Global analysis of crystal energy landscapes: applying the threshold algorithm to molecular crystal structures

Shiyue Yang and Graeme M. Day*
School of Chemistry, University of Southampton, Southampton, SO17 1BJ, United Kingdom
email: G.M.Day@soton.ac.uk

Abstract:
We describe the implementation of the Monte Carlo threshold algorithm for molecular crystals as a method to provide an estimate of the energy barriers separating crystal structures. By sampling the local energy minima accessible from multiple starting structures, the simulations yield a global picture of the crystal energy landscapes. This provides valuable information on the depth of the energy minima associated with crystal structures and adds to the information available from crystal structure prediction methods that are used for anticipating polymorphism. We present results from applying the threshold algorithm to four polymorphic organic molecular crystals, examine the influence of applying space group symmetry constraints during the simulations, and discuss the relationship between the structure of the energy landscape and the intermolecular interactions present in the crystals.

1 Introduction

Computational approaches for exploring the energy landscapes of molecular crystals continue to develop rapidly as applications of crystal structure prediction (CSP) methods expand beyond the main application of anticipating pharmaceutical polymorphs[1, 2, 3] into screening of co-crystals[4] and solvates[5], and incorporation of CSP into computer-guided discovery of functional materials[6, 7, 8, 9, 10].

CSP typically relies on an exploration for the local energy minima on the high-dimensional energy surface as a function of the structural variables that determine the packing of a molecule into a crystal [11]. The structures corresponding to each local energy minimum are usually considered as possible polymorphs, with the assumption that the lowest energy predicted structures correspond to the most likely candidates to be observed experimentally. One limitation of the output of such methods is that, while they provide the geometry and energy of each potential structure, no information is gained about the depth of each energy minimum, nor possible transition paths and energy barriers between structures. This is currently a limiting factor in the analysis of the results of CSP.

One reason for needing a more global picture of the crystal energy landscape is to distinguish between structures occupying deep and shallow energy minima, *ie.* are the energy barriers surrounding the structure large or small? An important observation that has been made is that structures corresponding to known polymorphs are often connected to multiple, shallow energy minima by small energy barriers[12]; traditional CSP methods would suggest each of these minima as possible alternative polymorphs, while a knownledge of small energy barriers separating such structures would show that they can merge into fewer distinct structures due to thermal energy at the temperature of interest. Thus,

not distinguishing between deep and shallow energy minima contributes to the over-prediction of polymorphism[13].

Another area of particular interest is the identification of crystal structures that do not correspond to the thermodynamically most stable structure, but occupy sufficiently deep energy basins to be isolable and kinetically stable. Knowing about such structures is important for anticipating polymorphism that could occur through crystallization routes where kinetics can lead the crystal structure away from the thermodynamically preferred, global energy minimum. One example of such a process is the desolvation of solvate crystal structures [14, 15], where solvent incorporated into the crystal structure stabilizes an alternative arrangement of molecules, so that removal of solvent leaves the structure in a metastable polymorph. Some recent studies[6, 8, 16] have identified very high energy polymorphs through desolvation of solvates. The importance of these structures is demonstrated by their very attractive properties, such as for high capacity for gas storage, selectivity for molecular separations[6] and high photocatalytic activity[8].

Molecular dynamics approaches have been applied to study the transitions between polymorphs[17] and, in the context of CSP, to identify structures that interconvert at non-zero temperatures[18, 19, 20, 21]. In this study, we present the implementation of the threshold algorithm, which is based on Monte Carlo sampling of the energy landscape[22], to molecular crystals, using an accurate force field with an atomic multipole description of electrostatics. The aim of the threshold algorithm is to find the lowest energy at which transitions are possible between local energy minima. By combining trajectories from multiple starting structures, a global picture of the connectivity of minima can be constructed. The threshold algorithm has previously been applied to fairly simple inorganic crystal structures, to investigate the energy landscape of MgF_2[23] and to study

the entropic stabilization of high energy phases of CaF2[24]. Here, we investigate its application to more complex molecular organic crystals, which are characterized by a balance of several types of intermolecular interactions, and discuss the insight that this method provides to help understand their polymorphism.

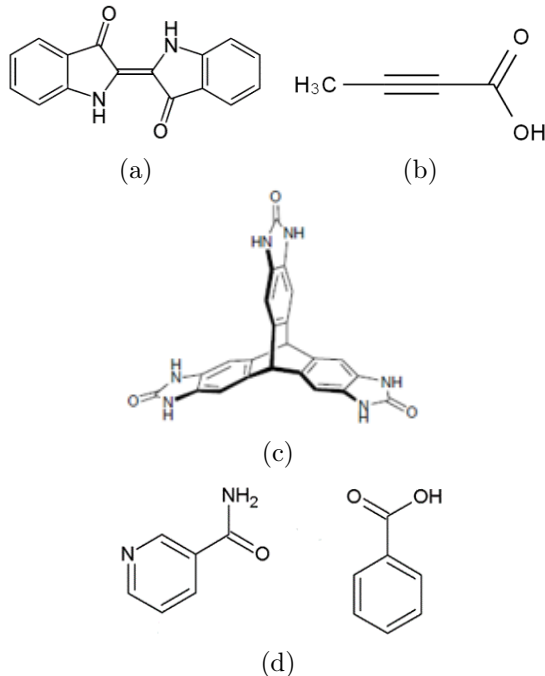# 2 Computational Details

2.1 Choice of Systems

Figure 1: Molecular crystal systems studied in this work. (a) indigo, (b) tetrolic acid, (c) TTBI and (d) the 1:1 co-crystal of nicotinamide and benzoic acid (GAZCES).

Four crystal systems (Fig.1), including single-component crystals and a co-crystal, were chosen for study, each of which has known polymorphism. All molecules have reasonably rigid molecular structures, so that rigid-molecule simulations should provide a realistic picture of the crystal energy landscape.

Indigo and tetrolic acid are both small, hydrogen bonding molecules, each with two known polymorphs. Both indigo polymorphs have the same space group symmetry and the same network of hydrogen bonds. In contrast, the polymorphs of tetrolic acid occupy different space groups and differ in hydrogen bonding. We also study the co-crystal formed between nicotinamide and benzoic acid, which is referred by the Cambridge Structural Database[25] reference code of its known crystal structure[26], GAZCES. The co-crystal system has both experimental structures in the same space group, $P2_1/c$, but with changes in the arrangement of hydrogen bonds. Triptycene trisbenzimidazolone (TTBI), has a more complex energy landscape and four experimen-

tally observed polymorphs distributed over a wide energy range.

These differences allow us to test how changes in the network of strong intermolecular interactions are reflected in the energy landscape. The systems with known structures in the same or in different space groups are used to assess sampling with and without symmetry constraints.

2.2 Threshold Algorithm and Disconnectivity Graph

The threshold algorithm was developed as a method for finding the energy barrier between structures without the requirement of energy gradient and Hessian matrix calculations[27, 22]. Initiated from a local minimum on the energy landscape, a Monte Carlo trial is generated by random local perturbations with the restriction that the single point (i.e. unminimized) energy of the perturbed structure is below a defined threshold energy, called the lid energy. All attempted moves that remain below the current lid energy are accepted and all moves that increase the energy above the lid energy are rejected. Thus, the trajectory can only explore a local pocket on the lattice energy surface and can never reach regions with energy higher than the lid. If the energy barrier between the current structure and another is higher than the lid, the transition between the two energy basins in which these structures are located cannot be sampled.

After a period of simulation, the lid energy is shifted to higher energy to increase the configurational space that is available to the trajectory, allowing transitions to new structures separated by energy barriers lower than the new energy lid. Therefore, when a trajectory visits the energy basin of a new local minimum, the energy barrier between the new and initial minima can be estimated as the current energy lid. An assumption here is that the step size of allowed perturbations is small enough that attempted moves cannot jump through energy barriers. The sampling under each energy lid needs, in principal, to be ergodic, although this is hindered by the required small step size.

From a sequence of pockets sampled with increasing lid energies, a tree structure, often called a disconnectivity graph, can be constructed to represent the energy landscape[28, 29, 30]. The disconnectivity graph condenses the continuous, high dimensional potential energy surface into the set of discrete local minima and information on the energy barriers that separate them. To construct the disconnectivity graph, the energy landscape is analyzed at a set of energies along the vertical axis. The nodes at a given energy, E_i, represent superbasins on the energy surface: the set of local minima that are connected by pathways below E_i. Moving up in energy, nodes are connected as higher energy pathways connect groups of superbasins, while moving down on the graph leads to disconnections until the end of each branch, corresponding to single local energy minima. The horizontal axis has no direct physical meaning and is introduced for visualization, so that structures connected by

lower energy barriers are grouped together. Vertices along the branches between nodes and minima are also for visualization only. A schematic of a one-dimensional potential energy surface and its associated disconnectivity graph is shown in Fig.2.
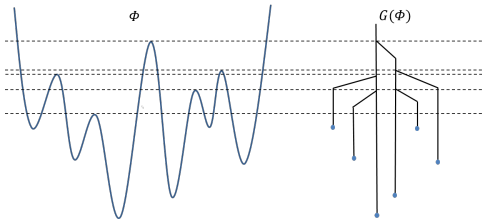

Figure 2: An example of tree-like representation $G(\Phi)$ of the energy landscape Φ.

In most simulations presented in this work, the lid energy was increased in increments of 5 kJ mol^{-1}, which is chosen as a balance between precision in the calculated energy barriers and the need to explore a wide energy range. For each threshold simulation, the lid energy was increased from the minimized energy of the initial structure. Sampling was started from multiple initial structures, usually corresponding to the energy minimized versions of observed polymorphs, which leads to different energy grids from different start points. When plotting the disconnectivity, we used a new energy grid starting from the lowest energy among all the initial structures, with the increment of 5 kJ mol^{-1}, to merge trajectories into one graph. Any lid energy from a single threshold algorithm was rounded up to the closest lid energy in the new grid.

The types of move allowed in the threshold simulations were molecular translations, rotations, perturbations of unit cell lengths and angles, as well as unit cell volume changes. Cutoffs on each move type (see ESI, Table S1) were chosen to give similar energy changes for different move types. Probabilities for each move type were set according to the proportion of the total number of degrees of freedom for each move type, in the same way as in our implementation of basin hopping for CSP.[31] The sampling at each lid energy and total lengths of simulations differs between systems, depending on the number of degrees of freedom and energy window that needed to be covered. In the current work, a fixed number of steps was performed at each lid energy. An adaptive schedule was investigated (see ESI) but due to the difficulty of choosing the convergence criteria, it did not improve the completeness of sampling.

We have studied rigid molecules in this work, so molecular geometries were constrained to their optimized geometries from density functional theory (DFT) calculations. Lattice energy calculations were performed with the DMACRYS software[32], using an empirically parametrized exp-6 intermolecular repulsion-dispersion potential with electrostatics described by atomic multipoles calculated from the DFT electron distribution (see ESI).

To put more emphasis on the connections between low energy structures, the disconnectivity graph is not presented for the whole energy range. The highest energy barrier between initial structures is taken as the upper limit and any local minima connected at lid energies higher than this upper limit are not presented on the disconnectivity graph. The disconnectivity graphs over the entire sampled energy range are presented in the ESI.

For comparison to the energy landscape generated from the threshold algorithm, CSP was performed for each molecule using quasi-random sampling (see ESI for details).

2.3 Identification of connections between trajectories

The threshold algorithm does not involve local optimization of structures in principle. However, we require a method to identify whether a perturbation has led to a new energy basin. For this, perturbed structures are locally energy minimized if the perturbation is accepted (ie. the unminimized energy was under the current lid energy). The minimized structures are compared to each other to identify where trajectories meet. Energy minimization is performed for every accepted perturbed structure because lattice energy landscapes are known to often contain many local minima around observed structures,[12] so every perturbation was assumed to potentially lead to a new local energy basin.

Two minima are considered to be connected at a given lid energy if the trajectories that start at or visit these minima are found to sample a common local minimum. Thus, the identification of identical structures (corresponding to the same local minimum) is an essential process to obtain the disconnectivity graph. We use a two-step strategy: 1) a fast initial screen for duplicates is performed by comparison of simulated X-ray diffraction patterns, followed by 2) checking of duplicates using the COMPACK algorithm,[33] which compares interatomic distances and angles within a cluster of 30 molecules taken from the compared crystal structures (see full details in the ESI).

2.4 Structural descriptors and data featuring

Two descriptors of structural similarity – atom-centered symmetry functions (ACSFs)[34] and the smooth overlap of atomic positions (SOAP)[35] – were used to investigate geometric clustering of crystal structures and their clustering into superbasins based on threshold simulations. Both approaches provide a measure of structural similarity based on comparison of local atomic environments.

ACSFs capture structural information from a series of radial and angular functions, which depend on neighbouring atom positions out to a cutoff radius R_c. We use ACSFs grouped by element, ie. the functions are evaluated separately for all pairs (for radial functions) and triples (angular functions) of elements. The spacing and width of ACSFs is chosen as in the ANI-1 neural network force field.[36]

In the SOAP kernel, the local region of each atom is described individually is represented by a sum of Gaussians

centered on all atoms within the local environment. The approach applied here to calculate the similarity between two structures based on similarity of their atomic environments is the regularized-entropy match (REMatch) kernel.

Full details of ACSF and SOAP are provided in the ESI.

Principal component analysis (PCA)[37] and the clustering method HDBSCAN*[38] are used to analyse the distribution of crystal structures in descriptor space, for comparison with the clustering into energy basins determined by the threshold simulations.

3   Results and Discussion

3.1   Sampling within a space group

We start with two examples where polymorphs exist with the same space group symmetry, so that a transformation between their corresponding local energy minima should be possible with space group symmetry constrained, *i.e.* Monte Carlo moves are only allowed which maintain the original symmetry. This is performed by perturbing the asymmetric unit of the crystal structure and applying symmetry-related perturbations to all other molecules in the unit cell. Constraints are also applied to unit cell parameters, where these are required to maintain space group symmetry.

The connections between structures found in this way exclude pathways that break symmetry, which might be lower in energy. However, the symmetry constraints simplify the simulation and we examine the picture of the landscape that we obtain with these constraints.

3.1.1   Indigo

Indigo (Fig. 1a) is known to form two polymorphs,[39, 40] named A and B, both containing layers of hydrogen bonded molecules. The structure of these layers is almost unchanged between polymorphs A and B, but their structures differ in the arrangement of these layers (see overlay, Fig. 3). Thus, the lowest energy pathway between polymorphs should not disrupt the hydrogen bonding and is expected to involve a relatively low energy barrier.
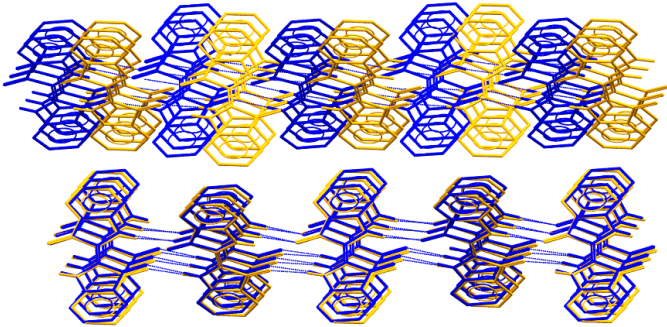

Figure 3: Overlay of the packing of polymorphs A (blue) and B (orange) of indigo. Intermolecular hydrogen bonds are shown as blue dashed lines. Non-hydrogen bonding hydrogen atoms are hidden.

Both polymorphs have space group symmetry $P2_1/c$ with half a molecule in the asymmetric unit of the unit cell (both molecules lying on centres of inversion) and thus two molecules in the unit cell. To allow free molecular translations, unconstrained by the position of crystallographic inversion centres, the most symmetric representation used for simulation of these structures was $P2_1$ with whole molecules in the asymmetric unit. Crystal structure prediction in $P2_1$ finds polymorphs A and B as the two lowest energy crystal structures (see Fig. S4), with B having a calculated lattice energy 4.0 kJ mol^{-1} below A. For comparison, the two polymorphs are reported to be nearly equi-energetic when evaluated using periodic DFT with a plane wave basis set and many-body dispersion correction.[41]

Monte Carlo simulations were started from the CSP structures matching both polymorphs, which were continued for 10 lid energies, incremented by 5.0 kJ mol^{-1} with 1,000 attempted steps under each lid (covering a total 50.0 kJ/mol energy window). From threshold simulations in space group $P2_1$, the first connection between polymorphs A and B is found when the lid is 26.0 kJ mol^{-1} above A (30.0 kJ mol^{-1} above B) (Fig. 4a). Below this energy, no other structures are connected to B, one slightly higher energy structure is connected to A and two additional structures connect to A and B at the same lid energy. All three of the additional crystal structures within the basin connected to A and B maintain the same hydrogen bond motif, but with greater differences in molecular orientation around the hydrogen bonds than between polymorphs A and B. No further structures are connected to the basin containing these five local energy minima (A, B and the three higher energy structures) until the lid energy is raised a further 15 kJ mol^{-1}. Globally, the results give a picture of an energy landscape that is funneled towards a small set of structures, all featuring the same favoured hydrogen bonding motif, with the two known polymorphs as the lowest energy local minima within this energy superbasin.

To investigate the effect of performing the threshold Monte Carlo simulation with different symmetry constraints, simulations were also performed with larger unit cells, containing four molecules in space group $P2_1/c$. Both known polymorphs can also be described with this symmetry. The resulting disconnectivity graph is shown in Fig. 4b. The overall picture of a single funnel towards polymorphs A and B is maintained, but a notably lower energy transition is found between the known polymorphs, when the lid is 11.0 kJ mol^{-1} above A, 15.0 kJ mol^{-1} above B. As a result, the connection between A and B occurs at a lower energy than their connection to any other local minimum on the landscape. While both indigo simulations yield useful information on the structure of the crystal energy landscape, the results confirm our expectation that the symmetry constraints applied during sampling can have an important influence on the details of the connectivity graph.

The 5 kJ mol^{-1} increments in threshold energy limit the precision with which the energy barrier between the two polymorphs can be estimated. Knowing that the transi-
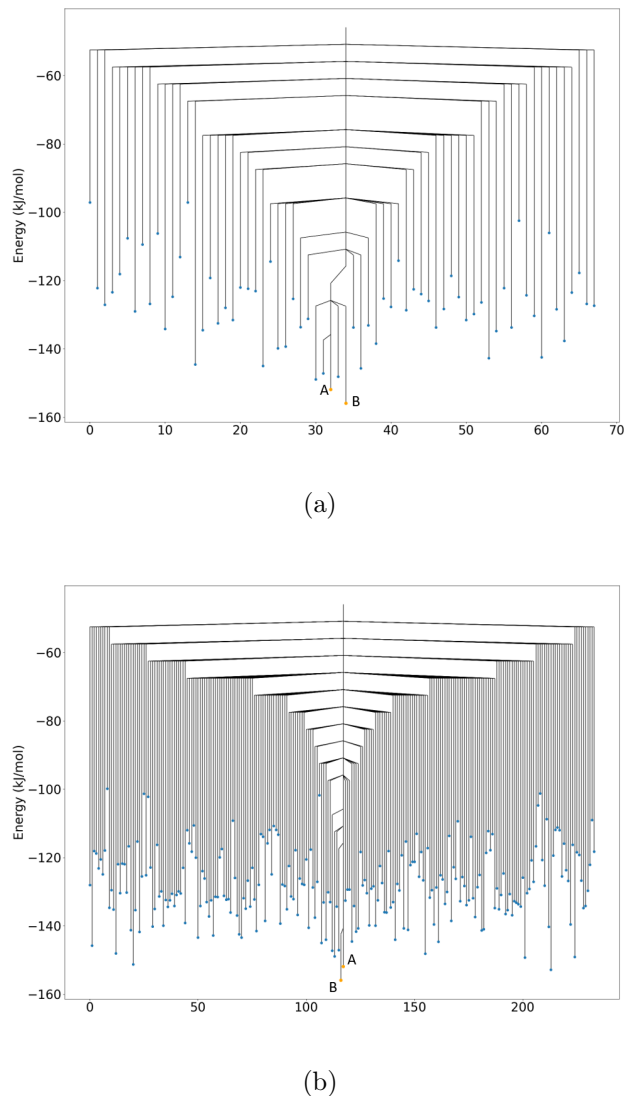
(a)

(b)

Figure 4: Disconnectivity graph from threshold simulations for indigo space group (a) $P2_1$ and (b) $P2_1/c$. Blue points are minimised structures. Orange points represent the initial structures.

that can be used to explore energy landscapes: an initial simulation with large energy threshold increments to capture the global structure of the energy landscape, followed by targeted re-sampling using smaller steps to refine the results in important regions of the energy landscape.

3.1.2 1:1 nicotinamide:benzoic acid co-crystal

As a second example, we studied the 1:1 nicotinamide:benzoic acid (GAZCES) co-crystal as a system with more degrees of freedom (due to two molecules in the asymmetric unit), but where the known polymorphs again have the same space group symmetry. Nicotinamide:benzoic acid is a highly polymorphic co-crystal; four polymorphs have been observed under mechanochemical co-crystallization conditions, but only polymorphs **I** and **II** have had their structures determined.[26] Both characterized polymorphs are in space group $P2_1/c$.

Unlike indigo, the polymorphs of this co-crystal differ in their hydrogen bonding (Fig. 5): polymorph **I** has nicotinamide doubly hydrogen bonded dimers connected by hydrogen bonds to benzoic acid molecules, while polymorph **II** contains an extended hydrogen-bonded nicotinamide chain with benzoic acid hydrogen bonding to the nicotinamide pyridyl nitrogen atoms at the edges of these chains.
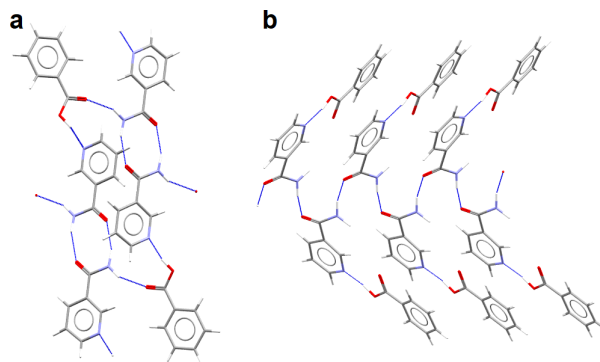

Figure 5: Hydrogen bonding in polymorphs a) **I** and b) **II** of the nicotinamide:benzoic acid co-crystal. Carbon atoms are grey, oxygen red, nitrogen blue and hydrogen white. Hydrogen bonds are shown as dashed blue lines.

tion occurs when the threshold is 15 kJ mol^{-1} or less above polymorph B, we re-sampled the landscape in space group $P2_1/c$ (four molecules in the unit cell), with smaller threshold increments of 1 kJ mol^{-1} and 1,000 steps under each lid energy (a five-fold increase in sampling per unit increase in the lid energy). Again, simulations were started from both polymorphs, for 15 increases in the lid energy. With this increased sampling, the energy barrier is now located when the threshold is 10 kJ mol^{-1} above polymorph B, 6 kJ mol^{-1} above A, and no other local energy minima are visited up to the highest energy threshold.

The negligible change with smaller lid energy steps and increased sampling gives us confidence that we have sampled the landscape sufficiently and the results illustrate a strategy

Threshold simulations were started from the structures from the CSP landscape (Fig. S6) corresponding to **I** and **II**. **I** and **II** have very similar calculated lattice energies (**I** has a calculated lattice energy 0.6 kJ mol^{-1} below **II**) and are located in the low energy region of the landscape. Due to the greater complexity of the landscape, 3,000 steps were performed at each value of the lid energy, which was increased in 5.0 kJ mol^{-1} steps up to 150.0 kJ mol^{-1} above the initial structures, i.e. 30 increases in the threshold energy and a total of 90,000 attempted perturbations from each starting structure.

The threshold simulations reveal a dual-funneled energy landscape with deep basins centred on polymorphs **I** and **II** (Fig. 6). Polymorph **I** is the lowest energy structure in its
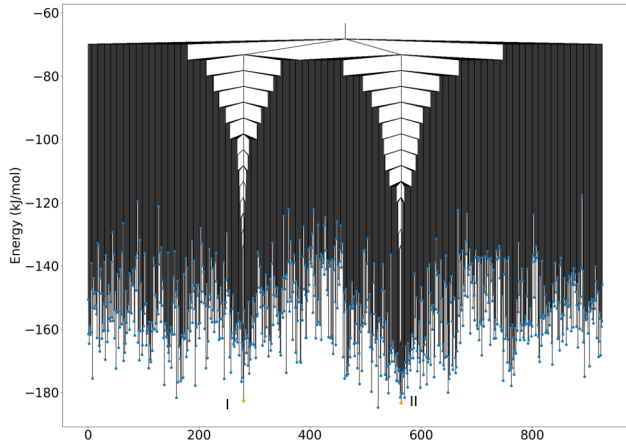
Figure 6: Disconnectivity graph from threshold simulations for the 1:1 nicotinamide:benzoic acid co-crystal in space group $P2_1/c$.

funnel (left of Fig. 6), while one lower energy structure is located in the funnel containing **II** (right of Fig. 6). Several lower energy structures are located by CSP in space group $P2_1/c$, but not sampled during the threshold simulations; these structures might lie outside of the two funnels, where little sampling has been performed. A complete picture of the energy landscape would require threshold sampling from some of the unobserved polymorphs as well as **I** and **II**.

The lowest energy connection between the funnels is approximately 120 kJ mol^{-1} above **I** and **II**. Thus, the threshold simulation is able to locate a pathway connecting these two very different crystal structures and the rearrangement in hydrogen bonding required to transform between them results in a high energy barrier. A lower energy connection between **I** and **II** might be found if space group symmetry constraints were removed from the Monte Carlo sampling, but it is unlikely that the global structure of the landscape would be changed. The funneled landscape gives an impression that polymorph selection will be strongly influenced by crystallization conditions, which could lead crystal growth into one funnel or the other, and that, once grown, interconversion of the polymorphs is unlikely without a large energetic input. Indeed, both **I** and **II** are reported to be stable for months once isolated.[26]

3.2 Sampling between space groups in a $P1$ cell

Simulations of the indigo and co-crystal examples were simplified by their polymorphs having the same space group symmetry. However, in practice, transformations between structures have no symmetry constraints. Therefore, we expect to obtain a better estimation of the actual energy barrier between structures by sampling with as many con-

straints as can practically be removed from the simulation. We also want to be able to analyse energy landscapes involving crystal structures with different symmetries. Here, we look at two examples involving known crystal structures with different space group symmetry. To remove the symmetry constraints, we use $P1$ unit cells. In the following examples, we construct $P1$ unit cells containing sufficient molecules to be able to represent all the known polymorph crystal structures, i.e., the lowest common multiple of Z (the number of formula units in the unit cell) for all known structures. Thus, the simulations presented below are not fully unconstrained - translational symmetry is imposed by the unit cell - but the models have sufficient flexibility to describe the polymorphs of interest. The approach could be extended to be able to visit more possible packing symmetries by expanding the unit cell to contain more molecules.
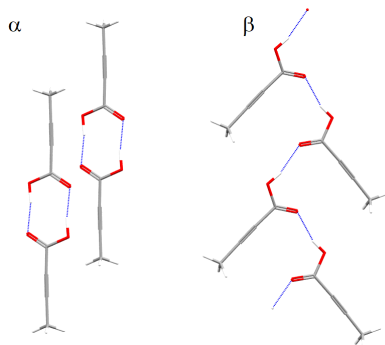
3.2.1 tetrolic acid

Figure 7: Hydrogen bonding in the two known polymorphs of tetrolic acid. Carbon atoms are grey, oxygen red and hydrogen white. Hydrogen bonds are shown as dashed blue lines.

Tetrolic acid has two known polymorphic forms: a triclinic structure in space group $P\bar{1}$ and a monoclinic structure in space group $P2_1$,[42] referred to as α and β, respectively. The two structures have different hydrogen bond motifs: the carboxylic acid groups form cyclic, doubly-hydrogen bonded dimers in α and infinite hydrogen bond chains in β (Fig. 7). Both crystal structures have two molecules in the unit cell, so simulations were performed using $P1$ unit cells containing two molecules. The number of degrees of freedom to sample this system was the same as the nicotinamide:benzoic acid co-crystal, so we applied the same number of Monte Carlo steps (3,000) at each lid energy, which was raised in 5 kJ mol^{-1} increments, starting simulations from the structures of α and β.

Threshold simulations reveal a similar energy landscape structure as found for the GAZCES co-crystal, with separate funnels containing structures α and β (Fig. 8). Because the simulations were run without space group constraints, a pathway could be found between the polymorphs. This connection was located when the lid energy was 40 kJ mol^{-1}
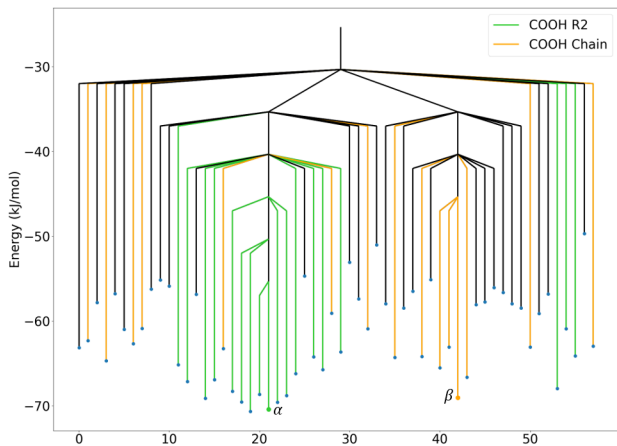
Figure 8: Disconnectivity graph from threshold simulations for tetrolic acid with two molecules in a $P1$ unit cell. Colors of edges represent the hydrogen bond motif of connected minimized structures where black edges are structures identified as having neither of the motifs. Hydrogen bond motifs were identified with Mercury[43], identifying hydrogen bonds as H...O contacts with a separation less than the sum of van der Waals radii - +0.1Å and an O-H...O angle greater than 125°.

above α, which has the slightly (1.4 kJ mol^{-1}) lower calculated lattice energy of the two known polymorphs. The relatively high energy barrier is unsurprising, considering the breaking of hydrogen bonds required to transform between α and β, along with substantial reorientation of the molecules.

To confirm that the structure of the landscape is determined largely by the hydrogen bond interactions, the local minima within each funnel were classified by their hydrogen bond motifs (Fig. 8). Indeed, all structures connected to α or β by barriers lower than 25 kJ mol^{-1} maintain the same hydrogen bonding as the starting structure, i.e. dimers within the α funnel and chains within the β funnel. At thresholds 30 and 35 kJ mol^{-1} above α, we start to see changes in hydrogen bond motifs: two structures with hydrogen bond chains within the α funnel and several crystal structures within both funnels that are not classified as either chains or dimers. Finally, at 40 kJ mol^{-1} above α, the two funnels are connected.

We now compare the results of the threshold sampling to the output from CSP on tetrolic acid, in this case restricting CSP to the space groups of the two known polymorphs ($P\bar{1}$ and $P2_1$). The resulting structures are shown in Figure 9 in the representation often used in CSP studies: plotting the energy vs density of all predicted structures. Structures sampled during threshold searches are also shown and labelled by whether they belong to the funnels around α, β or are disconnected from those superbasins at a lid energy 40
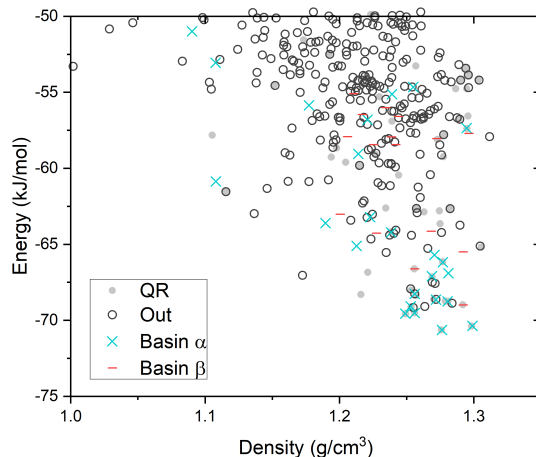

Figure 9: Lattice energy vs density plot of the landscape of predicted crystal structures of tetrolic acid. Local minima located after optimisation of structures from the threshold sampling are classified as belonging to either main basin (α and β) or not within one of these basins (black circles, labelled 'Out'). Crystal structure prediction results from quasi-random sampling (QR) are shown for comparison, with searches performed in space groups $P\bar{1}$ and $P2_1$.

kJ mol^{-1} above α. A first observation is that crystal structures in the α and β funnels occupy overlapping regions of energy-density space, so that the traditional CSP representation does not convey the important information about which structures belong to connected regions of the high-dimensional energy landscape. The disconnectivity graph conveys this information more clearly.

We also find that some structures are found in the threshold search, but not CSP, and *vice versa*. The threshold search is able to locate structures that are not accessible in the CSP because of lower symmetry constraints: some structures do not belong to either space group included in CSP. On the other hand, structures found by CSP, but not the threshold search indicate that the threshold sampling has not fully explored the energy landscape. However, we note that overlap between QR and threshold structures is very good in the important low energy region of the landscape.

3.2.2 TTBI

The final molecule investigated is TTBI (Fig. 1c), which has four known polymorphs[44, 6] that occupy high energy regions of the crystal structure landscape. The positions of the observed structures are indicated on the crystal energy landscape (Fig. 10), showing that the observed structures fall outside the usual energetic range of polymorphism.[45] The proposed explanation for the observation that these very high energy structures can be isolated as stable materials is that they occupy deep, isolated regions of the energy landscape, which is hinted at by the 'spikes' of structures
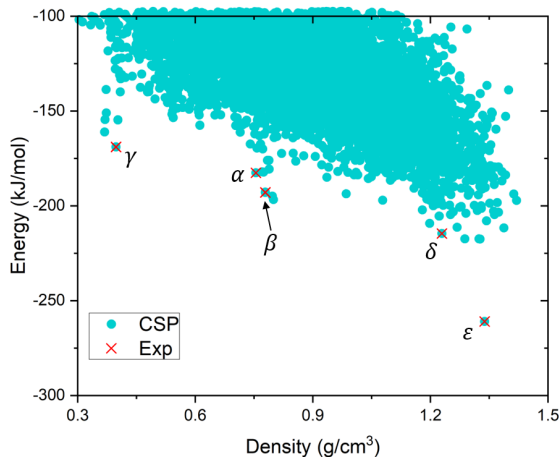
Figure 10: Lattice energy vs density plot of the landscape of predicted crystal structures of TTBI from CSP (quasi-random sampling), with predictions performed in $P1$ with four molecules in the unit cell. The experimentally observed structures are labelled α, β, γ and δ. The global energy minimum is labelled ε

Table 1: Lid energies at which transitions are located between γ, β, δ and ε in threshold sampling in space groups $P\bar{1}$ (upper right entries) and $P2_1/c$ (lower left entries). Energies in italics along the diagonal are the calculated lattice energies of the four structures. All energies are in kJ mol^{-1}.

	ε	δ	β	γ
ε	*-259.2*	-104.2	-104.2	-89.2
δ	-109.2	*-214.5*	-104.2	-89.2
β	-109.2	-114.2	*-192.1*	-89.2
γ	-59.2	-59.2	-59.2	*-169.7*

that fall below the bulk of structures on the energy-density landscape;[6] α, β and γ are low energy structures within these spikes. We apply the threshold algorithm to add information to the CSP landscape and improve our understanding of the observed polymorphism of TTBI.

Of the five important TTBI structures (the four observed polymorphs and the global energy minimum, which we label ε), four (all but α) have space group symmetries that could be sampled in either space group $P\bar{1}$ or $P2_1/c$ with one independent molecule ($Z' = 1$). The results of threshold simulations performed in $P\bar{1}$ and $P2_1/c$, starting trajectories from β, γ, δ and ε, are shown in Figure 11a and b. All searches used 5 kJ mol^{-1} increments in the lid energy and 1,000 steps were attempted at each lid energy.

The threshold simulations in $P\bar{1}$ and $P2_1/c$ yield disconnectivity graphs with similar structures and high energy barriers between the initial structures (lid energies at which transitions are found are summarised in Table 1). The connectivity structure shows a deep energy basin centered on ε with very few higher energy connected local minima until the energy lid is increased to very high energies. In the $P\bar{1}$ simulation, β, δ and ε are all connected at the same lid energy, $+155$ kJ mol^{-1} above ε ($+110$ kJ mol^{-1} above δ and 67 kJ mol^{-1} above β). The results are very similar in $P2_1/c$. As with the indigo polymorphs, there is a slight lowering of the lid energies at which several of the transitions occur in this larger unit cell: the connection between β and δ is lowered by 10 kJ mol^{-1} compared to the $P\bar{1}$ results, while their connection to ε is lowered by only 5 kJ mol^{-1}.

The basin containing the low density γ polymorph is separated by an even larger energy barrier from β, δ and ε. This energy barrier connecting the γ trajectory to the other

three structures shows the largest difference between results in the two space groups. The lid energy at which γ connects to the others is -89.2 and -59.2 kJ mol^{-1} in $P\bar{1}$ and $P2_1/c$, respectively, representing a barrier of 80 or 110 kJ mol^{-1} for the transformation of γ into the global energy minimum ε or one of the other known polymorphs. Despite the large quantitative difference found when sampling with different symmetry constraints, the same *qualitative* picture emerges: globally, γ is separated from the main pocket formed by ε, β and δ, occupying its own funnel on the connectivity graph. This finding agrees with the surprisingly good reported thermal stability of the γ polymorph, despite its exceptionally low density.[6]

The α polymorph of TTBI crystallizes in space group $P4_2/m$ with four molecules in the unit cell[44] and cannot be represented in the unit cells included in the $P\bar{1}$ and $P2_1/c$ simulations. To include α, we performed a further set of simulations in a $P1$ unit cell containing four symmetry-independent molecules, starting trajectories from all five structures (α, β, γ, δ, ε) and attempting 5,000 steps per lid energy. Due to the wide range of initial energies, the number of lids varied between trajectories, with the longest simulation started from ε involving 250,000 total Monte Carlo steps.

The resulting disconnectivity graph from the $P1$ threshold simulation is shown is figure 11c, where the branches corresponding to each local energy minimum are colored by the density of the corresponding crystal structure. Apart from now including α, a difference with respect to the $P\bar{1}$ and $P2_1/c$ results is that the lower symmetry allows more distinct local minima to be identified within the superbasins corresponding to the known polymorphs; this is particularly noticeable around δ, where 17 structures are connected to δ at lower energy barriers than the connection of the δ superbasin to those of ε, α and β. To more clearly visualise the relationships between the five polymorphs, a simplified disconnectivity graph is shown in figure 11d, in which all structures other than the five starting structures are hidden. The lid energies at which transitions are found between the five structures are summarised in Table 2.

As with the results from the higher symmetry simulations, we find the barrier separating the γ polymorph from other observed forms to be the highest; the connection between γ and the other polymorphs is found at the same lid energy
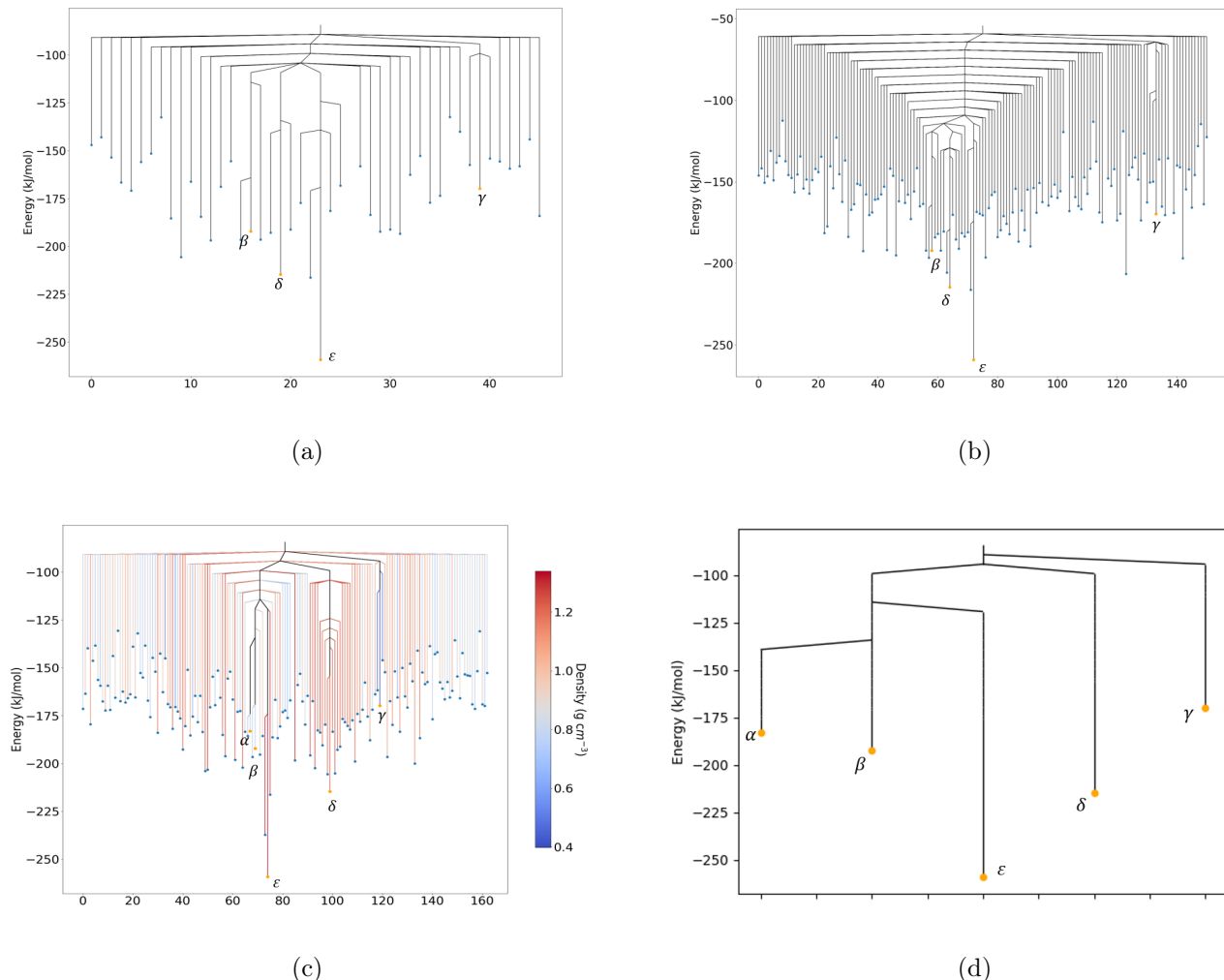
Figure 11: Disconnectivity graphs from threshold simulations for TTBI sampled in space groups: (a) $P\bar{1}$ (two molecules per unit cell), (b) $P2_1/c$ (four molecules per unit cell) and (c) $P1$ with four independent molecules per unit cell. (d) A simplified graph from the simulation in $P1$ with all local minima apart from the five landmark structures (α, β, δ, γ and ε) hidden. The branches in (c) are coloured to show the crystal density of each structure.

in the $P1$ simulation as the simulation with $P\bar{1}$ symmetry constraints in the smaller unit cell. This fulfills our expectation that sampling in $P1$ should find a transition path with an energy barrier at or lower than that found in space group-constrained trajectories.

The pair of polymorphs related by the lowest energy barrier is α and β, where the lid energy at which their trajectories meet is just under 50 kJ mol^{-1} above the calculated energy of α. The result agrees with molecular dynamics studies[6] in which the structures of all other observed TTBI polymorphs showed only fluctuations about their known crystal structures at 300K, apart from α, which partially transformed to β in a short, 500 ps, simulation. These structures also occupy the same 'spike' on the energy-density representation of the crystal structure landscape. We note that the disconnectivity graph tends to group structures of similar density; the lowest energy barriers tend to

be found between structures that are close in density.

3.3 Energy Landscape Featuring

The disconnectivity graph that is generated from the threshold simulations could be viewed as a form of clustering of local energy minima, grouping those that are related by the lowest energy barriers. The results for tetrolic acid demonstrate that there is a link between the basin structure on the crystal energy landscape and geometric features of the crystal structures - in this case, hydrogen bonding motif. The results for TTBI also suggest that crystal density differences explain some of the grouping of local energy minima into the superbasins on the energy landscape. Therefore, we asked the question whether structural descriptors commonly used in machine learning applications could provide a more general geometrical descriptor of structural similar-

Table 2: Lid energies at which transitions are located between γ, α, β, δ and ε in threshold sampling in space group $P1$ with four independent molecules in the unit cell (upper right entries). The entries in the lower left are the lowest energy lid at which transitions are found in the simulations in $P2_1/c$ and $P\bar{1}$. Energies in italics along the diagonal are the calculated lattice energies of the four structures. All energies are in kJ mol^{-1}.

	ε	δ	β	α	γ
ε	*-259.2*	-94.2	-114.2	-114.2	-89.2
δ	-109.2	*-214.5*	-94.2	-94.2	-89.2
β	-109.2	-114.2	*-192.1*	-134.2	-89.2
α	-	-	-	*-183.1*	-89.2
γ	-89.2	-89.2	-89.2	-	*-169.7*

ity that correlates with the clustering of local minima based on heights of energy barriers. If so, this could improve our physical understanding of the information gained by applying geometrical descriptors to CSP landscapes, which has started to gain attention for identifying families of related structures, with the goal of discovering structure-property relationships.[46, 47, 48] Furthermore, a successful geometrical clustering could replace the computationally expensive local energy minimization procedure to identify the basin to which each point on the Monte Carlo trajectory belongs.

We have taken the GAZCES co-crystal energy landscape in space group $P2_1/c$ as an example. This system was chosen for investigation due to the clear structure of the energy landscape and because of the sufficiently large number of structures in each basin from threshold simulations. For analysis, the total energy landscape was divided into three regions: two funnels, each containing an initial structure (polymorph **I** or **II**) and the structures outside the two basins. Two common descriptors for crystal systems - SOAP and atom-centered symmetry functions (ACSFs) - were applied with common data featuring methods.

We first investigated dimensionality reduction using PCA of the ACSFs and the SOAP kernel. In both cases, the descriptors were flattened over all atoms and, for symmetry functions, radial and angular functions were merged into one vector. With either descriptor, the eigenvalues corresponding to the first two principle components had twice the magnitude of the third. We plot the structures onto these first two principal components in Figure 12a (see ESI for the corresponding plot for ACSFs, Fig. S14). In neither case is there clear differentiation of structures corresponding to basins **I** and **II**, as identified by the threshold algorithm; the structures from both basins, and those outside of the two basins, overlap in this projection onto the first two principal components.

In case the overlap of basin **I** and basin **II** structures seen in the PCA visualization is due to the dimensional reduction, we also tested density-based clustering in the original, high dimensional descriptor space. The HDBSCAN* algorithm[38] was applied to the same dataset (the GAZCES
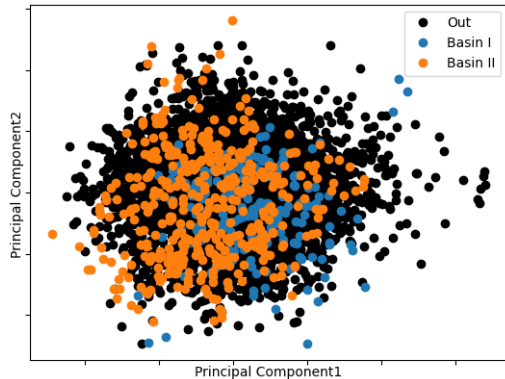
(a)
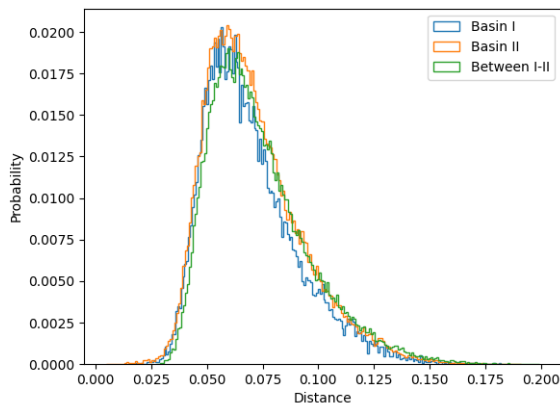
(b)

Figure 12: (a) The first two principle components of the SOAP REMatch kernel for structures for co-crystal GAZCES in space group $P2_1/c$, coloured according to the basin identified from threshold simulations. Structures labelled in black fall outside of the two main basins (see Fig. 6) (b) Distribution of pairwise dissimilarities between and within energy basins **I** and **II**.

co-crystal set of local energy minima in $P2_1/c$). However, trying different minimum cluster sizes and, for ACSFs, different measures of the pairwise distance between structures (see ESI), no clustering could be obtained that aligns with the grouping of structures from the threshold simulation results. Full details are given in the ESI. To understand these results, we examined the distributions of distances between structures within each basin and between basins. The distribution of distances between structures was found to be similar within and between energy basins (Figs. 12b, S16, S17, S18), indicating that geometrical similarity based on these atom-centered geometrical descriptors is not necessarily a good indicator of structures that are 'closer', in terms of energetic accessibility, on the energy landscape.

4 Conclusions

We have implemented the threshold algorithm to study the energy landscapes of molecular crystal structures with a force field energy model using atomic multipole electrostatics. The method has been applied to four molecular organic crystal systems to examine the energy barriers between known polymorphs and the global structure of their crystal energy landscapes, which we visualize using disconnectivity graphs.

The structures of the energy landscapes vary between the systems studied here, from a single funnel for the crystal structures of indigo, where hydrogen bonding is conserved between polymorphs, to multiple energy funnels when polymorphs differ in the arrangement of their strong intermolecular interactions. Thus, the threshold simulation results reinforce our chemical intuition and provide a quantification of the differences in energy barriers between different polymorphic systems.

Although the structures of the energy landscapes for the molecules studied here can be rationalized in terms of intermolecular interactions, we find that the grouping of crystal structures in energy basins is not reproduced by clustering or dimensionality reduction based on commonly-used structural descriptors. Thus, the results are complementary to unsupervised machine learning approaches that have been applied to the analysis of crystal structure landscapes.

The influence of space group constraints on energy barriers has been examined; for the systems studied here, the qualitative global picture of the energy landscape is not strongly influenced by imposing symmetry constraints, but the magnitude of energy barriers is affected. Therefore, space group-constrained simulations can be used to gain initial insight into crystal energy landscapes, but more computationally demanding, unconstrained simulations are the best route for more quantitative results.

Our implementation of the threshold algorithm is currently limited to rigid molecules and, thus, only requires evaluation of intermolecular interactions. The extension to flexible molecules would be needed to apply the method to systems where changes in molecular conformation between polymorphs are important. The extension should be straightforward and would require perturbations that involve intramolecular distorsion, and an energy model that includes the *inter-* and *intra*molecular energy contributions.

We believe that the method presented here is a powerful tool which, combined with crystal structure prediction methods, can provide a global picture of the energy landscapes of molecular crystals, and improving our understanding of polymorphism.

5 Acknowledgments

References

[1] S. L. Price, "The computational prediction of pharmaceutical crystal structures and polymorphism," *Advanced Drug Delivery Reviews*, vol. 56, no. 3, pp. 301–319, 2004. Pharmaceutical solid polymorphism in drug development and regulation.

[2] M.-A. Perrin, M. A. Neumann, H. Elmaleh, and L. Zaske, "Crystal structure determination of the elusive paracetamol form iii," *Chem. Commun.*, pp. 3181–3183, 2009.

[3] P. Zhang, G. P. F. Wood, J. Ma, M. Yang, Y. Liu, G. Sun, Y. A. Jiang, B. C. Hancock, and S. Wen, "Harnessing cloud architecture for crystal structure prediction calculations," *Crystal Growth & Design*, vol. 18, no. 11, pp. 6891–6900, 2018.

[4] D.-K. Bučar, G. M. Day, I. Halasz, G. G. Z. Zhang, J. R. G. Sander, D. G. Reid, L. R. MacGillivray, M. J. Duer, and W. Jones, "The curious case of (caffeine)·(benzoic acid): how heteronuclear seeding allowed the formation of an elusive cocrystal," *Chem. Sci.*, vol. 4, pp. 4417–4425, 2013.

[5] D. E. Braun, H. Oberacher, K. Arnhard, M. Orlova, and U. J. Griesser, "4-aminoquinaldine monohydrate polymorphism: prediction and impurity aided discovery of a difficult to access stable form," *CrystEngComm*, vol. 18, pp. 4053–4067, 2016.

[6] A. Pulido, L. Chen, T. Kaczorowski, D. Holden, M. A. Little, S. Y. Chong, B. J. Slater, D. P. McMahon, B. Bonillo, C. J. Stackhouse, A. Stephenson, C. M. Kane, R. Clowes, T. Hasell, A. I. Cooper, and G. M. Day, "Functional materials discovery using energy–structure–function maps," *Nature*, vol. 543, no. 7647, pp. 657–664, 2017.

[7] B. Rice, L. M. LeBlanc, A. Otero-de-la Roza, M. J. Fuchter, E. R. Johnson, J. Nelson, and K. E. Jelfs, "A computational exploration of the crystal energy and charge-carrier mobility landscapes of the chiral [6]helicene molecule," *Nanoscale*, vol. 10, pp. 1865–1876, 2018.

[8] C. M. Aitchison, C. M. Kane, D. P. McMahon, P. R. Spackman, A. Pulido, X. Wang, L. Wilbraham, L. Chen, R. Clowes, M. A. Zwijnenburg, R. S. Sprick, M. A. Little, G. M. Day, and A. I. Cooper, "Photocatalytic proton reduction by a computationally identified, molecular hydrogen-bonded framework," *J. Mater. Chem. A*, vol. 8, pp. 7158–7170, 2020.

[9] C. Y. Cheng, J. E. Campbell, and G. M. Day, "Evolutionary chemical space exploration for functional materials: computational organic semiconductor discovery," *Chem. Sci.*, vol. 11, pp. 4922–4933, 2020.

[10] I. Bier, D. O'Connor, Y.-T. Hsieh, W. Wen, A. M. Hiszpanski, T. Y.-J. Han, and N. Marom, "Crystal structure prediction of energetic materials and a twisted arene with genarris and gator," *CrystEngComm*, vol. 23, pp. 6023–6038, 2021.

[11] S. M. Woodley, G. M. Day, and R. Catlow, "Structure prediction of crystals, surfaces and nanoparticles," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 378, no. 2186, p. 20190600, 2020.

[12] E. C. Dybeck, D. P. McMahon, G. M. Day, and M. R. Shirts, "Exploring the multi-minima behavior of small molecule crystal polymorphs at finite temperature," *Crystal Growth & Design*, vol. 19, no. 10, pp. 5568–5580, 2019.

[13] S. L. Price, "Why don't we find more polymorphs?," *Acta Crystallographica Section B*, vol. 69, pp. 313–328, Aug 2013.

[14] D. E. Braun, J. A. McMahon, L. H. Koztecki, S. L. Price, and S. M. Reutzel-Edens, "Contrasting polymorphism of related small molecule drugs correlated and guided by the computed crystal energy landscape," *Crystal Growth & Design*, vol. 14, no. 4, pp. 2056–2072, 2014.

[15] R. M. Bhardwaj, J. A. McMahon, J. Nyman, L. S. Price, S. Konar, I. D. H. Oswald, C. R. Pulham, S. L. Price, and S. M. Reutzel-Edens, "A prolific solvate former, galunisertib, under the pressure of crystal structure prediction, produces ten diverse polymorphs," *Journal of the American Chemical Society*, vol. 141, no. 35, pp. 13887–13897, 2019. PMID: 31394896.

[16] P. Cui, D. P. McMahon, P. R. Spackman, B. M. Alston, M. A. Little, G. M. Day, and A. I. Cooper, "Mining predicted crystal structure landscapes with high throughput crystallisation: old molecules, new insights," *Chem. Sci.*, vol. 10, pp. 9988–9997, 2019.

[17] R. S. Hong, E. J. Chan, L. Vogt-Maranto, A. Mattei, A. Y. Sheikh, and M. E. Tuckerman, "Insights into the polymorphic structures and enantiotropic layer-slip transition in paracetamol form iii from enhanced molecular dynamics," *Crystal Growth & Design*, vol. 21, no. 2, pp. 886–896, 2021.

[18] W. T. M. Mooij, B. P. van Eijck, S. L. Price, P. Verwer, and J. Kroon, "Crystal structure predictions for acetic acid," *Journal of Computational Chemistry*, vol. 19, no. 4, pp. 459–474, 1998.

[19] E. Schneider, L. Vogt, and M. E. Tuckerman, "Exploring polymorphism of benzene and naphthalene with free energy based enhanced molecular dynamics," *Acta Crystallographica Section B*, vol. 72, pp. 542–550, Aug 2016.

[20] N. F. Francia, L. S. Price, J. Nyman, S. L. Price, and M. Salvalaglio, "Systematic finite-temperature reduction of crystal energy landscapes," *Crystal Growth & Design*, vol. 20, no. 10, pp. 6847–6862, 2020.

[21] N. F. Francia, L. S. Price, and M. Salvalaglio, "Reducing crystal structure overprediction of ibuprofen with large scale molecular dynamics simulations," *CrystEngComm*, vol. 23, pp. 5575–5584, 2021.

[22] J. Schön, H. Putz, and M. Jansen, "Studying the energy hypersurface of continuous systems-the threshold algorithm," *Journal of Physics: Condensed Matter*, vol. 8, no. 2, p. 143, 1996.

[23] M. A. C. Wevers, J. C. Schön, and M. Jansen, "Characteristic regions on the energy landscape of MgF2," *Journal of Physics A: Mathematical and General*, vol. 34, pp. 4041–4052, may 2001.

[24] J. C. Schön, M. A. C. Wevers, and M. Jansen, " entropically stabilized region on the energy landscape of an ionic solid," *Journal of Physics: Condensed Matter*, vol. 15, pp. 5479–5486, aug 2003.

[25] C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward, "The cambridge structural database," *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, vol. 72, no. 2, pp. 171–179, 2016.

[26] S. Lukin, T. Stolar, M. Tireli, M. Blanco, D. Babić, T. Friščić, K. Užarević, and I. Halasz, "Tandem in situ monitoring for quantitative assessment of mechanochemical reactions involving structurally unknown phases," *Chemistry A European Journal*, vol. 23, no. 56, pp. 13941–13949, 2017.

[27] J. C. Schön and M. Jansen, "First step towards planning of syntheses in solid-state chemistry: determination of promising structure candidates by global optimization," *Angewandte Chemie International Edition in English*, vol. 35, no. 12, pp. 1286–1304, 1996.

[28] O. M. Becker and M. Karplus, "The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics," *The Journal of chemical physics*, vol. 106, no. 4, pp. 1495–1517, 1997.

[29] A. Heuer, "Properties of a glass-forming system as derived from its potential energy landscape," *Physical review letters*, vol. 78, no. 21, p. 4051, 1997.

[30] D. Wales, M. Miller, and T. Walsh, "Archetypal energy landscapes," *Nature*, vol. 394, p. 758–760, 1998.

[31] S. Yang and G. M. Day, "Exploration and optimization in crystal structure prediction: Combining basin hopping with quasi-random sampling," *Journal of Chemical Theory and Computation*, 2021.

[32] S. L. Price, M. Leslie, G. W. Welch, M. Habgood, L. S. Price, P. G. Karamertzanis, and G. M. Day, "Modelling organic crystal structures using distributed multipole and polarizability-based model intermolecular potentials," *Phys. Chem. Chem. Phys.*, vol. 12, no. 30, pp. 8478–8490, 2010.

[33] J. A. Chisholm and S. Motherwell, "Compack: a program for identifying crystal structure similarity using distances," *Journal of applied crystallography*, vol. 38, no. 1, pp. 228–231, 2005.

[34] J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," *The Journal of chemical physics*, vol. 134, no. 7, p. 074106, 2011.

[35] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," *Physical Review B*, vol. 87, no. 18, p. 184115, 2013.

[36] J. S. Smith, O. Isayev, and A. E. Roitberg, "Ani-1: an extensible neural network potential with dft accuracy at force field computational cost," *Chemical science*, vol. 8, no. 4, pp. 3192–3203, 2017.

[37] I. T. Jolliffe, "Principal components in regression analysis," in *Principal component analysis*, pp. 129–155, Springer, 1986.

[38] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*, pp. 160–172, Springer, 2013.

[39] H. von Eller, "Sur le polymorphisme de l'indigo," *Bull.Soc.Chim.Fr.*, vol. 106, pp. 1433–1438, 1955.

[40] P. Süsse and A. Wolf, "A new crystalline phase of indigo," *Naturwissenschaften*, vol. 67, p. 453, 1980.

[41] T. Salzillo, S. d'Agostino, A. Rivalta, A. Giunchi, A. Brillante, R. Della Valle, N. Bedoya-Martínez, E. Zojer, F. Grepioni, and E. Venuti, "Structural, spectroscopic, and computational characterization of the concomitant polymorphs of the natural semiconductor indigo," *The Journal of Physical Chemistry C*, vol. 122, no. 32, pp. 18422–18431, 2018.

[42] V. Benghiat and L. Leiserowitz, "Molecular packing modes. part vi. crystal and molecular structures of two modifications of tetrolic acid," *J. Chem. Soc., Perkin Trans. 2*, pp. 1763–1768, 1972.

[43] C. F. Macrae, I. Sovago, S. J. Cottrell, P. T. Galek, P. McCabe, E. Pidcock, M. Platings, G. P. Shields, J. S. Stevens, M. Towler, *et al.*, "Mercury 4.0: From visualization to analysis, design and prediction," *Journal of applied crystallography*, vol. 53, no. 1, pp. 226–235, 2020.

[44] M. Mastalerz and I. M. Oppel, "Rational construction of an extrinsic porous molecular crystal with an extraordinary high specific surface area," *Angewandte Chemie International Edition*, vol. 51, no. 21, pp. 5252–5255, 2012.

[45] J. Nyman and G. M. Day, "Static and lattice vibrational energy differences between polymorphs," *CrystEngComm*, vol. 17, pp. 5154–5165, 2015.

[46] F. Musil, S. De, J. Yang, J. E. Campbell, G. M. Day, and M. Ceriotti, "Machine learning for the structure–energy–property landscapes of molecular crystals," *Chem. Sci.*, vol. 9, pp. 1289–1300, 2018.

[47] J. Yang, S. De, J. E. Campbell, S. Li, M. Ceriotti, and G. M. Day, "Large-scale computational screening of molecular organic semiconductors using crystal structure prediction," *Chemistry of Materials*, vol. 30, no. 13, pp. 4361–4371, 2018.

[48] C. Zhao, L. Chen, Y. Che, Z. Pang, X. Wu, Y. Lu, H. Liu, G. M. Day, and A. I. Cooper, "Digital navigation of energy–structure–function maps for hydrogen-bonded porous molecular crystals," *Nature Communications*, vol. 12, p. 817, 2021.