Supporting Information for

Reducing Overprediction of Molecular Crystal Structures *via* Threshold Clustering

Patrick W. V. Butler, Graeme M. Day

Graeme M. Day
E-mail: g.m.day@soton.ac.uk

This PDF file includes:

Supporting Information Text

1. Crystal Structure Prediction

The initial CSP landscapes were generated for benzene, acrylic acid, and resorcinol using our GLEE(1) program. For the CSPs of benzene and acrylic acid, we followed our previously described methodology based on rigid-body lattice optimisations using an empirically parametrised intermolecular atom-atom exp-6 potential combined with atomic multipole electrostatics (FIT+DMA), the parameters souced from the FIT(2) force field. The molecular geometries were optimised at the B3LYP/6-311G(d,p) level using Gaussian09(3) and held fixed through the rest of the search. Distributed, atom-centered multipoles up to hexadecapole were derived from the resulting electron density by a distributed multipole analysis and partial charges were fitted to the multipoles.(4, 5) A quasi-random search of the lattice packing space with one molecule in the asymmetric unit was then conducted in selected space groups. For benzene and acrylic acid the 25 most common space groups were searched. Valid structures were lattice energy minimised using the software packages PMIN(6) and DMACRYS(7) in a 3-stage protocol consisting of: PMIN at ambient pressure with partial charge electrostatics, FIT+DMA at 0.1 GPa with multipole electrostatics, and lastly the FIT+DMA once more at ambient pressure with multipole electrostatics. The search was terminated for both systems once 250,000 valid structures were generated. In each case, the structure set was clustered iteratively by comparison of simulated powder X-ray diffraction (pXRD) patterns generated by PLATON(8) followed by structural overlay comparisons using the CSD API with a molecular cluster size of 15 molecules and an RMSD cutoff in atomic positions of 0.3 Å.(9) The unique structures were then ranked by the lattice energy of the final optimization stage to yield the CSP landscapes.

In the case of resorcinol, the CSP landscape was generated by applying our recently developed flexible-molecule CSP protocol.* This protocol is largely similar to that described for the rigid systems, however, rather than searching the lattice packing space of a single conformation we instead search with a pre-calculated pool of rigid conformations. Structures are then generated by randomly selecting a conformation from the pool. For resorcinol, the pool was generated by fixing one of the hydroxyl group torsions in an anti position while stepping the other through 360 degrees in 40 degree increments. The space group of the experimental alpha and beta forms, $Pna2_1$, was then searched generating 10,000 valid structures. These were lattice energy minimized initially by the same 3-stage protocol, however, after clustering the unique structures were further optimised with D3 dispersion-corrected tight-binding DFT (DFTB+D3) as implemented in DFTB+(10) using the 3ob parameter set to allow the conformations to relax within the the lattice.(11)

*Manuscript in preparation

2. Threshold Clustering

For benzene, acrylic acid, and resocinol the lowest energy structures from the CSP landscapes were included in the threshold simulations. The structures were expanded to $P1$ cells to remove symmetry constraints on the sampling, with structures in higher symmetry, centered space groups, such that the corresponding $P1$ cells had more than 8 molecules in the cell, reduced to smaller primitive cells for computational efficiency. The Monte Carlo trajectories in all simulations employed the same core moveset consisting of molecular translations and rotations, cell length and angle changes, and cell volume changes. The resorcinol simulations included additionally torsional moves. The maximum move amounts were calibrated for each system by running short simulations and adjusting the cutoffs to achieve an absolute average energy change of around 1 kJ mol^{-1}. If necessary, these were refined further to produce an acceptance ratio of approximately 50-60%. The probability of choosing a given move was set equal to the proportion of the total degrees of freedom it represented. In the case of torsion moves the probability was increased three times this base value in order to better reflect the importance of this degree of freedom.

The number and length of Monte Carlo trajectories was set for each system based on the expected complexity of the energy surface following smaller preliminary simulations of the systems. For benzene 3 trajectories of 30,000 steps (15,000 with 2.5 kJ mol^{-1} lid then 15,000 with 5.0 kJ mol^{-1} lid) were initiated from each CSP structures. For acrylic acid and resorcinol the same number of trajectories were initiated per CSP structures but the length of the trajectories was 20,000 steps (single lid of 5.0 kJ mol^{-1}). In all simulations, the single point evaluations of the perturbed structures and the lattice energy minimizations of accepted structures were performed with the same energy model as used to calculated the CSP landscapes. For benzene and acrylic acid this was FIT+DMA and for resorcinol this was DFTB+D3. Following the simulations, the trajectories were combined by comparing pairs of energy minimized structures across all trajectories in an iterative procedure consisting of initial comparisons of simulated pXRD patterns generated by PLATON(8) followed by structural overlays using the CSD API(9) where matches were identified by root-mean-square differences (RMSD) in atomic positions of less than 0.3 Å using a molecular cluster of 15 molecules. In the case of resorcinol, due to the poor sensitivity of pXRD to hydrogen position, the equivalent structures from pXRD comparisons were checked to ensure they were the same conformation. From the clustering, overlapping trajectories are readily identifiable and the disconnectivity graph was then constructed assuming that all overlapping trajectories represent a single basin.

It was observed that the initial $P1$ minimizations at the start of the threshold simulations led to some minima coalescing. Consequently, the disconnecivity graphs appear to have fewer 'initial minima' than expected. For benzene the 5.0 kJ mol^{-1} disconnectivity graph indicates 90 distinct initial structures from the 100 CSP structures, for acrylic acid it is 97 from 100, and for resorcinol it is 45 from 50. In the case of benzene and resorcinol, this can be partly attributed to combining CSP landscapes generated by searching selected space groups with simulations in $P1$ because some minima in the space group symmetry may not be minima when the symmetry is removed (due to the additional degrees of freedom). However, this cannot be entirely the cause considering the same behaviour is observed for the resorcinol structures, which were optimised in $P1$ as part of the DFTB stage before the simulations. From this, it seems that some of the coalescing minima may then also be due to instability in the minimizer algorithms. Removing the duplicate trajectories of coalesced structures was not found to alter the results, which is expected since in order to coalesce the two structures must be within the same basin. Nevertheless, this does represent inefficient sampling, and thus ideally threshold clustering should be combined with CSP landscapes wherein the structures are true minima on the $P1$ energy surface.

3. Comparison of Threshold Clustering and MD-based Methodologies

To further expand on the comparison of methods for reducing CSP overprediction we have collected here the advantages and disadvantages of the proposed threshold clustering method and the established MD with enhanced sampling approach.

Firstly, as stated in the main text in terms of complexity of the workflow and software required, threshold clustering is much simpler than the MD approaches so far proposed. However, MD software is more well developed and accessible, which has practical benefits.

The enhanced sampling dynamics approaches, sch as metadynamics would be expected to more efficiently search high dimensional configurational spaces, such as that of highly flexible molecules, than random walkers. However, these typically rely on enhanced sampling along a collective variable (CV), which introduces sensitivity of the results to the choice of CV. Of course, this is not a like-for-like comparison and there are more efficient MC methods for exploring complex spaces that could be applied to improve the performance of threshold clustering.

Finally, threshold clustering is readily compatible with the static energy methods that are ubiquitous in the lattice energy approach of conventional CSP. While MD methods could in principal use these energy models, practically it is not an insignificant challenge, which is evidenced by the lack of examples in the literature. Overall, the performance of these methods will likely vary depending on the system and a perfect comparison will be challenging. Moreover, we do not expect such comparisons to be helpful in regard to advancing methods for reducing overprediction and thus such work is not in our future plans.

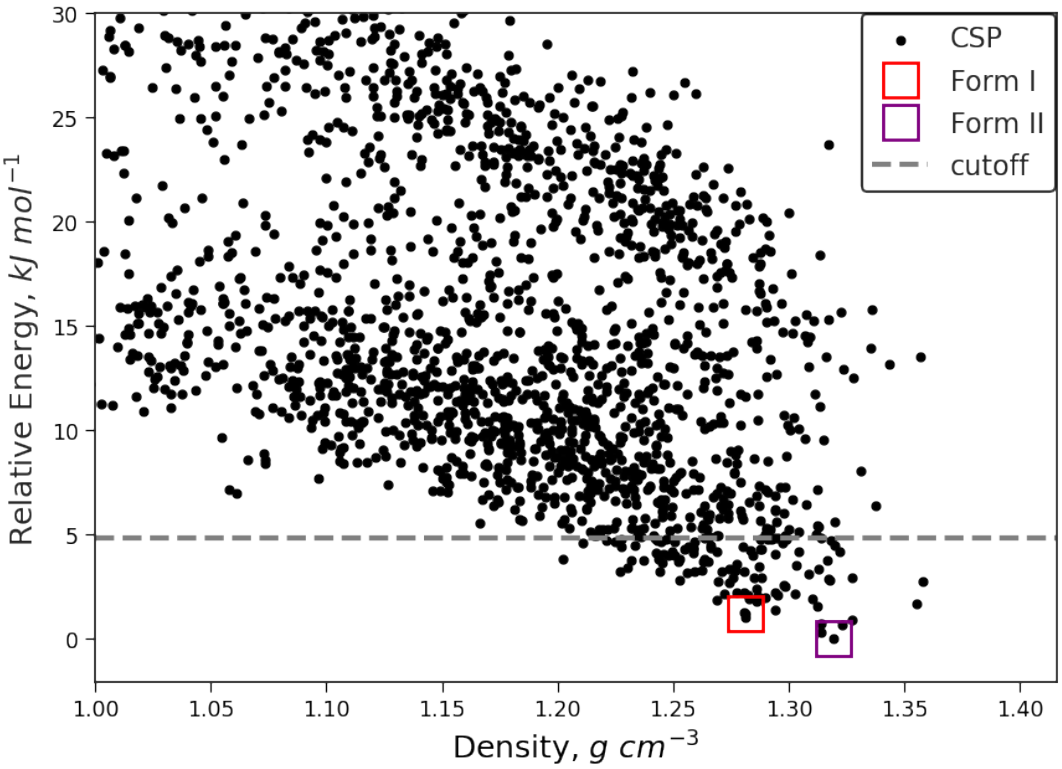

Fig. S1. Initial predicted landscape for acrylic acid. The matches to the experimental polymorphs and the energy cutoff for the structures included in threshold clustering simulations are shown.

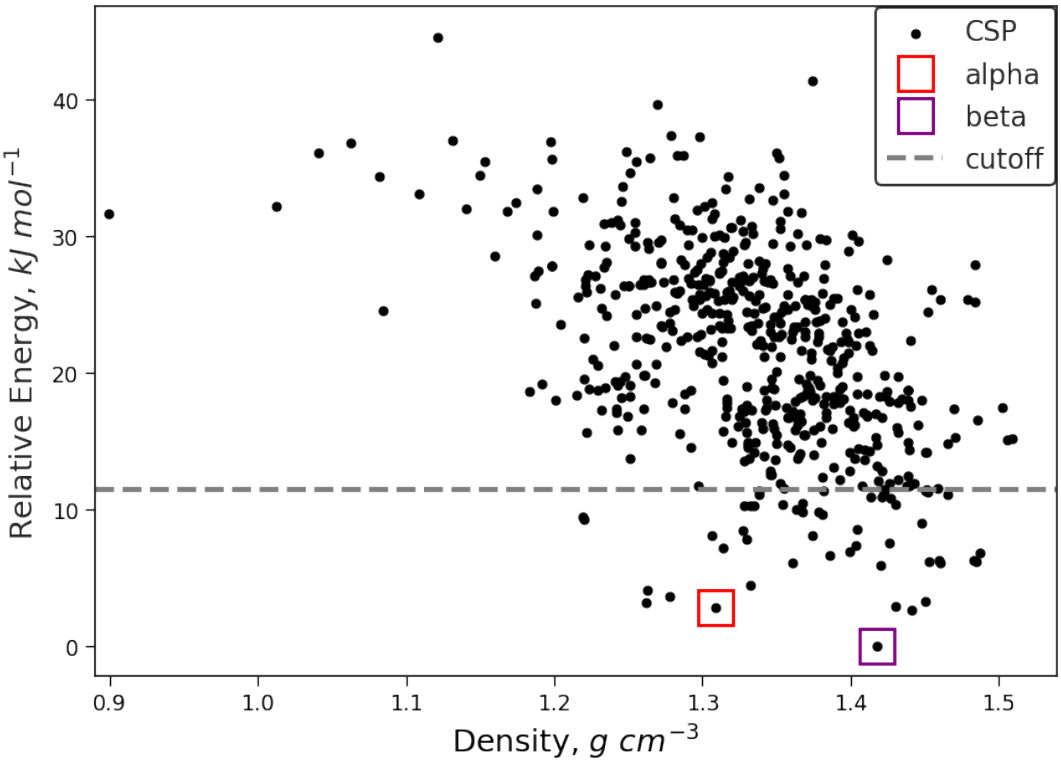

Fig. S2. Initial predicted landscape for resorcinol. The matches to the experimental polymorphs and the energy cutoff for the structures included in threshold clustering simulations are shown.

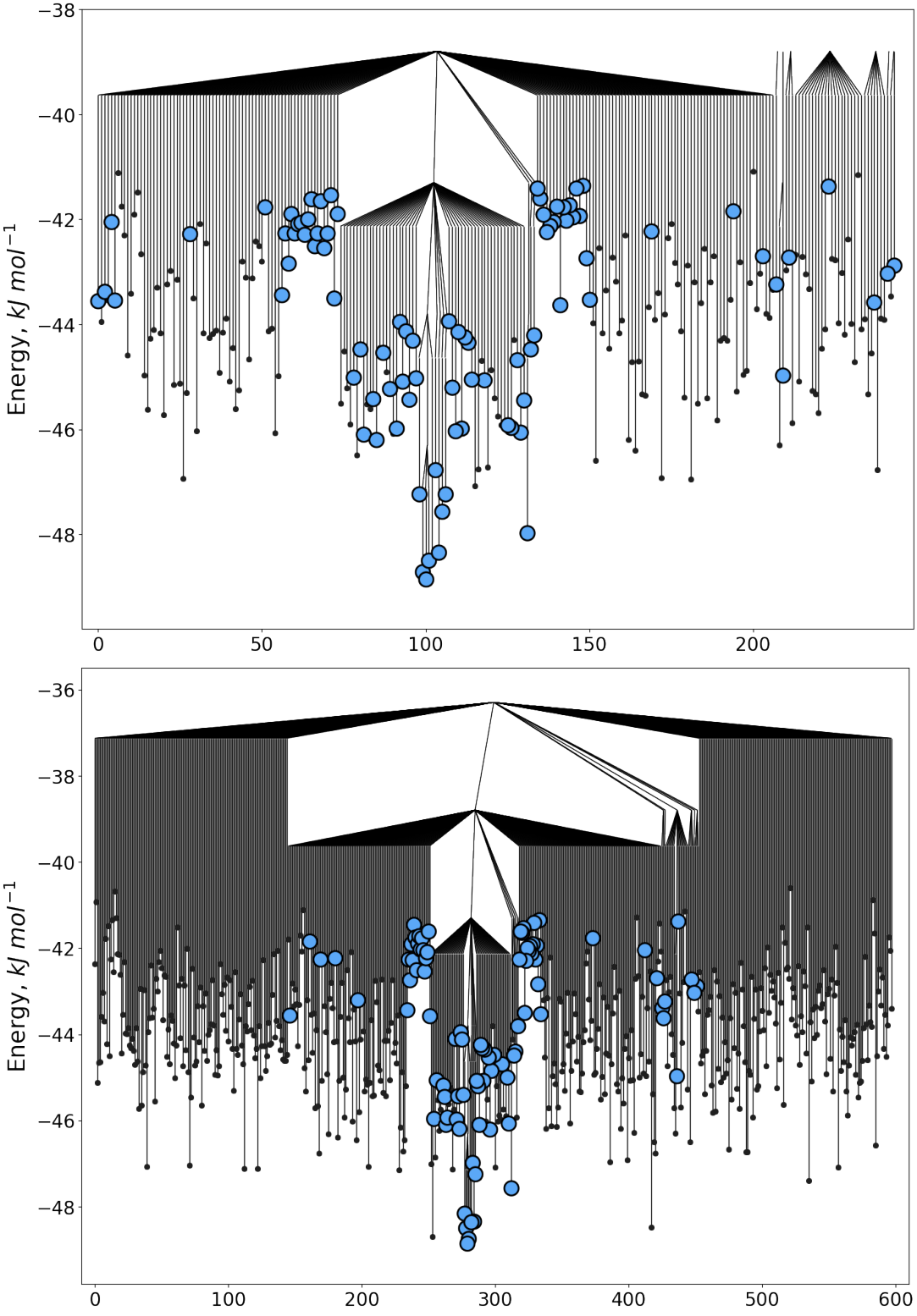

Fig. S3. Threshold clustering results of the lowest 100 structures predicted Benzene structures showing the disconnectivity graph under energy thresholds of 2.5 kJ mol^{-1} (above) and 5.0 kJ mol^{-1} (below). Initital Structures are coloured blue.

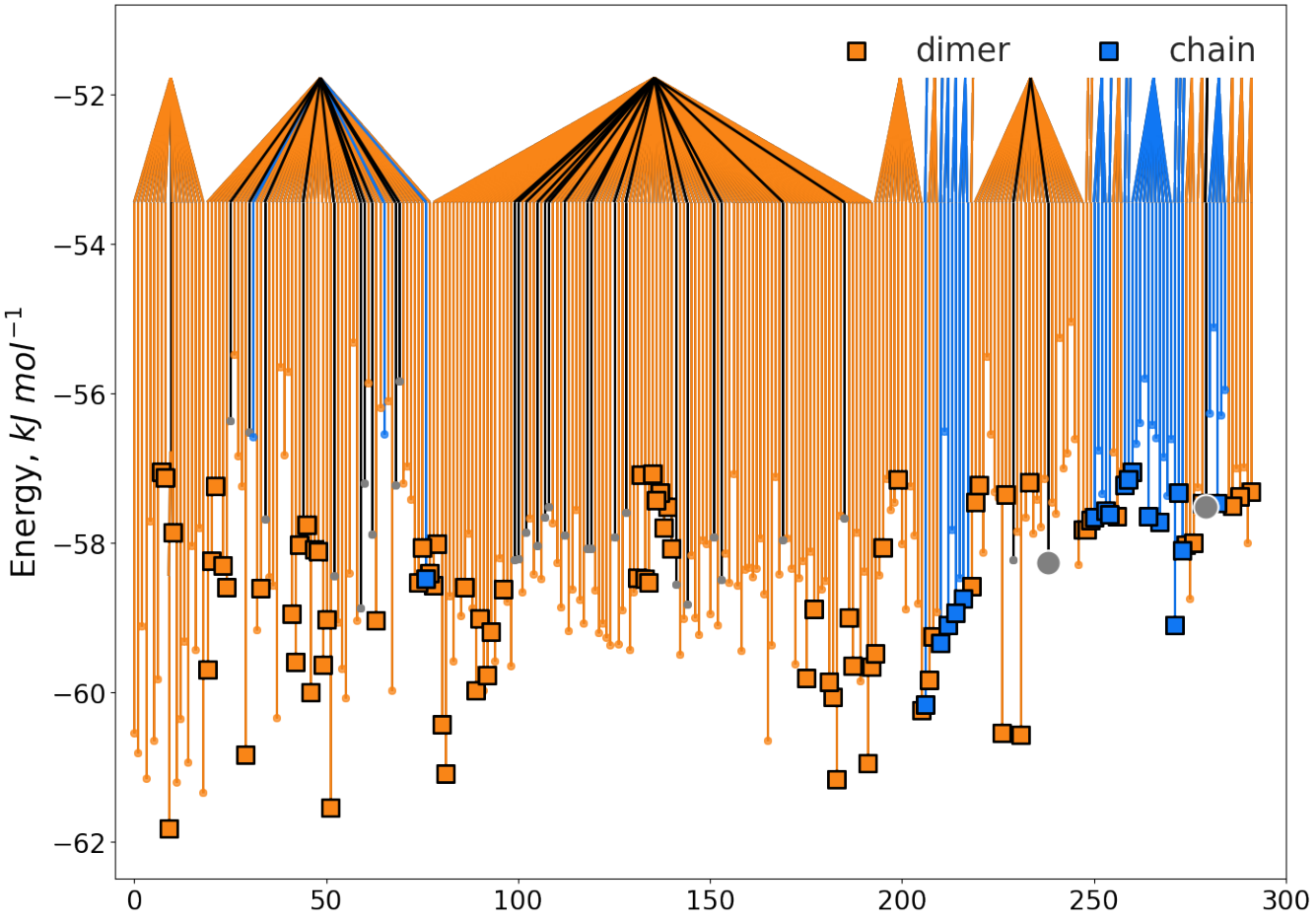

Fig. S4. Disconnectivity graph of the from the threshold simulations of the 100 lowest energy predicted acrylic acid structures. The structures are coloured by their corresponding hydrogen bonding motif. Structures that do not correspond to purely dimer or chain motifs are coloured grey.

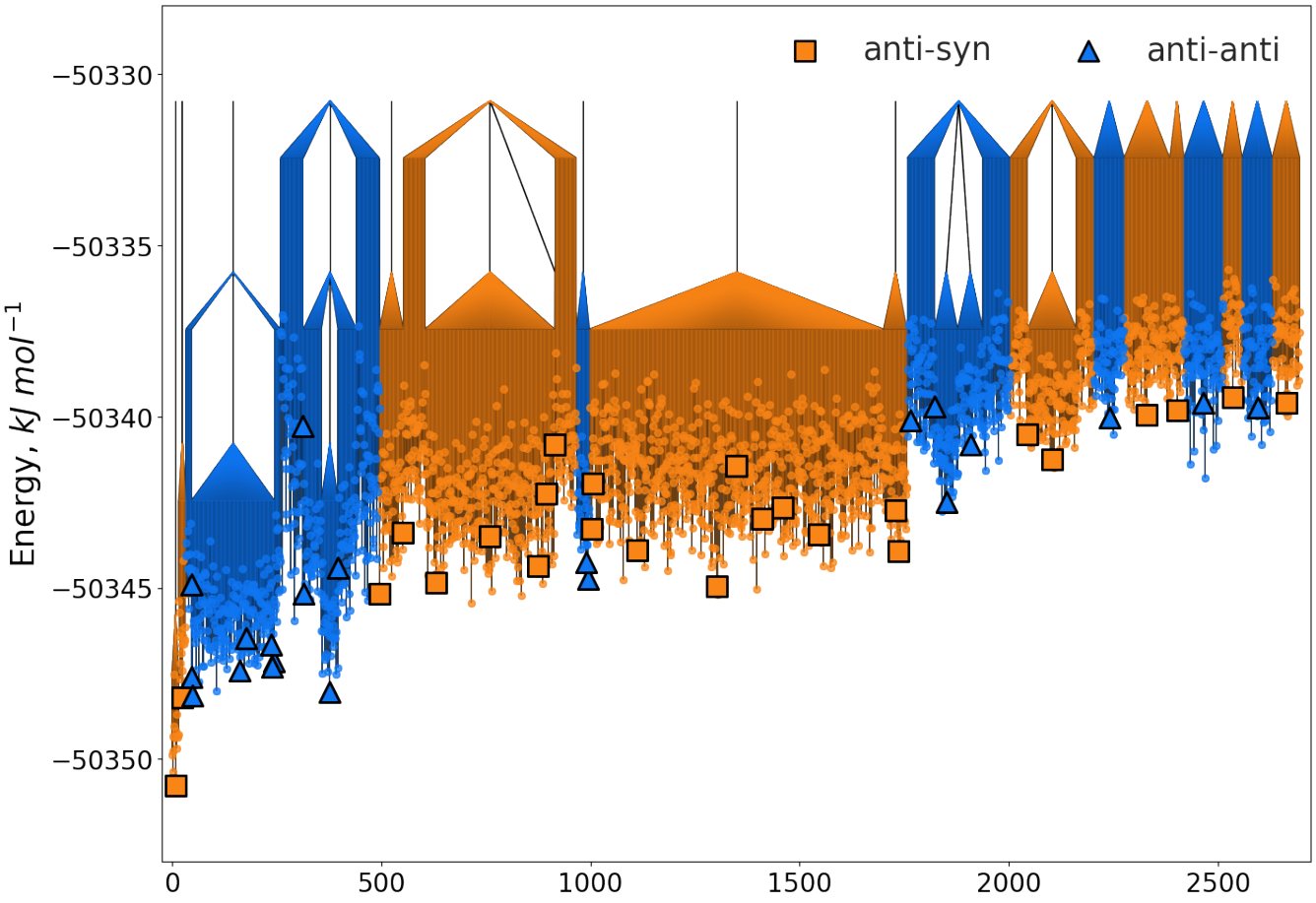

Fig. S5. Disconnectivity graph of the from the threshold simulations of the 100 lowest energy predicted acrylic acid structures. The structures are coloured depending whether closer to the anti-anti conformation or the anti-syn conformation.

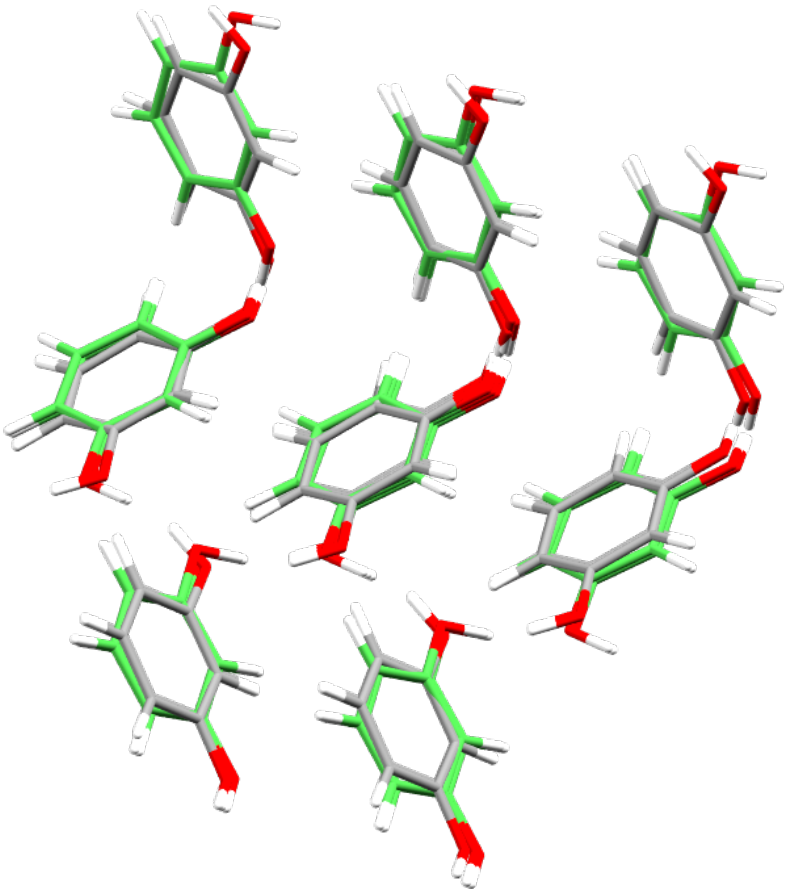

Fig. S6. Crystal structure overlay of the second and fourth lowest energy predicted resorcinol structures which have the same packing but different conformations. The conformation determines the direction of a hydrogen bonding chain.

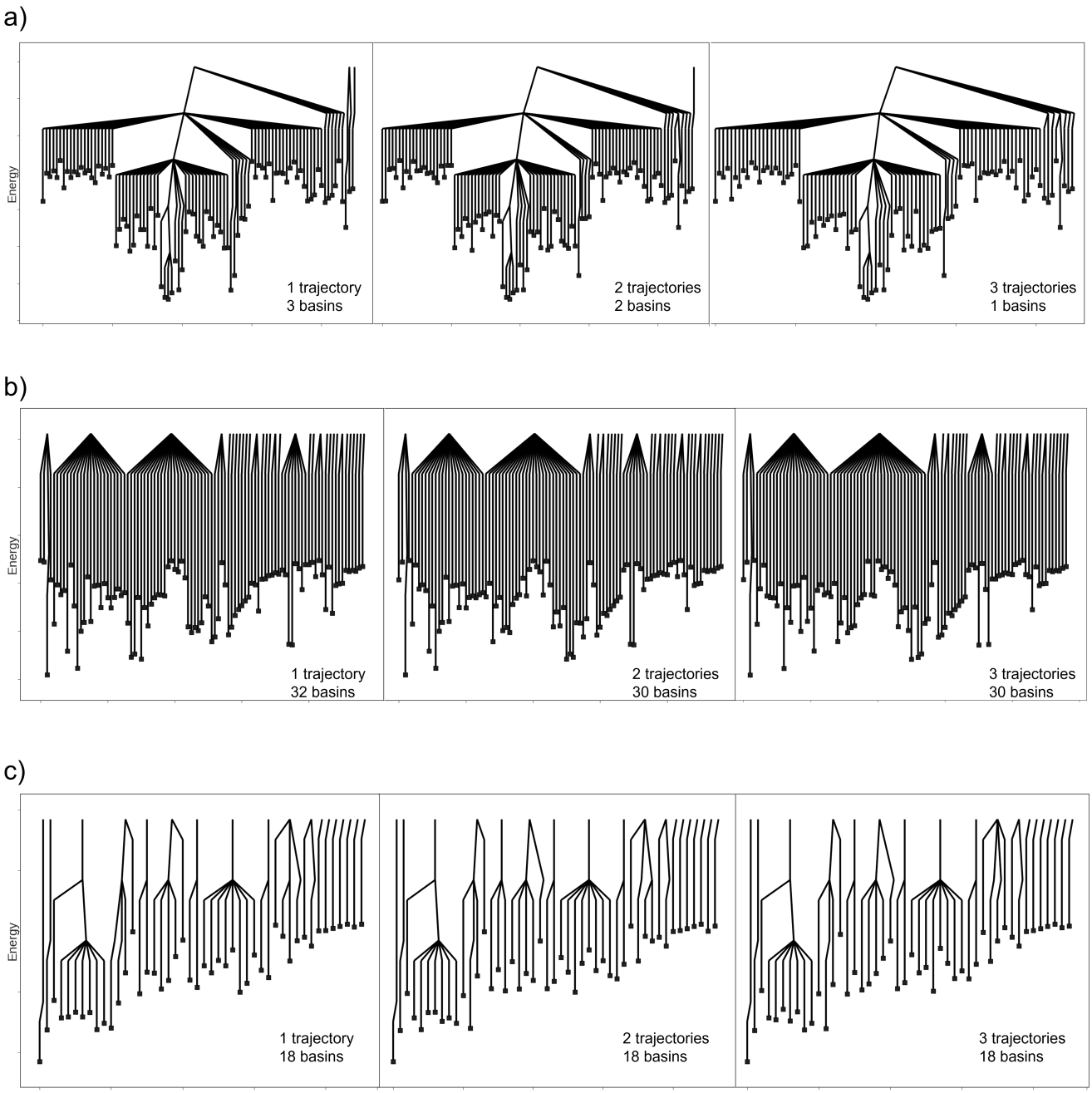

Fig. S7. Convergence of the connections between the initial structures for benzene (a), acrylic acid (b), and resorcinol (c) with increasing number of trajectories initiated per structure: one (left), two (middle), and three (right). MC Structures that were not part of the initial CSP have been omitted from the disconnectivity graphs for clarity.

Table S1. Breakdown of the cost in CPU hours to apply threshold clustering to each of the CSP landscapes. Considering the convergence presented above, we have provided an estimated total cost for running a single trajectory per initial structure.

	Benzene	**Acrylic Acid**	**Resorcinol**
Monte Carlo Simulations	2877	1905	2052
Lattice Energy Minimisations	14049	9111	47100
Clustering	272	309	893
Total	17198	11325	50045
Estimate for single trajectory	5732	3775	16681

References

1. DH Case, JE Campbell, PJ Bygrave, GM Day, Convergence Properties of Crystal Structure Prediction by Quasi-Random Sampling. *J. Chem. Theory Comput.* **12**, 910–924 (2016).
2. DE Williams, SR Cox, Nonbonded potentials for azahydrocarbons: The importance of the Coulombic interaction. *Acta Cryst B* **40**, 404–417 (1984).
3. MJ Frisch, et al., Gaussian09 Revision D.01 (Gaussian Inc.) (2013).
4. AJ Stone, Distributed Multipole Analysis: Stability for Large Basis Sets. *J. Chem. Theory Comput.* **1**, 1128–1132 (2005).
5. GG Ferenczy, Charges derived from distributed multipole series. *J. Comput. Chem.* **12**, 913–917 (1991).
6. JR Holden, Z Du, HL Ammon, Prediction of possible crystal structures for C-, H-, N-, O-, and F-containing organic compounds. *J. Comput. Chem.* **14**, 422–437 (1993).
7. SL Price, et al., Modelling organic crystal structures using distributed multipole and polarizability-based model intermolecular potentials. *Phys. Chem. Chem. Phys.* **12**, 8478–8490 (2010).
8. AL Spek, Single-crystal structure validation with the program PLATON. *J Appl Cryst* **36**, 7–13 (2003).
9. CR Groom, IJ Bruno, MP Lightfoot, SC Ward, The Cambridge Structural Database. *Acta Cryst B* **72**, 171–179 (2016).
10. B Hourahine, et al., DFTB+, a software package for efficient approximate density functional theory based atomistic simulations. *J. Chem. Phys.* **152**, 124101 (2020).
11. JG Brandenburg, S Grimme, Accurate Modeling of Organic Molecular Crystals by Dispersion-Corrected Density Functional Tight Binding (DFTB). *J. Phys. Chem. Lett.* **5**, 1785–1789 (2014).