STRUCTURAL SCIENCE CRYSTAL ENGINEERING MATERIALS

ISSN 2052-5206

Accurate and efficient representation of intramolecular energy in *ab initio* generation of crystal structures. II. Smoothed intramolecular potentials

Isaac J. Sugden, Claire S. Adjiman and Constantinos C. Pantelides

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

Received 11 January 2019
Accepted 27 April 2019

Edited by T. R. Welberry, Australian National University, Australia

Keywords: crystal structure prediction; computational chemistry; crystallography.

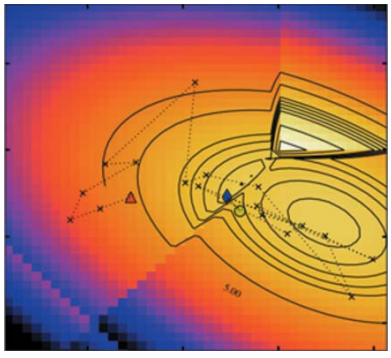
Supporting information: this article has supporting information at journals.iucr.org/b

The application of crystal structure prediction (CSP) to industrially relevant molecules requires the handling of increasingly large and flexible compounds. A revised model for the effect of molecular flexibility on the lattice energy that removes the discontinuities and non-differentiabilities present in earlier models (Sugden *et al.*, 2016), with a view to improving the performance of CSP is presented. The approach is based on the concept of computing a weighted average of local models, and has been implemented within the *CrystalPredictor* code. Through the comparative investigation of several compounds studied in earlier literature, it is shown that this new model results in large reductions in computational effort (of up to 65%) and in significant increases in reliability. The approach is further applied to investigate, for the first time, the computational polymorphic landscape of flufenamic acid for $Z' = 1$ structures, resulting in the successful identification of all three experimentally resolved polymorphs within reasonable computational time.

1. Introduction

Crystal structure prediction (CSP) methods have seen significant progress in the past decade, with molecules of a size and flexibility relevant to pharmaceutical industry now within the practical reach of existing approaches and software (Price *et al.*, 2016). This is shown by the increasing size and complexity of the target molecules studied in the series of blind tests organized by the Cambridge Crystallographic Data Centre (CCDC) (Reilly *et al.*, 2016; Bardwell *et al.*, 2011; Day *et al.*, 2009).

To date, most approaches to CSP have been based on the assumption that the crystal structures most likely to be observed are the ones corresponding to the global minimum in the lattice energy and those local minima with a lattice energy within a few kJ mol^{-1} of the global minimum energy. Thus, a key step in CSP is the generation of structures that are low-lying local minima in the crystal lattice energy landscape. While no method yet offers a guarantee of success, all methods that have been found to achieve frequent success in finding the experimentally resolved structures within these low energy minima (Case *et al.*, 2016; Habgood *et al.*, 2015; Price *et al.*, 2010; Neumann *et al.*, 2008; Karamertzanis & Pantelides, 2005) have been based on a multi-stage methodology: in the first ('global search') stage, a lower-accuracy energy model is used to search the energy landscape efficiently and globally. This is followed by one or more 'refinement' stages, at each of which the most promising structures from the previous stage are re-minimized using a more accurate, but also more expensive, energy model. This successive treatment has proven to be generally reliable in achieving sufficient accuracy whilst



covering the search space in a computationally tractable manner.

The above strategy has been deployed successfully in the blind tests and in the prediction of the crystal structures of pharmaceutically relevant molecules (Braun *et al.*, 2016, 2017; Price *et al.*, 2016). However, in view of the need to tackle ever larger and more flexible molecules of pharmaceutical interest, further improvements to the efficiency of the algorithms are required. To address this need, this paper proposes a lattice energy model that improves the efficiency and reliability of the global search stage, thereby enabling higher dimensional problems to be considered. Of particular interest in this work is the *CrystalPredictor* global search methodology developed by our research group (Karamertzanis & Pantelides, 2005, 2007; Habgood *et al.*, 2015; Sugden *et al.*, 2016) and currently being used by several research groups (Rice *et al.*, 2018; Braun & Griesser, 2018; Aina *et al.*, 2017) to generate initial lattice energy landscapes.

In *CrystalPredictor*, a large number, typically of the order of 10^5 to 10^6, of starting structures are generated by means of a search based on a low-discrepancy sequence (Sobol, 1967) and carried out across many space groups. Each such structure is characterized by a different unit cell geometry and molecular positions, orientations and conformations, and serves as the starting point for a local lattice energy minimization. The lattice energy model used for this purpose needs to be both accurate and inexpensive to allow the generation of a meaningful ranking of the putative crystal structures within reasonable computational effort. Developing such a model is especially challenging for large flexible molecules where the intramolecular energy needs to be computed accurately (*i.e.* at the quantum mechanical (QM) level) as a function of relatively large numbers of conformational degrees of freedom.

1.1. Local approximate models

The most recent version of the *CrystalPredictor* algorithm (Sugden *et al.*, 2016) makes use of local approximate models (LAMs) for the description of the effects of molecular conformation. Initially proposed in the context of the *CrystalOptimizer* (Kazantsev *et al.*, 2011) refinement algorithm, this approach allows the global search to make use of intramolecular energy values of near-QM accuracy within reasonable computational effort.

The LAM approach partitions the conformational degrees of freedom in a molecule into dependent and independent variables. The latter are those that are directly affected by intermolecular interactions within the crystallographic environment; they typically comprise flexible torsion angles and in some cases, bond angles, that have a significant impact on the lattice energy. On the other hand, the dependent degrees of freedom undergo small adjustments (*e.g.* most bond angles and bond lengths) or have a limited impact on the overall lattice energy (*e.g.* a torsion angle describing a methyl group rotation); they are assumed to adjust themselves to the values of the independent degrees of freedom so as to minimize the intramolecular energy of the molecule.

The LAMs are generated prior to the global search phase by performing isolated QM calculations at a set of reference points $\left\{\theta_k^{\rm ref},\ k = 1, .., K\right\}$ in the space of the independent degrees of freedom θ, and then constructing local approximants to the results of these calculations. This produces explicit expressions for the intramolecular energy, the dependent degrees of freedom and the quantities (*e.g.* atomic charges or multipoles) characterizing the electrostatic potential. Being simple (*e.g.* quadratic or linear) functions of θ, these approximants provide a computationally efficient way of computing the corresponding properties during the global search.

In general, any LAM k is accurate only in the vicinity of the reference point $\theta_k^{\rm ref}$ from which it has been generated. In view of the high cost of QM calculations, a key consideration is the selection of the set $\left\{\theta_k^{\rm ref},\ k = 1, .., K\right\}$ so as to accurately cover the entire θ-domain of interest with the minimum number of points, K. The original LAM implementation (Habgood *et al.*, 2015) made use of a sufficiently fine uniform grid over the θ-domain. More recently, the first part of this paper (Sugden *et al.*, 2016) has proposed an adaptive algorithm that starts from a coarse grid of reference points and adds new 'off-grid' ones only in crystallographically relevant regions of conformational space where the error in the intramolecular energy values predicted by the LAMs is less than a required threshold. This significantly reduces the number of required points (and therefore, QM calculations) for molecules with significant flexibility (*i.e.* many independent degrees of freedom, θ). For example, a reduction of approximately 70% was achieved in the cost of the global search stage for molecule (XXVI) from the sixth blind test (2-chloro-*N*-{2′-[(2-chlorobenzoyl)amino]-1,1′-binaphthalen-2-yl}benzamide) (Wheeler & Hopkins, 2016).

1.2. Discontinuities introduced by LAMs

During the global search phase, conformation-dependent quantities at any point θ in the space of the independent conformational degrees of freedom may be computed via an evaluation of the LAM expressions corresponding to the reference point $\theta_k^{\rm ref}$ that is nearest to θ, with distance being measured, for example, via a standard Euclidean norm $||\theta - \theta_k^{\rm ref}||_2$.

Such evaluations are very efficient and, provided the reference points have been selected correctly, they provide near-QM accuracy in the computed quantities and also their partial derivatives with respect to θ. However, the values obtained for two points in θ-space that are arbitrarily close to each other may be different if they appear on either side of the boundary between two adjacent LAMs. This complication potentially poses severe challenges to the gradient-based algorithms that are employed for lattice energy minimization. More specifically, in cases where successive iterates make use of different LAMs, the algorithm may exhibit slow convergence, oscillation and, very often, failure to converge. This effect may be quite serious especially if, in an effort to increase accuracy, we make use of larger numbers of LAMs, which

makes it more probable than two successive iterates will be on either side of a LAM boundary. For instance, 85% of the minimizations in the molecule (XXVI) investigation mentioned above failed to terminate normally as a result of such discontinuities.

It is possible to devise *ad hoc* approaches in an attempt to address the above issue. For example, one may increase the number of candidate structures generated in the hope that at least one of them will lie sufficiently closely to, and within the same LAM as each low-energy minimum. Another approach is to relax the requirement for strict local minima to be determined, *e.g.* by treating the final points of failed local minimizations as potentially successful ones, thereby allowing them to pass to the subsequent refinement stage provided they are of sufficiently low energy. However, this shifts a higher computational burden onto the final stage as the structures to be refined will be further from the true minima and will be greater in number. Overall, neither of these approaches is really satisfactory and they come at a significantly increased computational cost.

In this paper, we propose a weight-averaged LAM approach that eliminates the discontinuities mentioned above. The new approach is presented in Section 2, and is illustrated via its application to the ROY molecule {5-methyl-2-[(2-nitrophenyl)amino]-3-thiophenecarbonitrile} (Yu, 2010; Vasileiadis *et al.*, 2012). Section 3 applies the approach to CSP investigations involving three larger molecules, including, for the first time, flufenamic acid {*N*-[3-(trifluoromethyl)phenyl]-anthranilic acid, *N*-(α,α,α-trifluoro-*m*-tolyl)anthranilic acid}.

2. A continuous and differentiable LAM-based intramolecular potential

In this section, we briefly summarize the LAMs currently being used in *CrystalPredictor*, and then introduce the concept of weight-averaged LAMs. We then gain some insight as to the effects of this modification by applying it to the case of the ROY molecule.

2.1. LAMs for effects of molecular conformation

As described in the first part of this paper (Sugden *et al.*, 2016), the LAMs used by *CrystalPredictor* for describing the effects of the molecular conformation at given values of the independent configurational degrees of freedom θ in the vicinity of a reference point $\theta_k^{\rm ref}$ are as follows.

(i) For the intramolecular energy difference $\Delta U^{\rm intra} \equiv \min_{\bar\theta} U^{\rm intra}(\bar\theta, \theta) - U^{\rm intra,gas}$:

$$\Delta U_k^{\rm intra}(\theta) = \Delta U^{\rm intra}(\theta_k^{\rm ref}) + \mathbf{b}_k^T(\theta - \theta_k^{\rm ref}) + \frac{1}{2}(\theta - \theta_k^{\rm ref})^T \mathbf{C}_k (\theta - \theta_k^{\rm ref}). \quad (1)$$

(ii) For the dependent conformational degrees of freedom $\bar\theta$

$$\bar\theta_k(\theta) = \bar\theta(\theta_k^{\rm ref}) + \mathbf{A}_k(\theta_k - \theta_k^{\rm ref}), \quad (2)$$

where $\Delta U^{\rm intra}(\theta_k^{\rm ref})$ and $\bar\theta(\theta_k^{\rm ref})$ are obtained from the results of a QM calculation minimizing lattice energy at fixed values $\theta_k^{\rm ref}$ of the independent degrees of freedom. $\mathbf{A}_k$ and $\mathbf{C}_k$ are constant matrices and $\mathbf{b}_k$ are constant vectors; expressions for them are given in Part I of this paper.

The description of intermolecular electrostatic interactions in *CrystalPredictor* is in terms of atomic charges. Within the range of validity of LAM k, these are held constant at values derived from the QM electronic density at the corresponding reference point $\theta_k^{\rm ref}$:

$$q_k(\theta) = q(\theta_k^{\rm ref}). \quad (3)$$

2.2. Weight-averaged LAMs

As has already been mentioned, at any particular point θ, the current version of *CrystalPredictor* makes use of the expressions (1) to (3) for the LAM k whose reference point $\theta_k^{\rm ref}$ is closest to θ. This is the source of the discontinuities introduced as point θ crosses the boundaries between adjacent LAMs.

Instead of using a single LAM for any particular point θ, here we propose to employ weighted sums of the contributions of all LAMs, of the form:

$$\Delta U^{\rm intra,wa}(\theta) = \left[\sum_{k=1}^{K} w_k(\theta)\right]^{-1} \sum_{k=1}^{K} w_k(\theta)\Delta U_k^{\rm intra}(\theta) \quad (4)$$

$$\bar\theta^{\rm wa}(\theta) = \left[\sum_{k=1}^{K} w_k(\theta)\right]^{-1} \sum_{k=1}^{K} w_k(\theta)\bar\theta_k(\theta) \quad (5)$$

$$q^{\rm wa}(\theta) = \left[\sum_{k=1}^{K} w_k(\theta)\right]^{-1} \sum_{k=1}^{K} w_k(\theta) q(\theta_k^{\rm ref}), \quad (6)$$

where the weights $w_k(\theta)$ are positive functions of θ defined as exponentially decaying functions of the square of the distance of θ from the reference point $\theta_k^{\rm ref}$:

$$w_k(\theta) = \exp\left(-\frac{1}{s^2}||\theta - \theta_k^{\rm ref}||_2^2\right), \quad (7)$$

where s is a positive smoothing parameter.

The superscript wa on the functions defined in expressions (4) to (6) indicates that they are weighted averages that are applied throughout the θ-domain of interest, essentially ignoring the boundaries between LAMs. Therefore, no discontinuity is encountered anywhere in this domain. In principle, any point θ receives contributions from all LAMs, but these contributions decrease exponentially with increasing distance. In practice, given an appropriate choice of the smoothing parameter, s, only the nearest LAMs have non-negligible contributions. Our current implementation makes use of a default value of s = 0.4.

The partial derivatives of $\Delta U^{\rm intra,wa}$, $\bar\theta^{\rm wa}(\theta)$ and $q^{\rm wa}(\theta)$ with respect to the independent degrees of freedom θ are important quantities for the purposes of the local minimization of lattice energy. It can be shown that these gradients are given by expressions of the form:

$$\frac{\partial X^{\rm wa}}{\partial\theta}(\theta)=\left[\sum_{k=1}^{K}w_k(\theta)\right]^{-1}\sum_{k=1}^{K}\left\{w_k(\theta)\frac{\partial X_k}{\partial\theta}(\theta)+\left[X_k(\theta)-X^{\rm wa}(\theta)\right]\frac{\partial w_k}{\partial\theta}(\theta)\right\},\tag{8}$$

where X denotes any one of the three configuration-dependent quantities $\{\Delta U^{\rm intra},\bar{\theta},q\}$ and the gradients of the weighting functions $w_k(\theta)$ are given by:

$$\frac{\partial w_k}{\partial\theta}(\theta)=\frac{2}{s^2}w_k(\theta)(\theta-\theta_k^{\rm ref}).\tag{9}$$

Hereafter, the model based on a single LAM point (equations (1) to (3)) will be referred to as the 'LAM' model, whilst the smoothed potential model, described by equations (4) to (7) will be referred to as the weight-averaged LAM ('waLAM') model.

2.3. Illustrative example: ROY

In order to better understand the effects of the proposed approach, we use the much studied, small-dimensional ROY molecule. As shown in Fig. 1, this is modelled using two independent degrees of freedom corresponding to the torsion angles T1 and T2.

Using the adaptive LAM method detailed in Sugden *et al.* (2016), a grid comprising 38 LAM points is constructed in the range T1 $\in[-20^\circ,180^\circ]$ and T2 $\in[100^\circ,260^\circ]$. To illustrate the problems caused by discontinuities at the LAM boundaries, we use this grid to perform a local lattice energy minimization using the experimental form OP as the starting point. The progress of the optimization iterations is shown in Fig. 2(*a*) in terms of sequences of values of the lattice energy (left) and the norm of its gradients (right). It can be seen that the lattice energy decreases at each iteration until iteration 16. However, thereafter, the energy begins to oscillate between two distinct values, and the optimization algorithm is not able to decrease the gradient norm further. Eventually, the algorithm terminates unsuccessfully because it is not able to identify a search direction resulting in a decrease in the lattice energy even if the optimization step is reduced to the minimum allowable size.

On the other hand, as illustrated in Fig. 2(*b*), using the smoothed waLAMs allows the iterations to proceed smoothly until convergence is achieved after 18 iterations. Fast super-linear convergence is achieved during the last few iterations, as showed by the linear decrease of the logarithm of the gradient norm [see right part of Fig. 2(*b*)].

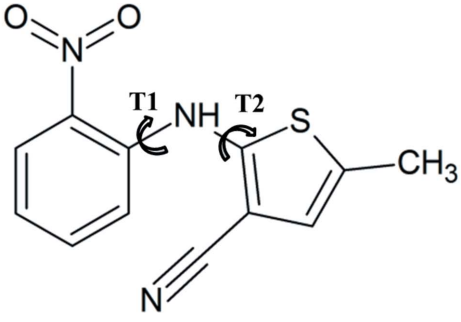


Figure 1
ROY molecule, with flexible torsions T1 and T2 indicated

Fig. 3 provides some insight on the origins of the different behaviour of the two models. The intramolecular energy surfaces across the range of values of the flexible torsions, as seen in Figs. 3(*a*) and 3(*b*), are qualitatively similar but the LAM model is slightly more jagged. In Figs. 3(*c*) and 3(*d*), we zoom in around the minimization paths corresponding to the iteration sequences shown in Fig. 2. We can see that the lattice energy minimum occurs very close to a LAM boundary. Consequently, the discontinuity in intramolecular energy exhibited by the original LAM model [Fig. 3(*c*)] results in a rather erratic behaviour, which prevents convergence to the total lattice energy minimum [Fig. 3(*e*)]. In contrast, as seen in Fig. 3(*d*), the waLAM model's continuous and differentiable intramolecular energy function allows the minimization to proceed to the lattice energy minimum [see Fig. 3(*f*)].

We now consider the effects of the discontinuities on the efficiency and effectiveness of the global search for the ROY molecule by using *CrystalPredictor* to perform lattice energy minimizations starting from 500 000 candidate structures. Of these, about 48 000 minimizations terminate at a point on the limits of the torsion angle range T1 $\in[-20^\circ,180^\circ]$ and T2 $\in[100^\circ,260^\circ]$ under consideration. Although these are correct local minima from the mathematical point of view, they are not true local lattice energy minima and are discarded from further consideration. More interestingly, the discontinuities at the boundaries of the original LAMs cause an additional 275 000 minimizations to fail to satisfy the convergence criterion, in the manner that has already been illustrated by Figs. 2 and 3. In our earlier studies, we considered the final points achieved by such minimizations as potentially close to true local minima, and therefore as putative structures for further refinement. Overall, this resulted in 6 539 structures being within 20 kJ mol^{-1} of the global minimum and, conse-

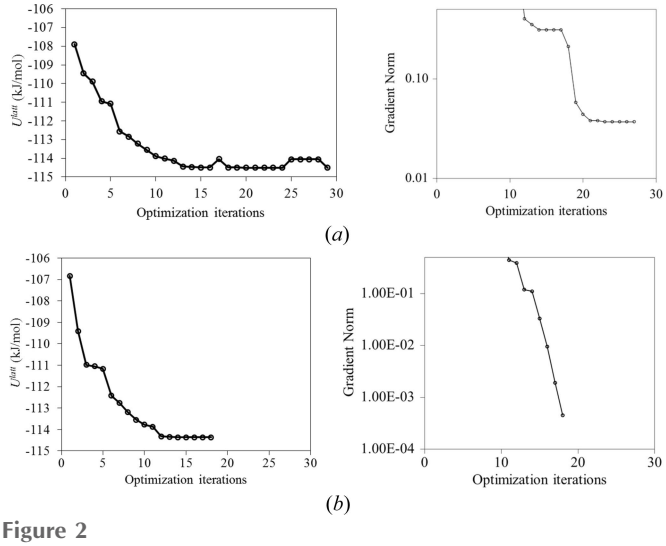


Figure 2
Progress of iterations in the minimization of experimental form OP of the molecule ROY, using the same LAM grid. (*a*) Iteration sequence using original LAMs. (*b*) Iteration sequence using waLAMs

quently, having to be refined by *CrystalOptimizer*. The latter successfully identified all experimentally known forms, including PO13 (Nyman *et al.*, 2019) which had not yet been resolved experimentally at the time of the original investigation. However, it is worth noting that, at least one of these forms (the YT04 one) would not have been identified without the *ad hoc* relaxation of the optimality requirements described above.

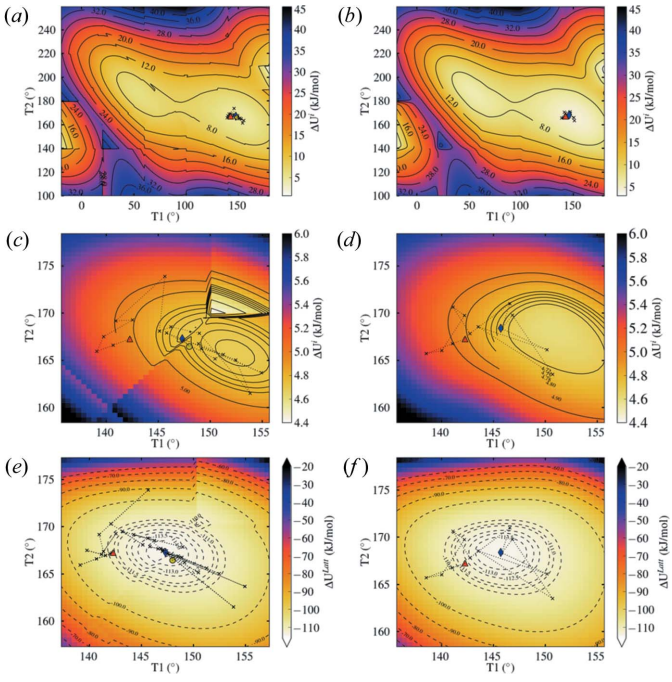


Figure 3
Minimization path for the OP (QAXMEH03) polymorph of ROY. (*a*) The intramolecular energy surface across the entire range of torsions T1 and T2 calculated using the LAM model and (*b*) using the waLAM model. (*c*) Intramolecular energy surface across the range of T1 and T2 spanned by the full minimization path, as seen in Fig. 2, for the LAM model, and (*d*) for the waLAM model; the red triangles and blue diamonds represent starting and finishing points respectively, whilst the yellow circle represents the higher-energy conformation that is sampled many times in the minimization using the LAM model. (*e*) Total lattice energy surface obtained with the LAM model with all optimization decision variables other than the flexible torsions (*i.e.* cell lengths and angles, and molecule's position and orientation) held at their final values, for the sake of visualization, and (*f*) the same surface with the waLAM model.

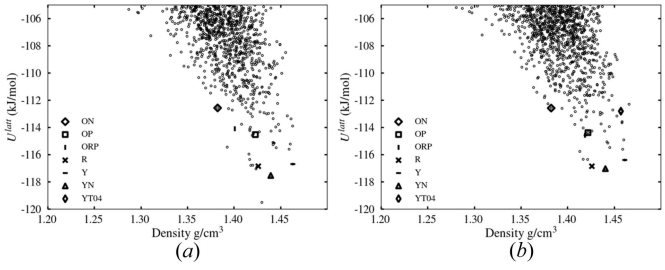


Figure 4
Polymorphic landscape for the ROY molecule determined by *CrystalPredictor* after 500 000 minimizations, under the (*a*) LAM and (*b*) waLAM models. The structures that match the experimentally resolved polymorphs are denoted by the symbols shown in the legend.

Table 1
Number of minimizations carried out before each experimental form of ROY is first identified by *CrystalPredictor* with the LAM and waLAM models.

	Minimizations before form is found	
Experimental form	LAM model	waLAM model
ON	2 292	90 576
OP	70 242	87 819
ORP	698 870	22 292
R	42 628	27 585
Y	185 845	24 737
YN	12 578	3 316
YT04	902 906	252 079
PO13	6 586	4 484

The new waLAM model eliminates all convergence failures caused by discontinuities. Because the corresponding minimizations terminate successfully within fewer iterations, the cost of the global search is reduced by approximately 36%. Moreover, only 2 277 unique structures occur within 20 kJ mol^{-1} of the global minimum and are therefore passed on to the refinement stage, thereby reducing the cost of the latter by about 65%.

The polymorphic landscapes obtained using the two models are shown in Fig. 4. It can be seen that the use of the waLAM model results in a sparser landscape, having eliminated many of the fictitious points (including one appearing as the global minimum) predicted using the original LAM model.

Table 1 reports the number of local minimizations required before each experimental form is identified. With the waLAM model, structures corresponding to all eight experimental forms are found within the first 260 000 minimizations of the global search. In contrast, the LAM model manages to identify only six of the experimental forms within this number of

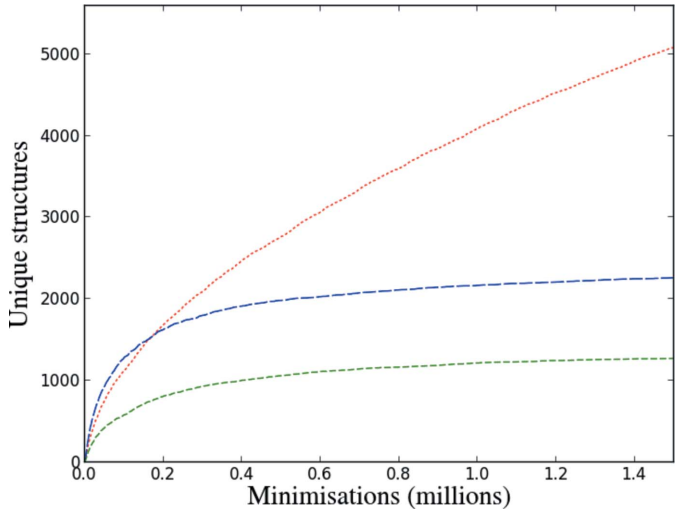


Figure 5
Number of unique structures for ROY found within +20 kJ mol^{-1} of the final global minimum as a function of the number of minimizations carried out by *CrystalPredictor* using the waLAM model (long dashed curve) and LAM model (short dashed curve). The dotted curve shows the number of unique structures if structures that fail to converge are also included for the LAM model.

minimizations; in fact, it requires a total of 902 906 minimizations before the remaining two forms are identified, and even then, the minimization leading to structure YT04 does not actually converge successfully.

Fig. 5 shows the number of unique structures identified as a function of the number of minimizations being carried out. It is evident that the waLAM model results in many more unique structures being identified within the same number of minimizations, unless one counts the structures from minimizations that fail to converge properly because of the discontinuities in the LAM model. Another interesting metric is the number of minimizations required for the global search to be deemed as complete, something considered to be the case if fewer than 50 new unique structures are generated in 50 000 successive minimizations. The search using the waLAM model reaches this point after 1.078 million minimizations whilst that using the original LAMs requires 1.316 million minimizations. Overall, these results illustrate the superiority of the waLAM model in terms of the efficiency and effectiveness of searching the lattice energy landscape.

3. Application of the waLAM model to more complex molecules

We now consider the application of the waLAM model to three challenging molecules and investigate the resulting impact on the performance of *CrystalPredictor* and on the cost of CSP studies.

3.1. Molecule (XXVI)

Molecule (XXVI) from the sixth blind test (Reilly *et al.*, 2016) involves seven independent configurational degrees of freedom, as shown in Fig. 6. The molecule was used in the first part of this work (Sugden *et al.*, 2016) to illustrate the ability of the adaptive LAM scheme to handle highly flexible molecules during the global search. In our investigation, the search was performed over the domains shown in the second column Table 2; the third column shows the values of these torsion angles in the known experimental form (Reilly *et al.*, 2016).

The use of the adaptive LAM algorithm ensured sufficient accuracy in the computation of the effects of molecular conformation using only 3 463 LAMs over the entire seven-dimensional space (Sugden *et al.*, 2016), thus rendering the investigation of this large molecule computationally tractable.

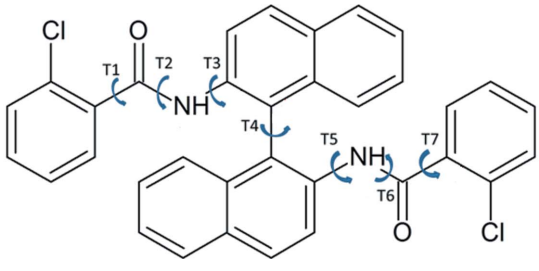


Figure 6
Molecular diagram of molecule (XXVI), with independent configurational degrees of freedom T1–T7 indicated.

Table 2
Search domain for CSP study on molecule (XXVI).

Independent degree of freedom, θ (*cf.* Fig. 6)	Search domain (°)	Experimental value in form 1 (°)
T1	[0, 360]	215.76
T2	[165, 195]	181.26
T3	[20, 260]	163.34
T4	[65, 125]	78.47
T5	[20, 260]	222.46
T6	[165, 195]	185.06
T7	[0, 360]	301.872

The global search involved one million minimizations. The 1 413 structures within 30 kJ mol^{-1} of the global minimum were taken through to the refinement stage, following which the experimental form was identified as the global minimum.

Notwithstanding the above success, it should be noted that about 47.5% of the minimization carried out during the global search resulted in solutions on the limits of the flexible angle domains listed in Table 2, and therefore do not correspond to local minima in the lattice energy surface. More importantly, an additional 50.5% of the minimizations actually failed to satisfy the convergence criterion due to discontinuities at LAM boundaries; these are clearly visible in Fig. 7(*a*) which shows a projection of the intramolecular energy surface on the T1–T7 plane. Overall, 1 283 of the 1 413 structures taken to the refinement stage actually resulted from such failed minimizations.

For the purposes of this paper, the set of 3 643 LAMs determined in Part I of this paper (Sugden *et al.*, 2016) via the adaptive LAM placement algorithm is recomputed making use of a modified molecular representation (*Z*-matrix) that results in more stable calculations for certain large molecules. As in Part I and in the interests of computational expedience, only one pass of the adaptive algorithm is applied. This results in some loss of accuracy, with some of the LAMs *k* predicting negative within their domain of applicability. Although this problem can be eliminated by allowing the adaptive algorithm to proceed to conclusion, here we use the restricted LAM set and simply replace any negative by zero. In any case, we use

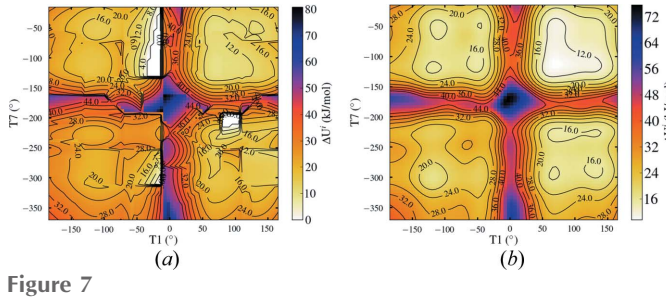


Figure 7
Intramolecular energy surfaces obtained through a 5° scan across the range of torsions T1 and T7 for molecule (XXVI) with (*a*) the LAM model and (*b*) the waLAM model. The remaining torsions are fixed at the midpoints of their ranges.

Table 3
CSP results for molecule (XXVI).

	CrystalPredictor global search involving 500 000 minimizations		
	LAM model	waLAM model	*CrystalOptimizer* refinement
Minimizations successfully converged at local lattice energy minima	7.8%	61.2%	N/A
Distinct structures within 20 kJ mol^{-1} of global minimum	123	86	
Structure corresponding to experimentally resolved polymorph			
Experimental RMSD$_{15}$ (Å)	0.345	0.345	0.33
Lattice energy (kJ mol^{-1})	−190.0596	−189.6025	−212.59
Lattice energy difference from global minimum (kJ mol^{-1})	+17.4519	+18.0371	0.0
Rank	85	56	1

exactly the same set of LAMs and the same approach with both the LAM and the waLAM models.

Fig. 7(*b*) shows the projection of the intramolecular energy approximation obtained using the waLAM model on the T1-T7 plane. A comparison with the corresponding surface obtained using the original LAM model [*cf.* Fig. 7(*a*)] clearly indicates that the waLAM model results in a much smoother surface.

Some key statistics for the global search carried out using the LAM and waLAM models are shown in Table 3. It can be seen that the waLAM model achieves an increase in the percentage of minimizations that converge to a true local lattice energy minimum from 7.8% to 61.2%. About 27% of the minimizations converge to points on the limits of the torsion angle range and do not, therefore, correspond to stable lattice energy minima; the remaining failures are primarily caused by the use of a restricted LAM set. The waLAM model also results in reducing the number of structures within the usual 20 kJ mol^{-1} cutoff of the global minimum from 123 to 86. The sparser nature of the lattice energy landscape obtained by the waLAM model is evident in Fig. 8, especially as far as the lowest energy structures (*e.g.* those within 10 kJ mol^{-1} of the global minimum) are concerned.

For both models, the structure corresponding to the experimentally resolved form is well within the 20 kJ mol^{-1} cutoff and is therefore carried over to the refinement stage using *CrystalOptimizer*. In both cases, this results in the experimental form being identified as the lowest ranked crystal structure, with a good reproduction of geometry (a RMSD$_{15}$ of 0.33 Å).

Overall, the key benefits arising from the use of the waLAM model for this particular molecule are increased reliability and greater confidence in the completeness of the polymorphic landscape, in addition to a reduced overall computational cost.

3.2. β-D-Glucose

Another challenging example that can be used to investigate the impact of the waLAM model is that of glucose, a highly flexible molecule with several intramolecular hydrogen bonds, as seen in Fig. 9. This molecule was considered in an earlier study (Habgood *et al.*, 2015) using a uniform LAM grid constructed over the five independent configurational degrees of freedom T1–T5 indicated in Fig. 9. Each of the torsion angles was considered to be fully flexible, potentially varying from 0° to 360°. With a grid spacing of ±30°, this resulted in a uniform grid comprising 7 776 (= 6^5) LAMs.

In order to make a valid comparison between the two LAM models, here we employ exactly the same set of LAMs, as well

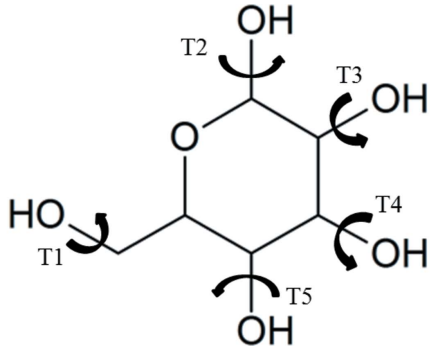


Figure 9
Molecular diagram of glucose, with independent configurational degrees of freedom T1–T5 indicated.

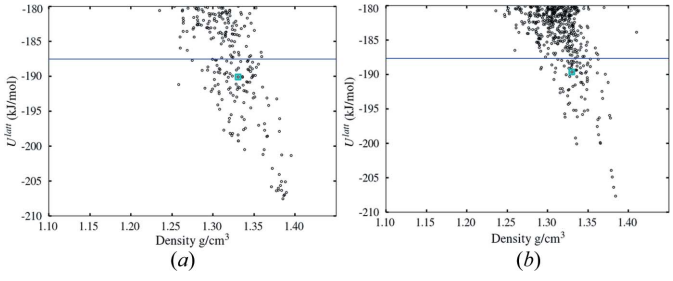


Figure 8
Polymorphic landscape for molecule (XXVI), after a global search involving 500 000 minimizations using (*a*) the LAM model and (*b*) the waLAM model. The open circles denote the structures generated, the open square the experimental structure and the blue line the 20 kJ mol^{-1} cutoff from the global minimum.

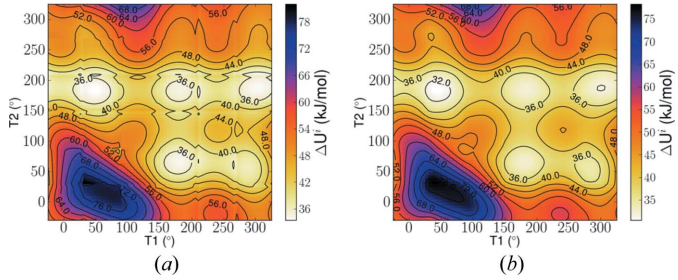


Figure 10
Variation of intramolecular energy surface for glucose over torsion angles T1 and T2, with the remaining torsions fixed at the midpoint of their ranges, obtained using (*a*) the LAM model and (*b*) the waLAM model.

as the same level of QM theory and repulsion-dispersion potential as in the original study (Habgood *et al.*, 2015). A comparison of the variation of intramolecular energy surfaces over T1 and T2, obtained using the LAM and waLAM models is shown in Fig. 10. As with the previous examples, despite the qualitative similarity, the discontinuities arising from the original LAM model are clearly visible.

The original study (Habgood *et al.*, 2015) performed a global search involving the generation and minimization of 500 000 candidate structures. About 25% of these minimizations ended up with minima on the limits of the torsion domains (*i.e.* with at least one of the torsion angles T1–T5 having a value of either 0° or 360°) and therefore did not correspond to local minima on the lattice energy boundaries. Another 18% of the minimizations failed to satisfy the convergence criterion due to discontinuities on LAM boundaries. In fact, the experimental form corresponded to a structure observed with rank 587, located at 36.4 kJ mol^{-1} above the global minimum. This very large gap is primarily due to a significant underestimation of the energy of the global minimum, again arising from the inability of the optimization algorithm to reach proper convergence due to discontinuities at the LAM boundaries[1]. As a result, the experimental form was successfully identified by the overall CSP algorithm only by setting the threshold of structures being passed from the global search stage (*CrystalPredictor*) to the refinement one (*CrystalOptimizer*) at an unusually high value of 40 kJ mol^{-1}. Overall, a total of 1 160 structures were taken through to the refinement stage.[1]

The use of the waLAM model for the global search leads to markedly improved performance. Thus, minimization failures due to discontinuities at LAM boundaries are almost eliminated, and a structure corresponding to the experimental form is identified with rank 8, just 3.96 kJ mol^{-1} above the global minimum. The application of the standard 20 kJ mol^{-1} cutoff for structures to be taken through to the refinement stage is therefore quite adequate. Overall, 408 structures were refined, a reduction of about 65% over the number refined in the original study. A comparison of the lattice energy landscapes identified by *CrystalPredictor* using the two LAM models is presented in Fig. 11.

Based on the recommendation by Cooper *et al.* (2008), the final refinement using *CrystalOptimizer* is carried out with the dielectric constant for PCM calculations set to 7.0 instead of the value of 3.0 used in the original investigation. As can be seen in Fig. 12, the experimental form is ranked second, only 0.77 kJ mol^{-1} above the global minimum.

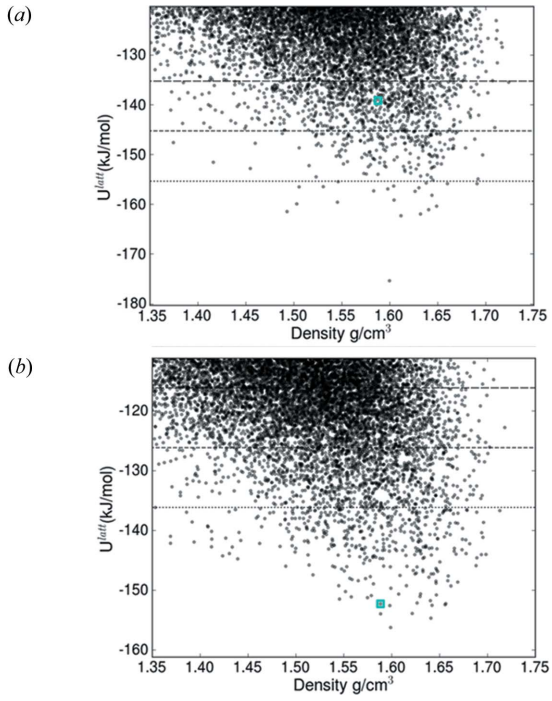


Figure 11
Polymorphic landscape generated with *CrystalPredictor* for glucose, using (*a*) the LAM model and (*b*) the waLAM model. The circles denote the generated structures and the square indicates the experimental structure. The dotted line indicates the 20 kJ mol^{-1} cutoff, the small dashed line the 30 kJ mol^{-1} cutoff and the large dashed line the 40 kJ mol^{-1} cutoff.

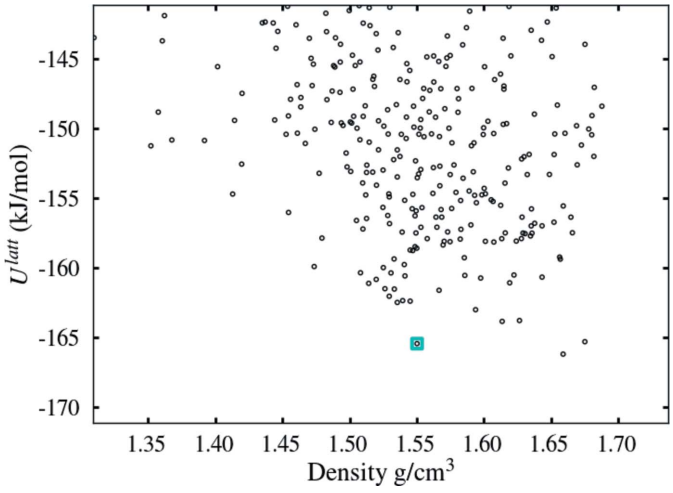


Figure 12
Polymorphic landscape generated with *CrystalOptimizer* for glucose. The square represents the experimental form.

3.3. Flufenamic acid

The new waLAM model is now used to undertake a CSP study for flufenamic acid, a highly polymorphic molecule with nine crystallographically-resolved crystal structures (López-Mejías *et al.*, 2012; Delaney *et al.*, 2014). For simplicity, the search reported here was performed in $Z' = 1$, giving three experimental targets (FPAMCA, FPAMCA11 and FPAMCA17). Following a combination of one-dimensional scans and searches for similar groups within the CSD using

[1] More specifically, this local minimum lies near the boundary between two adjacent LAMs, with each LAM indicating that the minimum lies within the *other* LAM's domain of applicability. This causes the optimization iterations to oscillate between the two LAMs without convergence. After a maximum number of iterations is reached, the *CrystalPredictor* algorithm forces convergence by fixing the LAM being used to one of the two adjacent ones (selected arbitrarily). However, the resulting local minimum is well outside that LAM's domain of applicability, which, in this case, leads to a severe error in the corresponding lattice energy.

Table 4
Ranges of flexible degrees of freedom in flufenamic acid and description of the uniform grid used to construct the initial LAM points.

	Uniform LAM grid		
Independent degree of freedom, θ (*cf.* Fig. 13)	Search domain (°)	Initial (coarse) grid spacing $\Delta\theta$ (°)	No. of grid points
T1	[−30, 30]	±10	3
T2	[60, 280]	±10	18
T3	[0, 360]	±10	11
T4	[120, 240]	±60	1

Conquest (Bruno *et al.*, 2002), the four torsions shown in Fig. 13 were designated as independent conformational degrees of freedom in the global search stage; their respective ranges are given in Table 4.

The data in Table 4 result in an initial coarse uniform grid of 594 LAMs. Following two passes of the adaptive LAM algorithm (Sugden *et al.*, 2016), the number of LAMs grows to 1 497. The corresponding QM calculations are performed at the M06/6-31(d,p) level of theory. The FIT potential is used to model repulsion/dispersion interactions (Williams & Cox, 1984; Coombes *et al.*, 1996; Beyer & Price, 2000; Cox *et al.*, 1981; Williams & Houpt, 1986). The fluorine repulsion/dispersion potential is taken from Williams & Houpt (1986) with cross interactions evaluated using geometric combining rules.

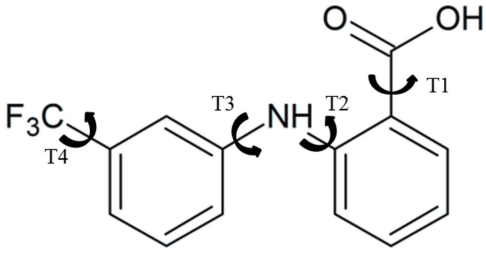


Figure 13
Molecular structure and flexible degrees of freedom for flufenamic acid.

Table 5
Energetic (kJ mol^{-1}) and geometric data (Å) for the Z' = 1 flufenamic acid targets generated by the *CrystalPredictor* and *CrystalOptimizer* algorithms.

	CrystalPredictor with waLAM model			*CrystalOptimizer*		
Polymorph	Rank	U^{latt}	$RMSD_{15}$	Rank	U^{latt}	$RMSD_{15}$
FPAMCA17	1	−123.83	0.215	1	−136.26	0.140
FPAMCA11	57	−116.95	0.482	9	−130.74	0.285
FPAMCA	539	−111.48	0.233	49	−127.41	0.141

Two million minimizations are performed within *CrystalPredictor* using the waLAM model. Of these, approximately 15.5% result in minima on the limits of the torsion angle domain which are discarded as they do not correspond to local minima of lattice energy. The application of the standard 20 kJ mol^{-1} cutoff results in 5 983 unique structures being taken through to the refinement stage; in fact, structures corresponding to all three experimental Z' = 1 forms are observed within +12 kJ mol^{-1} of the global minimum, as seen in Fig. 14.

The refinement of the 5 983 structures using *CrystalOptimizer* results in the polymorphic landscape presented in Fig. 15. As presented in Table 5, the three experimental forms are found with ranks 1, 9 and 49, with good geometric reproduction as showed by $RMSD_{15}$ values lower than 0.3 Å. However, at about +8.9 kJ mol^{-1} above the global minimum, the FPAMCA structure is higher in energy than is typically expected for an experimentally observed polymorph. For this structure, the short contacts between fluorine atoms on adjacent molecules uniquely dominate the interactions, which makes the structure most susceptible to uncertainty in the

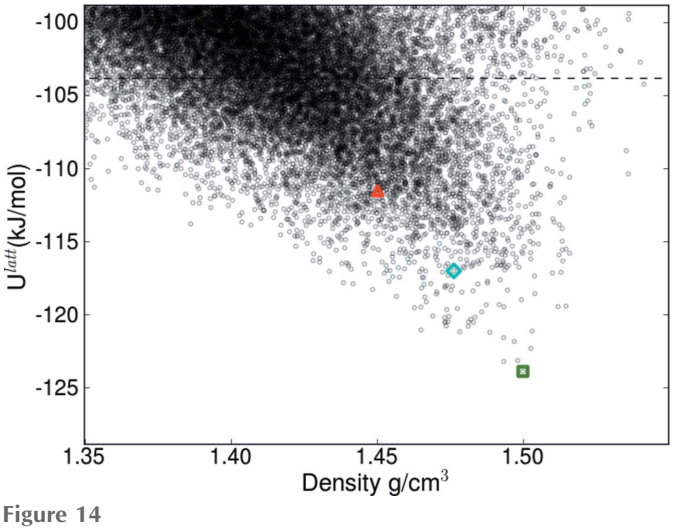


Figure 14
Polymorphic landscape of flufenamic acid after two million minimizations using *CrystalPredictor* with the waLAM model. The diamond denotes FPAMCA11, the square FPAMCA17 and the triangle FPAMCA. The dashed line indicates the 20 kJ mol^{-1} cutoff.

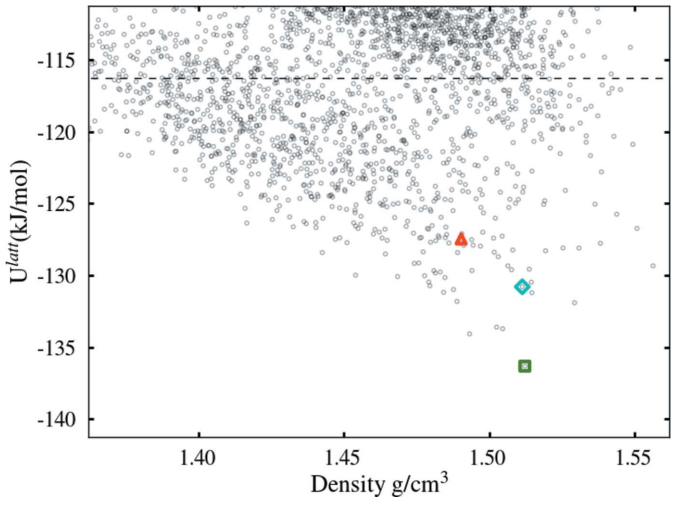


Figure 15
Polymorphic landscape of flufenamic acid following refinement with *CrystalOptimizer*. The diamond denotes FPAMCA11, the square FPAMCA17 and the triangle FPAMCA.

Table 6
Comparative computational performance statistics for global searches carried out with LAM and waLAM models.

System	Model	Successful minimizations resulting in local lattice energy minima (%)	No. of structures within 20 kJ mol^{-1} of global minimum	No. of minimizations per CPU hour
ROY	LAM	38	6539	465
	waLAM	82	2277	724
Glucose	LAM	57	1160	2143
	waLAM	76	408	2269
Molecule (XXVI)	LAM	7.8	123	88
	waLAM	61.2	86	115
Flufenamic acid	waLAM	83.9	5983	470

fluorine repulsion/dispersion potential. In this context, we note that the FIT parameters for fluorine were estimated by Williams & Houpt (1986) using data for rigid molecules containing carbon and fluorine only; they may, therefore, benefit from some further refinement using data for a wider range of fluorine-containing molecules.

4. Concluding remarks

The motivation for the development and implementation of the waLAM model was to improve the efficiency and reliability of the global search for a given system. Table 6 summarizes some relevant computational statistics for the systems considered in this paper.

The results demonstrate the superiority of the waLAM model on all metrics. The percentage of minimizations that (*a*) satisfy the mathematical convergence criteria and (*b*) result in local lattice energy minima is significantly increased. In fact, with the waLAM model, practically all minimizations satisfy the convergence criteria although some still yield points on the limits of the domain of the flexible degrees of freedom which do not correspond to local lattice energy minima.

Moreover, the waLAM model results in much fewer distinct structures which are within the 20 kJ mol^{-1} cutoff of the global energy minimum, and therefore need to be carried over to the refinement stage. As we have seen in the detailed discussions of these examples, this in no way affects the ability of the CSP algorithm to actually locate the experimentally known forms.

Finally, as indicated by the last column of Table 6, the use of the waLAM model allows more minimizations to be carried out per unit time during the global search. Whilst the evaluation of $\Delta U^{\rm intra}$, q and $\bar{\theta}$ via equations (4)–(7) is clearly more expensive than that associated with equations (1)–(3), this is more than compensated by a reduction in the time spent minimizing each structure due to the continuous and differentiable nature of the weight-averaged intramolecular energy function.

Using the proposed waLAM model, a CSP study was carried out, for the first time, on flufenamic acid. All three Z' = 1 experimental forms were found within the landscape. The two experimental forms for which the hydrogen bonding motif dominates were found within the lowest ten structures. The remaining experimental form was found as structure at rank 49, about 8.9 kJ mol^{-1} above the global minimum. We speculate that this may be due to the geometry of this form making it particularly susceptible to inaccuracies in the fluorine repulsion-dispersion potential parameters, including cross interactions. Based on recent experience (Gatsiou *et al.*, 2018), these parameters may benefit from being fitted to the geometry and sublimation energy of molecules similar to flufenamic acid. However, notwithstanding a possible future re-ranking of the low-energy structures, the investigation reported in this paper has generally been successful: all three experimental targets were identified as low-lying minima with good geometric reproduction, within a reasonable computational effort.

5. Data statement

Data underlying this article can be accessed on Zenodo at https://zenodo.org/record/1290769, and used under the Creative Commons Attribution licence.

Acknowledgements

We acknowledge the use of the Imperial HPC service for the quantum mechanical calculations: Imperial College Research Computing Service, DOI: 10.14469/hpc/2232. We would like to acknowledge the use of the *DMACRYS* software, from the group of Professor Sally Price at University College London.

Funding information

Funding for this research was provided by: Engineering and Physical Sciences Research Council (grant Nos. EP/J014958/1, EP/J003840/1, EP/P022561/1 and EP/P020194) and Eli Lilly and Company. For the bulk of the global search stage calculations, we are grateful for computational support from the UK Materials and Molecular Modelling Hub, which is partially funded by EPSRC (EP/P020194), for which access was obtained via the UKCP consortium and funded by EPSRC grant EP/P022561/1.

References

Aina, A. A., Misquitta, A. J. & Price, S. L. (2017). *J. Chem. Phys.* **147**, 161722.
Bardwell, D. A. *et al.* (2011). *Acta Cryst.* B**67**, 535–551.
Beyer, T. & Price, S. L. (2000). *J. Phys. Chem. B*, **104**, 2647–2655.
Braun, D. E., Gelbrich, T., Wurst, K. & Griesser, U. J. (2016). *Cryst. Growth Des.* **16**, 3480–3496.
Braun, D. E. & Griesser, U. J. (2018). *Front. Chem.* **6**, 31.
Braun, D. E., Kahlenberg, V. & Griesser, U. J. (2017). *Cryst. Growth Des.* **17**, 4347–4364.
Bruno, I. J., Cole, J. C., Edgington, P. R., Kessler, M., Macrae, C. F., McCabe, P., Pearson, J. & Taylor, R. (2002). *Acta Cryst.* B**58**, 389–397.
Case, D. H., Campbell, J. E., Bygrave, P. J. & Day, G. M. (2016). *J. Chem. Theory Comput.* **12**, 910–924.

Coombes, D. S., Price, S. L., Willock, D. J. & Leslie, M. (1996). *J. Phys. Chem.* **100**, 7352–7360.
Cooper, T. G., Hejczyk, K. E., Jones, W. & Day, G. M. (2008). *J. Chem. Theory Comput.* **4**, 1795–1805.
Cox, S. R., Hsu, L.-Y. & Williams, D. E. (1981). *Acta Cryst.* A**37**, 293–301.
Day, G. M. *et al.* (2009). *Acta Cryst.* B**65**, 107–125.
Delaney, S. P., Smith, T. M. & Korter, T. M. (2014). *J. Mol. Struct.* **1078**, 83–89.
Gatsiou, C. A., Adjiman, C. S. A. & Pantelides, C. C. (2018). *Faraday Discuss.* **211**, 297–323..
Habgood, M., Sugden, I. J., Kazantsev, A. V., Adjiman, C. S. & Pantelides, C. C. (2015). *J. Chem. Theory Comput.* **11**, 1957–1969.
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.
López-Mejías, V., Kampf, J. W. & Matzger, A. J. (2012). *J. Am. Chem. Soc.* **134**, 9872–9875.
Neumann, M. A., Leusen, F. J. J. & Kendrick, J. (2008). *Angew. Chem. Int. Ed.* **47**, 2427–2430.
Nyman, J., Yu, L. & Reutzel-Edens, S. M. (2019). *CrystEngComm*, **21**, 2080–2088.
Price, S. L., Braun, D. E. & Reutzel-Edens, S. M. (2016). *Chem. Commun.* **52**, 7065–7077.
Price, S. L., Leslie, M., Welch, G. W. A., Habgood, M., Price, L. S., Karamertzanis, P. G. & Day, G. M. (2010). *Phys. Chem. Chem. Phys.* **12**, 8478–8490.
Reilly, A. M. *et al.* (2018). *Acta Cryst.* B**72**, 439–459.
Rice, B., LeBlanc, L. M., Otero-de-la-Roza, A., Fuchter, M. J., Johnson, E. R., Nelson, J. & Jelfs, K. E. (2018). *Nanoscale*, **10**, 1865–1876.
Sobol, I. M. (1967). *USSR Comput. Math. Math. Phys.* **7**, 86–112.
Sugden, I., Adjiman, C. S. & Pantelides, C. C. (2016). *Acta Cryst.* B**72**, 864–874.
Vasileiadis, M., Kazantsev, A. V., Karamertzanis, P. G., Adjiman, C. S. & Pantelides, C. C. (2012). *Acta Cryst.* B**68**, 677–685.
Wheeler, K. A. & Hopkins, G. W. (2016). CSD Communication, deposition No. 1447529.
Williams, D. E. & Cox, S. R. (1984). *Acta Cryst.* B**40**, 404–417.
Williams, D. E. & Houpt, D. J. (1986). *Acta Cryst.* B**42**, 286–295.
Yu, L. A. (2010). *Acc. Chem. Res.* **43**, 1257–1266.