Acta Cryst B
STRUCTURAL SCIENCE
CRYSTAL ENGINEERING
MATERIALS
ISSN 2052-5206

Accurate and efficient representation of intramolecular energy in *ab initio* generation of crystal structures. Part III: partitioning into torsional groups

Isaac J. Sugden, David H. Bowskill, Benjamin I. Tan, Yizu Zhang, Claire S. Adjiman* and Constantinos C. Pantelides

Department of Chemical Engineering, Sargent Centre for Process Systems Engineering, Institute for Molecular Science and Engineering, Imperial College London, London SW7 2AZ, United Kingdom. *Correspondence e-mail: c.adjiman@imperial.ac.uk

Received 18 June 2024
Accepted 15 October 2024

Edited by R. Černý, University of Geneva, Switzerland

Keywords: crystal structure prediction; computational chemistry; flexible molecules; energy partitioning.

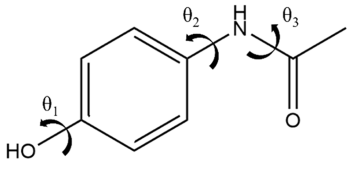
Supporting information: this article has supporting information at journals.iucr.org/b

We present an approach to reduce this computational cost substantially, based on the partitioning of the molecule into geometrically separated torsional groups, with the dependence of the intramolecular energy and atomic point charges and dependent degrees of freedom on molecular conformation being computed as a linear combination of the contributions of these groups. This can lead to large savings in computational cost without a significant impact on accuracy, as demonstrated in the cases of *N*-acetyl-*para*-aminophenol (paracetamol) and methyl 4-hydroxybenzoate (methyl paraben). The approach is also applied successfully to two larger molecules, benzyl [4-(4-methyl-5-[(4-methylphenyl)sulfonyl]-1,3-thiazol-2-yl)phenyl]carbamate (molecule XX from the fifth CSP blind test) and (2*S*)-2-[4-(3-fluorobenzyloxy)benzylamino]propionamide (safinamide), for which we conduct the first reported CSP study. In both cases, the use of torsional groups results in over 99% reduction in computational cost, which enables the generation of an initial CSP landscape with high-quality structures found within the standard cutoff of 20 kJ mol^{-1} for progression to refinement.

1. Introduction

Crystal structure prediction (CSP) methods seek to provide a short, yet complete, list of experimentally realizable crystal structures for a given molecule or set of molecules, solely based on knowledge of the relevant molecular connectivity diagram(s). The *ab initio* determination of the crystalline structures that a molecule can form could be very impactful for several industrial sectors, enabling the faster development of manufacturing processes and de-risking the production, distribution and storage of crystalline products. For the pharmaceutical industry, for instance, a key motivation is avoiding the catastrophic withdrawal of a drug resulting from the appearance of a previously unknown, less soluble, polymorph, as was the case for ritonavir (Chemburkar *et al.*, 2000), DPC 961 (Rietveld & Céolin, 2015) and rotigotine (Rietveld & Céolin, 2015; European Medicines Agency, 2005). There are likely many more examples of solid form issues affecting drug product development that are not in the public domain, as pharmaceutical companies are motivated to exhaustively explore the solid form landscape by these cases (Lee *et al.*, 2011).

The field of crystal structure prediction has witnessed a great deal of progress in recent years, as shown by the increasing size and flexibility of targets in the blind tests



OPEN ACCESS
Published under a CC BY 4.0 licence

organized by the Cambridge Crystallographic Data Centre (CCDC) (Day *et al.*, 2009; Bardwell *et al.*, 2011; Reilly *et al.*, 2016). In CSP methods, it is generally assumed that viable crystals are low-energy local minima with respect to the cell variables [lengths and angles, molecule position(s) and orientation(s)], and any internal (conformational) degrees of freedom. In most cases, given the small contribution of entropic effects to the total free energy, coupled with the high cost of evaluating entropic contributions, lattice energy is assumed to suffice for the ranking of crystal structures (Bowskill *et al.*, 2021; Francia *et al.*, 2020; Abraham & Shirts, 2020).

Most successful CSP methodologies use a multistage approach (Day *et al.*, 2009; Bowskill *et al.*, 2021), in which a large set of candidate structures is initially generated using a relatively simple energy model, followed by successive refinements of the most promising (lowest energy) candidates carried out using increasingly accurate models. This stage-wise framework has been widely adopted because, for the molecules that have typically been studied, covering the full breadth of possible crystal structures necessitates 10^5–10^6 minimizations in the initial global search step, which would render the use of the most accurate and expensive energy models prohibitively expensive. It is, therefore, essential that the initial global search uses a lattice energy model that is computationally efficient enough to perform millions of minimizations, whilst being sufficiently accurate to ensure that the low-energy forms are not excluded from the next refinement step. Several approaches have been proposed to derive a lattice energy model, or potential, all of which include some degree of customization to the molecule(s) of interest (Bowskill *et al.*, 2021).

The lattice energy model used at this stage could include a general force field, such as the Dreiding (Mayo *et al.*, 1990), Compass (Sun *et al.*, 1998) or Charm (MacKerell *et al.*, 2000) forcefields, as was common in the first blind test of crystal structure prediction (Lommerse *et al.*, 2000). This approach has not generally been successful in blind tests, as the energy model is not tailored to the molecule of interest, and the subtle differences in stability between crystals structures are not described accurately enough for relevant crystal structures to be progressed to the more computationally intense ranking stages. The improvement of individual force field parameters by tailoring with Quantum Mechanical (QM) calculations has a long precedent (Schmidt *et al.*, 2007), and extending this methodology to using QM calculations on crystal structures, and extending this methodology to using QM calculations on all the low energy crystal structures, often termed the Ψ_{crys} method (Price, 2018), has been shown to be a successful approach under blind test conditions. The program *GRACE* (Neumann *et al.*, 2008) for instance, has been very successful (Reilly *et al.*, 2016), and makes use of tailor-made force fields in which the traditional force field terms (angle bends, intramolecular van der Waals terms *etc*.) are parameterized by QM calculations. A similar approach is used in the more recent methodology adopted in *XtalPi* (Zhang *et al.*, 2018). Such approaches require access to very significant high performance computing resources [*e.g.* in the sixth blind test, groups making use of periodic DFT methods routinely, required millions CPU hours for a given molecule (Reilly *et al.*, 2016)] and, due to this high cost, can often only be deployed with relatively simple levels of theory/basis set, leading to potential inaccuracies or incomplete global searches (Nyman *et al.*, 2019). Another class of methods, often termed the Ψ_{mol} method (Price, 2018), consists in tailoring some of the force field parameters to QM calculations on the isolated molecule, while using transferable parameters for some terms in the force field. Such an approach, initially adopted by van Eijck *et al.* (2001) and in *DMACRYS* (Price *et al.*, 2010), has been further developed in the *CrystalPredictor* suite of global search codes (Karamertzanis & Pantelides, 2005, 2007; Habgood *et al.*, 2015; Sugden *et al.*, 2016, 2019) and in the rigid-molecule global search code GLEE (Case *et al.*, 2016; Yang & Day, 2022). These approaches have been successfully applied in some challenging global searches (Beran *et al.*, 2022; Kazantsev, Karamertzanis, Adjiman, Pantelides *et al.*, 2011). *CrystalPredictor II* (Habgood *et al.*, 2015; Sugden *et al.*, 2016, 2019) has been applied successfully to a wide range of systems (Wade *et al.*, 2022; Tchoń *et al.*, 2021; Schmidt *et al.*, 2021; Racher *et al.*, 2023; Pawlak *et al.*, 2021; Braun *et al.*, 2019; Shunnar *et al.*, 2020; Braun *et al.*, 2021), including flexible molecules and co-crystals (Sugden *et al.*, 2022). In the latter case, it has proven to be especially efficient as isolated-molecule QM calculations can be reused in modelling a co-crystal, thereby avoiding numerous calculations. Nevertheless, the applicability of *CrystalPredictor II* to molecules with more than six to seven flexible torsions is limited as the computational cost increases exponentially with the number of such torsions. To address this, this manuscript focuses on developments for the program *CrystalPredictor*, improving its efficiency further without compromising its accuracy. In the *CrystalPredictor II* energy model there are no traditional forcefield terms beyond repulsion/dispersion, and minor internal degrees of freedom such as angle bends are updated as a function of the major degrees of freedom such as torsions, as will be described.

In a variation of an approach first suggested by van Eijck *et al.* (2001), *CrystalPredictor II* uses isolated-molecule QM calculations to create customized models for the intramolecular energy, selected (dependent) conformational variables and the electrostatic potential as functions of a set of independent degrees of freedom that usually comprises the most flexible torsion angles. The concept of independent/dependent variables is a mathematical construct, rather than a chemical description, and reduces the computational cost without compromising accuracy. These tailored models are referred to as Local Approximate Models (LAMs) (Kazantsev *et al.*, 2010) and are initially constructed by performing QM calculations at fixed values of the independent degrees of freedom (LAM reference points) placed on a uniform grid (Habgood *et al.*, 2015). Conformationally dependent properties (energy, electrostatic potential, dependent degrees of freedom) are then approximated via Taylor expansions around the nearest LAM reference point. It has been shown that, within their

area of validity, LAMs provide near-QM accuracy for the conformationally dependent properties (Sugden *et al.*, 2016), at a much reduced computational cost relative to a full QM calculation. Recent advances have allowed the bulk of the computational effort to be focused in areas of crystallographic relevance through an adaptive LAM algorithm (Part I: Sugden *et al.*, 2016), and have led to improvements in efficiency and accuracy through the introduction of a smoothed LAM potential (Part II: Sugden *et al.*, 2019).

However, a difficulty with this approach is that the number of grid points required rises exponentially with the number of independent degrees of freedom. Even a relatively coarse uniform grid can become very expensive to evaluate for molecules with five to six flexible torsions and completely impractical for more flexible molecules. For example, Ritonavir has 22 independent degrees of freedom (Chemburkar *et al.*, 2000); an extremely coarse grid with just three points in each dimension would involve approximately 31 billion grid points, each requiring an isolated-molecule QM calculation. Moreover, in practice, a much finer grid might be required to achieve a sufficiently good level of accuracy in energy evaluations.

In many pharmaceutically relevant molecules, however, the flexible torsion angles are often separated geometrically in *para* or *meta* positions of rigid benzyl groups. For example, a search of the Crystal Structure Database (CSD; Groom *et al.*, 2016) (November 2021 version) reveals that 170549, or 46.08%, of 370142 di-substituted benzene moieties were in the *para* arrangement, with 23467 (6.34%) in the *meta* and 176126 (47.58%) in the *ortho* arrangements. Our hypothesis is that, for compounds with *para* or *meta* arrangements, the cross interactions between torsion angles with such geometrical separation are likely to be negligible. As we will see later in this paper, this hypothesis makes the investigation of larger and more flexible molecules tractable. It should be pointed out that this approximation would not hold for molecular connectivities without geometric separation, such as aliphatic chains.

Partitioning flexible degrees of freedom into (approximately) non-interacting groups was a concept previously used in *CrystalPredictor I* which, instead of LAMs, used restricted Hermite interpolants to approximate the intramolecular energy and electrostatic potential. While this was successful for several systems (Francia *et al.*, 2021; Iuzzolino, 2018; Kazantsev, Karamertzanis, Adjiman, Pantelides *et al.*, 2011), the applicability of this approach had been limited by the requirement that the molecule's dependent degrees of freedom , *i.e.* its bond lengths, bond angles and non-flexible torsions, be fixed to their *in vacuo* values. This often leads to conformations with unnecessarily high intramolecular energy values. Furthermore, because the use of restricted interpolants can lead to a sharp decrease in the quality of the approximation as the number of degrees of freedom increases, there is a practical limit of three torsions per group to maintain accuracy.

The above shortcomings are addressed by the more recent LAM-based approaches employed by *CrystalPredictor II*. However, the number of flexible degrees of freedom that can be handled is still limited by computational cost. This is the issue that the current paper aims to address via the introduction of torsional angle groups. In Section 2, we present the algorithmic framework, briefly re-examining the concept of local approximate models (LAMs) and their effect on conformationally dependent properties, and outlining the changes that need to be introduced to the lattice energy model to account for torsional group partitioning. In Section 3, we investigate the impact of the use of partitioning on the accuracy of the LAMs, using methylparaben and molecule XX of the fifth CSP blind test (Bardwell *et al.*, 2011) as test cases. In Section 4, the approach is tested in the context of CSP by conducting global searches for four molecules with the modified *CrystalPredictor II* algorithm and analysing the results. The findings are summarized in Section 5.

2. LAMs with partitioning

The global search stage of CSP as implemented in *CrystalPredictor II* is preceded by (i) the identification of the main flexible torsions (independent degrees of freedom) in the isolated molecule(s) of interest; (ii) the generation of a database of LAM reference points, each involving an isolated-molecule QM minimization for the corresponding fixed values of the independent degrees of freedom, followed by the generation of a local approximate model for the corresponding optimal values of the remaining molecular geometry variables (the dependent degrees of freedom), the electrostatic potential of the molecule(s) and the intramolecular energy. During the global search itself, a large number of candidate crystal structures are generated using a quasi-random sequence, and each one of them is used as a starting point for a local minimization of the lattice energy using a gradient-based algorithm. The pre-generated LAMs are used to calculate the intramolecular energy contributions to the lattice energy, the atomic positions within the unit cell, which are needed to compute all intermolecular energy contributions, and the atomic charges used in the electrostatic energy contributions.

The focus of the current paper is in achieving a significant reduction in the cost of LAM generation for large, flexible molecules. We therefore present this step in more detail next.

2.1. Standard formulation of LAMs

In order to construct a LAM, the molecule's conformational degrees of freedom are partitioned into a vector of l independent degrees of freedom, θ, and a vector of M dependent degrees of freedom, $\bar{\theta}$, where θ contains the flexible torsions that have the largest impact on the intramolecular energy $U^{\rm intra}$. The intramolecular energy for given values of θ is then determined via the solution of an *in vacuo* energy minimization,

$$\Delta U^{\rm intra}(\theta) = \min_{\bar{\theta}} U^{\rm intra}(\bar{\theta}; \theta) - U^{\rm vac} \qquad (1)$$

where the semicolon indicates that the independent degrees of freedom θ are kept constant during the optimization, and $U^{\rm vac}$ is the minimum *in vacuo* energy of the molecule, obtained by solving the following problem:

$$U^{\rm gas} = \min_{\theta,\bar{\theta}} U^{\rm intra}\left(\theta, \bar{\theta}\right), \tag{2}$$

where the values of the optimization variables at the solution are denoted by $\theta^{\rm vac}$, $\bar{\theta}^{\rm vac}$.

In addition to computing the minimum intramolecular energy for the given values of θ, the QM minimization (2) also yields the corresponding values of the dependent degrees of freedom and the electrostatic potential field.

The above QM calculations are performed at a set of $N_{\rm ref}$ distinct reference values of θ, denoted by $\theta_l^{\rm ref}$, $l = 1, \ldots, N_{\rm ref}$. The combination of a second-order Taylor expansion of the intramolecular energy and the optimality conditions of equation (2) leads to the following LAMs (Kazantsev, Karamertzanis, Adjiman & Pantelides, 2011) that are valid in the vicinity of reference point $\theta_l^{\rm ref}$:

(*a*) intramolecular energy difference, $\Delta U^{\rm intra}$,

(*b*) dependent conformational degrees of freedom, $\bar{\theta}$ and

(*c*) point charges, q.

Details of each LAM are provided in the supporting information.

More recently, a smoothed LAM (Sugden *et al.*, 2019) based on a weighted average of the LAM expressions across several reference points has been introduced to obtain continuously differentiable approximants for the intramolecular energy, dependent degrees of freedom and point charges. The developments presented here are equally applicable to LAMs based on a single reference point or on a weighted average.

Regardless of the approach chosen, all reference points used in *CrystalPredictor II* are pre-computed using a uniform (Habgood *et al.*, 2015) or adaptive (Sugden *et al.*, 2016) grid and stored in a database, which can be accessed during local lattice energy minimizations. Where a conformation is indicated to have a lower energy than the current $U^{\rm vac}$, the $\theta^{\rm vac}$, $\bar{\theta}^{\rm vac}$ and $U^{\rm vac}$ are updated. As can be seen in equations (1) and (2), at each reference point, a QM minimization is carried out and the Hessian matrix at the solution is computed. As a result, the generation of the LAM database is often the most expensive component of the global search, and the exponential dependence of the number of LAM points on the number of independent degrees of freedom has a large impact on overall computational cost.

2.2. Formulation of LAMs with torsional groups

In the proposed approach, the vector of independent degrees of freedom θ is further partitioned into N_G subvectors θ_g, $g = 1, \ldots, N_G$, referred to as torsional groups, such that each torsion belongs to exactly one group. Each such group is assumed to have an additive effect on the deviations of the intramolecular energy, the values of the dependent degrees of freedom and the point charges from their corresponding values *in vacuo* molecular conformation. Consequently, the deviation in any one of these quantities brought about by varying torsion angles in two different groups can be computed by summing the deviations incurred when the torsion angles in each group are varied independently.

To apply the above principle during the global search stage of the CSP algorithm, we need to derive a separate set of LAMs at a separate set of reference points for each torsional group g while keeping all other torsional groups $g' \neq g$ at their *in vacuo* values.

More specifically, for each torsional group, $g = 1, \ldots, N_G$, we perform the following steps:

(1) We select a set of $N_{g,\,\rm ref}$ reference points, $\theta_{g,l}^{\rm ref}$, $l = 1, \ldots, N_{g,\,\rm ref}$. These are completely independent of the reference points for any other torsional group.

(2) At each reference point $l = 1, \ldots, N_{g,\rm ref}$, we perform an isolated-molecule QM conformational energy minimization:

$$\Delta U_{g,l}^{\rm intra}\left(\theta_{g,l}^{\rm ref}\right) = \min_{\bar{\theta}} U^{\rm intra}(\bar{\theta}; \theta_{g,l}^{\rm ref}, \theta_{g'}^{\rm vac},\ g' \neq g) - U^{\rm vac} \tag{3}$$

with respect to the dependent degrees of freedom $\bar{\theta}$ while fixing the independent degrees of freedom θ as follows:

(*a*) for group g, at the corresponding reference values $\theta_{g,l}^{\rm ref}$

(*b*) for all other groups, $g' = 1, \ldots, N_G$, $g' \neq g$, at the corresponding *in vacuo* conformational values $\theta_{g'}^{\rm vac}$.

(3) From the solution of minimization problem (3), we

(*a*) obtain the values of the dependent degrees of freedom, $\bar{\theta}_{g,l}$

(*b*) determine the corresponding point charges $q_{g,l}$.

(4) By analogy to (3), (6) and (8), we construct the following LAMs for the deviations of intramolecular energy, dependent degrees of freedom and point charges from the corresponding *in vacuo* values $U^{\rm vac}$, $\bar{\theta}^{\rm vac}$ and $q^{\rm vac}$, respectively:

$$\begin{aligned}\Delta U_{g,l}^{\rm intra}(\theta_g) = {} & \Delta U_{g,l}^{\rm intra}\left(\theta_{g,l}^{\rm ref}\right) + b_{g,l}^T\left(\theta_g - \theta_{g,l}^{\rm ref}\right) \\ & + \tfrac{1}{2}\left(\theta_g - \theta_{g,l}^{\rm ref}\right)^T C_{g,l}\left(\theta_g - \theta_{g,l}^{\rm ref}\right)\end{aligned} \tag{4}$$

$$\Delta\bar{\theta}_{g,l}\left(\theta_g\right) = \left(\bar{\theta}_{g,l} - \bar{\theta}^{\rm vac}\right) + A_{g,l}\left(\theta_g - \theta_{g,l}^{\rm ref}\right) \tag{5}$$

$$\Delta q_{g,l}\left(\theta_g\right) = q_{g,l} - q^{\rm vac} \tag{6}$$

where we have introduced the vector:

$$b_{g,l} = \left[\frac{\partial U^{\rm intra}}{\partial\theta_g}\right] \tag{7}$$

and the matrices:

$$C_{g,l} = \left[\frac{\partial^2\Delta U^{\rm intra}}{\partial\theta_g^2}\right] - \left[\frac{\partial^2\Delta U^{\rm intra}}{\partial\theta_g\partial\bar{\theta}}\right]\left[\frac{\partial^2\Delta U^{\rm intra}}{\partial\bar{\theta}^2}\right]^{-1}\left[\frac{\partial^2\Delta U^{\rm intra}}{\partial\theta_g\partial\bar{\theta}}\right]^T \tag{8}$$

$$A_{g,l} = -\left[\frac{\partial^2\Delta U^{\rm intra}}{\partial\bar{\theta}^2}\right]^{-1}\left[\frac{\partial^2\Delta U^{\rm intra}}{\partial\theta_g\partial\bar{\theta}}\right]^T \tag{9}$$

where all partial derivatives are computed at the solution of the minimization problem (3).

Overall, the above algorithm requires $\sum_{g=1}^{N_G} N_{g,\rm ref}$ QM energy minimizations at step (2), plus one more for the initial

(and subsequent if lower energy conformations are found) *in vacuo* conformation(s) [see equation (1)], and results in the construction and storage of $\sum_{g=1}^{N_G} N_{g,\mathrm{ref}}$ LAMs at step (4). By comparison, without the additivity principle, we would need $1 + \prod_{g=1}^{N_G} N_{g,\mathrm{ref}}$ QM energy minimizations, which would be a significantly higher number for many practical problems.

Once constructed, the LAMs can be used during the global search stage to evaluate the corresponding quantities at any given set of values of the independent degrees of freedom, θ, based on the following algorithm:

Given the vector θ:

(1) For each torsion group $g = 1, \ldots, N_G$

(*a*) determine the values of the corresponding subvector θ_g

(*b*) determine the reference point $\theta_{g,l}$ that is nearest to θ_g

(*c*) evaluate expressions (4)–(6) to determine, $\Delta U_{g,l}^{\mathrm{intra}}(\theta_g)\Delta\bar{\theta}_{g,l}(\theta_g)$, $\boldsymbol{\Delta} q_{g,l}(\theta_g)$.

(2) Apply the additivity principle to determine the required values of the intramolecular energy, the dependent degrees of freedom and the point charges:

$$\Delta U^{\mathrm{intra}}(\theta) = \sum_{g=1}^{N_G} \Delta U_g^{\mathrm{intra}}(\theta_g) \tag{10}$$

$$\bar{\theta}(\theta) = \bar{\theta}^{\mathrm{vac}} + \sum_{g=1}^{N_G} \bar{\theta}_g(\theta_g), \tag{11}$$

$$q(\theta) = q^{\mathrm{vac}} + \sum_{g=1}^{N_G} q_g(\theta_g). \tag{12}$$

As an illustration, Fig. 1 considers a molecule with two independent degrees of freedom, θ_1 and θ_2, that have been trivially separated into two (N_G = 2) groups comprising one angle each; torsional ranges are given in the range 0–360° for sequencing clarity. The *in vacuo* values, $\left(\theta_1^{\mathrm{vac}}, \theta_2^{\mathrm{vac}}\right)$ = (243°, 126°), are indicated by a red symbol. A uniform spacing of 60° is chosen to specify six reference points per group ($N_{\mathrm{ref},1}$ = $N_{\mathrm{ref},2}$ = 6) located at 30°, 90°, 150°, 210°, 270° and 330°, respectively, as indicated by the open squares and circles. Overall, before embarking on the global search, we would need to perform 13 QM energy minimizations, corresponding to these 12 reference points plus one more for the *in vacuo* conformation. We would then store the 13 sets of LAMs in terms of the corresponding quantities $\bar{\theta}^{\mathrm{vac}}$, q^{vac}, $\Delta U_{g,l}^{\mathrm{intra}}(\theta_{g,l}^{\mathrm{ref}})$, $\bar{\theta}_{g,l}$, $q_{g,l}$, $b_{g,l}$, $C_{g,l}$,$A_{g,l}$ [*cf.* equations (4)–(9)].

Now consider a situation during the global search stage where we need to evaluate the lattice energy for a crystal structure involving a molecular conformation with θ = (104°, 235°)$^{\mathrm{T}}$. The closest reference points are $\theta_{1,2}^{\mathrm{ref}}$ = 90° in the first torsional group and $\theta_{2,4}^{\mathrm{ref}}$ = 210° in the second one. The values of interest can be computed by evaluating the lower-dimensional LAMs generated at these points using equations (10), (11) and (12), and then combining these using equations in supporting information.

Finally, we note that the proposed approach does not require the reference points to be uniformly spaced. Therefore, these points can be generated by running an adaptive LAM generation algorithm (Sugden *et al.*, 2016) to convergence in each group g independently, with all flexible torsions belonging to other groups $g' \neq g$ held at their *in vacuo* values.

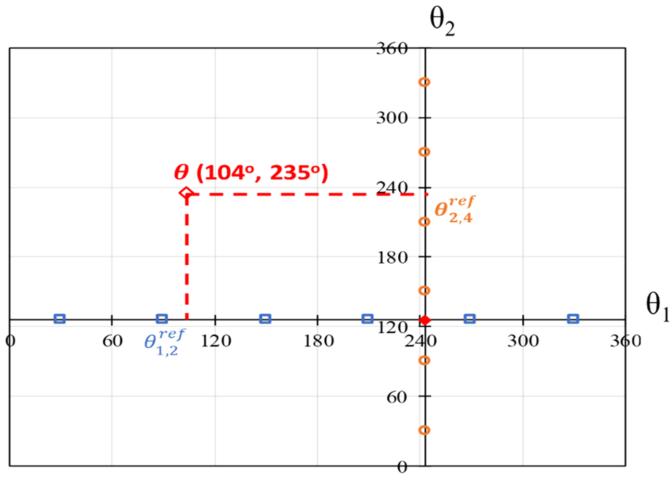


Figure 1
Illustration of the generation of reference points using the torsional group partitioning scheme. Two groups are considered, each with one independent degree of freedom: θ_1 (blue) for group 1 and θ_2 (orange) for group 2. The *in vacuo* minimum is shown in a filled red diamond at the intersection of the axes. The projected reference points are shown as blue squares for group 1 and orange triangles for group 2. The reference points in the two groups that would be used in the LAMs that are relevant to a particular point θ = (104°, 235°)^T in the space of the independent conformational degrees of freedom (see text for details) are shown. Torsions are given in °.

2.3. Impact of partitioning of torsional groups on LAM accuracy

2.3.1. Accuracy of LAMs for methyl paraben

The torsional group partitioning algorithm is illustrated on methyl paraben, a small molecule from the GRAS list (Burdock & Carabin, 2004) with a significant degree of separation between the independent degrees of freedom, as can be seen in Fig. 2. Because the atoms in the hydroxyl and ester groups are in *para* positions on the central benzene ring, it can be expected that the effect of changes in the value of θ_1 (hydroxyl) on the conformationally dependent variables ($\Delta U^{\mathrm{intra}}$, point charges, and dependent degrees of freedom) will be nearly independent of the effect of any concurrent changes in θ_2 or θ_3 (ester). The same approximation cannot be applied to θ_2 and θ_3 however, since changes in the value of θ_2

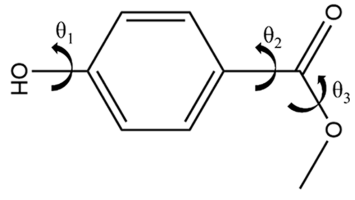


Figure 2
Molecular diagram of methyl paraben, with flexible torsions being indicated by curved arrows, and set at arbitrary values in this 2D diagram.

Table 1
Average errors in the conformationally dependent properties of methyl paraben at 110 crystallographically relevant points using the non-partitioned and partitioned LAM schemes, relative to QM-calculated values.

LAM generation scheme	Intramolecular energy (kJ mol^{-1})	Bond lengths (Å)	Bond angles (°)	Torsion angles (°)	Charge (e)
Non-partitioned	1.68	0.0004	0.09	0.5	0.002
Partitioned	1.39	0.0004	0.07	0.5	0.002

will have a direct impact on the strains involved in changing the value of θ_3. Thus, two torsional groups are defined, $G_1 = \{1\}$ and $G_2 = \{2, 3\}$.

To construct the LAM databases, reference points are generated using a uniform spacing of 60° for all independent degrees of freedom across the range (0°, 360°), with the first grid point at 30°; therefore, there are six points in each direction. Two LAM databases are then generated:

(1) LAM database without torsional group partitioning: in this case, there are a total of $6 \times 6 \times 6 = 216$ reference points; 217 QM calculations are required, including the *in vacuo* calculation.

(2) LAM database with torsional group partitioning: in this case, there are six reference points for G_1 and $6 \times 6 = 36$ reference points for G_2, giving a total of 43 QM calculations, including the *in vacuo* calculation.

Overall, we note that the introduction of partitioning reduces the number of QM calculations by approximately 80%. However, the key question is whether this has a materially adverse effect on the accuracy of the predictions based on these LAMs. We assess this via the following procedure.

(1) We generate 2000 random points from the uniform probability distribution over the space of the flexible torsions and evaluate the intramolecular energy $\Delta U^{\rm intra}$ at each point by performing a QM calculation [*cf.* equation (1)].

(2) We discard all points that are found to be outside the region of crystallographic relevance *i.e.* where $\Delta U^{\rm intra} > 20.0$ kJ mol^{-1}. This leaves 110 points.

(3) For these 110 crystallographically relevant points, we evaluate the LAMs for both the partitioned and non-partitioned schemes and compare the results with the corresponding exact values. The average errors are given in Table 1. We note that both LAM generation schemes result in very similar prediction errors. Moreover, the latter are remarkably low considering that the LAMs were generated on a uniform grid with only six points for each torsional angle.

Additionally, Fig. S2 shows a parity plot of the $\Delta U^{\rm intra}$ values predicted with the partitioned scheme at the 216 reference points used in the non-partitioned scheme. The R^2 value is 0.9985, indicating that the hypothesis that there is very little interaction across the two torsional groups is justified.

Overall, it can be concluded that, for the case of methyl paraben, the approximation introduced by the additivity assumption does not lead to any significant loss in the quality of the model.

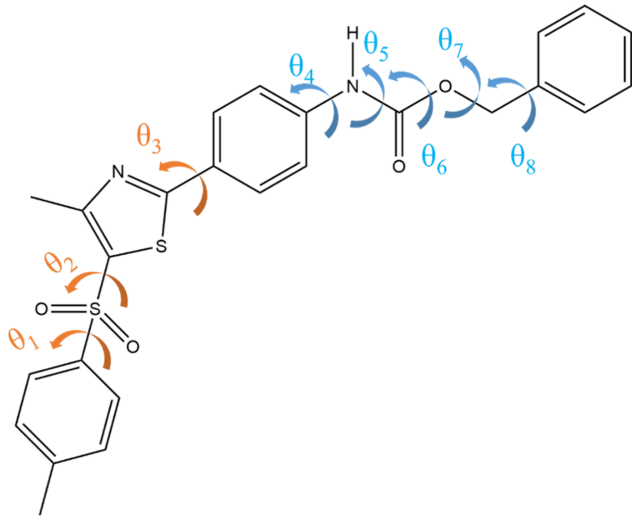


Figure 3
Molecular diagram of molecule XX, with flexible torsions indicated by arrows. Torsions 1–3 (orange) are assigned to torsional group 1, and 4–8 (blue) to torsional group 2.

2.3.2. Accuracy of LAMs for molecule XX

A further assessment of the validity of the partitioning approximation is made by considering a much larger compound with greater flexibility, namely molecule XX from the fifth blind test (Bardwell *et al.*, 2011). As shown in Fig. 3, there are eight independent degrees of freedom, θ_1 to θ_8, with θ_6 expected to vary only slightly around 180°. This set of torsions is partitioned into two groups by taking advantage of the physical separation arising from the presence of the central benzyl ring: $G_1 = \{1, 2, 3\}$ and $G_2 = \{4, 5, 6, 7, 8\}$. Group G_1 could conceivably be further partitioned into two separate groups, containing θ_1 and θ_2 on the one hand, and θ_3 on the other hand. However, this is unnecessary as a group containing three torsions is computationally manageable.

For this compound, it is computationally intractable to cover the entire range of flexibility of the eight independent degrees of freedom with a non-partitioned LAM generation scheme. Therefore, for the purposes of comparison with the partitioned scheme, a narrower range of flexibility that extends by ±30° around a point neighbouring the experimental values of the independent degrees of freedom is considered. The selected ranges are shown in Table 2. A uniform grid is set up with 30° increments in all independent degrees of freedom, with the reference points positioned away from the edges of the grid (+15° from the lower bound and −15° from the upper bound), so that two points are needed in each direction. For the purpose of investigating LAM accuracy and in view of the limited flexibility of θ_6, only a single grid point (at 180°) is employed for this torsion angle. This constraint is lifted during the CSP study of this molecule presented in Section 3.3.

Table 2
Experimental values and range of torsion values used to assess the quality of the proposed approximation for molecule XX.

The grid midpoint is determined by choosing a value that includes the experimental value and the expected end point from minimizations using the standard LAM database, rounded to the nearest 10° angle. The upper and lower bounds are then set at ±30° of this midpoint and two reference points are introduced for each torsion, except for θ_6, where the bounds are set at ±15° and one reference point is used.

Torsion	Experimental value (°)	Grid midpoint (°)
θ_1	107.04	110
θ_2	104.74	90
θ_3	167.52	180
θ_4	1.09	0
θ_5	176.41	180
θ_6	185.72	180
θ_7	254.18	270
θ_8	261.95	270

Overall, the above approach leads to a manageable number of QM calculations for both LAM generation schemes:

(*a*) non-partitioned scheme: 2^7 reference points, giving a total of 129 QM calculations (including the one for the *in vacuo* conformation);

(*b*) partitioned scheme: $2^3 = 8$ reference points for G_1 and $2^4 = 16$ points for G_2, giving a total of 25 QM calculations (including the one for the *in vacuo* conformation).

Overall, even with the significantly reduced ranges of torsional angle flexibility, the use of the partitioning scheme results in an approximately 80% reduction in the number of QM calculations. The remaining computational parameters are given in the supporting information.

To investigate the limits of the partitioning scheme, we also consider an extreme case where we partition the independent degrees of freedom into eight groups, each containing just one torsion. Once again, two reference points are used for each torsion, other than for θ_6 for which only one is used, yielding a total of only 16 QM calculations (including the one for the *in vacuo* conformation).

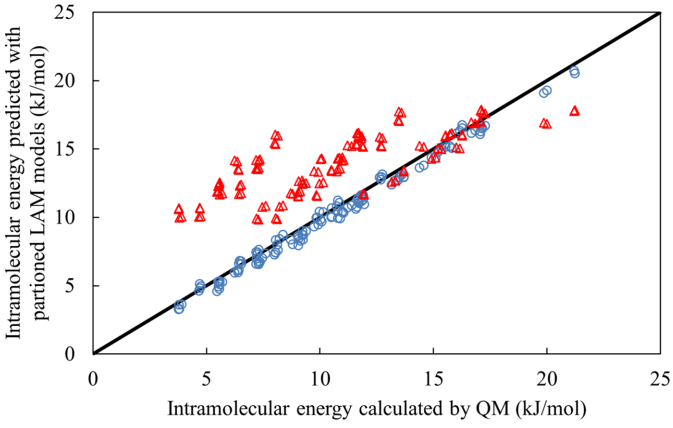


Figure 4
Parity plot for the intramolecular energy predicted by the LAM with torsional group partitioning versus intramolecular energy computed quantum mechanically for molecule XX. The blue circles correspond to the partitioning of the eight flexible torsions into two groups and the red triangles to full partitioning into eight independent groups.

Three lattice energy minimizations are then performed, respectively using the non-partitioned, two-group and eight-group LAM databases; all three minimizations use the experimental values of the independent degrees of freedom as the initial point. The $RMSD_{15}$ between the minimum-energy crystal structure produced by the non-partitioned scheme and those resulting from the two-group and eight-group schemes are 0.0710 Å and 0.219 Å, respectively. This indicates that the partitioning into two groups has a very small impact on the resulting crystal structure. However, completely ignoring all interactions between adjacent torsions leads to an unacceptable change in accuracy. This is further confirmed by examining how the two partitioned databases perform at approximating ΔU^{intra} at the 129 QM-calculated reference points in the non-partitioned LAM database, as shown in Fig. 4. It can be seen that a high degree of accuracy is achieved in the two-group case [0.47 kJ mol^{-1} average absolute deviation (AAD) and 1.1 kJ mol^{-1} maximum error] while the eight-group partitioning exhibits a significant decrease in accuracy (3.55 kJ mol^{-1} AAD and 8.0 kJ mol^{-1} maximum error).

3. Application of torsional group partitioning to CSP

The proposed partitioning scheme has been implemented within the *CrystalPredictor II* software (Habgood *et al.*, 2015) for structure generation and global search. In this section, we investigate the effect of partitioning on CSP studies of four molecules: paracetamol, methyl paraben, molecule XX and safinamide. Following the approach highlighted in Pantelides *et al.* (2014), our CSP studies include a global search step followed by one refinement step to ascertain whether the experimental forms are found as low-lying energy minima. Further refinements to improve the reliability of the relative energy rankings are possible (Bowskill *et al.*, 2021) [as reported, for example in the sixth blind test (Reilly *et al.*, 2016) and in a recent study of the ROY molecule (Beran *et al.*, 2022)], but these are not undertaken here. Our focus is on examining whether the partitioning into torsional groups leads to a model that is sufficiently accurate to generate a global search landscape that leads to the successful identification of experimentally known polymorphs upon further refinement. Where the original non-partitioned scheme is computationally

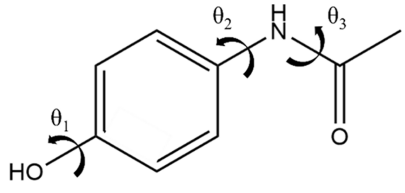


Figure 5
Molecular diagram of paracetamol, with flexible torsions indicated by arrows.

Table 3
Uniform LAM grid for investigation of paracetamol.

Torsion	Grid lower bound (°)	Grid upper bound (°)	Grid increment (°)	Number of distinct torsion values
θ_1	−30	30	±15	3
θ_2	120	240	±15	5
θ_3	120	240	±15	5

tractable, we also consider the relative reduction in computational cost achieved via partitioning.

3.1. Paracetamol

The new methodology is first applied to a small flexible active pharmaceutical ingredient (API), paracetamol (see Fig. 5). The three flexible torsions are separated geometrically by a *para* arrangement in a benzene ring, with θ_1 on one side of the ring, and θ_2 and θ_3 on the other. This suggests that the flexible degrees of freedom can be partitioned into two groups: $G_1 = \{1\}$ and $G_2 = \{2, 3\}$.

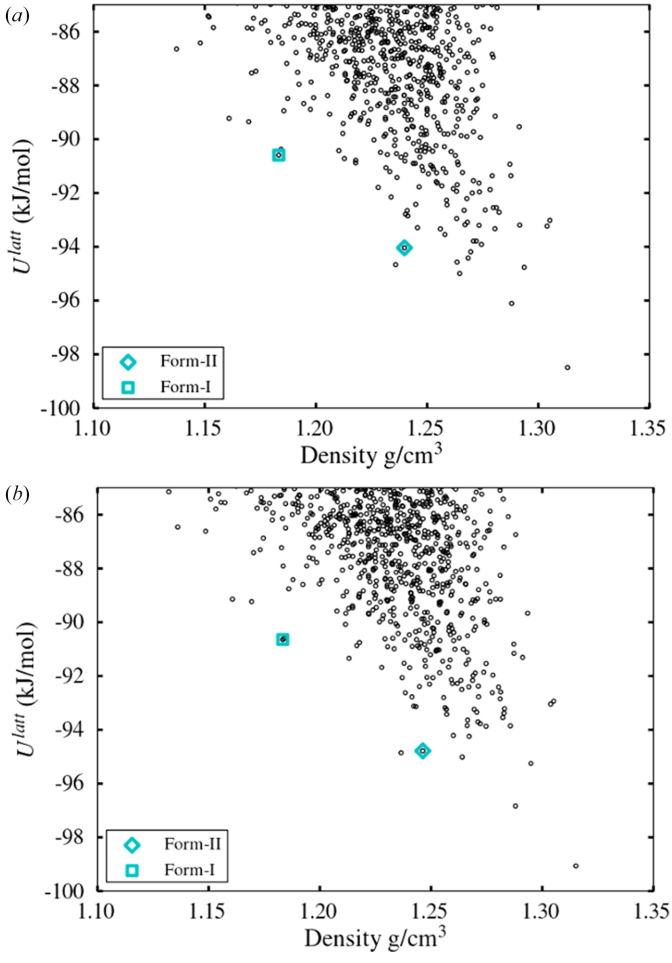


Figure 6
Polymorphic landscapes (lattice energy U^{latt} versus density of computed structures) after global search stage with *CrystalPredictor*, for paracetamol, in the (*a*) non-partitioned and (*b*) partitioned schemes.

Two parallel investigations are performed without and with partitioning, in both cases using a uniform grid as summarized in Table 3, that covers a narrow range of conformational space around the known experimental conformation, as a proof of concept. The total number of QM calculations are $76 = 3 \times 5 \times 5 + 1$ and $29 = 3 + 5 \times 5 + 1$, respectively. Thus, partitioning results in a 62% decrease in the number of these calculations.

Following the global search stage, the polymorphic landscape for the non-partitioned and partitioned schemes can be compared in Fig. 6. The main consideration at this stage is that the landscape should include structures that converge to the experimental crystal structures upon subsequent refinement with the more accurate, but also more expensive, energy model used in *CrystalOptimizer* (Kazantsev *et al.*, 2010), and that these structures should be low enough to be selected for such refinement, *e.g.* within 20 kJ mol^{-1} of the global minimum or within the lowest 1000 structures. For this simple case, it is clear that the two schemes result in nearly identical landscapes that include the experimental forms well within the 20 kJ mol^{-1} cutoff. This can be seen in more detail in Table 4.

Following refinement with *CrystalOptimizer* of all structures within 20 kJ mol^{-1} of the global minimum from the global search, the landscape shown in Fig. S2 is obtained with the structures generated by the non-partitioned LAM scheme. Polymorph I is observed as the global minimum structure, with an $RMSD_{15}$ of 0.1933 Å relative to the experimental structure, whilst polymorph II is seen at rank 9, 3.23 kJ mol^{-1} above the global minimum, with an $RMSD_{15}$ of 0.2487 Å. The refined landscape obtained with the structures generated by the partitioned LAM scheme are practically identical. This illustrates that the global search stage simply needs to be of sufficient accuracy to provide the refinement stage with good starting points from which to generate an accurate and complete list of viable structures. This is evidently achieved with both partitioning schemes.

The computational costs of the two approaches are compared in Table 5. A reduction of 63% is observed in the generation of the LAM database with the new scheme, in line with the reduction in the number of QM calculations. However, as LAM generation is relatively inexpensive for this small system, the overall CSP costs remain similar.

3.2. Methyl paraben

The next CSP study is conducted on methyl paraben (*cf.* Section 2.3.1), a compound with three known polymorphs. It is of a similar size to paracetamol, but we now consider the use of the adaptive LAM generation algorithm within *CrystalPredictor II* (Sugden *et al.*, 2016), as this is more representative of the approach one would typically take when investigating a new compound. By concentrating the reference points in those regions of conformational space that are of crystallographic relevance, adaptive LAM generation can greatly reduce the number of QM calculations required to study a molecule compared to a uniform grid of the same accuracy. Moreover, the adaptive algorithm can readily be applied to a partitioned

Table 4
Lattice energies, rank and $RMSD_{15}$ for paracetamol and methyl paraben using the non-partitioned and partitioned schemes.

	Non-partitioned LAM scheme			Partitioned LAM scheme		
Experimental form (CSD refcode)	U^{latt} (kJ mol^{-1})	Rank	$RMSD_{15}$ (Å)	U^{latt} (kJ mol^{-1})	Rank	$RMSD_{15}$ (Å)
			Paracetamol			
Monoclinic polymorph I (HXACAN57)	7.90	77	0.398	8.42	91	0.406
Orthorhombic polymorph II (HXACAN08)	4.45	9	0.491	4.28	6	0.527
			Methyl paraben			
1-111 (CEBGOF03)	–	1	0.251	–	1	0.215
1-112 (CEBGOF04)	1.45	5	0.218	1.68	5	0.306
1-107 (CEBGOF05)	4.38	15	0.110	3.18	17	0.205

Table 5
Approximate CPU cost of each stage of a CSP study for paracetamol and methyl paraben using non-partitioned and partitioned LAM databases.

Approximate computational cost (CPU hour)		
	Paracetamol	
Stage	Non-partitioned LAM scheme	Partitioned LAM scheme
LAM generation	390	145
Global search	2000	2000
Refinement	18000	18000

	Methyl paraben	
Stage	Non-partitioned LAM database	Partitioned LAM database
LAM database generation	100	65
Global search	1250	1250

set of degrees of freedom to reduce computational effort still further.

In the non-partitioned case, starting from a uniform grid with grid increments of 120.0° (*i.e.* 28 reference points), convergence of the adaptive LAM generation algorithm is reached after a total of 46 LAM points have been evaluated ($\Delta^* = 5$ kJ mol^{-1}, $\Delta^{**} = 20$ kJ mol^{-1}). Using the same grid increments, a partitioned LAM scheme based on the partitioning introduced in Section 2.3.1 initially contains three and nine LAM points in groups 1 and 2, respectively. The adaptive LAM generation algorithm is run independently for each group, and convergence is reached after a total of six and 23 reference points are generated in groups 1 and 2, respectively. Therefore, including the *in vacuo* configuration, there are 47 reference points in the non-partitioned case and 30 in the partitioned case. Thus, partitioning results in a 36% decrease in the number of QM calculations.

Following the global search stage, all three experimental polymorphs are found by both LAM schemes within 5 kJ mol^{-1} of the global minimum (see Fig. 7) with an accurate geometrical representation of the experimentally observed polymorphs, as indicated by the $RMSD_{15}$ values listed in Table 4. The impact of partitioning on CPU time is shown in Table 5. As in the case of paracetamol, the use of partitioning results in a significant reduction in the CPU time required for LAM generation, but the effect on the total CPU time is small.

3.3. Molecule XX

As the two molecules considered in Sections 3.1 and 3.2 had only three independent conformational degrees of freedom, the corresponding CSP studies were tractable even without the use of the partitioning scheme presented in this paper. In this section, we consider the application of the partitioning scheme to molecule XX of the fifth blind test (*cf.* Section 2.3.2) which has eight flexible torsion angles. In this case, a CSP study using a standard non-partitioned LAM scheme would be computationally intractable.

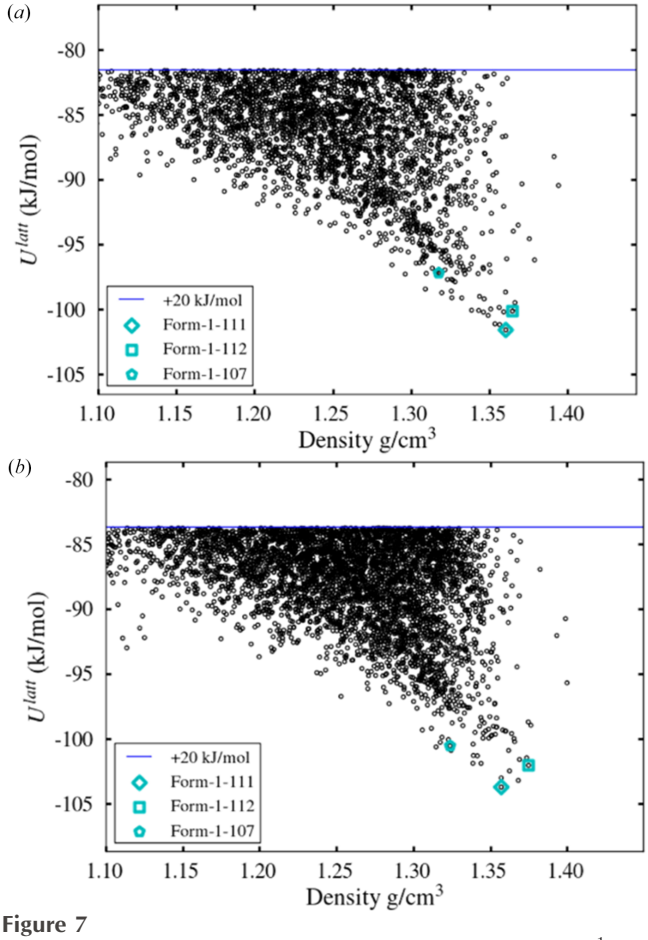


Figure 7
Polymorphic landscapes for methyl paraben within 20 kJ mol^{-1} of the global minimum after global search stage with *CrystalPredictor II* in the (*a*) non-partitioned and (*b*) partitioned schemes.

Table 6
Description of partitioned LAM database for molecule XX.

Torsional group	Torsion angle	Initial uniform grid [Lower bound, increment, upper bound] (°)	Number of reference points	Number of LAM points	Number of adaptive LAM points
1	θ_1	[60, 120, 300]	3	27	307
	θ_2	[−30, 120, 210]	3		
	θ_3	[60, 120, 300]	3		
2	θ_4	[−120, 120, 120]	3	162	8299
	θ_5	[0, 180, 180]	2		
	θ_6	[60, 120, 300]	3		
	θ_7	[30, 120, 270]	3		
	θ_8	[−30, 120, 210]	3		

Successful CSPs of this molecule were carried out in the fifth blind test (Bardwell *et al.*, 2011; Kazantsev, Karamertzanis, Adjiman, Pantelides *et al.*, 2011), including one by our research group using *CrystalPredictor I*. As has already been mentioned, the latter uses restricted Hermite interpolants instead of LAMs, which practically limits the number of independent torsions that can be handled within a single torsional group to three. To address this limitation, torsion θ_4 was fixed at 0° and the flexibility in θ_5 was approximated by fixing this torsion to two distinct values, namely either 0 or 180° (*cis* or *trans* configurations) and performing separate searches for each option. Further, the restricted Hermite interpolant scheme necessitates a uniform grid, requiring a large number of QM calculations. To make this tractable, the computational cost was reduced by using a combination of one-dimensional scans and Crystal Structure Database analysis to identify the most likely ranges of values of the flexible torsions and the corresponding energetically meaningful regions of conformational space. A separate search was then performed for each such region. In total, this led to eight separate searches. The group of Professor Graeme Day at the University of Southampton was also successful in predicting the experimental form of molecule XX by conducting multiple (48) rigid-molecule searches to identify initial structures (Price *et al.*, 2010).

However, while the use of separate global searches proved successful in this particular case, the assumptions made to narrow the search space cannot generally be relied upon. In contrast, the partitioning approach presented in this paper combined with the adaptive LAM generation algorithm allows for full coverage of conformational space within a single global search. The LAM generation is initiated with a uniform grid for each torsional group (see Section 2.3.2) with increments of 120°, except for θ_5, for which an increment of 180° is used, as shown in Table 6. The subsequent application of the adaptive LAM generation algorithm results in 8461 LAMs for group 1 and 334 LAMs for group 2, *i.e.* a total number of 8796 QM calculations (including the one required for establishing the *in vacuo* conformation). A non-partitioned database of equivalent accuracy would require more than 2.8 million (∼8461 × 334) QM calculations, which would be prohibitively expensive even with modern computer resources.

The global search landscape is shown in Fig. 8(*a*). The experimental form is at rank 9 and 4.54 kJ mol^{-1} above the global minimum. This indicates that the energetic representation provided by the model is sound, whilst the $RMSD_{15}$ of 0.378 Å confirms the quality of the geometric representation.

Finally, the 411 structures that are found to be within the usual 20 kJ mol^{-1} cutoff are refined with *CrystalOptimizer* using the same level of theory as in the original study (Kazantsev, Karamertzanis, Adjiman, Pantelides *et al.*, 2011). The polymorphic landscape is seen in Fig. 8(*b*) where the experimental form is, once again, observed as the global minimum. The $RMSD_{15}$ comparison between the experimentally observed and predicted structure is 0.186 Å. Overall, these results confirm the reliability of the proposed partitioning scheme.

The CPU cost of the CSP study is reported in Table 7. As can be seen, the dominant component is now the LAM database generation and the savings afforded by the partitioning are therefore critical to the success of the study. We note that the overall cost shown here is larger than that of the original study. This is in large part due to the fact that the full

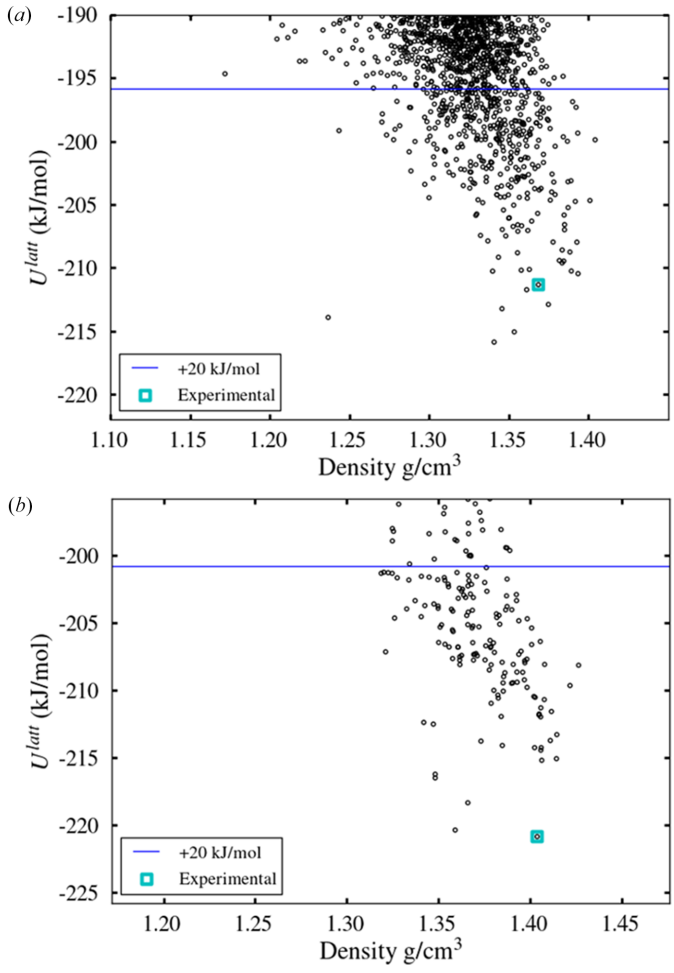


Figure 8
Polymorphic landscape for molecule XX, after (*a*) global search and (*b*) refinement.

Table 7
Approximate CPU cost of each stage of a CSP study for molecule XX, using a partitioned LAM database.

CPU cost equates to 3–6 months on standard HPC architecture.

Stage	Approximate computational cost (CPU hour)
LAM database generation	700000
CrystalPredictor global search	60000
CrystalOptimizer refinement	90000

search space is explored in our current work, whereas only a subset of the conformational space was investigated previously.

3.4. Safinamide

As a final assessment of the performance of the proposed approach, we carry out a CSP study of safinamide, a drug that is used to treat Parkinson's disease (Leuratti *et al.*, 2013). To our knowledge, this is the first time that such a study is being reported for this molecule in the open literature. The safinamide molecule involves 41 atoms and seven torsional degrees of freedom, as seen in Fig. 9. It has two known conformational polymorphs (Ravikumar & Sridhar, 2010; Cruz-Cabeza & Bernstein, 2014). Form I (CSD refcode TUWFIB) features an unusual amine N—H···F contact, whilst in form II (TUWFIB01) the hydrogen bonding scheme satisfies all of the NH donors/acceptors, which is indicative of stability ranking This example represents a sighted test, in that experimental polymorphs are known before the CSP investigation, but illustrates the application for a given molecule before a polymorphic screen.

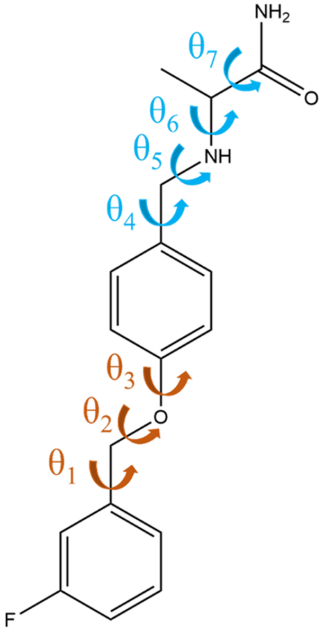


Figure 9
Molecular diagram of safinamide, with flexible degrees of freedom indicated. Group 1 (θ_1–θ_3) degrees of freedom are shown in orange, whilst group 2 (θ_4–θ_7) is shown in blue.

The independent degrees of freedom in the safinamide molecule are geometrically separated either side of a benzene moiety, making it an ideal, if challenging, test case for the proposed algorithm. More specifically, we employ two torsional groups: group 1 with torsions θ_1 to θ_3, and group 2 with torsions θ_4 to θ_7, assuming complete flexibility in all cases. Starting from a uniform grid with three increments of 120° spanning 360° degrees in each torsion and applying the adaptive LAM algorithm until convergence, we construct LAM sets containing 1148 and 4121 reference points for groups 1 and 2, respectively. An equivalent non-partitioned database would require approximately 4.7 million LAMs.

The polymorphic landscape after a global search of 500000 minimizations is shown in Fig. 10. A structure matching form I is found 13.5 kJ mol^{-1} above the global minimum at rank 572, and a structure matching form II is found at 15.5 kJ mol^{-1} above the global minimum at rank 1165. Although high in energy, both forms are well within the normal cutoff of 20 kJ mol^{-1}, with good geometric representation: the $RMSD_{15}$ for form I is 0.355 Å, and that for form II is 0.347 Å. There are 8250 structures within 20 kJ mol^{-1} of the global minimum in total.

Following refinement of the 3000 lowest-energy structures, form I is observed at rank 150, 10.97 kJ mol^{-1} above the

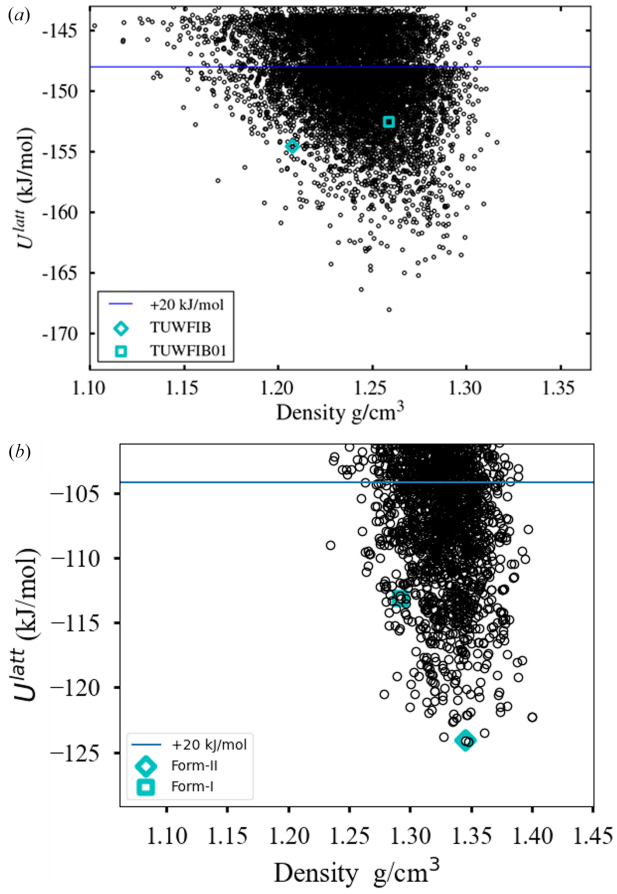


Figure 10
(*a*) Safinamide polymorphic landscape after global search with partitioned LAMs. (*b*) Safinamide polymorphic landscape after refinement of the 3000 lowest-energy structures identified by the global search.

Table 8
Approximate CPU cost of each stage of a CSP for safinamide, using a partitioned LAM database.

Stage	Approximate computational cost (CPU hour)
LAM database generation	240000
CrystalPredictor global search	4000
CSO-FM refinement	140000

global minimum, whilst form II is commensurate with the global minimum (+0.08 kJ mol^{-1}, rank 2), as seen in Fig. 10(*b*). The geometric representation is satisfactory for both forms, with an $RMSD_{15}$ for form II of 0.2232 Å, and for form I of 0.4142 Å. This provides some confidence that the two structures will be identified as low-lying minima upon further refinement using even more accurate lattice energy models, such as those computed via periodic DFT calculations (Hafner & Kresse, 1997). Once again, Table 8 shows the LAM generation to be a significant cost and the use of partitioning brings this down significantly, to be of the same order of magnitude as the first refinement step. The fact that these initial steps lead to the identification of the two experimental forms within the top 150 structures within less than 400000 CPU hours is encouraging for such a large molecule. Within a pharmaceutical context, an indication that an experimentally known form is commensurate with the thermodynamic global minimum would provide confidence in the completeness of the polymorphic screen.

4. Concluding remarks

We have presented an approach for the calculation of the lattice energy of putative crystal structures based on a partitioning of the molecule's independent conformational degrees of freedom into torsional groups. The approach can significantly reduce the number of isolated molecule QM calculations required to conduct a global search of the potential crystal structures of organic molecules. It is applicable to molecules where there is sufficient geometric separation between the torsional groups. In such cases, the combined effects of the deviations of torsional angles belonging to different groups on the intramolecular energy, molecular conformation and electronic charge distribution are approximately additive.

The validity of the additivity approximation was demonstrated using systems which are computationally tractable even without the partitioning. These included two small molecules (paracetamol and methyl paraben) and a larger one (molecule XX from the fifth blind test) with restricted ranges of variation of the flexible torsions.

The partitioning approach was further tested on two systems of such size and flexibility that would otherwise be considered as practically intractable: molecule XX with the full ranges of variation of the flexible torsion angles, and safinamide. The global search landscapes generated in both cases were found to be of high quality, with all known experimental structures ranked low enough in energy to be considered for further refinement. Further work will include using the method on ever larger systems, with more rotatable torsions, to understand where further improvements to efficiency are required for practical applications

5. Related literature

The following references are cited in the supporting information for this article: Bowskill (2021), Frisch *et al.* (2016), Hu *et al.* (2007) and Williams & Cox (1984).

Acknowledgements

We would like to gratefully acknowledge the use of the *DMACRYS* software, from the group of Professor Sally Price at University College London. We would also like to acknowledge Tomasz Pawlak from the Centre of Molecular and Macromolecular Studies, Polish Academy of Sciences, Lodz, Poland for useful discussions regarding the safinamide molecule. Funding for this research was also provided Eli Lilly and Company. All computations reported in this paper were performed using the facilities of the Imperial College Research Computing Service, DOI: 10.14469/hpc/2232. We would like to acknowledge that this work used the Cirrus UK National Tier-2 HPC Service at EPCC (https://www.cirrus.ac.uk) funded by the University of Edinburgh and EPSRC (EP/P020267/1) and we are grateful to the UK Materials and Molecular Modelling Hub, which is partially funded by EPSRC (EP/P020194/1 and EP/T022213/1) for computational resources to conduct second (VASP periodic DFT-d) refinement calculations.

Data availability

All data accompanying this paper are available at https://zenodo.org/records/10433580.

Funding information

The following funding is acknowledged: Engineering and Physical Sciences Research Council (grant No. EP/J014958/1; grant No. EP/J003840/1; grant No. EP/P022561/1; grant No. EP/P020194).

References

Abraham, N. S. & Shirts, M. R. (2020). *J. Chem. Theory Comput.* **16**, 6503–6512.

Bardwell, D. A., Adjiman, C. S., Arnautova, Y. A., Bartashevich, E., Boerrigter, S. X. M., Braun, D. E., Cruz-Cabeza, A. J., Day, G. M., Della Valle, R. G., Desiraju, G. R., van Eijck, B. P., Facelli, J. C., Ferraro, M. B., Grillo, D., Habgood, M., Hofmann, D. W. M., Hofmann, F., Jose, K. V. J., Karamertzanis, P. G., Kazantsev, A. V., Kendrick, J., Kuleshova, L. N., Leusen, F. J. J., Maleev, A. V., Misquitta, A. J., Mohamed, S., Needs, R. J., Neumann, M. A., Nikylov, D., Orendt, A. M., Pal, R., Pantelides, C. C., Pickard, C. J.,

Price, L. S., Price, S. L., Scheraga, H. A., van de Streek, J., Thakur, T. S., Tiwari, S., Venuti, E. & Zhitkov, I. K. (2011). *Acta Cryst.* B**67**, 535–551.
Beran, G. J., Sugden, I. J., Greenwell, C., Bowskill, D. H., Pantelides, C. C. & Adjiman, C. S. (2022). *Chem. Sci.* **13**, 1288–1297.
Bowskill, D. H. (2021). PhD thesis. Imperial College London, UK.
Bowskill, D. H., Sugden, I. J., Konstantinopoulos, S., Adjiman, C. S. & Pantelides, C. C. (2021). *Annu. Rev. Chem. Biomol. Eng.* **12**, 593–623.
Braun, D. E., Hald, P., Kahlenberg, V. & Griesser, U. J. (2021). *Cryst. Growth Des.* **21**, 7201–7217.
Braun, D. E., Vickers, M. & Griesser, U. J. (2019). *Mol. Pharm.* **16**, 3221–3236.
Burdock, G. A. & Carabin, I. G. (2004). *Toxicol. Lett.* **150**, 3–18.
Case, D. H., Campbell, J. E., Bygrave, P. J. & Day, G. M. (2016). *J. Chem. Theory Comput.* **12**, 910–924.
Chemburkar, S. R., Bauer, J., Deming, K., Spiwek, H., Patel, K., Morris, J., Henry, R., Spanton, S., Dziki, W., Porter, W., Quick, J., Bauer, P., Donaubauer, J., Narayanan, B. A., Soldani, M., Riley, D. & McFarland, K. (2000). *Org. Process Res. Dev.* **4**, 413–417.
Cruz-Cabeza, A. J. & Bernstein, J. (2014). *Chem. Rev.* **114**, 2170–2191.
Day, G. M., Cooper, T. G., Cruz-Cabeza, A. J., Hejczyk, K. E., Ammon, H. L., Boerrigter, S. X. M., Tan, J. S., Della Valle, R. G., Venuti, E., Jose, J., Gadre, S. R., Desiraju, G. R., Thakur, T. S., van Eijck, B. P., Facelli, J. C., Bazterra, V. E., Ferraro, M. B., Hofmann, D. W. M., Neumann, M. A., Leusen, F. J. J., Kendrick, J., Price, S. L., Misquitta, A. J., Karamertzanis, P. G., Welch, G. W. A., Scheraga, H. A., Arnautova, Y. A., Schmidt, M. U., van de Streek, J., Wolf, A. K. & Schweizer, B. (2009). *Acta Cryst.* B**65**, 107–125.
European Medicines Agency (2005). *Scientific Conclusions and Grounds for the Suspension of the Marketing Authorisation of Hexavac presented by the EMEA*. https://www.ema.europa.eu/en/documents/scientific-conclusion/hexavac-epar-scientific-conclusion_en.pdf.
Francia, N. F., Price, L. S., Nyman, J., Price, S. L. & Salvalaglio, M. (2020). *Cryst. Growth Des.* **20**, 6847–6862.
Francia, N. F., Price, L. S. & Salvalaglio, M. (2021). *CrystEngComm*, **23**, 5575–5584.
Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Petersson, G. A., Nakatsuji, H., Li, X., Caricato, M., Marenich, A. V., Bloino, J., Janesko, B. G., Gomperts, R., Mennucci, B., Hratchian, H. P., Ortiz, J. V., Izmaylov, A. F., Sonnenberg, J. L., Williams, Ding, F., Lipparini, F., Egidi, F., Goings, J., Peng, B., Petrone, A., Henderson, T., Ranasinghe, D., Zakrzewski, V. G., Gao, J., Rega, N., Zheng, G., Liang, W., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Throssell, K., Montgomery Jr., J. A., Peralta, J. E., Ogliaro, F., Bearpark, M. J., Heyd, J. J., Brothers, E. N., Kudin, K. N., Staroverov, V. N., Keith, T. A., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A. P., Burant, J. C., Iyengar, S. S., Tomasi, J., Cossi, M., Millam, J. M., Klene, M., Adamo, C., Cammi, R., Ochterski, J. W., Martin, R. L., Morokuma, K., Farkas, O., Foresman, J. B. & Fox, D. J. (2016). *Gaussian16*. Rev. C.01. Gaussian Inc., Wallingford, CT, USA.
Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. (2016). *Acta Cryst.* B**72**, 171–179.
Habgood, M., Sugden, I. J., Kazantsev, A. V., Adjiman, C. S. & Pantelides, C. C. (2015). *J. Chem. Theory Comput.* **11**, 1957–1969.
Hafner, J. & Kresse, G. (1997). *Properties of Complex Inorganic Solids*, edited by A. Gonis, A. Meike & P. E. A. Turchi, pp. 69–82. Boston: Springer.
Hu, H., Lu, Z. & Yang, W. (2007). *J. Chem. Theory Comput.* **3**, 1004–1013.
Iuzzolino, L. (2018). PhD thesis. University College London.
Karamertzanis, P. G. & Pantelides, C. C. (2005). *J. Comput. Chem.* **26**, 304–324.
Karamertzanis, P. G. & Pantelides, C. C. (2007). *Mol. Phys.* **105**, 273–291.
Kazantsev, A. V., Karamertzanis, P. G., Adjiman, C. S. & Pantelides, C. C. (2011). *J. Chem. Theory Comput.* **7**, 1998–2016.
Kazantsev, A. V., Karamertzanis, P. G., Adjiman, C. S., Pantelides, C. C., Price, S. L., Galek, P. T., Day, G. M. & Cruz-Cabeza, A. J. (2011). *Int. J. Pharm.* **418**, 168–178.
Kazantsev, A. V., Karamertzanis, P. G., Pantelides, C. C. & Adjiman, C. S. (2010). *Comput. Aided Chem. Eng.* **28**, 817–822.
Lee, A. Y., Erdemir, D. & Myerson, A. S. (2011). *Annu. Rev. Chem. Biomol. Eng.* **2**, 259–280.
Leuratti, C., Sardina, M., Ventura, P., Assandri, A., Müller, M. & Brunner, M. (2013). *Pharmacology*, **92**, 207–216.
Lommerse, J. P. M., Motherwell, W. D. S., Ammon, H. L., Dunitz, J. D., Gavezzotti, A., Hofmann, D. W. M., Leusen, F. J. J., Mooij, W. T. M., Price, S. L., Schweizer, B., Schmidt, M. U., van Eijck, B. P., Verwer, P. & Williams, D. E. (2000). *Acta Cryst.* B**56**, 697–714.
MacKerell, A. D. Jr, Banavali, N. & Foloppe, N. (2000). *Biopolymers*, **56**, 257–265.
Mayo, S. L., Olafson, B. D. & Goddard, W. A. (1990). *J. Phys. Chem.* **94**, 8897–8909.
Neumann, M. A., Leusen, F. J. & Kendrick, J. (2008). *Angew. Chem. Int. Ed.* **47**, 2427–2430.
Nyman, J., Yu, L. & Reutzel-Edens, S. M. (2019). *CrystEngComm*, **21**, 2080–2088.
Pantelides, C. C., Adjiman, C. S. & Kazantsev, A. V. (2014). In *Prediction and Calculation of Crystal Structures: Methods and Applications*, edited by S. Atahan-Evrenk & A. Aspuru-Guzik, pp. 25–28. Springer.
Pawlak, T., Sudgen, I., Bujacz, G., Iuga, D., Brown, S. P. & Potrzebowski, M. J. (2021). *Cryst. Growth Des.* **21**, 3328–3343.
Price, S. L. (2018). *Proc. R. Soc. London Ser. A*, **474**, 20180351.
Price, S. L., Leslie, M., Welch, G. W., Habgood, M., Price, L. S., Karamertzanis, P. G. & Day, G. M. (2010). *Phys. Chem. Chem. Phys.* **12**, 8478–8490.
Racher, F., Petrick, T. L. & Braun, D. E. (2023). *Cryst. Growth Des.* **23**, 4638–4654.
Ravikumar, K. & Sridhar, B. (2010). *Acta Cryst.* C**66**, o317–o320.
Reilly, A. M., Cooper, R. I., Adjiman, C. S., Bhattacharya, S., Boese, A. D., Brandenburg, J. G., Bygrave, P. J., Bylsma, R., Campbell, J. E., Car, R., Case, D. H., Chadha, R., Cole, J. C., Cosburn, K., Cuppen, H. M., Curtis, F., Day, G. M., DiStasio Jr, R. A., Dzyabchenko, A., van Eijck, B. P., Elking, D. M., van den Ende, J. A., Facelli, J. C., Ferraro, M. B., Fusti-Molnar, L., Gatsiou, C.-A., Gee, T. S., de Gelder, R., Ghiringhelli, L. M., Goto, H., Grimme, S., Guo, R., Hofmann, D. W. M., Hoja, J., Hylton, R. K., Iuzzolino, L., Jankiewicz, W., de Jong, D. T., Kendrick, J., de Klerk, N. J. J., Ko, H.-Y., Kuleshova, L. N., Li, X., Lohani, S., Leusen, F. J. J., Lund, A. M., Lv, J., Ma, Y., Marom, N., Masunov, A. E., McCabe, P., McMahon, D. P., Meekes, H., Metz, M. P., Misquitta, A. J., Mohamed, S., Monserrat, B., Needs, R. J., Neumann, M. A., Nyman, J., Obata, S., Oberhofer, H., Oganov, A. R., Orendt, A. M., Pagola, G. I., Pantelides, C. C., Pickard, C. J., Podeszwa, R., Price, L. S., Price, S. L., Pulido, A., Read, M. G., Reuter, K., Schneider, E., Schober, C., Shields, G. P., Singh, P., Sugden, I. J., Szalewicz, K., Taylor, C. R., Tkatchenko, A., Tuckerman, M. E., Vacarro, F., Vasileiadis, M., Vazquez-Mayagoitia, A., Vogt, L., Wang, Y., Watson, R. E., de Wijs, G. A., Yang, J., Zhu, Q. & Groom, C. R. (2016). *Acta Cryst.* B**72**, 439–459.
Rietveld, I. B. & Céolin, R. (2015). *J. Pharm. Sci.* **104**, 4117–4122.
Schmidt, J. A., Weatherby, J. A., Sugden, I. J., Santana-Bonilla, A., Salerno, F., Fuchter, M. J., Johnson, E. R., Nelson, J. & Jelfs, K. E. (2021). *Cryst. Growth Des.* **21**, 5036–5049.
Schmidt, M. U., Dinnebier, R. E. & Kalkhof, H. (2007). *J. Phys. Chem. B*, **111**, 9722–9732.
Shunnar, A. F., Dhokale, B., Karothu, D. P., Bowskill, D. H., Sugden, I. J., Hernandez, H. H., Naumov, P. & Mohamed, S. (2020). *Chem. A Eur. J.* **26**, 4752–4765.

Sugden, I., Adjiman, C. S. & Pantelides, C. C. (2016). *Acta Cryst.* B**72**, 864–874.
Sugden, I. J., Adjiman, C. S. & Pantelides, C. C. (2019). *Acta Cryst.* B**75**, 423–433.
Sugden, I. J., Braun, D. E., Bowskill, D. H., Adjiman, C. S. & Pantelides, C. C. (2022). *Cryst. Growth Des.* **22**, 4513–4527.
Sun, H., Ren, P. & Fried, J. R. (1998). *Comput. Theor. Polym. Sci.* **8**, 229–246.
Tchoń, D., Bowskill, D., Sugden, I., Piotrowski, P. & Makal, A. (2021). *J. Mater. Chem. C.* **9**, 2491–2503.
van Eijck, B. P., Mooij, W. T. & Kroon, J. (2001). *J. Comput. Chem.* **22**, 805–815.
Wade, J., Salerno, F., Kilbride, R. C., Kim, D. K., Schmidt, J. A., Smith, J. A., LeBlanc, L. M., Wolpert, E. H., Adeleke, A. A., Johnson, E. R., Nelson, J., Mori, T., Jelfs, K. E., Heutz, S. & Fuchter, M. J. (2022). *Nat. Chem.* **14**, 1383–1389.
Williams, D. E. & Cox, S. R. (1984). *Acta Cryst.* B**40**, 404–417.
Yang, S. & Day, G. M. (2022). *Commun. Chem.* **5**, 86.
Zhang, P., Wood, G. P., Ma, J., Yang, M., Liu, Y., Sun, G., Jiang, Y. A., Hancock, B. C. & Wen, S. (2018). *Cryst. Growth Des.* **18**, 6891–6900.