocess communication being handled *via* the LAM implementation (http://www.lam-mpi.org) of the MPI protocol.[48] This allows the execution of the code on a wide range of multiprocessor machines. The case studies reported in the next section were performed on a 48-processor Beowulf cluster. For small organic molecules of the type considered in this article, this allows the examination of several tens of thousands of candidate structures per day.

Case Studies

To illustrate the various features of the algorithm described in the previous section and its performance, we will now use it to study the set of molecules shown in Figure 2.

Testing Methodology

The algorithm of the last section is applied to each molecule twice, first with $G = 1$, and then with $G = 2$.

Modeling of Intermolecular Interactions

Van der Waals interactions were modeled using the intermolecular potential due to William:[53]

$$U_{ii'}^{\text{vdw}}(r) = (A_iA_{i'})^{1/2}e^{-(B_i+B_{i'})r/2} - \frac{(C_iC_{i'})^{1/2}}{r^6} \qquad (11)$$

To maintain consistency with the methodology used by Williams to fit the parameters in the above potential, all hydrogen atoms are displaced by 0.1 Å along the bond and towards the atom to which they are connected.

Electrostatic interactions were modeled with a distributed charge model comprising atomic charges and, where necessary, satellite charges. The magnitudes of all charges and the positions of the satellite charges were determined by fitting the model to the molecular electrostatic potential on a dense geodesic grid[54] computed at the SCF 6-31G** level. An *ab initio* optimized molecular conformation was also obtained as part of the same calculation. The electrostatic potential was sampled on a geodesic grid starting at a distance of 1.4 times the van der Waals radii and involving 16–32 layers separated by distances of 0.025–0.05 times the van der Waals radii. For the molecules studied here, this results in more than 5000 points being used for the fitting of the distributed charge model.

As is well known, the SCF wave function overestimates the molecular dipole by approximately 10% due to the lack of electron correlation. This inaccuracy has led some researchers to apply an empirical scaling factor of 0.9 to charges computed *via* SCF. However, to maintain consistency with the type of wave function used in the derivation of the Williams potential parameters, we did not use an MP2 wave function nor did we apply a scaling correction to the computed charges.

A Posteriori *Analysis of Low-Energy Crystal Structures*

We used the PLATON software[55] to verify that all low-energy minima allocated to the same cluster by step 3 of our algorithm had the same simulated powder diffraction pattern and the same space group symmetry.

PLATON was also used to determine the space group of each low-energy minimum and the number of molecules Z' in the asymmetric unit. For some molecules, a molecular symmetry element coincides with a symmetry relation present in the space group, thus leading to a space group of higher symmetry than the one used for the generation of the initial candidate structure at step 1 of the algorithm.

The clusters were examined for similarities by visual inspection and, in the case of hydrogen bonded crystals, by analyzing the hydrogen bonding patterns and their corresponding graph sets.[56,57]

Validation of Molecular and Potential Models

The accuracy of the *ab initio*-determined molecular conformation and the model of intermolecular interactions was checked by comparing the known experimental structure with the "minimized experimental" structure. The latter is obtained by performing a local minimization starting from an initial guess generated from the experimental structure.

It is emphasized that the experimentally known crystal structures were used only for the purposes of *a posteriori* checking of some aspects of our calculations. No information derived from these experimental structure was employed by our algorithm for determining the crystal structures of the molecules being studied.

Generation of Initial Point for Experimental Crystal Minimization. The minimization of the experimental crystal requires an initial point ("guess") for the numerical optimization code. The generation of this guess involves replacing each molecule in the experimentally determined asymmetric unit with the quantum mechanically optimized molecule. The position of the center of mass, $\mathbf{r}_j$ and the orientation (ϕ_j, θ_j, ψ_j) of each molecule j in the asymmetric unit is determined by solving the following optimization problem separately for each molecule j in the asymmetric unit:

$$\min_{\mathbf{r}_j,\phi_j,\theta_j,\psi_j} \sum_{i=1}^{N^S} w_i \|\mathbf{r}_{ji} - \mathbf{r}_{ji}^{\text{exp}}\|^2, \quad \forall j = 1, \ldots, Z' \qquad (12)$$

where $\mathbf{r}_{ji} = \mathbf{A}^T(\phi_j, \theta_j, \psi_j)\mathbf{r}_{ji}^{bf} + \mathbf{r}_j$ and $\mathbf{r}_{ji}^{\text{exp}}$, $i = 1, \ldots, N^S$ are the positions of the atoms of molecule j in the *ab initio* optimized and experimental asymmetric units, respectively. The weights w_i can be adjusted so as to bias the optimization towards more accurate matching of the positions of certain atoms. For example, if the experimental crystal structure is determined by means of X-ray crystallography, then there is significant uncertainty in the positions of the hydrogens and these should, therefore, be given very low or zero weights.

Validation of the Ab Initio *Molecular Conformation.* The first step in our validation procedure is to compare the *ab initio* molecular conformation obtained by quantum mechanical energy minimization to the conformations that have been measured in the experimental crystals. For crystal structures involving more than one molecule in the asymmetric unit (i.e., $Z' > 1$), there will generally be a different such conformation for each molecule.

The solution of the optimization problem 12 will have already positioned and rotated the *ab initio* molecular conformation to match as closely as possible the experimentally measured one in the asymmetric unit. We now compute the molecular conformation differences:

$$\delta_j^{\text{MC}} \equiv \sqrt{\frac{\sum_{i\in\mathcal{B}} \|\tilde{\mathbf{r}}_{ji} - \mathbf{r}_{ji}^{\text{exp}}\|^2}{|\mathcal{B}|}}, \quad j = 1, \ldots, Z' \qquad (13)$$

where $\tilde{\mathbf{r}}_{ji}$ denote the atom positions in the *ab initio* conformation at the solution of 12, $\mathcal{B}$ is a subset of the atoms in the molecule and $|\mathcal{B}|$ its cardinality. As has already been mentioned, for experimental structures determined *via* X-ray measurements, it may be appropriate to omit hydrogen atoms from the set $\mathcal{B}$.

Large differences δ_j^{MC} may indicate that the quantum mechanical calculations are subject to large errors, for example, because of the omission of electron correlation effects. More likely for the relatively small molecules studied here, they may indicate that the assumption of molecular rigidity is not justified, which leads to distortions of the molecular conformation in the crystal from the corresponding *in vacuo* conformation.

Comparison of Experimental and Minimized Experimental Crystals. Assuming the molecular conformation differences δ_j^{MC} are acceptably small, the initial positions $\tilde{\mathbf{r}}_j$ determined by solving problem 12 are used as the starting point for a local minimization of lattice energy. This leads to the *minimized experimental structure,* which is characterized by the quantities $\mathbf{R}^*$, $\hat{\mathbf{r}}_j^*$, ϕ_j^*, θ_j^*, ψ_j^*, $j = 1, \ldots, Z$.

A comparison of the minimized experimental structure with the original experimental structure may provide some valuable *a posteriori* insight regarding the accuracy of the model of intermolecular interactions used for a particular molecule. In this article, the following three measures are used to quantify the difference between these two structures:

1. Translational difference of centers of mass of molecules in the asymmetric unit:

$$\delta_j^T \equiv \|\mathbf{R}^*(\hat{\mathbf{r}}_j^* - \hat{\tilde{\mathbf{r}}}_j)\|, \quad \forall j = 1, \ldots, Z' \qquad (14)$$

where ˜ denotes the *ab initio* optimized molecular conformation optimally pasted into the experimental crystal by solving problem 12, ˆ denotes normalized coordinates, and * refers to the minimized experimental crystal.

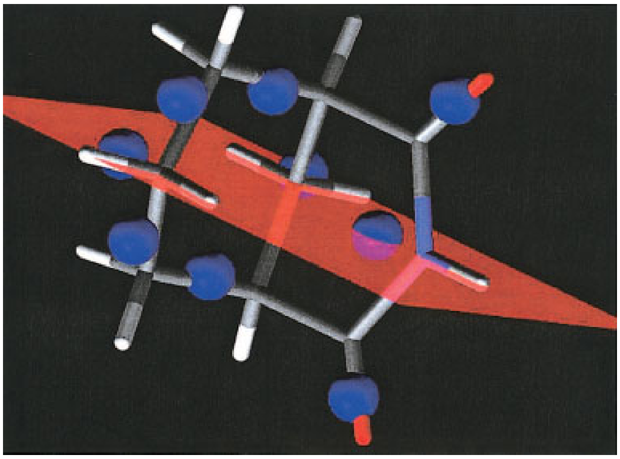


Figure 3. Satellite charge model for 3-aza-bicyclo(3.3.1)nonane-2,4-dione. Spheres correspond to satellite charges; charges are also placed on all atoms. All charge magnitudes and positions of the satellite charges are optimally determined. The molecular plane of symmetry is also shown for reference.

2. Rotational difference of molecules in the asymmetric unit:

$$\delta_j^R \equiv \cos^{-1}\left(\frac{\text{trace}(\mathbf{A}_j^*\tilde{\mathbf{A}}_j^T) - 1}{2}\right), \quad \forall j = 1, \ldots, Z' \qquad (15)$$

3. Mean difference of atomic positions in molecules in the asymmetric unit:

$$\bar{\delta}^A \equiv \sqrt{\frac{1}{Z'}\sum_{j=1}^{Z'}\frac{1}{N_j^A}\sum_{i=1}^{N_j^A}\|\mathbf{r}_{ji}^* - \tilde{\mathbf{r}}_{ji} + (\tilde{\mathbf{R}} - \mathbf{R}^*)\hat{\tilde{\mathbf{r}}}_j\|^2} \qquad (16)$$

where N_j^A is the number of atoms in molecule j.

3-Aza-bicyclo(3.3.1)nonane-2,4-dione

3-Aza-bicyclo(3.3.1)nonane-2,4-dione was one of the molecules studied in the context of the second blind test,[19] held by the Cambridge Crystallographic Data Centre. The blind test rules required the search to be carried out within the most frequently occurring space groups with $Z' \leq 1$.

Molecular Modeling

Our distributed charge model comprises charges placed on the 22 atomic nuclei supplemented by an additional nine satellite charges—one associated with each carbon and nitrogen atom, and is shown in Figure 3. The molecular symmetry (point group Cs) was exploited in the derivation of this model. The error in the electrostatic field description was RMS = 0.309 kcal mol^{-1} (RRMS = 3.063%), which is significantly better than what can be achieved using atomic charges only (RMS = 0.844 kcal mol^{-1} and RRMS = 7.813%).

Crystal Structure Prediction

We performed 103,919 and 66,228 minimizations with $G = 1$ and $G = 2$ crystallographically independent molecules, respectively. Every single one of these local minimizations was successful. This provides a good indication of the robustness of our formulation and algorithm even in the case of the optimization of structures with two crystallographically independent molecules with their larger number of decision variables.

The stationary points obtained during the global optimization are shown in Figure 4. Each point in this graph corresponds to a local stationary point, with its position on the graph indicating the corresponding lattice energy and volume per molecule. The space group shown is the one used for generating the candidate initial guess; the final space group for this point, as determined by PLATON may be different.

The number of new distinct stationary points identified during the $G = 1$ search is shown as a function of the number of local minimizations carried out in Figure 5. It can be seen that, after the first 30,000 local minimizations, no new point is found within 5 kJ mol^{-1} of the global minimum. On the other hand, the search continues to identify new points within 7.5 kJ mol^{-1} of the global minimum until approximately 70,000 local minimizations. In the $G = 2$ search (not shown in the figure), the number of unique stationary points within 5 kJ mol^{-1} from the global minimum stopped growing after approximately 50,000 minimizations, but no such plateau was observed for points within 7.5 kJ mol^{-1} of the global minimum; however, encouragingly, most stable points generated in the $G = 1$ search were also found as supercells or in lower symmetry space groups in the $G = 2$ search (cf. Fig. 4).

The above statistics indicate that the energy surface is particularly ragged, which makes it unlikely that a search involving only a few hundreds or even thousands of structures will manage to identify all low energy minima. This may partly explain some of the failures (reported in the second blind test[19]) to identify the experimental crystal despite the fact that the potential models employed appeared to be sufficiently accurate (as indicated, e.g., by the close agreement between the experimental and minimized experimental structures).

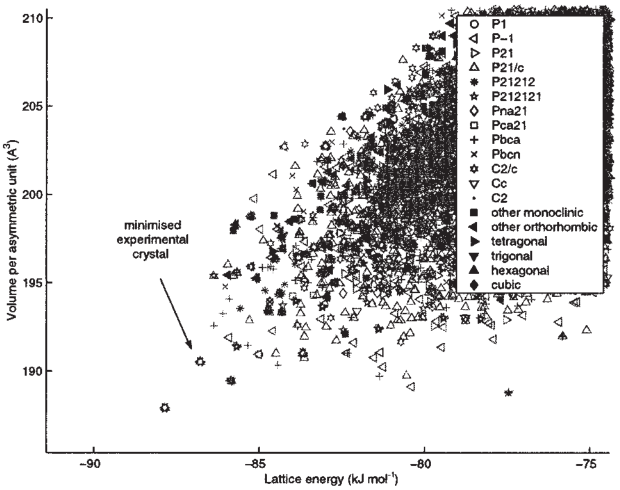


Figure 4. Global optimization of lattice energy for 3-aza-bicyclo(3.3.1)nonane-2,4-dione, black points $G = 1$, gray points $G = 2$.

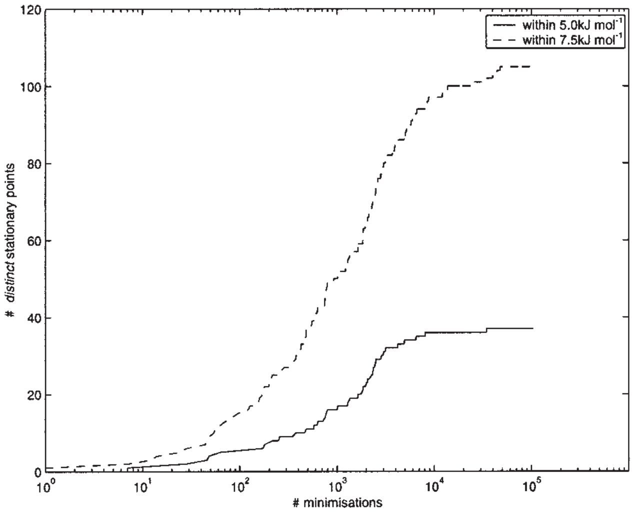


Figure 5. Number of distinct stationary points identified during the $G = 1$ search for 3-aza-bicyclo(3.3.1)nonane-2,4-dione.

Overall, steps 1, 2, and 3 of our algorithm identified 157 and 177 unique stationary points (clusters) within 5 and 7.5 kJ mol^{-1} of the global minimum, respectively. The application of step 4 to the 20 most stable clusters detected that 5 of these were actually saddle points of the energy surface, and could be perturbed and reminimized to lead to more stable crystals of lower symmetry. The 10 most stable clusters identified overall are presented in Table 2.

The global minimum had energy −87.845 kJ mol^{-1} and was identified 65 times from minimizations carried out in P2$_1$/c with $G = 1$, and in P2$_1$ and P2$_1$/c with $G = 2$. The minimized experimental crystal was ranked second, and was found 132 times starting from initial guesses in P2$_1$/c with $G = 1$, and in P$\bar{1}$, P2$_1$, and P2$_1$/c with $G = 2$.

The third minimum belongs to space group Pbca, and has two molecules in the asymmetric unit. Its hydrogen bonding motif comprises chains of finite length with a dimeric structure connected to two other molecules *via* the unsatisfied oxygen acceptor in each of the molecules. Table 2 contains additional $Z' = 2$ structures of low lattice energy, which indicates the increased complexity of the energy landscape.

The global minimum is characterized by hydrogen bonded dimers (cf. Fig. 6) while the experimental structure has a catemeric hydrogen bonding motif. Interestingly, a search of the CSD for molecules that contain the CH—CO—NH—CO—CH group in a ring system, with no other strong hydrogen bond donors or acceptors, revealed both dimeric and catemeric hydrogen bonding motifs.[19] Also, a search for glutarimide moieties substituted with 4-connected C atoms at the 3- and 5-positions of the ring produced five hits (BAHFIZ, LERDIF, PIVFIJ, RERYES, YUFYED), all containing centrosymmetric or pseudocentrosymmetric hydrogen-bonded dimers in the respective crystal structures (see the contribution by Dunitz and Schweizer in the supplementary material of the second blind test article[19]).

On the other hand, another publication[58] studied 88 entries in the CSD that are somewhat related to 3-aza-bicyclo(3.3.1)nonane-2,4-dione. Of these, four molecules were considered to bear strong similarities to this molecule, and three of these formed crystals with catemeric hydrogen-bonded motifs, suggesting that the latter may be kinetically favored.

The different conclusions of the above studies provide a measure of the difficulty of predicting the occurrence of a catemeric hydrogen bonding structure based on a statistical analysis of pre-

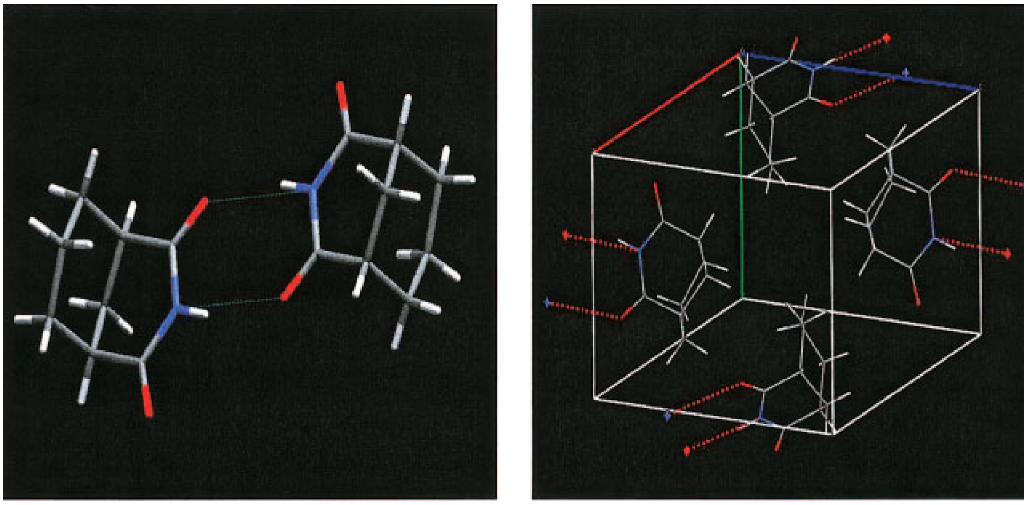
Table 2. Lowest Energy Minima of 3-Aza-bicyclo(3.3.1)nonane-2,4-dione.

				Conventional cell							
Rank	Times found	U (kJ mol^{-1})	$\hat{V}$[b] (Å^3)	Space group	Z'	l_1 (Å)	l_2 (Å)	l_3 (Å)	ω_1	ω_2	ω_3	Hydrogen bonding motif
1	65	−87.845	187.90	P2$_1$/c	1.000	10.402	7.503	9.891	90.00°	103.20°	90.00°	Dimers
2[a]	132	−86.784	190.51	P2$_1$/c	1.000	9.281	10.644	7.741	90.00°	94.78°	90.00°	Infinite chains
3	2	−86.379	192.55	Pbca	2.000	11.945	11.035	23.374	90.00°	90.00°	90.00°	Finite chains
4	14	−86.379	195.40	P2$_1$2$_1$2$_1$	2.000	7.392	10.882	19.435	90.00°	90.00°	90.00°	Infinite chains
5[c]	333	−86.375	195.38	Pc	4.000	10.884	7.387	19.564	90.00°	96.46°	90.00°	Infinite chains
		−85.674	*195.56*	*Pbca*	*1.000*	*10.793*	*7.478*	*19.386*	*90.00°*	*90.00°*	*90.00°*	
6[c]	1	−86.344	193.37	P2$_1$/c	4.000	10.751	7.571	38.051	90.00°	92.69°	90.00°	Infinite chains
		−86.106	*193.23*	*Pbca*	*2.000*	*10.713*	*7.611*	*37.919*	*90.00°*	*90.00°*	*90.00°*	
7[c]	4	−86.244	195.68	P$\bar{1}$	4.000	7.350	10.861	19.703	84.43°	89.87°	89.87′	Infinite chains
		−85.790	*195.38*	*P2$_1$/c*	*2.000*	*7.435*	*19.571*	*10.742*	*90.00°*	*90.22°*	*90.00°*	
8[c]	83	−86.111	191.85	P$\bar{1}$	4.000	7.735	10.584	18.846	90.00°	95.84°	90.00°	Infinite chains
		−85.844	*189.44*	*C2/c*	*1.000*	*19.222*	*10.302*	*8.036*	*90.00°*	*107.75°*	*90.00°*	
9	1	−86.028	194.77	Pbcn	2.000	21.199	7.372	19.941	90.00°	90.00°	90.00°	Infinite chains
10	2	−85.949	196.01	P2$_1$/c	2.000	10.877	7.331	20.768	90.00°	108.80°	90.00°	Infinite chains

[a] Corresponds to the minimized experimental crystal structure.

[b] Volume per molecule.

[c] Determined by the reminimization without space group constraints of the saddle point of the lattice energy surface shown in the next row.



(a) Hydrogen bonded dimer observed at the global minimum

(b) Crystal structure

Figure 6. Global minimum of 3-aza-bicyclo(3.3.1)nonane-2,4-dione. (a) Hydrogen-bonded dimer observed at the global minimum. (b) Crystal structure.

viously experimentally resolved crystal structures. In particular, there seems to be no universally applicable way of quantifying molecular similarity. Of course, it is also possible that a low-energy polymorph of 3-aza-bicyclo(3.3.1)nonane-2,4-dione, involving hydrogen bonded dimers, also exists but has not yet been observed experimentally.

Our results are in good agreement with those by Mooij,[19] who reported that the minimized experimental structure was ranked second, while the global minimum also belonged to space group $P2_1/c$ with lattice constants similar to the ones we found (cf. Table 2). Further evidence of the similarity of the two global minima structures was provided by a comparison of the corresponding simulated X-ray diffraction patterns.

Validation

The molecular conformation difference δ_1^{MC} is 0.037 Å when the hydrogen atoms are excluded, and 0.098 Å when they are included.

A comparison of the experimental and minimized experimental crystal structures is shown in Table 3 and in Figure 7. The minimized experimental structure was also computed using a distributed charge model without satellite charges (see bottom two rows of Table 3). The beneficial effects of introducing satellite charges in terms of improving the accuracy of prediction are evident mostly in the reduction of the δ_1^R error.

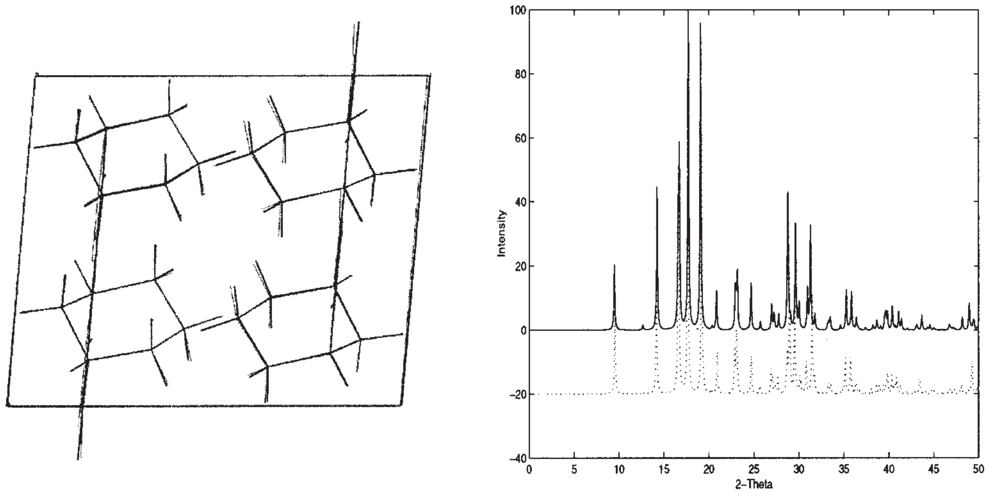
Allopurinol

The prediction of the crystal structure of allopurinol, a drug used for the treatment of gout, has already been studied in the literature.[44] It is a planar heterocyclic molecule with one six-membered and one five-membered ring. Each ring has one hydrogen bond donor NH and one acceptor N. There is also an additional carbonyl acceptor O.

Table 3. Comparison of Experimental and Minimized Experimental Crystal Structures for 3-Aza-bicyclo(3.3.1)nonane-2,4-dione.

	Lattice geometry					H-bonds		Asymmetric Unit		
	l_1 (Å)	l_2 (Å)	l_3 (Å)	ω_2	$\hat{V}$ (Å^3)	N(H) . . . O (Å)	U (kJ mol^{-1})	δ_1^T (Å)	δ_1^R (°)	$\bar{\delta}^A$ (Å)
Exper.	7.705	10.606	9.338	95.03°	190.04	2.974	−86.339[a]/−81.895[b]			
Min. exp.[a]	7.741	10.644	9.281	94.78°	190.505	3.034	−86.784			
Difference[a]	0.47%	0.36%	−0.61%	−0.25°	0.25%	0.060	−0.446	0.0315	1.1479	0.0047
Min. exp.[b]	7.749	10.606	9.338	95.01°	191.80	3.077	−83.137			
Difference[b]	0.57%	1.45%	−1.09%	−0.02°	0.93%	0.103	−1.243	0.0422	2.3596	0.0845

[a] Using atomic and satellite charges, RMS = 0.309 kcal mol^{-1}, RRMS = 3.063%.
[b] Using atomic charges only, RMS = 0.844 kcal mol^{-1}, RRMS = 7.813%.



(a) Superposition of minimised experimental (grey) and experimental (black) crystals

(b) Comparison of the simulated X-ray diffraction pattern ($\lambda = 1.5418$Å) of the minimised experimental (continuous line) and experimental (dotted) structures

Figure 7. Comparison of experimental and minimized experimental crystal for 3-aza-bicyclo(3.3.1)nonane-2,4-dione. The "experimental" crystal results are obtained using an *ab initio* molecular conformation, optimally positioned on the experimental structure using the procedure described in text. (a) Superposition of minimized experimental (gray) and experimental (black) crystals. (b) Comparison of the simulated X-ray diffraction pattern ($\lambda = 1.5418$ Å) of the minimized experimental (continuous line) and experimental (dotted) structures.

Molecular Modeling

The error in the description of the electrostatic field using only atomic charges was RMS = 1.117 kcal mol^{-1} (RRMS = 8.192%). The introduction of one, optimally placed, satellite charge associated with each nitrogen and each carbon atom reduced these significantly to 0.202 kcal mol^{-1} and 1.479%, respectively.

The experimental crystal of allopurinol belongs to the $P2_1/c$ space group and is formed by stacked sheets that present little overlap. Each molecule in a sheet is connected to three other molecules *via* N(H) $\cdots$ N hydrogen bonding; also, one CH hydrogen atom is in close contact with the oxygen atom.

Crystal Structure Prediction

The results of the global optimization are presented in Figure 8 and Table 4. We performed 66,540 minimizations with $G = 1$ and 66,096 with $G = 2$, identifying 32 unique stationary points within 5 kJ mol^{-1} of the global minimum. Both searches reached a clear plateau in the number of distinct stationary points, with no new points being identified within 7.5 kJ mol^{-1} of the global minimum over several thousands of minimizations before the end of the search. Moreover, the $G = 2$ search identified most of the $G = 1$ minima, which provides a further indication that a reasonably wide exploration of the $G = 2$ space has been achieved.

Step 4 of the algorithm applied to the 20 most stable clusters detected negative eigenvalues in the Hessian matrices of clusters 16 and 18, both of which were generated in C2/c with $G = 2$. Their reminimization without symmetry constraints led to lower energy structures, but even these did not rank among the 10 most stable ones presented in Table 4.

Both the global minimum and the experimental crystal belong to the same space group, have the same number of molecules in the asymmetric unit and are characterized by the same hydrogen bonding graph set. Moreover, both involve infinite nonpolar chains formed by N1(H1) $\cdots$ N3 hydrogen bonds (see Fig. 2 for atom numbering). Every molecule in the chains is connected *via* two N4(H4) $\cdots$ N2 hydrogen bonds to a molecule in an adjacent chain. Adjacent chains in the same sheet alternate in direction in both structures. The main difference between the two structures is that the molecules in the global minimum do not form perfect sheets. Instead, some of them have their molecular plane tilted with respect to the others in the same sheet, with the oxygen atom being in close contact to two (rather than one) hydrogen atoms, leading to slightly lower lattice energy.

We have also found three stable crystals with $Z' = 2$, in $Pna2_1$ (second minimum), $P2_1/c$ (third minimum), and C2/c (fifth minimum) that are not characterized by hydrogen bonded sheets. Although their hydrogen bond motif resembles that of the minimized experimental structure, involving the same hydrogen bonding donors and acceptors, the infinite nonpolar chains formed by N1(H1) $\cdots$ N3 hydrogen bonds are connected by N4(H4) $\cdots$ N2 hydrogen bonds to form three-dimensional networks. The third and fifth minima comprise 10-membered hydrogen bonded rings pass-

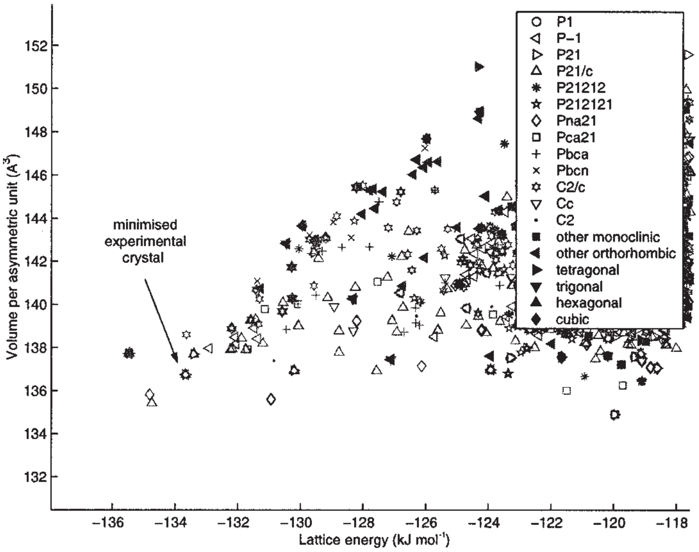


Figure 8. Global optimization of lattice energy for allopurinol, black points $G = 1$, gray points $G = 2$.

ing through each other. Although the structural complexity of these structures is such that their growth may be kinetically hindered, it demonstrates that, for $G \geq 2$, the potential energy surface is characterized by a large number of low energy minima not easily identified by limited search algorithms.

The minima with ranks 6–10 contain sheets with antiparallel N1(H1) · · · N3 hydrogen-bonded chains linked with double N4(H4) · · · N2 hydrogen bonds. Their main differences from the experimental structure is the stacking of the adjacent sheets.

In conclusion, the stable structures predicted by our algorithm indicate that the experimental crystal of allopurinol will be characterized by infinite nonpolar chains formed by N1(H1) · · · N3 hydrogen bonds. The only acceptors that are not satisfied are the oxygen atoms, despite each of them being in close contact with one or two methylene hydrogen atoms. There are several distinct packing motifs of these chains characterized by similar lattice energy and density. In particular, there is the possibility of forming sheets, jagged sheets, or three-dimensional hydrogen bonding networks. In the case of sheet structures, we identified a variety of stackings that can also differ in the directionality of the nonpolar chains on adjacent sheets.

Validation

The molecular conformation difference δ_1^{MC} is 0.0195 Å when the H atoms are excluded; if all atoms are considered, this rises to 0.0664 Å.

A quantitative comparison of the unit cells for the experimental and minimized experimental structures is shown in Table 5. The translational and rotational differences are $\delta_1^T = 0.1146$ Å and δ_1^R

Table 4. Lowest Energy Minima of Allopurinol.

				Conventional cell								
Rank	Times found	U (kJ mol^{-1})	$\hat{V}$[b] (Å^3)	Space group	Z'	l_1 (Å)	l_2 (Å)	l_3 (Å)	ω_1	ω_2	ω_3	Hydrogen bonding motif
1	124	−135.469	137.72	P2_1/c	1.000	3.705	10.864	14.103	90.00°	103.99°	90.00°	Jagged sheets
2	5	−134.813	135.81	Pna2_1	2.000	21.077	3.718	13.864	90.00°	90.00°	90.00°	3D
3	2	−134.742	135.41	P2_1/c	2.000	17.069	3.740	21.019	90.00°	126.16°	90.00°	3D
4[a]	69	−133.667	136.73	P2_1/c	1.000	3.711	14.297	10.407	90.00°	97.83°	90.00°	Sheets
5	2	−133.639	138.59	C2/c	2.000	35.482	3.648	21.386	90.00°	126.77°	90.00°	3D
6	48	−133.405	137.69	P2_1/c	1.000	4.634	14.431	8.271	90.00°	95.23°	90.00°	Sheets
7	3	−132.915	137.96	P$\bar{1}$	2.000	3.788	10.905	13.624	98.53°	94.97°	94.48°	Sheets
8	1	−132.236	137.90	P2_1/c	2.000	8.158	14.464	9.401	90.00°	95.96°	90.00°	Sheets
9	24	−132.193	138.89	Pnma	1.000	11.851	6.486	14.457	90.00°	90.00°	90.00°	Sheets
10	2	−132.177	137.95	P2_1/c	2.000	7.417	14.486	10.405	90.00°	99.20°	90.00°	Sheets

[a] Corresponds to the minimized experimental crystal structure.
[b] Volume per molecule.

Table 5. Comparison of Experimental and Minimized Experimental Crystal Structures for Allopurinol.

	Lattice geometry						Hydrogen bond lengths	
	a (Å)	b (Å)	c (Å)	ω_2	$\hat{V}$ (Å^3)	U (kJ mol^{-1})	N1(H1) · · · N3[b] (Å)	N4(H4) · · · N2[b] (Å)
Exper.	3.683	14.685	10.318	97.47°	138.33	−128.388[a]	2.874	2.881
Min. exp.[a]	3.711	14.297	10.407	97.83°	136.73	−133.666	2.757	2.719
Difference[a]	0.75%	−2.64%	−0.86%	0.36°	−1.15%	−5.278	−0.117	−0.136

[a] Using the satellite charge model, RMS = 0.331 kcal mol^{-1}, RRMS = 3.063%.
[b] Labeling of atoms as in Figure 2.

= 5.1383°, respectively, while the mean atomic position difference is $\bar{\delta}^A$ = 0.1950 Å. Our results regarding lattice energies and the deviation between the experimental and minimized experimental crystal are in good agreement with those obtained earlier[44] using distributed multipoles.

For this particular molecule, electrostatic interactions are dominant, accounting for about 75% of the total lattice energy. This means that even small inaccuracies can have significant impact in our ability to model the solid state. Consequently, the introduction of additional satellite charges for the description of the electrostatic interactions is very important in this case. An attempt to minimize the experimental crystal using a model with only atomic charges was found to lead to a break in symmetry and large errors in the lattice lengths (7.5%) and angles (18.0°).

1,3,4,6,7,9-Hexa-azacycl(3.3.3)azine

This molecule is particularly interesting in that the experimentally observed crystal structure has more than one molecule in its asymmetric unit (Z' = 2).

Molecular Modeling

The molecule formed part of the training set for the development of the Williams potential. The comparison between the reported[53] experimental and minimized experimental crystal structures indicates that the description of van der Waals interactions is reasonably accurate.

The electrostatic interactions were modeled using the procedure stated earlier, with the electrostatic field being sampled at 10,032 points positioned on a geodesic mesh. A distributed charge model based only on atomic charges was found to exhibit errors of RMS = 0.558 kcal mol^{-1} and RRMS = 4.823%. Introducing satellite charges associated with all carbon atoms and all nitrogen atoms (with the exception of the one at the inversion center) reduced these errors to 0.172 kcal mol^{-1} and 1.490%, respectively. The maximum error in the charge magnitudes was estimated as 0.015 atomic charge units. In this case, the derivation of the optimal site charge model was facilitated by the high molecular symmetry (D3h), which implies that, in reality, there are only four independent satellite charges.

Crystal Structure Prediction

The crystal structure prediction algorithm minimized 22,739 candidate structures with G = 1 and 48,099 with G = 2. The energy surface appears to be considerably simpler to those for the previous two case studies, with only 3 and 14 distinct structures identified within 5 and 7.5 kJ mol^{-1} of the global minimum, respectively. No new distinct low-energy stationary points were identified during the last 15,000 minimizations in either the G = 1 or the G = 2 searches; this indicates that the latter are reasonably complete.

The results of the global minimization are shown in Figure 9 and in Table 6. Because of the high molecular symmetry of this molecule, the same crystal packings can be achieved in different space groups, even involving different numbers of molecules in the asymmetric unit. Thus, many clusters contained local minima produced by minimizations performed in different space groups and/or values of G. For example, the lowest energy structure was found 1002 times starting from initial guesses generated in several different space groups, including P$\bar{1}$ (G = 1), P2$_1$ (G = 1), P2$_1$/c (G = 1), and P6$_1$ (G = 1). It is characterized by parallel sheets at distance 3.022 Å with no stacking of the molecules.

The minimized experimental structure was ranked second, and was identified 81 times. In contrast to the global minimum, this structure is not characterized by parallel planes; instead, two of the closest neighbors of each molecule have their planes perpendicular to it (T-shaped configuration). This structure is almost 4 kJ mol^{-1} less stable than the global minimum.

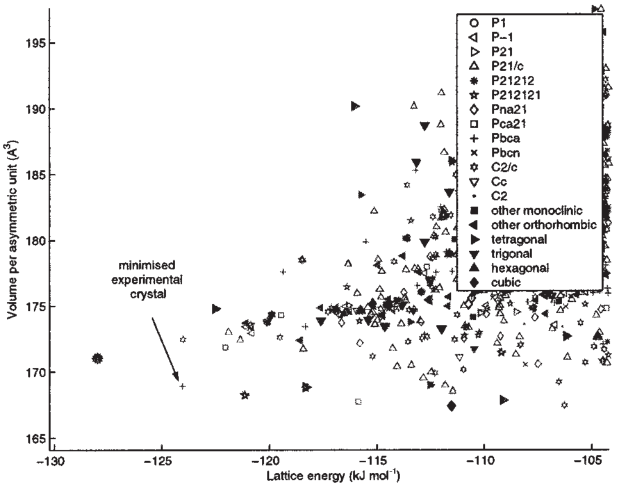


Figure 9. Global optimization of lattice energy for 1,3,4,6,7,9-hexa-azacycl(3.3.3)azine, black points G = 1, gray points G = 2.

Table 6. Lowest Energy Minima of 1,3,4,6,7,9-Hexa-azacycl(3.3.3)azine.

			Conventional cell								
Rank	Times found	U (kJ mol^{-1})	$\hat{V}$[b] (Å^3)	Space group[c]	Z'	l_1 (Å)	l_2 (Å)	l_3 (Å)	ω_1 (°)	ω_2 (°)	ω_3 (°)
1	1002	−127.990	171.03	P6$_3$/m	0.167	8.082	8.082	6.044	90.00°	90.00°	120.00°
2[a]	81	−124.044	168.90	Pbca	2.000	13.855	7.106	27.452	90.00°	90.00°	90.00°
3	1	−124.012	172.46	C2/c	2.000	13.975	8.060	24.570	90.00°	94.42°	90.00°
4	78	−122.515	174.78	I4$_1$/a	1.000	17.407	17.407	9.230	90.00°	90.00°	90.00°
5[c]	15	−122.174	172.85	P$\bar{1}$	4.000	8.045	8.057	24.682	85.10°	88.07°	60.18°
		−121.905	*173.00*	*P2$_1$/c*	*2.000*	*8.062*	*24.618*	*8.036*	*90.00°*	*119.79°*	*90.00°*
6	10	−122.043	171.85	Pca2$_1$	2.000	14.500	14.041	6.753	90.00°	90.00°	90.00°
7	49	−121.351	172.42	P2$_1$/c	2.000	15.118	6.946	14.807	90.00°	90.00°	117.49°
8	1339	−121.140	168.24	P2$_1$2$_1$2$_1$	1.000	5.042	7.265	18.373	90.00°	90.00°	90.00°
9	15	−121.097	173.70	Iba2	2.000	28.227	14.468	6.805	90.00°	90.00°	90.00°
10	21	−121.089	173.43	R$\bar{3}$	0.667	8.049	8.049	37.096	90.00°	90.00°	120.00°

[a] Corresponds to the minimized experimental crystal structure.

[b] Volume per molecule.

[c] Determined by the reminimization without space group constraints of the saddle point of the lattice energy surface shown in the next row.

Validation

When no hydrogens are included in the set $\mathcal{B}$ [cf. eq. (13)], the molecular conformation differences δ_j^{MC} are 0.047 Å and 0.043 Å for the two molecules in the asymmetric unit, respectively. When the hydrogens are included in the set $\mathcal{B}$, these increase to 0.116 Å and 0.082 Å, respectively.

The results of the minimization of the experimental crystal are presented in Table 7. Local minimizations were carried out in P1 and in the space group Pbca (to which the experimentally observed crystal belongs), and both were found to lead to the same minimum. The lattice energies reported for the experimental crystal are those computed using our distributed charge model. The largest error in lattice lengths was 1.65% for l_1 (cf. Table 7).

Table 7 also shows results obtained for a molecular model that includes atomic charges only. In this case, the mean positional change of the asymmetric unit caused by the minimization $\bar{\delta}^A$ [cf. eq. (16)] was 0.1959 Å compared with 0.1602 Å when additional optimally placed satellite charges are introduced.

Triethylenediamine

Molecular Model

In this case, we introduce one satellite charge associated with each carbon and each nitrogen atom. Atomic charges lead to an error in the description of the electrostatic field of RMS = 1.930 kcal mol^{-1} (RRMS = 32.256%); the satellite charge model significantly reduces this to RMS = 0.266 kcal mol^{-1} (RRMS = 4.440%).

Crystal Structure Prediction

We performed 40,000 and 46,959 minimizations with one (G = 1) and two (G = 2) crystallographically independent molecules. Of these, only two initial guesses failed to reach a valid minimum. The search in the G = 1 space appears to be complete, as no new stationary point within 7.5 kJ mol^{-1} was identified during the last 15,000 minimizations. The computationally more demanding G = 2 search is less exhaustive, but the number of unique stationary

Table 7. Comparison of Experimental and Minimized Experimental Crystal Structures of 1,3,4,6,7,9-Hexa-azacycl(3.3.3)azine.

	Lattice geometry					Asymmetric unit		
	l_1 (Å)	l_2 (Å)	l_3 (Å)	$\hat{V}$ (Å^3)	U (kJ mol^{-1})	δ_1^T/δ_2^T (Å)	δ_1^R/δ_2^R (°)	$\bar{\delta}^A$ (Å)
Exper.	7.225	27.193	13.858	170.17	−122.544[a]/−123.311[b]			
Min. exp.[a]	7.105	27.452	13.855	168.90	−124.042			
Difference[a]	−1.65%	0.95%	−0.02%	−0.74%	−1.498	0.1299/0.0959	3.8405/1.9670	0.1602
Min. exp.[b]	7.199	27.357	13.680	168.39	−125.402			
Difference[b]	−0.36%	0.60%	−1.28%	−1.04%	−2.090	0.1740/0.0295	5.8508/2.7447	0.1959

[a] Using atomic and satellite charges, RMS = 0.172 kcal mol^{-1}, RRMS = 1.490%.

[b] Using atomic charges only, RMS = 0.558 kcal mol^{-1}, RRMS = 4.823%.

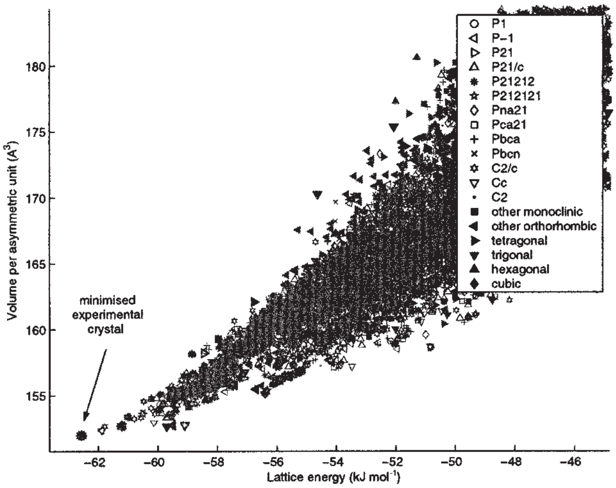


Figure 10. Global optimization of lattice energy for triethylenediamine, black points $G = 1$, gray points $G = 2$.

points identified also seems to have reached a plateau. Overall, we identified 245 and 867 unique stationary points within 5 and 7.5 kJ mol^{-1} from the global minimum, respectively.

The results of the global optimization are shown in Figure 10 and Table 8. The global minimum had energy of −62.569 kJ mol^{-1} and was identified 1088 times starting from candidate structures in many space groups with both $G = 1$ and $G = 2$ because of the high molecular symmetry. The global minimum corresponds to the experimental crystal, and is approximately 0.3 kJ mol^{-1} more stable than the second most stable crystal.

When examined with PLATON, the third lowest energy cluster formed by our code was found to contain structures belonging to two different space groups (shown as two separate rows in Table 8); although they have equal energy, density, and closest contacts, and in both cases the N–N axes of the molecules were parallel to the c axis, their powder patterns were clearly distinct. This illustrates the difficulties associated with the clustering of similar structures in crystal structure prediction—a problem that is aggravated further when one carries out a more extensive search in a large number of space groups with more than one crystallographically independent molecule.

Minima 2 and 5 were obtained by the reminimization without space group constraints of saddle points obtained with two crystallographically independent molecules in P2$_1$/c and C2/c, respectively. In both cases, the examination of the final crystals with PLATON revealed triclinic cells. In particular, the second minimum had slightly higher energy than that of the global minimum. Its visual inspection showed that the molecules have their N–N axes aligned and are forming layers similar to the experimental. However, the spatial arrangement of the layers in the two cases is different.

Validation

The molecular conformation difference δ_1^{MC} is 0.0229 Å for the nonhydrogen atoms, and 0.1158 Å when all atoms are included.

The details of the minimization of the experimental structure in space group P$\bar{1}$ ($Z' = 1$) are shown in Table 9. Minimizations in both P1 ($Z' = 2$) and in P$\bar{1}$ ($Z' = 1$) lead to the same minimum. The experimental crystal P6$_3$/m ($Z' = 0.167$) can also be seen as P$\bar{1}$ ($Z' = 1$). The use of an optimal satellite charge model leads to smaller deviations between the experimental and the minimized experimental crystals.

Table 8. Lowest Energy Minima of Triethylenediamine.

				Conventional cell							
Rank	Times found	U (kJ mol^{-1})	$\hat{V}$[b] (Å^3)	Space group[c]	Z'	l_1 (Å)	l_2 (Å)	l_3 (Å)	ω_1 (°)	ω_2 (°)	ω_3 (°)
1[a]	1088	−62.569	151.97	P6$_3$/m	0.167	6.056	6.056	9.569	90.00	90.00	120.00
2[c]	3	−62.227	152.15	P1	16.000	10.481	6.051	38.545	90.00	95.20	90.00
		−60.796	*153.23*	*C2/c*	*2.000*	*6.069*	*10.454*	*38.647*	*90.00*	*90.80*	*90.00*
3[d]	17	−61.886	152.34	P$\bar{6}$2c	0.333	6.046	6.046	19.248	90.00	90.00	120.00
				R$\bar{3}$	0.667	6.046	6.046	57.752	90.00	90.00	120.00
4	2	−61.788	152.66	P$\bar{3}$c1	0.667	6.054	6.054	38.483	90.00	90.00	120.00
5[c]	1	−61.535	152.81	P$\bar{1}$	4.000	6.051	6.051	38.718	85.52	85.52	60.00
		−61.224	*152.58*	*P2$_1$/c*	*2.000*	*6.033*	*38.571*	*6.051*	*90.00*	*119.90*	*90.00*
6	869	−61.209	152.69	R$\bar{3}$c	0.167	6.035	6.035	29.048	90.00	90.00	120.00
7	230	−61.007	153.35	P$\bar{3}$c1	0.333	6.051	6.051	19.345	90.00	90.00	120.00
8	1	−60.569	153.32	R$\bar{3}$c	0.667	6.035	6.035	116.662	90.00	90.00	120.00
9	27	−60.501	153.65	P6$_3$	0.667	6.045	6.045	19.421	90.00	90.00	120.00
10	58	−60.477	154.54	C2/c	2.000	20.481	12.269	12.104	90.00	125.61	120.00

[a] Corresponds to the minimized experimental crystal structure.

[b] Volume per molecule.

[c] Determined by the reminimization without space group constraints of the saddle point of the lattice energy surface shown in the next row.

[d] Cluster contains two distinct types of crystals, shown in the two rows here.

Table 9. Comparison of Experimental and Minimized Experimental Crystal for Triethylenediamine.

	Lattice geometry								Asymmetric Unit		
	l_1 (Å)	l_2 (Å)	l_3 (Å)	ω_1	ω_2	ω_3	$\hat{V}$ (Å^3)	U (kJ mol^{-1})	δ^T (Å)	δ^R (°)	$\bar{\delta}^A$ (Å)
Exper.	6.080	6.080	9.340	90.00°	90.00°	120.00°	149.51	−61.722[a]/−55.800[b]			
Min. exp.[a]	6.056	6.056	9.569	90.00°	90.00°	120.00°	151.97	−62.569			
Difference[a]	−0.39%	−0.39%	2.45%	0.00°	0.00°	0.00°	1.65%	−0.847	0.0012	1.5570	0.0467
Min. exp.[b]	6.047	6.047	9.716	90.00°	90.00°	120.00°	153.83	−58.793			
Difference[b]	−0.55%	−0.55%	4.03%	0.00°	0.00°	0.00°	2.89%	−2.993	0.0012	8.4677	0.2541

[a]Using atomic and satellite charges, RMS = 0.266 kcal mol^{-1}, RRMS = 4.440%.

[b]Using atomic charges only, RMS = 1.930 kcal mol^{-1}, RRMS = 32.256%.

Concluding Remarks

This article has presented a systematic methodology for the identification of low-energy crystal structures formed by rigid molecules. The key elements of the proposed approach are an automatically derived description of the electrostatic field[25] that is both accurate and efficient; the use of a novel search procedure for systematically searching the domain of possible structures, and the use of this procedure in a parallelized computer environment to examine many tens of thousands of candidate structures; the mathematical formulation of the optimization problem and the accurate computation of a twice-differentiable lattice energy function in such a way that practically no failures occur during the local energy minimization; and the use of a state-of-the-art local optimization code coupled with analytically evaluated partial derivatives, tight optimization tolerances, and *a posteriori* checks for saddle points so that only true local minima are detected. Although, from the theoretical point of view, some of the above considerations may appear to be more interesting than others, our experience indicates that, in practice, they are all essential for ensuring a robust and reliable crystal structure prediction algorithm.

Overall, the proposed methodology appears to be successful. In all cases studied and despite the very extensive search carried out, the experimentally observed structure is among the few most stable predicted structures. This is primarily an indication of the accuracy of the model of intermolecular interactions employed. Of course, the low-energy crystal structures need to be further examined, taking into account other aspects, such as kinetic, thermodynamic, and mechanical stability factors.

The theoretical presentation and the practical results in this article were based on the assumption that all molecules in the unit cell are identical. The extension to cover crystals involving different types of molecule in the unit cell (e.g., cocrystals, hydrates, and solvates) is straightforward.

Despite its generality, the proposed methodology still relies on the use of external information for the description of the van der Waals interactions. Although this is available for many molecules, this is not always the case. Moreover, even when available, this information has often been derived by fitting minimized experimental crystal structures to observed experimental structures for a range of molecules; this was, for example, the approach adopted in deriving[53] the potentials that we used in our work. However, the parameters obtained in any such fitting exercise will also depend on the description of the electrostatic interactions used in the molecular model. Thus, some refitting may be appropriate to derive values that are consistent with the more accurate electrostatic descriptions used here.

Finally, a key assumption underpinning the work presented here is molecular rigidity. Fortunately, it is possible to extend the approach[24] to cover flexible molecules in which intramolecular energy is a function of internal degrees of freedom such as flexible torsion and bond angles. In this case, the molecular conformation of each molecule in the asymmetric unit has to be determined simultaneously with the position and orientation of the molecule, as well as the overall unit cell geometry.

References

1. Gdanitz, R. J. Curr Opin Solid State M 1998, 3, 414.
2. Verwer, P.; Leusen, F. J. J. In Reviews in Computational Chemistry; Lipkowitz, K. B.; Boyd, D. B., Eds.; John Wiley & Sons: New York, 1998, vol 12, chap 7.
3. Holden, J. R.; Du, Z. Y.; Ammon, H. L. J Comput Chem 1993, 14, 422.
4. Allen, F. H. Acta Crystallogr 2002, B58, 380.
5. Allen, F. H.; Motherwell, W. D. S. Acta Crystallogr 2002, B58, 407.
6. Busing, W. R. WMIN, a Computer Program to Model Molecules and Crystals in Terms of Potential Energy Functions; Report ORNL-5747, Oak Ridge National Laboratory, Oak Ridge, TN, 1981.
7. Willock, D. J.; Price, S. L.; Leslie, M.; Catlow, C. R. J Comput Chem 1995, 16, 628.
8. Stone, A. J.; Alderton, M. Mol Phys 1985, 56, 1047.
9. Gavezotti, A. J Am Chem Soc 1991, 113, 4622.
10. Dzyabchenko, A. V.; Agafonov, V.; Davydov, V. A. J Phys Chem A 1999, 103, 2812.
11. Dzyabchenko, A. V. PMC (Version July 2001). User's Guide; Karpov Institute of Physical Chemistry: Moscow, 2001.
12. Schmidt, M. U.; Englert, U. J Chem Soc Dalton Trans 1996, 10, 2077.
13. van Eijck, B. P.; Kroon, J. J Comput Chem 1999, 20, 799.
14. Mooij, W. T. M.; van Duijneveldt, F. B.; van Duijneveldt–van de Rijdt, J. G. C. M.; van Eijck, B. P. J Phys Chem A 1999, 103, 9872.
15. Mooij, W. T. M.; van Eijck, B. P.; Kroon, J. J Phys Chem A 1999, 103, 9883.
16. Lii, J. H.; Allinger, N. L. J Phys Org Chem 1994, 7, 591.

17. Mooij, W. T. M.; van Eijck, B. P.; Kroon, J. J Am Chem Soc 2000, 122, 3500.
18. van Eijck, B. P.; Mooij, W. T. M.; Kroon, J. J Comput Chem 2001, 22, 805.
19. Motherwell, W. D. S.; Ammon, H. L.; Dunitz, J. D.; Dzyabchenko, A.; Erk, P.; Gavezzotti, A.; Hoffmann, D. W. M.; Leusen, F. J. J.; Lommerse, J. P. M.; Mooij, W. T. M.; Price, S. L.; Scheraga, H.; Schweizer, B.; Schmidt, M. U.; van Eijck, B. P.; Verwer, P.; Williams, D. E. Acta Crystallogr 2002, B58, 647.
20. Gavezotti, A. Faraday Discuss 1997, 106, 63.
21. Gavezotti, A.; Filippini, G.; Kroon, J.; van Eijck, B. P.; Klewinghaus, P. Chem Eur J 1997, 3, 893.
22. van Eijck, B. P. J Comput Chem 2001, 22, 816.
23. Beyer, T.; Day, G. M.; Price, S. L. J Am Chem Soc 2001, 123, 5086.
24. Karamertzanis, P. G.; Pantelides, C. C., in prep, 2005.
25. Karamertzanis, P. G.; Pantelides, C. C. Mol Sim 2004, 30, 413.
26. Piermarini, G. J.; Mighell, A. D.; Weir, C. E.; Block, S. Science 1969, 165, 1250.
27. Thiéry, M. M.; Léger, J. M. J Chem Phys 1988, 89, 4255.
28. Ciabini, L.; Santoro, M.; Bini, R.; Schettino, V. J Chem Phys 2001, 115, 3742.
29. Gibson, K. D.; Scheraga, H. A. J Phys Chem 1995, 99, 3752.
30. van Eijck, B. P.; Kroon, J. J Phys Chem B 1997, 101, 1096.
31. van Eijck, B. P.; Kroon, J. Acta Crystallogr 2000, B56, 535.
32. Theodorou, D. N.; Suter, U. W. Macromolecules 1985, 18, 1467.
33. Karamertzanis, P. G. Prediction of Crystal Structure of Molecular Solids, Ph.D. thesis, University of London, 2004.
34. Catti, M. Acta Crystallogr 1978, A34, 974.
35. Kitaigorodsky, A. J. Molecular Crystals and Molecules; Academic Press: New York, 1973.
36. van Eijck, B. P.; Mooij, W. T. M.; Kroon, J. Acta Crystallogr 1995, B51, 99.
37. Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. Numerical Recipes; Cambridge University Press: Cambridge, 1992.
38. Sobol', I. M. Comput Math Math Phys 1967, 7, 86.
39. L'Ecuyer, P. Commun ACM 1988, 31, 742.
40. Kempster, C. J. E.; Lipson, H. Acta Crystallogr 1972, B28, 3674.
41. Levitt, M. Annu Rev Biophys Bioeng 1982, 11, 251.
42. Krivy, I.; Gruber, B. Acta Crystallogr 1976, A32, 297.
43. Lepage, Y. J Appl Crystallogr 1982, 15, 255.
44. Price, S. L.; Wibley, K. S. J Phys Chem A 1997, 101, 2198.
45. Lommerse, J. P. M.; Motherwell, W. D. S.; Ammon, H. L.; Dunitz, J. D.; Gavezzotti, A.; Hoffmann, 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. Acta Crystallogr 2000, B56, 697.
46. Dzyabchenko, A. V. Acta Crystallogr 1994, B50, 414.
47. van Eijck, B. P.; Kroon, J. J Comput Chem 1997, 18, 1036.
48. Gropp, W.; Lusk, E.; Skjellum, A. Using MPI; The MIT Press: Cambridge, MA, 1999.
49. Howie, R. A.; Skakle, J. M. S. Acta Crystallogr 2001, E57, o822.
50. Prusiner, P.; Sundaralingam, M. Acta Crystallogr 1972, B28, 2148.
51. Hosmane, R. S.; Rossman, M. A.; Leonard, N. J. J Am Chem Soc 1982, 104, 5497.
52. Sauvajol, J. L. J Phys C Solid Stat Phys 1980, 13, 927.
53. Williams, D. E. J Comput Chem 2001, 22, 1154.
54. Spackman, M. A. J Comput Chem 1996, 17, 1.
55. Spek, A. L. PLATON, A Multipurpose Crystallographic Tool; Utrecht University: Utrecht, The Netherlands, 2003.
56. Etter, M. C. Acc Chem Res 1990, 23, 120.
57. Etter, M. C.; MacDonald, J. C.; Bernstein, J. Acta Crystallogr 1990, B46, 256.
58. Sarma, J. A. R. P.; Desiraju, G. R. Cryst Growth Des 2002, 2, 93.