Machine-Learnt Fragment-Based Energies for Crystal Structure Prediction

David McDonagh, Chris-Kriton Skylaris, and Graeme M. Day*

School of Chemistry, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom

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

Abstract

Crystal structure prediction involves a search of a complex configurational space for local minima corresponding to stable crystal structures, which can be performed efficiently using atom-atom force fields for the assessment of intermolecular interactions. However, for challenging systems, the limitations in the accuracy of force fields prevents a reliable assessment of the relative thermodynamic stability of potential structures, while the cost of fully quantum mechanical approaches can limit applications of the methods. We present a method to rapidly improve force field lattice energies by correcting two-body interactions with a higher level of theory in a fragment-based approach, and predicting these corrections with machine learning. Corrected lattice energies with commonly used density functionals and second order perturbation theory (MP2) all significantly improve the ranking of experimentally known polymorphs where the rigid molecule model is applicable. The relative lattice energies of known polymorphs are also found to systematically improve with the fragment corrections. Predicting two-body interactions with atom-centered symmetry functions in a Gaussian process is found to give highly accurate results using as little as 10-20% of the data for training, reducing the cost of the energy correction by up to an order of magnitude.

The machine learning approach opens up the possibility of more widespread use of fragment-based methods in crystal structure prediction, whose increased accuracy at a low computational cost will benefit applications in areas such as polymorph screening and computer-guided materials discovery.

1 Introduction

The prediction of crystal structures given only basic connectivity information of a molecule, such as a chemical diagram, is an important challenge for computational chemistry.[1,2] The motivations for crystal structure prediction (CSP) are three-fold. Many molecules do not produce single crystals suitable for structure determination using conventional methods, and direct structure solution from powder X-ray diffraction patterns is not always possible. The computational generation of likely structures can provide a useful starting point for structure determination from incomplete data: examples include structure determination from powder X-ray diffraction[3], electron diffraction[4] and solid state NMR[5] data. There is strong interest in CSP in the pharmaceutical industry, where different crystal forms of a molecule (polymorphs) can have different properties, so that the anticipation of polymorphism is important for maintaining strict property control[6], as well as having intellectual property implications[7]. In a broader context, the prediction of crystal structures could have extensive ramifications for materials science. The reliable prediction of solid-state phases from molecular information creates an opportunity for *in silico* design, where materials with desirable properties can be identified before entering the laboratory[8]. In addition to the economic benefits of guiding experimental work towards the most promising candidate molecules, a computation-led approach can enable the exploration of untouched areas of chemical space.[9,10]

CSP is typically approached as an exploration of the high dimensional lattice energy surface for low energy local minima. To exhaustively explore the space of possible crystalline forms, it is often necessary to generate tens or hundreds of thousands of trial crystal structures for

a given molecule[11], and a metric is required for identifying structures that are of interest. In the most general case, candidate crystal structures are ranked by their geometry-optimized, static lattice energies, under the assumption that crystal structure is largely driven by thermodynamics[1,2]. Hence, the lowest energy predicted structure is assumed to be the most likely experimental structure, and alternative structures within a small lattice energy range are assumed to be potential polymorphs. Calculating accurate lattice energies for large numbers of crystal structures is one of the key challenges in CSP, due to the computational expense of calculating sufficiently accurate energies to rank predicted crystal structures: the lowest energy predicted crystal structures, and pairs of observed polymorphs, are often separated by less than a kJ/mol.[12] For organic molecular crystals, the two most widely adopted approaches to ranking predicted structures[13] are anisotropic, multipole-based atom-atom force fields[14,15], and dispersion-corrected periodic Density Functional Theory (DFT-D) methods[16]. The latter are typically found to provide higher accuracy[17], but at a cost that is restrictive for larger systems.

Alternative approaches to traditional force fields have recently been growing in popularity, where highly flexible potentials are optimized entirely on known data under the umbrella term of machine learning[18,19]. Applications of machine learning in chemistry have grown rapidly in the last decade[20], with successful results in potential energy surface prediction[21,22], energy corrections between levels of theory[23,24], predicting chemical properties[25], and promising indications of predicting chemical phenomena that are not explicitly defined in the dataset[26].

Most pertinently, machine learning potentials have been shown to approximate DFT-D lattice energies for crystal structure landscapes to a good degree of accuracy[27–29]. In the context of CSP, machine learning potentials have typically be trained on data generated using generalised gradient approximation (GGA) DFT functionals, as these have largely remained the standard for accuracy[13] where hybrid functionals or post Hartree-Fock wavefunction methods

applied directly to the solid state remain too expensive. However, there are known limitations to GGA functionals[30], such as delocalization error, which leads to the overstabilization of charge separation[31], potentially resulting in serious errors for systems such as acid-base co-crystals or salts[32]. Furthermore, hybrid functionals also fail to reproduce the experimental stability order for some challenging polymorphic systems[33,34]. These discrepancies highlight the need for different energy models to give reliable methods for the full spectrum of organic molecular crystal structures. Machine learning approaches could offer a means to expand the range of available energy models used in CSP without a large increase in computational cost.

The approach we adopt here, as opposed to using a fully periodic formalism, is to calculate the energy of a crystal structure as a sum of many-body terms,

$$\sum_i E_i + \sum_{i<j} E_{ij} + \sum_{i<j<k} E_{ijk} + ..., \tag{1}$$

where the first term is the sum of monomer energies, the second the sum of dimer energies, and so on. Such fragment-based methods are now a common method to approximating more expensive periodic levels of theory[35–38]. Of particular relevance is the Hybrid-Many Body Interaction (HMBI)[37,39,40] scheme, which employs polarizable force fields for many-body and long-range pairwise interactions, combined with quantum mechanical (QM) methods such as Møller–Plesset perturbation theory or Coupled Cluster theory for short-range pairwise interactions.

Fragment-based approaches are attractive for CSP as they allow for the introduction of more accurate energy models without the limitations of GGA DFT-D methods. Incorporating fragment-based methods into CSP remains expensive, however, owing to the many two-body terms that must be calculated. To overcome this high computational cost, we investigate using a machine learning method to map force field two-body interaction energies

to a higher level of theory. The prediction of many-body terms has been demonstrated for finite systems[41,42], indicating that machine learning methods can readily be applied to this task.

We first outline the fragment-based method used is the study, which combines anisotropic atom-atom force fields with a range of QM methods. Following similar work in comparing energy models for CSP[43], the energy model is tested by monitoring the ranking of observed crystal structures within CSP structure sets for a range of molecules when calculating each two-body term explicitly. We then describe a machine learning approach to learning the difference between force field and QM dimer energies, which reduces the cost of the fragment based QM correction. The results of our machine learnt energy model are assessed by comparison to explicit QM calculations.

2 Computational Details

2.1 Test Set Molecules

The fragment-based correction to force fields was tested on ten organic molecules chosen to assess a variety of intermolecular interactions. These were selected from the X23 dataset[44] of crystal structures with reliable experimentally determined sublimation energies, or molecules that are known from previous studies to be problematic for force field energy models[43]. So that our evaluation is focused on the intermolecular energy model, molecules were selected where the gas phase geometry optimization of the isolated molecule is in good agreement with the molecular geometry in the experimental crystal structure. This allows us to apply a rigid molecule approach during CSP and means that no model for the intramolecular contribution to relative lattice energies is required. Application of the fragment approach to flexible molecules, where the molecular geometry is influenced by intermolecular interactions in its crystal structures, will require methods for modelling the energy associated with

flexible degrees of freedom within a molecule and will need to describe the dependence of intermolecular interactions on molecular conformation; these issues are not addressed in the current work.

The selected molecules are shown in Fig. 1 along with the Cambridge Structural Database (CSD) reference codes[45] of their known crystal structures. Five of these molecules are known to be polymorphic, so that our test set contains 16 experimentally determined crystal structures.

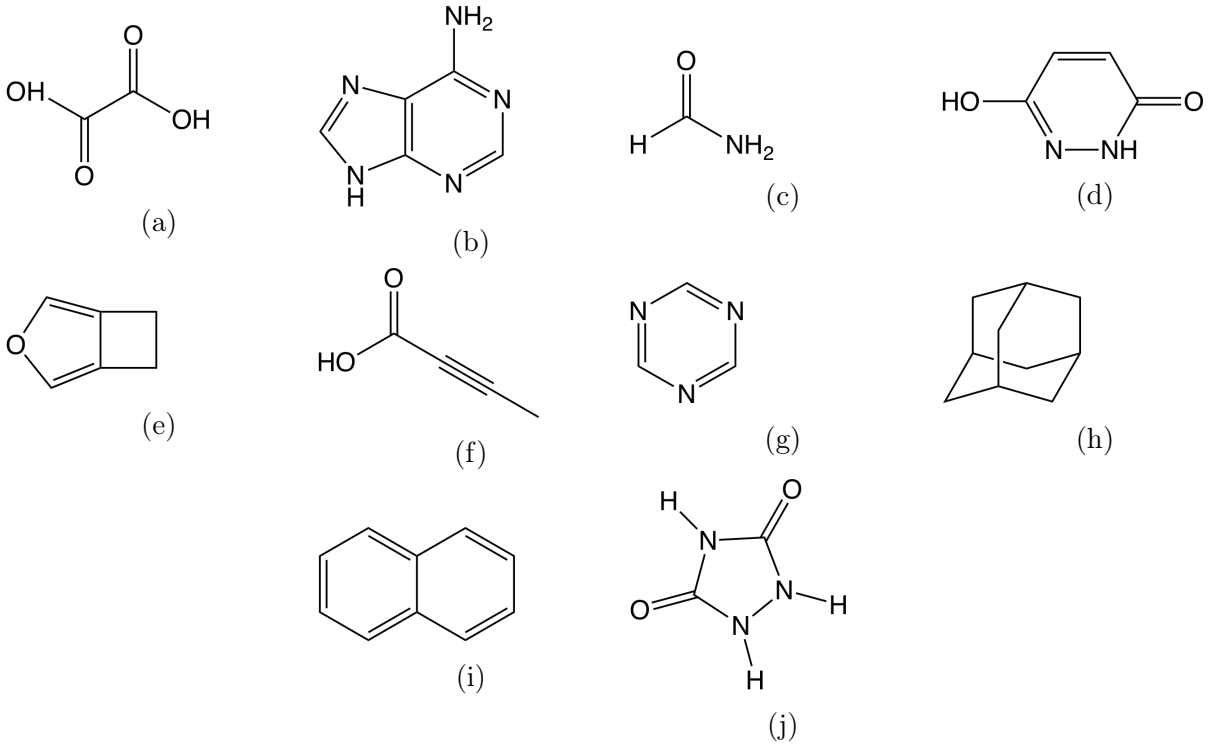


Figure 1: The 10 selected species for this study, with 16 experimentally known crystal forms (CSD reference codes[45] for each form are given in parentheses): (a) oxalic acid α (OXALAC05[46]) and β (OXALAC07[47]) polymorphs; (b) adenine polymorphs I (KOBFUD[48]) and II (KOBFUD01[49]); (c) formamide (FORMAM02[50]); (d) maleic hydrazide monoclinic (MALEHY01[51]), triclinic (MALEHY10[52]), and MH3 monoclinic (MALEHY12[53]) polymorphs; (e) 3,4-cyclobutylfuran metastable orthorhombic (XULDUD[54]) and monoclinic (XULDUD01[54]) polymorphs; (f) tetrolic acid α (TETROL[55]) and β (TETROL01[55]) polymorphs; (g) 1,3,5-triazine (TRIZIN03[56]); (h) adamantane (ADAMAN08[57]); (i) naphthalene (NAPHTA30[58]); and (j) urazole (KOXRIY[59]).

2.2 Crystal Structure Prediction

For each species, molecular geometry optimizations were carried out in Gaussian 09[60] using the B3LYP hybrid functional[61] and the 6-311G** Pople basis set. With this as a fixed molecular geometry, structures were generated using a quasi-random sampling method using the GLEE (Global Lattice Energy Explorer) code[11]. Each space group considered is sampled separately, by generating trial structures with unit cell dimensions, molecular positions and orientations sampled using a low discrepancy method. Approximately 10,000 valid (successfully lattice energy minimized) structures were generated in each of the 11 most common space groups ($P2_1/c$, $P2_12_12_1$, $P\bar{1}$, $P2_1$, $Pbca$, $C2/c$, $Pna2_1$, Cc, $Pca2_1$, $C2$, $P1$) with one independent molecule in the crystallographic asymmetric unit ($Z' = 1$). Approximately 85% of known molecular organic crystal structures are observed to crystallize with $Z' \leq 1$ in one of these space groups. For adenine, the known space group of form (II)[49] ($Fdd2$) was also included. All generated trial crystal structures were geometry-optimized using the crystal structure modelling code DMACRYS[15] with the molecular geometry kept fixed. Intermolecular interactions were evaluated using the **FIT+DMA** anisotropic force field[14], which has the form

$$E_{MN}^{intermolecular} = \sum_{i,k} A^{cd} exp(-B^{cd} r_{ik}) - C^{cd} r_{ik}^{-6} + E_{ik}^{elec}, \quad (2)$$

where i and k refer to atoms with type c and d from molecules M and N at a distance r_{ik}, respectively. The parameters A^{cd}, B^{cd} and C^{cd} of the FIT model were determined by empirical parameterization against experimental crystal structures and sublimation enthalpies[14,62,63]. The term E_{ik}^{elec} describes electrostatic interactions between atoms i and k from atom-centered multipoles up to hexadecapole on all atoms, which were derived from a distributed multipole analysis of the B3LYP/6-311G** charge density[64]. Charge–charge, charge–dipole and dipole–dipole interactions were calculated with Ewald summation. All other interactions were calculated between whole molecules with a centre of mass separation of less than 25Å.

Previous work indicates that crystal structures of interest are typically within 10 kJ/mol of the global minimum of crystal structure landscapes[12,43]. However, as many of the molecules studied here were chosen specifically as examples where force fields perform poorly in ranking the observed structures, we retained structures within a wider (20 kJ/mol) energy window of each landscape for further calculations.

Duplicate crystal structures were removed by first clustering within space groups, comparing any pair of structures with nearly identical lattice energies and densities (within 1 kJ/mol and 0.05 g/cm^3) using dynamic time-warping comparisons[65] of simulated X-ray diffraction patterns calculated using PLATON[66]. Structures were then clustered further within space groups using the COMPACK algorithm, where arrays of interatomic distances within finite clusters of 30 molecules from each crystal structure were compared[67]; structures were identified as duplicates if all 30 molecules matched, and the 30-molecule clusters could be overlaid with an RMSD in atomic positions of below 0.3 Å. A third clustering was performed between space groups using the dynamic time-warping comparison of X-ray diffraction patterns, to obtain a final data set.

The known crystal structures for each molecule (Fig. 1) were identified in a given landscape by comparing the predicted structures to the experimental crystal structure using COMPACK (as implemented in Mercury[68]) with 30-molecule clusters.

2.3 Fragment-Based Lattice Energy Model

The energy ranking of the sets of force field predicted crystal structures for each molecule has been assessed using a 'single-point' energy correction: a re-calculation of the lattice energy using a fragment-based correction to the force field energy, without further structural re-optimization. The fragment-based energy model tested in this study has a similar form

to the HMBI scheme[37,39,40], where the lattice energy is calculated as a force field (*ff*) lattice energy, with significant two-body terms corrected to a higher level (*hl*) of theory,

$$E'_{latt} = E^{(ff)}_{latt} + \sum_i \left(E^{(hl)}_{cm,i} - E^{(ff)}_{cm,i} \right) S, \tag{3}$$

where $E^{(ff)}_{latt}$ is the force field (**FIT+DMA**) total lattice energy, and $E^{(hl)}_{cm,i}$ and $E^{(ff)}_{cm,i}$ are the two-body interaction energies calculated between the i^{th} component and a central, reference molecule (cm) at the higher (QM) and force field levels of theory, respectively. S is a cutoff function that switches from 0, 1, such that

$$S = \begin{cases} 1 & \text{for } R < R_c - t \\ \\ 0 & \text{for } R > R_c, \end{cases} \tag{4}$$

where R is the nearest intermolecular atom-atom distance of a dimer, and R_c is a critical distance within which energies are replaced. To prevent discontinuities at the transition between the two levels of theory, cubic splines were used to interpolate between levels of theory in a buffer region between $R_c - t$ and R_c. In this work, we applied a buffer region width of $t = 0.5$ Å.

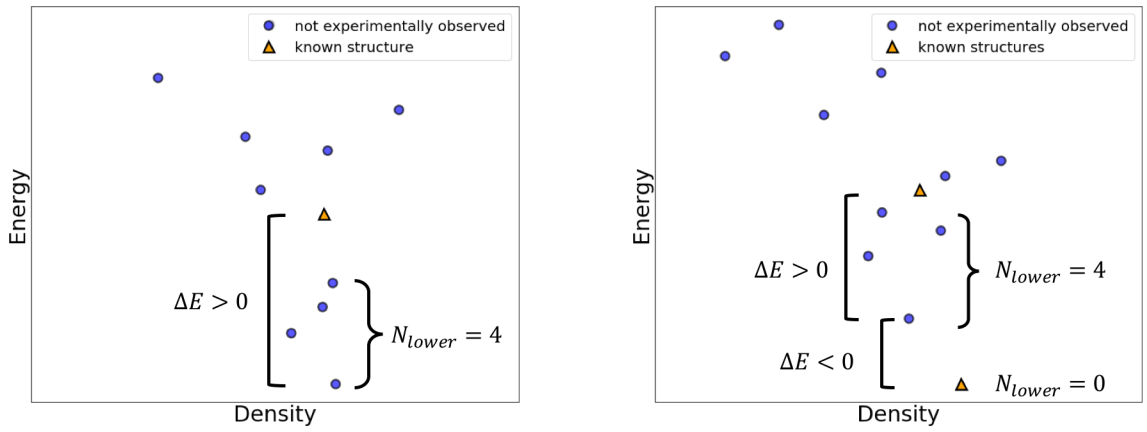
To identify two-body interactions that contribute significantly to the lattice energy, the variance of the sum of the force field dimer energies with R_c is used as a guide. For each crystal structure, a rolling average of the sum of dimer energies is retained with the addition of each shell of interacting molecules. When the addition of a new shell (taken as an increase in R_c of 0.5 Å) results in a variance of less than 0.1 kJ/mol, the energy is assumed to be sufficiently converged, with interactions beyond this distance left at the force field level of theory. To reduce the chance of identifying convergence erroneously, a minimum R_c of 8 Å was used, calculated as the nearest intermolecular atom-atom distance from the reference molecule. Symmetrically equivalent dimers were identified by comparing the centroid-centroid dis-

tances, and the distances of each atom in the two molecules to the inter-centroid midpoint.

The QM dimer-corrected energy model was applied to all crystal structures within 20 kJ/mol of the global minimum for each molecule. To investigate how **FIT+DMA** CSP landscapes change using dimer corrections at different levels of quantum mechanical theory, second order Møller-Plesset perturbation theory[69] (MP2) and the hybrid B3LYP functional[61] were selected as higher levels of theory, along with the PBE[70] functional as a reference for GGA methods. In both DFT methods, dispersion corrections were added using Grimme's D3 correction[71]. The fragment approach opens the possibility for corrections at any affordable QM level of theory and it has been shown that wavefunction methods beyond MP2 are required for chemical accuracy (in particular for π-stacking dispersion interactions)[40]. The significant increase in cost, and scaling with molecular size, for methods such as CCSD(T) makes them less viable for calculating two-body interactions for the entire low energy region of crystal structure landscapes. Therefore, we have focussed initially on the lower cost QM methods that could provide a significant improvement in energy ranking, while still being comparable with force field calculations for efficiency when combined with a machine learning approach. For each unique dimer, the PBE-D3/6-31+G**, B3LYP-D3/6-31+G** and MP2/6-31+G** interaction energy was evaluated as the difference between the dimer and monomer energies, where basis set superposition error was treated with a counterpoise correction[72]. All QM calculations were performed using Gaussian 09[60]. Using these energies, the energy difference $E_{cm,i}^{(hl)} - E_{cm,i}^{(ff)}$ was evaluated for all dimers within the defined distance cut-off, giving the crystal structure landscape at the **FIT+DMA** force field geometry, but with dimer-corrected energies. Henceforth, we refer to these landscapes as **FIT+DMA+PBE**, **FIT+DMA+B3LYP**, and **FIT+DMA+MP2**.

Following earlier work on the assessment of force field methods for CSP[43], each energy model was assessed by using two measures of the position of the known crystal structures amongst

the energy-ranked predicted structures, ΔE and N_{lower} (see Fig. 2). N_{lower} is a measure of the energetic *rank* of the known crystal structure and is defined as the number of predicted structures not observed experimentally that are lower in energy than the observed structure. The relative energy (ΔE) of a known crystal structure is defined as the energy difference between its calculated energy and the energy of the lowest energy predicted structure that is not observed experimentally. Under the assumption that observed crystal structures always correspond to the lowest energy possible structures, and that all polymorphs of the studied molecules have been found, the reported polymorphs should correspond to the lowest energy structures within the CSP structure sets, resulting in $N_{lower} = 0$ and $\Delta E < 0$ for all polymorphs of each molecule.



(a) (b)

Figure 2: Examples of CSP energy-density plots illustrating N_{lower} and ΔE (a) for a molecule with one known structure, and (b) for a molecule with two known polymorphs. Each point corresponds to a distinct predicted crystal structure.

Deviations from $N_{lower} = 0$ and $\Delta E < 0$ indicate either additional, kinetic, structure-determining effects or deficiencies in the model used to rank the lattice energies. The success of CSP by global lattice energy minimization, and the systematic improvements in results that have been observed with improved accuracy of lattice energy calculations,[13] support the view that minimization of the lattice energy is a dominant influence in crystallization. Nev-

ertheless, kinetics of crystal nucleation and growth can be important in determining crystal structure, particularly in the presence of strong intermolecular interactions; the importance of kinetic effects seems to increase with molecular size, as conformational rearrangement and the templating effects of solvent within crystal structures can lead to experimental realization of high energy, metastable crystal structures.[8,73,74] An understanding of kinetic influences on crystal structure should be based on an accurate understanding of the relative thermodynamic stability of competing crystal structures, which is the focus of the current work. The computation of vibrational contributions to the free energy is not treated in this work, as this contribution to free energy differences between polymorphs is usually small compared to the lattice energy difference[12,75]. This does limit the accuracy of all discussed energy models, but is an effect that is expected to become more significant in applications to flexible molecules.[76,77]

To further assess the accuracy of the fragment-based energy models, we examined calculated polymorph lattice energy differences for three molecules, oxalic acid, tetrolic acid and adenine, where the polymorph stability order and, for oxalic acid and adenine, the measured energy difference is known. To extend this comparison to popular solid state DFT-D methods, periodic DFT-D calculations were carried out in the VASP software package[78–81], using the PBE exchange-correlation functional[70] with Grimme D3 dispersion[82] (PBE-D3). DFT-D geometry optimizations were performed using a two step process of first optimizing atomic positions with fixed lattice parameters, before then optimizing all degrees of freedom; this has been found to improve convergence for molecular crystals[83,84]. All calculations used a plane-wave energy cutoff of 500 eV with a maximum k-point spacing of 0.05 Å^{-1}, using the projector-augmented wave (PAW) method[85] and standard pseudopotentials[86]. Tolerances for energy convergence were set to 10^{-7} eV per atom, and force tolerances for geometry optimization to 3×10^{-2} eV/Å.

2.4 Machine Learning of Dimer Energy Corrections

Atom-centered symmetry functions[87] were used to describe two-body interactions, including the recent modification to allow for high resolution[22]. Here the local environment of the ith atom in a system of N atoms is described by a radial terms

$$G_{i_a}^R = \sum_{j \neq i}^{N} \exp\left(-\eta_R (R_{ij} - R_{s_a})^2\right) f_c(R_{ij}), \tag{5}$$

where R_{ij} is the distance between atoms i and j, η_R and R_{s_k} are hyperparameters affecting the localization and shift of each Gaussian function, respectively, and $f_c(R_{ij})$ is a cutoff function, such that

$$f_c(R_{ij}) = \begin{cases} \frac{1}{2} cos(\frac{\pi R_{ij}}{R_r}) + \frac{1}{2} \quad for \quad R_{ij} \leq R_r \\ 0 \quad for \quad R_{ij} > R_r, \end{cases} \tag{6}$$

for a predefined radial cutoff R_r, and b angular terms

$$G_{i_b}^A = 2^{1-\zeta} \sum_{j,k \neq i}^{N} (1 + cos(\theta_{ijk} - \theta_{s_b}))^{\zeta} \exp\left(-\eta_A \left(\frac{R_{ij} + R_{ik}}{2} - R_{s_b}\right)^2\right) f_c(R_{ij}) f_c(R_{ik}), \tag{7}$$

where θ_{ijk} is the angle between atoms i, j and k, $f_c(R_{ij})$ are $f_c(R_{ij})$ cutoff functions with an angular cutoff R_a, and θ_{s_l}, η_A and ζ are hyperparameters. To combine these as a descriptor for a dimer, the symmetry functions are arranged in a vector. Although this removes invariance to permutation, the machine learning model is only trained on a specific landscape with all dimers containing the same molecules, and so maintaining consistent ordering of atoms for every dimer is trivial. The descriptor is also not invariant to permutation of the molecules themselves, however this effect is negligible as only unique dimers are used in the data set. The oxalic acid **FIT+DMA+MP2** landscape was selected to test the machine learning approach, as this shows the most dramatic changes relative to **FIT+DMA**, which

should be the most challenging for machine learning methods to capture. For comparison, we also tested the approach on the **FIT+DMA+MP2** landscape of maleic hydrazide, as a second polymorphic molecule where the fragment-based energy corrections are large.

Hyperparameter values were initially selected to give an even spacing of Gaussians out to the largest intermolecular nearest atom-atom distance for radial terms, and to sample $-\pi$ to π for the angular terms. The number of symmetry functions and hyperparameters were then optimized using the the entire dataset of both species at the **FIT+DMA** level of theory, where the number of symmetry functions was reduced to the minimum number at which all dimers were fully distinguishable with nearby neighbours in descriptor space displaying similar **FIT+DMA** dimer energies. Three radial and two angular terms were used to describe each atom in the oxalic acid and maleic hydrazide dimers (giving a total of 80 symmetry functions per oxalic acid dimer and 120 symmetry functions per maleic hydrazide dimer). The final values were angular (R_a) and radial (R_r) cutoffs of 9.66 Å and 9.53 Å, respectively, radial R_s terms of 1 Å, 2.8665 Å, and 4.733 Å, angular θ_s terms of 0 and π rad each with R_s values of 2.0 Å, η_R and η_A values of 2.8 Å^{-2} and 2.5 Å^{-2}, respectively, and a ζ value of 4.0. Both landscapes have dimer corrections with intermolecular nearest atom-atom distances out to approximately 9.3 Å, and so the resolution of weakly interacting dimers is expected to be much lower than those with small intermolecular distances. However, given the relatively small data sets, and as these terms are expected to only make minor contributions to the energy, focus was given to accurately resolving strongly interacting dimers without significantly increasing the size of the descriptor space.

A Gaussian process model was built using the Scikit-learn Python package[88], using a radial basis function (squared exponential) kernel with a white noise kernel to avoid overfitting. To evenly sample the descriptor space, training data were selected using the max-min algorithm[89].

3 Results and Discussion

3.1 Crystal Structure Ranking using the Fragment-Based Approach

Accurate energetic ranking of crystal structures is vital in CSP, as polymorphs are known to often differ by only a few kJ/mol.[12] It is generally understood that a more accurate energy model leads to improved (i.e. lower) ranking of experimentally observed crystal structures, which in turn makes it easier to identify novel structures of interest[2,90] perhaps as undiscovered polymorphs or realizable materials with attractive properties. Given a crystal structure landscape with lattice energies calculated using a force field, the fragment-based model described here can be seen as one of the simplest initial improvements to the energy model, as energetically significant two-body force field interactions are replaced with a higher level of theory without any re-optimization of the geometry.

All of the known crystal structures were successfully identified in the **FIT+DMA** landscapes of the ten molecules, with the exception of the monoclinic form of 3,4-cyclobutylfuran. This molecule featured in the first blind test of CSP[54] and is known to present difficulties for force field methods. Optimizing the experimental crystal structure from the CSD using **FIT+DMA** leads to a large structural change when using the DFT-optimized molecular geometry, where only 11 out of 30 molecules are found to match when comparing the optimized and experimental structures using the crystal structure comparison tool in Mercury[68]. Using a similar force field[91], Price *et al.* found that this polymorph was sensitive to the position of the hydrogen atom interaction site, finding a RMSD of 5.9 to 6.6% in lattice parameters when comparing the experimental structure to their optimized version[54]. The experimental crystal structure is reproduced more satisfactorily when optimized with **FIT+DMA** using the experimental molecular geometry. However, to be consistent with the methodology used for all other molecules, we do not consider this polymorph further in the current study. The results for monoclinic 3,4-cyclobutylfuran highlight a limitation of the single-point energy

correction approach, which relies on the force field to produce accurate structures. Thus, we plan to implement geometry optimization on the dimer-corrected lattice energy surface in future work.

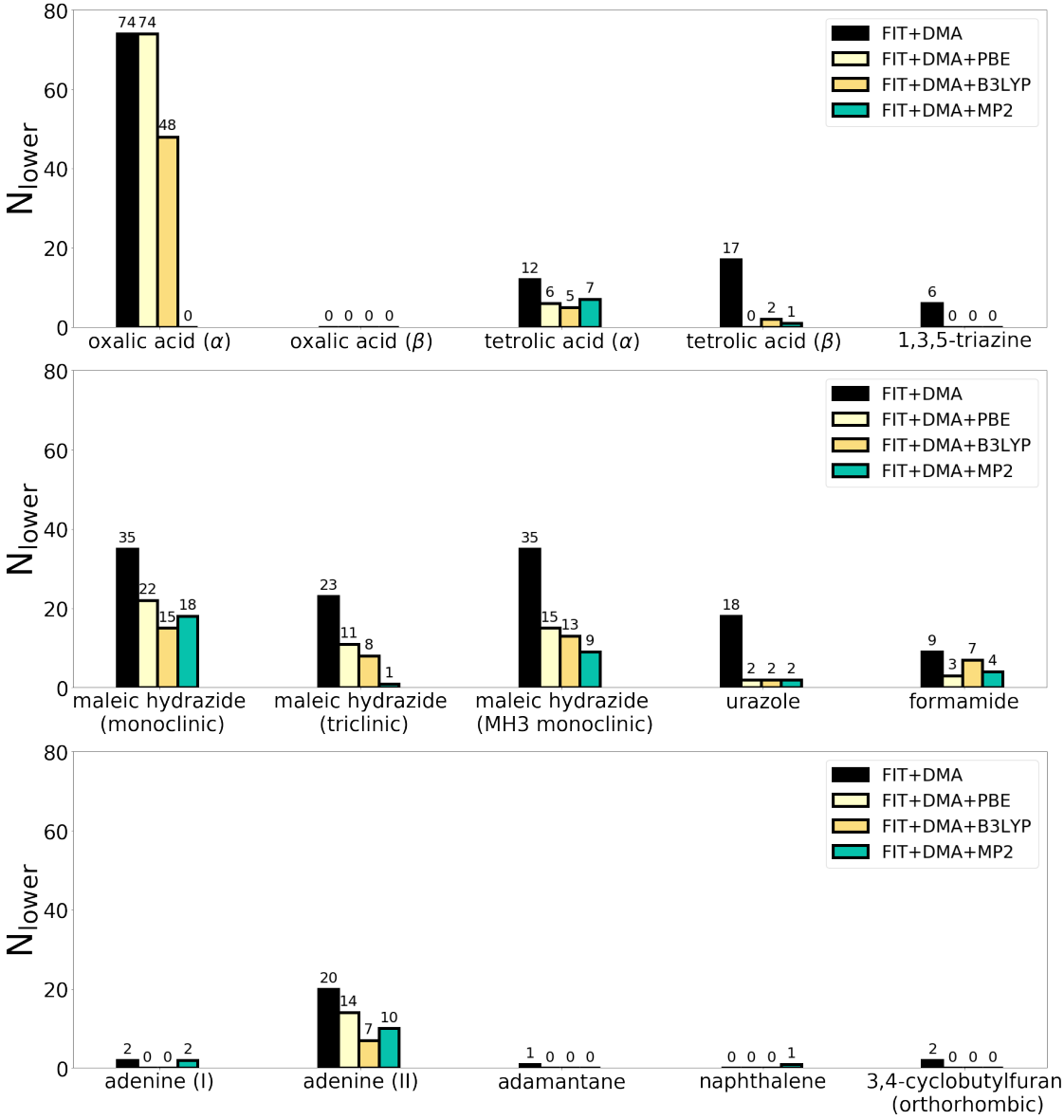


Figure 3: Differences in ranking (N_{lower}) of experimentally known structures in crystal structure landscapes calculated with **FIT+DMA** (black), **FIT+DMA+PBE** (tan), **FIT+DMA+B3LYP** (brown), and **FIT+DMA+MP2** (cyan).

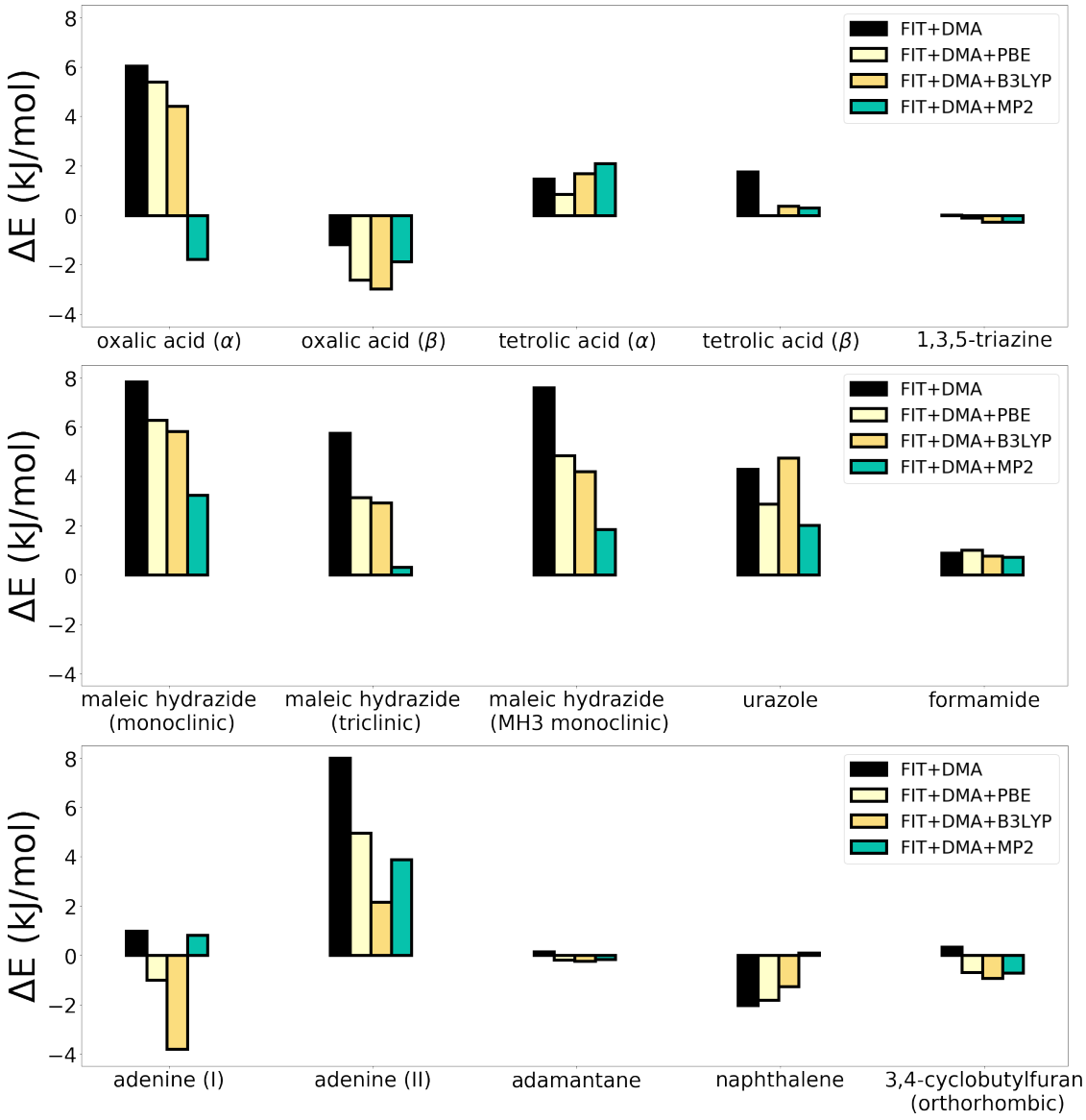


Figure 4: Comparisons of relative energies, ΔE, of known crystal structures to the lowest energy structure in the crystal structure landscape that is not experimentally observed using **FIT+DMA** (black), **FIT+DMA+PBE** (tan), **FIT+DMA+B3LYP** (brown), and **FIT+DMA+MP2** (cyan) energy models.

N_{lower} values for the other 15 observed crystal structures are shown in Fig. 3 for the four energy models - **FIT+DMA** and the three dimer-corrected models - with ΔE comparisons given in Fig. 4.

It should first be noted that the majority of molecules studied here were selected on the

basis that they are known to be challenging for force field methods. This is reflected in the large N_{lower} values for many of the structures at the force field (**FIT+DMA**) level: there are many energetically-competitive crystal structures in these landscapes that are not distinguished successfully from the experimentally known forms by the force field energy model. Thus, the **FIT+DMA** results shown here are not representative of typical results with empirically parameterized, atomic multipole-based force fields, but represent situations where improvements are needed beyond such models.

Across all ten molecules, both N_{lower} and ΔE are generally found to improve with all three QM dimer-corrected energy models, although there are some significant differences in the results, depending on the electronic structure method used in the dimer energy corrections.

We start with the model using the lowest cost QM dimer correction: the GGA DFT functional PBE. Using the **FIT+DMA+PBE** model, the ranking (N_{lower}) either remains the same (for the α form of oxalic acid) or improves for *all* experimentally observed crystal structures compared to **FIT+DMA** (Fig. 3). We also observe a comparative lowering of ΔE in every case aside from negligible increases in ΔE for formamide and naphthalene (Fig. 4). In some instances, this leads to experimental ranking much closer to the global minimum of the landscape (e.g. urazole, the two tetrolic acid polymorphs, and all three maleic hydrazide polymorphs). As importantly, for the systems where **FIT+DMA** already performs well (such as naphthalene, adamantane, and the β form of oxalic acid), the **FIT+DMA+PBE** model maintains the good energetic ranking and ΔE. Thus, this relatively inexpensive correction offers a clear improvement in the reliability of CSP for these molecules, by better distinguishing observed from unobserved crystal structures based on the calculated lattice energy.

The **FIT+DMA+B3LYP** model essentially amplifies the improvements found when using

FIT+DMA+PBE. Applying the hybrid B3LYP functional for dimer energy corrections leads to greater improvements in ranking compared to **FIT+DMA+PBE** in nearly every instance - only the β form of tetrolic acid and formamide have slightly worse N_{lower} values compared to **FIT+DMA+PBE**, but are still improvements over the uncorrected force field results. ΔE is also generally improved with **FIT+DMA+B3LYP** relative to **FIT+DMA+PBE**, again with a few exceptions (Fig. 4).

The **FIT+DMA+MP2** model continues the trend towards improving results from the **FIT+DMA** model. The ΔE results highlight cases where the **FIT+DMA+MP2** model out-performs the DFT-based models: for the three polymorphs of maleic hydrazide, urazole, and the α form of oxalic acid. ΔE using **FIT+DMA+MP2** has increased relative to the **FIT+DMA+B3LYP** results for both forms of adenine, but the N_{lower} results show that this is caused by a small number of predicted, but unobserved structures having low energies with **FIT+DMA+MP2**. The largest improvements that we observe between the wavefunction-based MP2 dimer-corrected model and the DFT-based models are in the ranking, N_{lower}, for a few of the molecules. The most dramatic change is for the α polymorph of oxalic acid, which remained high on the energy landscape with all other models, but with **FIT+DMA+MP2** is almost equi-energetic with β oxalic acid, as the two lowest energy crystal structures; this is in agreement with the experimentally determined relative stabilities for the polymorphs of oxalic acid, as discussed below. The changes in the overall distribution of crystal structures and position of the known polymorphs with all four energy models is shown for oxalic acid in Fig. 5. We also see important improvements for the triclinic and MH3 monoclinic forms of maleic hydrazide (Fig. 4).

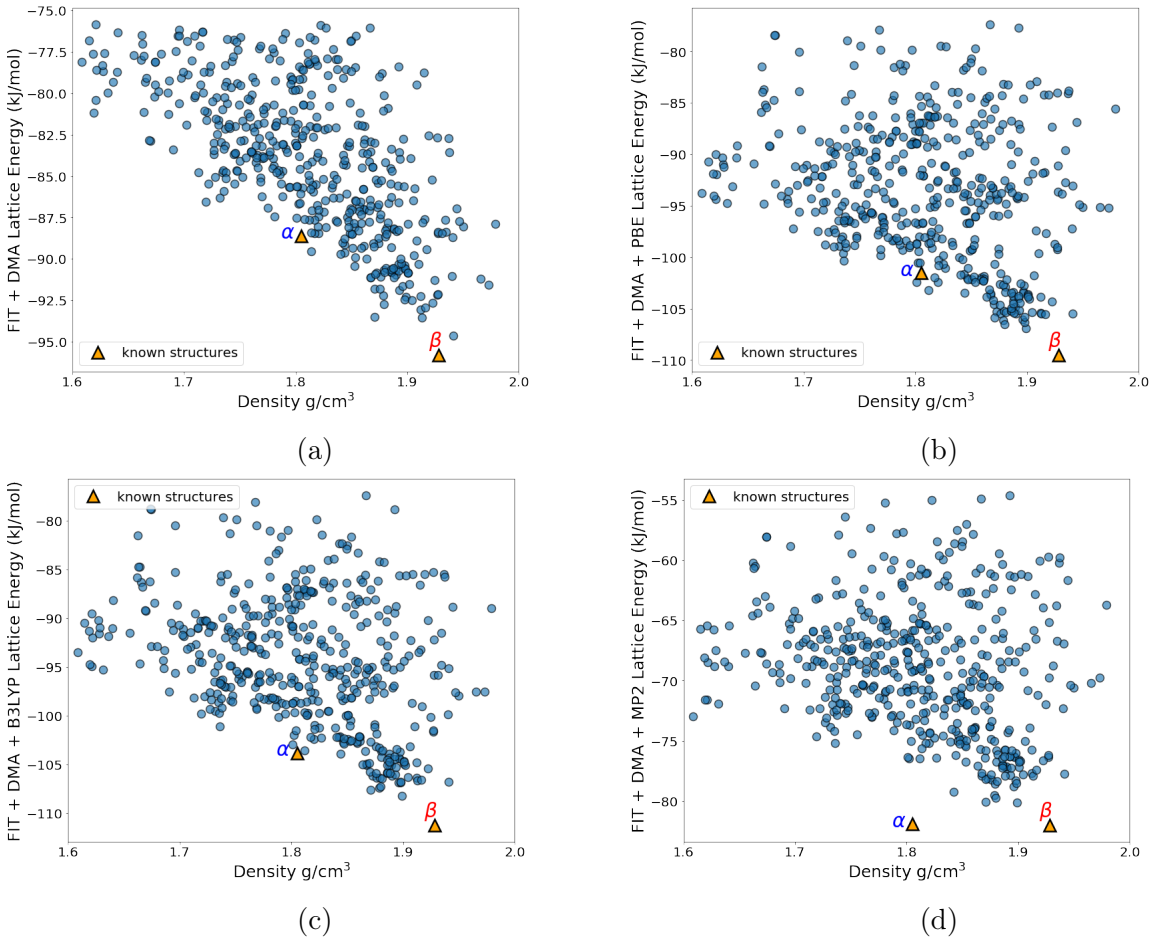


Figure 5: The low energy region of the crystal structure landscape of oxalic acid using (a) **FIT+DMA**, (b) **FIT+DMA+PBE**, (c) **FIT+DMA+B3LYP**, and (d) **FIT+DMA+MP2**. The lattice energy range shown in each part includes all of the structures taken from the lowest 20 kJ/mol of the **FIT+DMA** landscape. Each blue data point corresponds to a predicted crystal structure, with predicted crystal structures corresponding to the known polymorphs shown as orange triangles.

3.2 Polymorph Energy Differences: Comparison to Experiment

The overall results demonstrate that all three QM fragment-corrected methods improve upon the relative lattice energies calculated with the **FIT+DMA** force field. The MP2 dimer corrected model, **FIT+DMA+MP2**, provides the best overall results and, for some structures, shows important differences from the DFT fragment-corrected energy models. For the known polymorphs of oxalic acid, for example, both **FIT+DMA+PBE** and **FIT+DMA+B3LYP** predict the β form as the global minimum, whereas the α form is much higher in the land-

scape, while **FIT+DMA+MP2** predicts α and β as the two lowest energy structures (Fig. 5). The polymorphs of oxalic acid are known as a challenging system for force field methods,[92] partly due to the differences in hydrogen bonding in the two forms, with cyclic carboxylic acid dimer interactions in the β form (Fig. 6c), compared to corrugated layers in α, stabilized by a combination of short and long hydrogen bonds to both the carbonyl and O-H oxygen atoms (Fig. 6b). The differences between models can be further assessed by comparison to the polymorphic relative stabilities determined by experimental methods and solid state quantum mechanical calculations.

The measured relative stabilities of the oxalic acid polymorphs[93] show that α is the more stable form. After correction of the measured polymorph enthalpy difference for lattice vibrational contributions, the measured lattice energy difference is approximately 0.2 kJ/-mol.[44,94]. The correct stability order is only predicted correctly by the **FIT+DMA+MP2** model, which predicts the α form to be 0.02 kJ/mol more stable than β. This model also gives good agreement with fully relaxed solid state DFT models[95] (Fig. 6), despite only being a single point correction at the **FIT+DMA** geometry. To investigate the influence of the modest basis set size (6-31+G**) used in our calculations, the energies of the two oxalic acid polymorphs were also calculated using our fragment-corrected model with MP2 dimer interaction energies evaluated at the MP2/6-311++G** level of theory (shown in Fig. 6). The larger basis set was found to change the relative lattice energies slightly, increasing the lattice energy difference to 0.31 kJ/mol, now in excellent agreement with experiment.

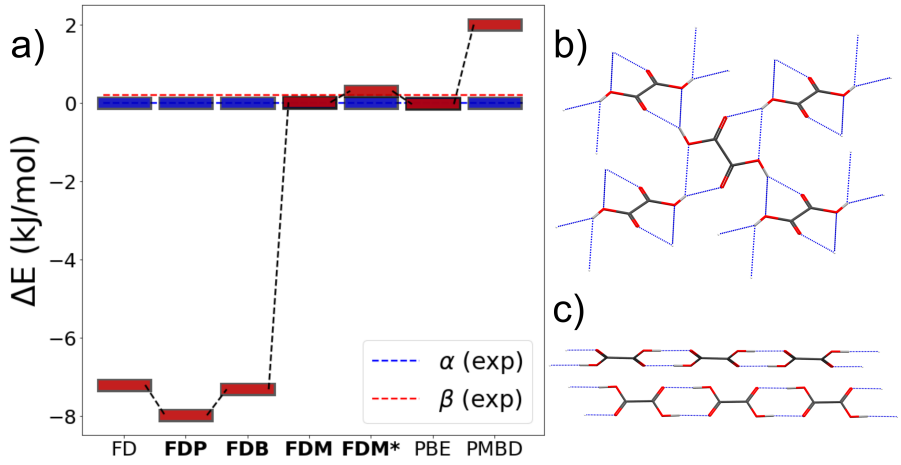


Figure 6: a) Relative lattice energies of the known b) α and c) β oxalic acid polymorphs calculated using the **FIT+DMA** (FD), **FIT+DMA+PBE** (FDP), **FIT+DMA+B3LYP** (FDB), **FIT+DMA+MP2** (FDM) models. Results are also shown using **FIT+DMA+MP2** with a 6-311++G** basis set (FDM*), PBE-D3 (PBE), and PBEh-MBD (PMBD) - a hybrid DFT functional combined with many-body dispersion (results from Marom et al. (2013)[95]). Lattice energies of the β form (red) are given relative to the α form (blue), with experimentally known stabilties[44,93] shown as the dashed red and blue lines. All corrections to **FIT+DMA** are single-point corrections at the **FIT+DMA** geometry (shown in bold), whereas PBE-D3 refers to full re-optimization from the **FIT+DMA** geometry. Hydrogen bonds in b) and c) are shown as dashed blue lines.

The tetrolic acid polymorphs also differ in hydrogen bonding, with the triclinic α polymorph forming cyclic hydrogen bond dimers and the monoclinic β polymorph forming hydrogen bond chains (Fig. 7b and c). The β polymorph is known to be the more stable form below $56-58$°C, but we are not aware of a measured energy difference between the polymorphs. As our calculations exclude the effects of temperature, the calculated lattice energy difference should agree with the low temperature order. The **FIT+DMA** force field gives the incorrect stability order, with β calculated to be 0.29 kJ/mol less stable than α. In this case, all of the fragment-corrected models correct the stability order (Fig. 7), predicting β to be more stable than α by between 0.82 kJ/mol (**FIT+DMA+PBE**) and 1.84 kJ/mol (**FIT+DMA+MP2**). These results are in good agreement with the relative energy calculated using the popular solid state PBE-D3 method, which calculates the α form to be 0.93 kJ/mol less stable than β, but much smaller than the energy difference calculated

with PBEh-MBD[95]. Calculations for **FIT+DMA+MP2** were again repeated with a larger basis set for the MP2 dimer interaction energies, where only minor changes in energy were observed.

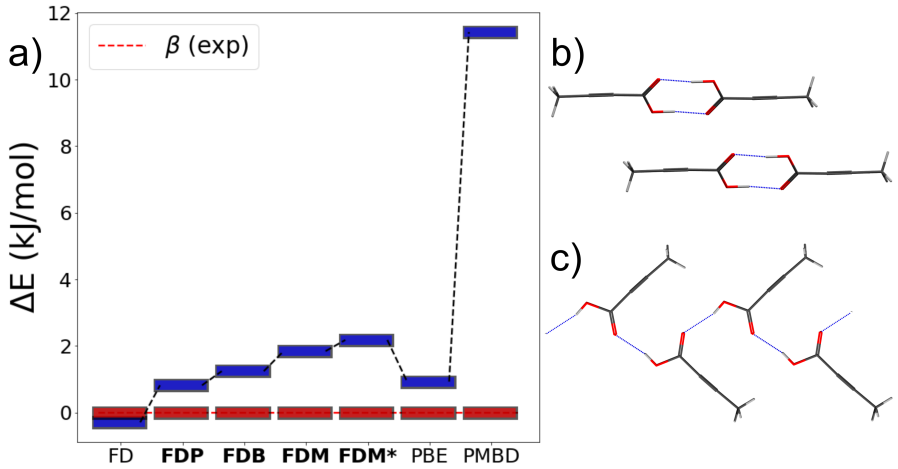


Figure 7: a) Relative lattice energies of the known b) α and c) β tetrolic acid polymorphs calculated using the **FIT+DMA** (FD), **FIT+DMA+PBE** (FDP), **FIT+DMA+B3LYP** (FDB), **FIT+DMA+MP2** (FDM) models. Results are also shown using **FIT+DMA+MP2** with a 6-311++G** basis set (FDM*), PBE-D3 (PBE), and PBEh-MBD (PMBD) - a hybrid DFT functional combined with many-body dispersion (results from Marom et al. (2013)[95]). The β form is known to be more stable than α.[55] Lattice energies calculated for the α form (blue) are given relative to the β form (red). All corrections to **FIT+DMA** are single-point corrections at the **FIT+DMA** geometry (shown in bold), whereas PBE-D3 refers to full re-optimization from the **FIT+DMA** geometry. Hydrogen bonds in b) and c) are shown as dashed blue lines.

The most notable exception to the trend of **FIT+DMA+MP2** providing the best results is for adenine, whose two polymorphs form similar hydrogen bonded sheet structures that differ in the arrangement of some of the hydrogen bonds (Fig. 8). Both forms have a significantly improved (i.e. lowered) ΔE value in the **FIT+DMA+B3LYP** model, with small improvements in ranking, compared to **FIT+DMA+MP2** (Fig. 4). However, estimations of relative polymorph free energies based on solubility data[49] indicate that form I is 1.1 ± 0.3 kJ/mol more stable than form II at room temperature, which is reproduced most closely by the **FIT+DMA+MP2** lattice energies (Fig. 8). All methods are found to overestimate the relative stability of form I, including fully optimized PBE-D3. The comparison between

the experimental free energy difference and calculated lattice energy difference ignores the lattice vibrational contribution to the free energy difference between polymorphs, but this was found to be only about 0.5 kJ/mol for adenine[49], so would not change these conclusions.

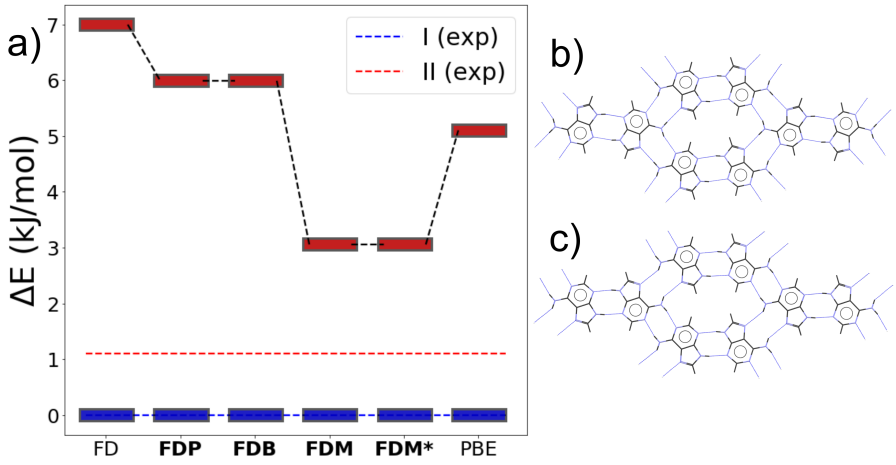


Figure 8: a) Relative lattice energies of the known b) polymorph I and c) polymorph II of adenine calculated using the **FIT+DMA** (FD), **FIT+DMA+PBE** (FDP), **FIT+DMA+B3LYP** (FDB), **FIT+DMA+MP2** (FDM) models. Results are also shown using **FIT+DMA+MP2** with a 6-311++G** basis set (FDM*) and PBE-D3 (PBE). Lattice energies of form II (red) are given relative to the form I (blue), with the experimentally determined free energy difference[49] shown as the dashed red and blue lines. All corrections to **FIT+DMA** are single-point corrections at the **FIT+DMA** geometry (shown in bold), whereas PBE-D3 refers to full re-optimization from the **FIT+DMA** geometry. All methods are found to overestimate the relative stability of form I, with **FIT+DMA+MP2** showing the best agreement with experiment. Hydrogen bonds in b) and c) are shown as dashed blue lines.

The comparisons to experimental polymorph relative stability data indicate that, despite some small exceptions when using N_{lower} and ΔE as metrics, the MP2 corrections provide the most reliable improvement to the lattice energies of force field crystal structure landscapes. Thus, **FIT+DMA+MP2** is our preferred method. In all three cases, using a larger basis for the MP2 dimer interaction energies leads to only small changes in relative lattice energies, which suggests that the 6-31+G** basis set is sufficiently large for the calculation of relative lattice energies.

4 Gaussian Process Learning of Dimer Energy Corrections

Our results have demonstrated that a single-point QM fragment-based correction leads to improved energetic ranking of observed crystal structures within CSP landscapes. The additional computational cost associated with this correction depends on the QM method used (MP2 is the most promising of those tested thus far), as well as the molecular size and number of dimers required to correct the lattice energy for all low energy crystal structures on the landscape. The number of crystal structures and dimer interaction energies required to correct the lowest 20 kJ/mol of each landscape is listed in Table 1; the number of dimers varies widely between molecules, largely due to the differences in numbers of low energy crystal structures on their landscapes. Average converged distances are found to be similar for most species, indicating that the minimum cutoff is perhaps overly conservative in many cases (a violin plot in the ESI highlights the differences between species in more detail). A clear exception to this is formamide, which has a far greater spread of convergence values, probably owing to its large dipole moment and resulting strong, long-range intermolecular interactions.

Table 1: The number of dimer calculations required to convert the lowest 20 kJ/mol of each **FIT+DMA** crystal structure landscape to a QM fragment-corrected model. Maximum and average convergence distances for the set of CSP structures are given in Å.

Molecule	No. Crystal Structures	No. Unique Dimers	Max (Avg.) Converged Distances
oxalic acid	476	16,379	9.2 (8.1)
adenine	86	2690	8.8 (8.2)
formamide	499	27,498	13.2 (9.0)
maleic hydrazide	388	12,522	9.3 (8.2)
3,4-cyclobutylfuran	765	22,466	8.9 (8.1)
tetrolic acid	568	19,114	8.9 (8.1)
adamantane	146	3,431	9.0 (8.3)
triazine	270	7,645	8.9 (8.2)
naphthalene	451	11,102	9.0 (8.2)
urazole	468	16,797	9.1 (8.2)

A comparison of the cost associated with dimer calculations compared to periodic PBE-D3 single-point calculations (see Table S3 in the ESI) shows that the relative cost varies considerably between structures and with the level of theory used for the dimer calculations. For example, the **FIT+DMA+PBE** dimer calculations require more CPU time than periodic PBE-D3 for the oxalic acid and tetrolic acid crystal structures, but less for the larger adenine structures. In practice, it is the inherent parallelizability of the fragment-based approach provides an advantage over fully periodic QM calculations in terms of computing time because each dimer interaction energy can be assessed simultaneously. In this work, dimer calculations were typically run over 160-240 processors, running each dimer calculation across two processors, reducing the computation time by over two orders of magnitude. Despite the advantage of nearly perfect parallelizability, the cost of the fragment-corrected calculations is large compared to the force field energy evaluations and could become prohibitive if we wanted to optimize crystal structures on the fragment-corrected lattice energy surface, which would require the QM dimer interaction energy calculations at each optimization step. Furthermore, the time taken for each calculation could become problematic for larger species, especially for MP2 or higher levels of QM theory. The data-rich nature

of the problem makes machine learning an attractive approach to lower the computational expense. Thus, we have investigated using a machine learning model to learn the dimer energy corrections - the difference between force field and QM interaction energies - as a function of the dimer geometry. We use the oxalic acid and maleic hydrazide landscapes to evaluate this approach, as these show the largest ranking and relative energy changes between **FIT+DMA** and **FIT+DMA+MP2** models, so should be the most challenging targets for machine learning the energy corrections.

Gaussian Process (GP) models were fitted separately for each molecule, to a training set of dimers from the low energy predicted crystal structures, selected to be maximally separated in descriptor space, using atomic symmetry functions as structural descriptors of the dimers and the max-min algorithm[89] for selection of the dimers included in the training set (for which MP2 calculations are performed). Our previous work[27] demonstrated that such an approach to training point selection yields improved results compared to random selection of training data. As the purpose of the GP model is to reduce the computation time associated with the QM dimer energy correction, only training fractions of 40% and lower were considered when training the models; we refer to the resulting models as **FIT+DMA+GP** hereafter.

Mean signed error (MSE) and mean absolute error (MAE) in the **FIT+DMA+GP** model, relative to **FIT+DMA+MP2**, are shown for the dimer energy corrections and total lattice energies for both crystal structure landscapes in Fig. 9. Here the lattice energy is calculated as before using equation (3), where the energy correction ($E_{cm,i}^{(MP2)} - E_{cm,i}^{(FIT+DMA)}$) for training set dimers are added explicitly, and dimer energy differences from all other (test set) dimers are replaced with the predicted energy differences from the machine learning model.

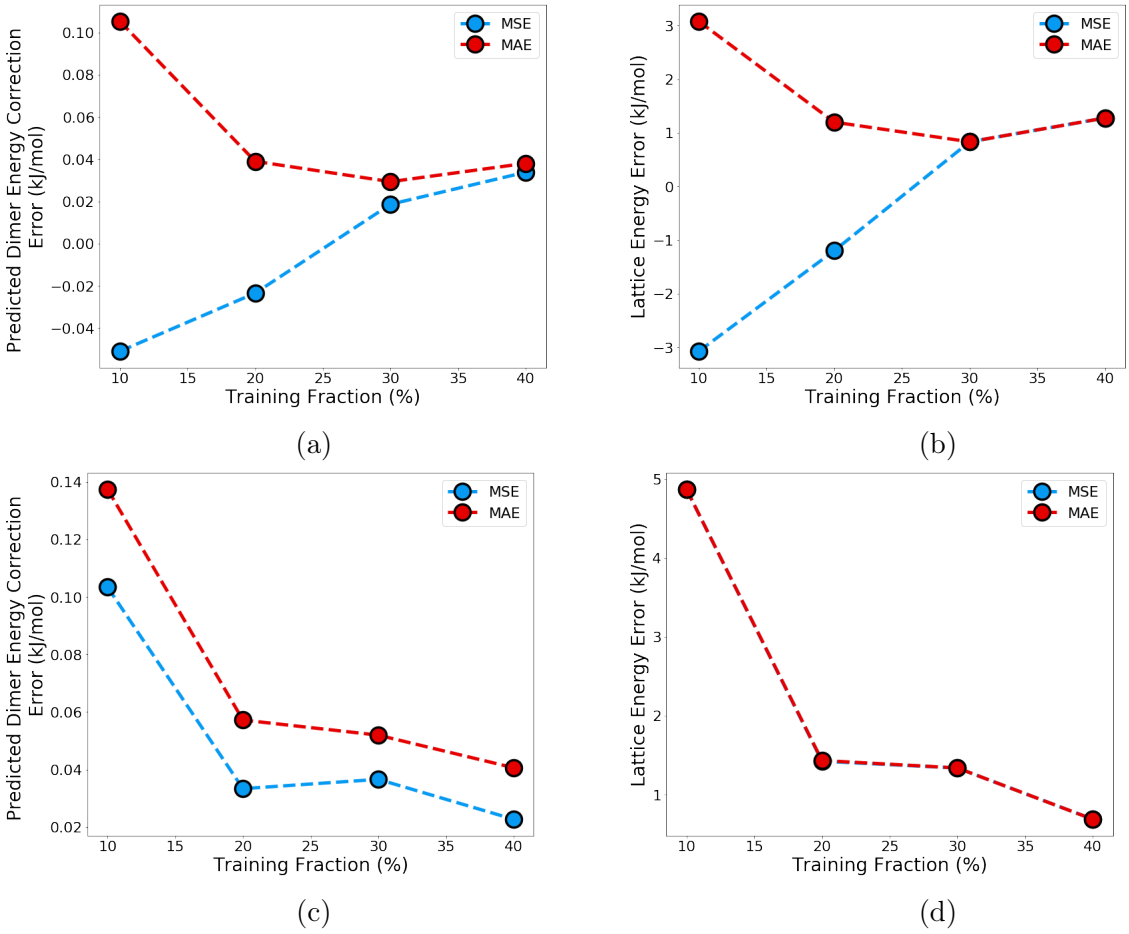


Figure 9: MSE (blue) and MAE (red) at different training fractions for the Gaussian Process model relative to explicit calculations of interaction energy differences between MP2 and FIT+DMA for (a) all oxalic acid dimers not used to train the Gaussian process, (b) the resulting errors in the total oxalic acid lattice energies (relative to **FIT+DMA+MP2**), (c) all maleic hydrazide dimers not used to train the Gaussian Process, (d) the resulting errors in the total maleic hydrazide lattice energies (relative to **FIT+DMA+MP2**) - in this case the MSE and MAE are very similar and overlap.

Errors in the dimer energy corrections (the difference between MP2 and FIT+DMA) show the largest decrease from 10% to 20% training fractions, at which point MAE errors are approximately 0.04 and 0.06 kJ/mol for oxalic acid and maleic hydrazide dimers, respectively. The errors continue to decrease, albeit slowly, at higher training fractions, apart from a slight increase for oxalic acid between 30% and 40%. These dimer correction errors are very small, even at a 10% training fraction, considering the total energetic range of the dimer error corrections, which span the range -0.71 to +4.58 kJ/mol for oxalic acid and -4.06 to +8.50 kJ/mol for maleic hydrazide. The dimer energy prediction errors translate directly to errors in the lattice energies (measured relative to the fully corrected **FIT+DMA+MP2** landscape); for both molecules, we see that lattice energy errors are just over 1 kJ/mol at 20% training fractions (Fig. 9). Thus, the magnitude of the errors introduced by machine learning the dimer energy corrections are comparable to typical energy differences between polymorphs[12] and to the energy differences between low energy predicted structures on the oxalic acid landscape (see Fig. 5). However, the MSE shows that the errors in dimer energy correction, and hence in the total lattice energies, are largely systematic; the systematic nature of the errors is also clear in scatter plots of predicted vs calculated energy corrections (shown in the ESI).

The excellent performance of the Gaussian process predictions, and the resulting systematic errors in the predicted lattice energies, are partly due to the selection of training data by the max-min algorithm. The atomic symmetry function descriptors decay at longer distances, and so the greatest diversity in descriptor space is found at short intermolecular separations, where the differences between MP2 and FIT+DMA energies are greatest. The max-min algorithm iteratively selects for dimers distinct to those already in the training set, and so at low training fractions those dimers with significant energy differences (and short intermolecular separations) will form the majority of the training data (figures showing the selected dimers at different training fractions are given in the ESI). The low number of dimers with

long intermolecular separations selected by the max-min algorithm at small training fractions result in many of these dimers in the test set being predicted with similar energy differences. These contributions are small, but the large number of dimers at longer separations adds up to a systematic shift of the lattice energies. These systematic errors cancel in the calculation of relative lattice energies and, so, are unimportant for the ranking of crystal structures. We also note that, at higher training fractions, the selection of training data continues to be iteratively selected by intermolecular distance (shown in the ESI), indicating that the resolution in the descriptor for these highly separated dimers is still sufficient to distinguish them.

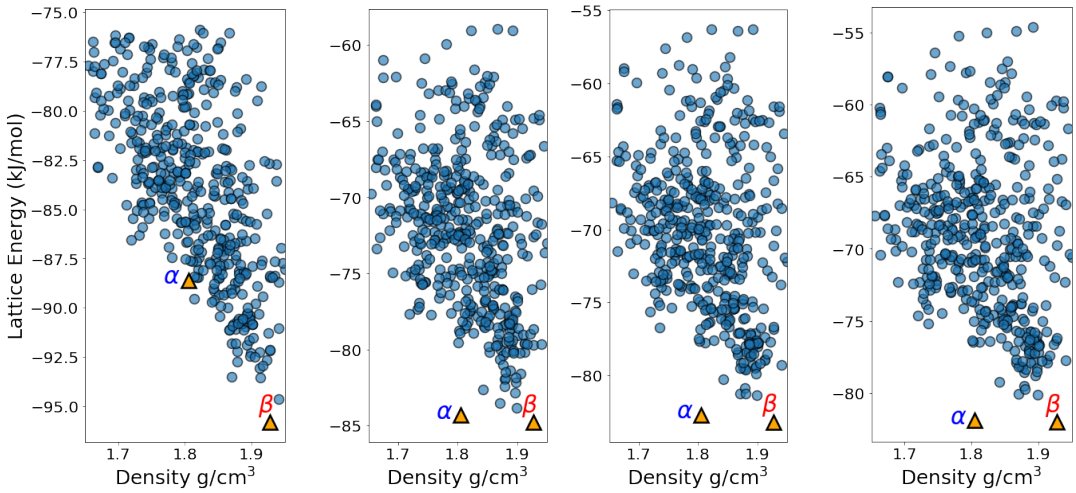
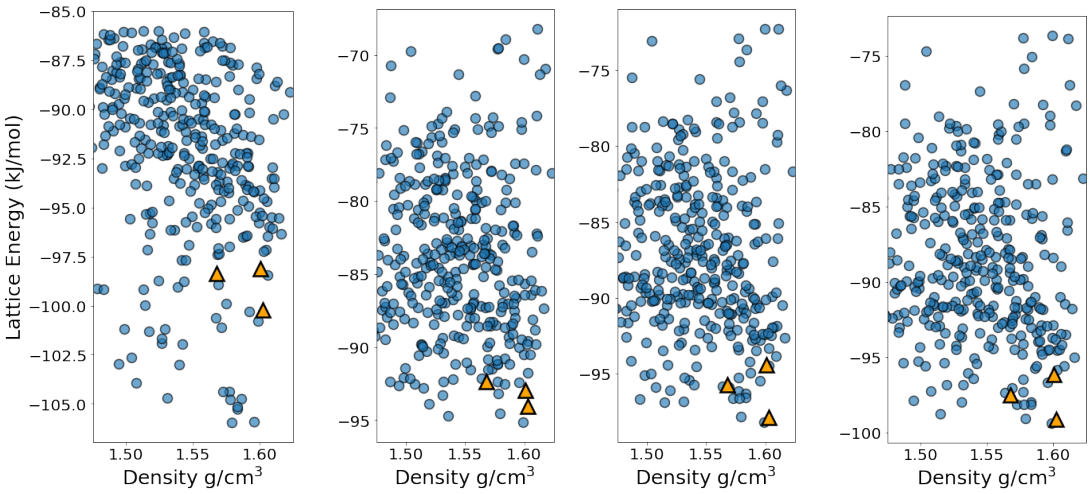


Figure 10: Lattice energy vs density for predicted crystal structures of oxalic acid using (a) the **FIT+DMA** force field, (b) the **FIT+DMA+GP** machine learnt model using 10% training data, (c) the **FIT+DMA+GP** model using 20% training data, and (d) the **FIT+DMA+MP2** (target) landscape. In (b) 87% of the dimers in the α polymorph and 90% in β polymorph are predicted using the **FIT+DMA+GP** model, whereas in (c) 77% and 74% are predicted for the α and β polymorphs, respectively (as chosen deterministically by the max-min algorithm[89]).

The performance of the machine learnt models can be seen more clearly in the resulting lattice energy landscapes, which are presented for oxalic acid from the **FIT+DMA+GP** models trained with 10% and 20% these training fractions in Figs 10b and 10c. Comparison to the target **FIT+DMA+MP2** landscape (Fig. 10d) shows that the model trained

with 10% of the dimers introduces the most important corrections from the fragment-based model and that the ranking of low energy structures has largely been resolved once 20% of dimers are used to train the GP model. The Kendall rank correlation of lattice energies from **FIT+DMA+GP** models vs **FIT+DMA+MP2** increases from $\tau = 0.870$ (10% training) to 0.956 (20% training), compared to $\tau = 0.606$ for the comparison between uncorrected **FIT+DMA** and **FIT+DMA+MP2** rankings. This demonstrates that the ranking of predicted structures is almost unaffected by the errors in the **FIT+DMA+GP** model trained with 20% of dimers. The rank correlation makes only small further increases at higher training set sizes (shown in the ESI). Training fractions of 10% and 20% reduce the number of required MP2 calculations from 16,379 for the full **FIT+DMA+MP2** model to 1,678 and 3,276, respectively. This 80 to 90% saving in time required for the MP2 correction will be increased if we also use the GP model for lattice energy optimization.

The lattice energy landscape for maleic hydrazide makes a similar progression (Fig. 11), but here the differences between **FIT+DMA+GP** trained with 10% of dimers and the **FIT+DMA+MP2** landscape are more important. This is due to higher errors for maleic hydrazide than oxalic acid at small training fractions; these can be explained by the smaller dataset of 12,522 unique dimers compared to 16,379 in the oxalic acid set (table 1) and the larger range of dimer energy corrections that must be learnt for maleic hydrazide; this reflects a greater range of dimer geometries for maleic hydrazide, whose hydrogen bond donors and acceptors can be combined in a range of motifs. It should also be noted that the descriptor space for maleic hydrazide has considerably more dimensions than oxalic acid.



(a) (b) (c) (d)

Figure 11: Lattice energy vs density for the predicted crystal structures of maleic hydrazide using (a) the **FIT+DMA** force field, (b) the **FIT+DMA+GP** machine learnt model using 10% training data, (c) the **FIT+DMA+GP** landscape using 20% training data, and (d) the **FIT+DMA+MP2** (target) landscape, where the known structures marked as orange triangles are (from left to right in terms of density) the MH3 monoclinic form, the monoclinic form, and the triclinic form. In (b) 96%, 98%, and 96% of the dimers in the known polymorphs are predicted for the MH3 monoclinic, monoclinic, and triclinic forms, respectively. In (c) 80%, 86%, and 74% are predicted for the three polymorphs, respectively (again as chosen deterministically by the max-min algorithm[89]).

The rank correlation of crystal structures between the **FIT+DMA+GP** and **FIT+DMA +MP2** models is only slightly lower than for oxalic acid ($\tau = 0.853$ for the GP model trained with 10% of dimers, $\tau = 0.943$ at 20%, compared to $\tau = 0.501$ comparing **FIT+DMA** to **FIT+DMA+MP2**). This demonstrates that the ranking of most predicted structures is correct at the lowest training fraction. However, it is clear (see Fig. 11b) that, using the GP model trained with 10% of dimers, the relative energies of the known polymorphs is incorrect. This may be because almost all of the dimer energy corrections for the three observed polymorphs have been predicted by the GP model (i.e. are not in the training set). In particular, the stability of the monoclinic polymorph, which is least stable of the three observed forms using **FIT+DMA+MP2**, is exaggerated. For this structure, only 2% of dimers are in the 10% training set, so errors related to the GP model will be higher for this structure than the average. Comparatively, in the oxalic acid 10% landscape (Fig. 10b),

both known polymorphs have a smaller fractions of 87% and 90% of the dimers predicted by the **FIT+DMA+GP** model for α and β, respectively.

The landscape predicted using the **FIT+DMA+GP** model using 20% of dimers as training data (Fig. 11c) gives a faithful reproduction of the **FIT+DMA+MP2** landscape, including the relative stabilities of the three observed crystal structures. Only the absolute lattice energies are slightly underestimated with this model, for which only 2,504 of the 12,522 unique dimers have been calculated using MP2. As with oxalic acid, the systematic error is absent at higher training fractions. In both cases the prediction errors with the **FIT+DMA+GP** models are comparable to previous work on machine learning lattice energies for CSP [27] and highlight that the target landscape can be accurately resolved at a significant reduction in computational cost.

5 Conclusions

Fragment-based models are known to provide complementary approaches to periodic calculations, which can be essential when the computational cost of hybrid functionals or post-Hartree Fock, wavefunction methods become computationally unfeasible, and GGA methods are not sufficiently accurate. This work has demonstrated an approach to rapidly improve anisotropic, atomic multipole-based force field lattice energies by correcting two-body interactions to a quantum chemical level of theory, and the application of this approach to structure prediction of organic molecular crystals. By replacing significant force field two-body interactions with PBE-D3, B3LYP-D3, or MP2 dimer interaction energies, without further geometry optimization, we observe improvements to the ranking of experimentally known crystal structures within the sets of predicted structures in almost all cases, and good ranking of experimental structures is maintained in cases where the force field performs well. Results systematically improve when replacing two-body interactions with a GGA method

to that with a hybrid functional, and the largest improvements are observed when MP2 is used for the dimer energy corrections. The MP2 fragment-corrected model also provides polymorph lattice energy differences that are in excellent agreement with experimental measurements, and with the results of fully periodic, dispersion-corrected DFT methods. Thus, the QM fragment-corrected force field lattice energy models will increase our confidence in the energetic ordering of predicted crystal structures, which will improve the reliability of crystal structure prediction methods in applications such as pharmaceutical polymorph screening and computer-guided materials discovery. Furthermore, the approach is generalizable to any quantum chemical method, such as higher level correlated wavefunction methods and larger basis sets, potentially offering further improvements than those demonstrated here.

An attractive feature of the fragment-based approach is the inherent parallelizibility of the energy correction. Furthermore, we have shown that the difference between force field and QM dimer interaction energies can be predicted using a Gaussian process, with the dimer geometries described using atomic symmetry functions. Results for two of the most challenging molecules in our study - where differences between force field and MP2 relative energies are greatest - show that lattice energies of the MP2 dimer-corrected model can be predicted to within 1-1.5 kJ/mol by training the Gaussian process model on 20% of the dimer energies from a CSP landscape, and that the resulting machine learnt model reproduces the ranking of structures on the fully MP2-corrected landscape. Thus, a 5-fold reduction in the computational cost of the QM correction can be achieved using a simple machine learning approach.

The machine learning approach helps minimize the additional computational expense involved with the QM energy corrections, which is particularly important for future applications to larger molecules. However, the conformational flexibility of larger molecules will introduce additional challenges that have not been addressed here: all molecules studied in this work have been modelled using their gas phase geometries, which are undistorted by

crystal packing forces. For species where this is not a realistic approach, the machine learning will need to be extended to include some or all intramolecular degrees of freedom, either within the same Gaussian Process and within the same descriptor used for the dimer energy corrections or through an additional intramolecular energy model. Conformational flexibility will also significantly increase the number and variety of possible two-body interactions, likely presenting a greater challenge to achieve similar levels of prediction accuracy. The work presented here provides a basis for these developments and an efficient, accurate, and systematically improvable alternative to the increasingly popular periodic DFT approach to crystal structure prediction.

6 Acknowledgments

We are grateful for support from the EPSRC Centre for Doctoral training, Theory and Modelling in Chemical Sciences, under grant EP/L015722/1. We acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton. We also acknowledge the ARCHER UK National Supercomputing Service which was accessed via the UK's HPC Materials Chemistry Consortium membership, which is funded by the EPSRC (EP/L000202). Finally, we are grateful to the UK Materials and Molecular Modelling Hub for computational resources, which is partially funded by EPSRC (EP/P020194/1)

All data supporting this study are openly available from the University of Southampton repository at https://doi.org/10.5258/SOTON/D0814

SI Paragraph: The SI includes details on the selection of molecules, tables of N_{lower} and ΔE for each structure, figures showing convergence of the fragment correction, dimer and total lattice energy correlation plots from the machine learning, further details on the training

data selection for machine learning, plots of the rank correlation coefficient and comparison of computational costs. This information is available free of charge via the Internet at http://pubs.acs.org

References

(1) Price, S. L. Computed Crystal Energy Landscapes for Understanding and Predicting Organic Crystal Structures and Polymorphism. *Acc. Chem. Res.* **2009**, *42*, 117–126.

(2) Day, G. M. Current Approaches to Predicting Molecular Organic Crystal Structures. *Crystallogr. Rev.* **2011**, *17*, 3–52.

(3) Perrin, M.-A.; Neumann, M. A.; Elmaleh, H.; Zaske, L. Crystal Structure Determination of the Elusive Paracetamol Form III. *Chem. Commun.* **2009**, 3181–3183.

(4) Eddleston, M. D.; Hejczyk, K. E.; Bithell, E. G.; Day, G. M.; Jones, W. Determination of the Crystal Structure of a New Polymorph of Theophylline. *Chem. Eur. J.* **2013**, *19*, 7883–7888.

(5) Baias, M.; Dumez, J.-N.; Svensson, P. H.; Schantz, S.; Day, G. M.; Emsley, L. De Novo Determination of the Crystal Structure of a Large Drug Molecule by Crystal Structure Prediction-Based Powder NMR Crystallography. *J. Am. Chem. Soc.* **2013**, *135*, 17501–17507.

(6) Chemburkar, S. R.; Bauer, J.; Deming, K.; Spiwek, H.; Patel, K.; Morris, J.; Henry, R.; Spanton, S.; Dziki, W.; Porter, W.; Quick, J.; Bauer, P.; Donaubauer, J.; Narayanan, B. A.; Soldani, M.; Riley, D.; McFarland, K. Dealing with the Impact of Ritonavir Polymorphs on the Late Stages of Bulk Drug Process Development. *Org. Process Res. Dev.* **2000**, *4*, 413–417.

(7) Cabri, W.; Ghetti, P.; Pozzi, G.; Alpegiani, M. Polymorphisms and Patent, Market, and Legal Battles: Cefdinir Case Study. *Org. Process Res. Dev* **2007**, *11*, 64–72.

(8) Pulido, A.; Chen, L.; Kaczorowski, T.; Holden, D.; Little, M. A.; Chong, S. Y.; Slater, B. J.; McMahon, D. P.; Bonillo, B.; Stackhouse, C. J.; Stephenson, A.; Kane, C. M.; Clowes, R.; Hasell, T.; Cooper, A. I.; Day, G. M. Functional Materials Discovery Using Energy-Structure-Function Maps. *Nature* **2017**, *543*, 657–664.

(9) Campbell, J. E.; Yang, J.; Day, G. M. Predicted Energy-Structure-Function Maps for the Evaluation of Small Molecule Organic Semiconductors. *J. Mater. Chem. C* **2017**, *5*, 7574–7584.

(10) Yang, J.; De, S.; Campbell, J. E.; Li, S.; Ceriotti, M.; Day, G. M. Large-Scale Computational Screening of Molecular Organic Semiconductors Using Crystal Structure Prediction. *Chem. Mater.* **2018**, *30*, 4361–4371.

(11) Case, D. H.; Campbell, J. E.; Bygrave, P. J.; Day, G. M. Convergence Properties of Crystal Structure Prediction by Quasi-Random Sampling. *J. Chem. Theory Comput.* **2016**, *12*, 910–924.

(12) Nyman, J.; Day, G. M. Static and Lattice Vibrational Energy Differences Between Polymorphs. *CrystEngComm* **2015**, *17*, 5154–5165.

(13) Reilly, A. M.; Cooper, R. I.; Adjiman, C. S.; Bhattacharya, S.; Boese, A. D.; Brandenburg, J. G.; Bygrave, P. J.; Bylsma, R.; Campbell, J. E.; Car, R.; Case, D. H.; Chadha, R.; Cole, J. C.; Cosburn, K.; Cuppen, H. M.; Curtis, F.; Day, G. M.; DiStasio, R. A.; Dzyabchenko, A.; Van Eijck, B. P.; Elking, D. M.; Van Den Ende, J. A.; Facelli, J. C.; Ferraro, M. B.; Fusti-Molnar, L.; Gatsiou, C. A.; Gee, T. S.; De Gelder, R.; Ghiringhelli, L. M.; Goto, H.; Grimme, S.; Guo, R.; Hofmann, D. W.; Hoja, J.; Hylton, R. K.; Iuzzolino, L.; Jankiewicz, W.; De Jong, D. T.; Kendrick, J.; De Klerk, N. J.; Ko, H. Y.; Kuleshova, L. N.; Li, X.; Lohani, S.; Leusen, F. J.; Lund, A. M.; Lv, J.;

Ma, Y.; Marom, N.; Masunov, A. E.; McCabe, P.; McMahon, D. P.; Meekes, H.; Metz, M. P.; Misquitta, A. J.; Mohamed, S.; Monserrat, B.; Needs, R. J.; Neumann, M. A.; Nyman, J.; Obata, S.; Oberhofer, H.; Oganov, A. R.; Orendt, A. M.; Pagola, G. I.; Pantelides, C. C.; Pickard, C. J.; Podeszwa, R.; Price, L. S.; Price, S. L.; Pulido, A.; Read, M. G.; Reuter, K.; Schneider, E.; Schober, C.; Shields, G. P.; Singh, P.; Sugden, I. J.; Szalewicz, K.; Taylor, C. R.; Tkatchenko, A.; Tuckerman, M. E.; Vacarro, F.; Vasileiadis, M.; Vazquez-Mayagoitia, A.; Vogt, L.; Wang, Y.; Watson, R. E.; De Wijs, G. A.; Yang, J.; Zhu, Q.; Groom, C. R. Report on the Sixth Blind Test of Organic Crystal Structure Prediction Methods. *Acta Crystallogr., Sect. B: Struct. Sci.* **2016**, *72*, 439–459.

(14) Coombes, D. S.; Price, S. L.; Willock, D. J.; Leslie, M. Role of Electrostatic Interactions in Determining the Crystal Structures of Polar Organic Molecules. A distributed Multipole Study. *J. Phys. Chem.* **1996**, *100*, 7352–7360.

(15) Price, S. L.; Leslie, M.; Welch, G. W.; Habgood, M.; Price, L. S.; Karamertzanis, P. G.; Day, G. M. Modelling organic crystal structures using distributed multipole and polarizability-based model intermolecular potentials. *Phys. Chem. Chem. Phys.* **2010**, *12*, 8478–8490.

(16) Neumann, M. A.; Perrin, M.-A. Energy Ranking of Molecular Crystals Using Density Functional Theory Calculations and an Empirical van der Waals Correction. *J. Phys. Chem. B* **2005**, *109*, 15531–15541.

(17) Nyman, J.; Pundyke, O. S.; Day, G. M. Accurate Force Fields and Methods for Modelling Organic Molecular Crystals at Finite Temperatures. *Phys. Chem. Chem. Phys.* **2016**, *18*, 15828–15837.

(18) Handley, C. M.; Popelier, P. L. Potential Energy Surfaces Fitted by Artificial Neural Networks. *J. Phys. Chem. A* **2010**, *114*, 3371–3383.

(19) Behler, J. Representing Potential Energy Surfaces by High-Dimensional Neural Network potentials. *J. Phys. Condens. Matter* **2014**, *26*, 183001.

(20) Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep Learning for Computational Chemistry. *J. Comput. Chem.* **2017**, *38*, 1291–1307.

(21) Behler, J. Constructing High-Dimensional Neural Network Potentials: A Tutorial Review. *Int. J. Quantum Chem.* **2015**, *115*, 1032–1050.

(22) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8*, 3192–3203.

(23) Shen, L.; Wu, J.; Yang, W. Multiscale Quantum Mechanics/Molecular Mechanics Simulations with Neural Networks. *J. Chem. Theory Comput.* **2016**, *12*, 4934–4946.

(24) Gao, T.; Li, H.; Li, W.; Li, L.; Fang, C.; Li, H.; Hu, L.; Lu, Y.; Su, Z. M. A Machine Learning Correction for DFT Non-Covalent Interactions Based on the S22, S66 and X40 Benchmark Databases. *J. Cheminformatics* **2016**, *8*, 24.

(25) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O. A.; Müller, K. R.; Tkatchenko, A. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326–2331.

(26) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*.

(27) Musil, F.; De, S.; Yang, J.; Campbell, J. E.; Day, G. M.; Ceriotti, M. Machine Learning for the Structure-Energy-Property Landscapes of Molecular Crystals. *Chem. Sci.* **2018**, *9*, 1289–1300.

(28) Tong, Q.; Xue, L.; Lv, J.; Wang, Y.; Ma, Y. Accelerating CALYPSO Structure Prediction by Data-Driven Learning of a Potential Energy Surface. *Faraday Discuss.* **2018**, *211*, 31–43.

(29) Deringer, V. L.; Proserpio, D. M.; Csanyi, G.; Pickard, C. J. Data-Driven Learning and Prediction of Inorganic Crystal Structures. *Faraday Discuss.* **2018**, *211*, 45–59.

(30) Nyman, J.; Yu, L.; Reutzel-Edens, S. M. Accuracy and Reproducibility in Crystal Structure Prediction: the Curious Case of ROY. *CrystEngComm* **2019**, DOI: 10.1039/C8CE01902A.

(31) Becke, A. D. Perspective: Fifty Years of Density-Functional Theory in Chemical Physics. *J. Chem. Phys.* **2014**, *140*, 18A301.

(32) LeBlanc, L. M.; Dale, S. G.; Taylor, C. R.; Becke, A. D.; Day, G. M.; Johnson, E. R. Pervasive Delocalisation Error Causes Spurious Proton Transfer in Organic Acid-Base Co-Crystals. *Angew. Chem. Int. Ed.* **2018**, 14906–14910.

(33) Wen, S.; Beran, G. J. O. Crystal Polymorphism in Oxalyl Dihydrazide: Is Empirical DFT-D Accurate Enough? *Journal of Chemical Theory and Computation* **2012**, *8*, 2698–2705.

(34) Wen, S.; Beran, G. J. O. Accidental Degeneracy in Crystalline Aspirin: New Insights from High-Level ab Initio Calculations. *Crystal Growth & Design* **2012**, *12*, 2169–2172.

(35) Collins, M. A.; Bettens, R. P. A. Energy-Based Molecular Fragmentation Methods. *Chem. Rev.* **2015**, *115*, 5607–5642.

(36) Raghavachari, K.; Saha, A. Accurate Composite and Fragment-Based Quantum Chemical Models for Large Molecules. *Chem. Rev.* **2015**, *115*, 5643–5677.

(37) Wen, S.; Nanda, K.; Huang, Y.; Beran, G. J. O. Practical Quantum Mechanics-Based

Fragment Methods for Predicting Molecular Crystal Properties. *Phys. Chem. Chem. Phys.* **2012**, *14*, 7578–7590.

(38) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. Fragmentation Methods: A Route to Accurate Calculations on Large Systems. *Chem. Rev.* **2012**, *112*, 632–672.

(39) Beran, G. J. O. Approximating Quantum Many-Body Intermolecular Interactions in Molecular Clusters Using Classical Polarizable Force Fields. *J. Chem. Phys.* **2009**, *130*, 164115.

(40) Beran, G. J. O.; Nanda, K. Predicting Organic Crystal Lattice Energies with Chemical Accuracy. *J. Phys. Chem. Lett.* **2010**, *1*, 3480–3487.

(41) Cai, Z.; Liu, J. Approximating Quantum Many-Body Wave Functions Using Artificial Neural networks. *Phys. Rev. B* **2018**, *97*, 035116.

(42) Yao, K.; Herr, J. E.; Toth, D. W.; McKintyre, R.; Parkhill, J. The TensorMol-0.1 Model Chemistry: A Neural Network Augmented with Long-Range Physics. *Chem. Sci.* **2018**, *9*, 2261–2269.

(43) Day, G. M.; Sam Motherwell, W. D.; Jones, W. Beyond the Isotropic Atom Model in Crystal Structure Prediction of Rigid Molecules: Atomic Multipoles Versus Point Charges. *Cryst. Growth Des.* **2005**, *5*, 1023–1033.

(44) Reilly, A. M.; Tkatchenko, A. Understanding the Role of Vibrations, Exact Exchange, and Many-Body van der Waals Interactions in the Cohesive Properties of Molecular Crystals. *J. Chem. Phys.* **2013**, *139*, 024705.

(45) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr., Sect. B: Struct. Sci.* **2016**, *72*, 171–179.

(46) Thalladi, V. R.; Nüsse, M.; Boese, R. The Melting Point Alternation in α,ω-Alkanedicarboxylic Acids. *J. Am. Chem. Soc.* **2000**, *122*, 9227–9236.

(47) Bhattacharya, S.; Saraswatula, V. G.; Saha, B. K. Thermal Expansion in Alkane Diacids - Another Property Showing Alternation in an Odd-Even Series. *Cryst. Growth Des.* **2013**, *13*, 3651–3656.

(48) Mahapatra, S.; Nayak, S. K.; Prathapa, S. J.; Guru Row, T. N. Anhydrous Adenine: Crystallization, Structure, and Correlation with Other Nucleobases. *Cryst. Growth Des.* **2008**, *8*, 1223–1225.

(49) Stolar, T.; Lukin, S.; Požar, J.; Rubčić, M.; Day, G. M.; Biljan, I.; Jung, D. Š.; Horvat, G.; Užarević, K.; Meštrović, E.; Halasz, I. Solid-State Chemistry and Polymorphism of the Nucleobase Adenine. *Cryst. Growth Des.* **2016**, *16*, 3262–3270.

(50) Stevens, E. Low-Temperature Experimental Electron Density Distribution of Formamide. *Acta Crystallogr., Sect. B: Struct. Sci* **1978**, *B34*, 544–551.

(51) Katrusiak, A. A New Polymorph of Maleic Hydrazide. *Acta Crystallogr., Sect. C: Cryst. Struct. Commun.* **1993**, *49*, 36–39.

(52) Cradwick, P. D. Crystal Structure of the Growth Inhibitor, 'Maleic Hydrazide'(1,2-Dihydropyridazine-3,6-Dione). *J. Chem. Soc., Perkin Trans. 2* **1976**, *0*, 1386–1389.

(53) Katrusiak, A. Polymorphism of Maleic Hydrazide. I. *Acta Crystallogr., Sect. B: Struct. Sci* **2001**, *57*, 697–704.

(54) Lommerse, J. P.; Motherwell, W. D.; Ammon, H. L.; Dunitz, J. D.; Gavezzotti, A.; Hofmann, D. W.; Leusen, F. J.; Mooij, W. T.; Price, S. L.; Schweizer, B.; Schmidt, M. U.; Van Eijck, B. P.; Verwer, P.; Williams, D. E. A Test of Crystal Structure Prediction of Small Organic Molecules. *Acta Crystallogr., Sect. B: Struct. Sci* **2000**, *56*, 697–714.

(55) Benghiat, V.; Leiserowitz, L. Molecular Packing Modes. Part VI. Crystal and Molecular Structures of Two Modifications of Tetrolic Acid. *J. Chem. Soc., Perkin Trans. 2* **1972**, *0*, 1763–1768.

(56) Markwell, A. J. The Crystal and Molecular Structure of Tetrabutylammonium Tetraphenylaurate(III). *J. Organomet. Chem.* **1985**, *293*, 257–263.

(57) Amoureux, J. P.; Foulon, M. Comparison Between Structural Analyses of Plastic and Brittle Crystals. *Acta Crystallogr., Sect. B: Struct. Sci* **1987**, *43*, 470–479.

(58) Capelli, S. C.; Albinati, A.; Mason, S. A.; Willis, B. T. Molecular Motion in Crystalline Naphthalene: Analysis of Multi-Temperature X-ray and Neutron Diffraction Data. *J. Phys. Chem. A* **2006**, *110*, 11695–11703.

(59) Belaj, F. Structure of 1,2,4-Triazolidine-3,5-Dione (Urazole) at 105 K. *Acta Crystallogr., Sect. C: Cryst. Struct. Commun.* **1992**, *48*, 1088–1090.

(60) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09, Revision D.01. 2009; `http://www.gaussian.com/index.htm`.

(61) Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(62) Cox, S. R.; Hsu, L.-Y.; Williams, D. E. Nonbonded Potential Function Models for Crystalline Oxohydrocarbons. *Acta Crystallographica Section A* **1981**, *37*, 293–301.

(63) Williams, D. E.; Cox, S. R. Nonbonded Potentials for Azahydrocarbons: the Importance of the Coulombic Interaction. *Acta Crystallogr., Sect. B: Struct. Sci* **1984**, *40*, 404–417.

(64) Stone, A. J. Distributed Multipole Analysis: Stability for Large Basis Sets. *J. Chem. Theory Comput.* **2005**, *1*, 1128–1132.

(65) Sakoe, H.; Chiba, S. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Trans Acoust.* **1978**, *26*, 43–49.

(66) Spek, A. L. Structure Validation in Chemical Crystallography. *Acta Crystallogr., Sect D: Biol. Crystallogr.* **2009**, *65*, 148–155.

(67) Chisholm, J. A.; Motherwell, S. COMPACK: A Program for Identifying Crystal Structure Similarity Using Distances. *J. Appl. Crystallogr.* **2005**, *38*, 228–231.

(68) Macrae, C. F.; Bruno, I. J.; Chisholm, J. A.; Edgington, P. R.; McCabe, P.; Pidcock, E.; Rodriguez-Monge, L.; Taylor, R.; Van De Streek, J.; Wood, P. A. Mercury CSD 2.0 - New Features for the Visualization and Investigation of Crystal Structures. *J. Appl. Crystallogr.* **2008**, *41*, 466–470.

(69) Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46*, 618–622.

(70) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(71) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(72) Boys, S. F.; Bernardi, F. The Calculation of Small Molecular Interactions by the Differences of Separate Total Energies. Some Procedures with Reduced Errors. *Mol. Phys.* **1970**, *19*, 553–566.

(73) Braun, D. E.; McMahon, J. A.; Koztecki, L. H.; Price, S. L.; Reutzel-Edens, S. M. Contrasting Polymorphism of Related Small Molecule Drugs Correlated and Guided by the Computed Crystal Energy Landscape. *Crystal Growth & Design* **2014**, *14*, 2056–2072.

(74) McMahon, D. P.; Stephenson, A.; Chong, S. Y.; Little, M. A.; Jones, J. T. A.; Cooper, A. I.; Day, G. M. Computational modelling of solvent effects in a prolific solvatomorphic porous organic cage. *Faraday Discuss.* **2018**, *211*, 383–399.

(75) Gavezzotti, A.; Filippini, G. Polymorphic Forms of Organic Crystals at Room Conditions: Thermodynamic and Structural Implications. *J. Am. Chem. Soc.* **1995**, *117*, 12299–12305.

(76) Hoja, J.; Reilly, A. M.; Tkatchenko, A. First-Principles Modeling of Molecular Crystals: Structures and Stabilities, Temperature and Pressure. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2017**, *7*, e1294.

(77) Hoja, J.; Ko, H.-Y.; Neumann, M. A.; Car, R.; DiStasio, R. A.; Tkatchenko, A. Reliable and practical computational description of molecular crystal polymorphs. *Science Advances* **2019**, *5*, eaau3338.

(78) Kresse, G.; Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **1993**, *47*, 558–561.

(79) Kresse, G.; Hafner, J. Ab Initio Molecular-Dynamics Simulation of the Liquid-Metal–Amorphous-Semiconductor Transition in Germanium. *Phys. Rev. B* **1994**, *49*, 14251–14269.

(80) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B* **1996**, *54*, 11169–11186.

(81) Kresse, G.; Furthmüller, J. Efficiency of ab Initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set. *Comput. Mater. Sci.* **1996**, *6*, 15–50.

(82) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

(83) van de Streek, J.; Neumann, M. A.; IUCr, Validation of Experimental Molecular Crystal Structures with Dispersion-Corrected Density Functional Theory Calculations. *Acta Crystallogr., Sect. B: Struct. Sci.* **2010**, *66*, 544–558.

(84) Taylor, C. R.; Day, G. M. Evaluating the Energetic Driving Force for Cocrystal Formation. *Cryst. Growth Des.* **2018**, *18*, 892–904.

(85) Blöchl, P. E. Projector Augmented-Wave Method. *Phys. Rev. B* **1994**, *50*, 17953–17979.

(86) Kresse, G.; Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys. Rev. B* **1999**, *59*, 1758–1775.

(87) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(88) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Duchesnay, M. P. É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(89) Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11*, 137–148.

(90) Price, S. L. Predicting Crystal Structures of Organic Compounds. *Chem. Soc. Rev.* **2014**, *43*, 2098–2111.

(91) Willock, D. J.; Price, S. L.; Leslie, M.; Catlow, C. R. A. The Relaxation of Molecular Crystal Structures Using a Distributed Multipole Electrostatic Model. *J.Comp. Chem.* **1995**, *16*, 628–647.

(92) Nobeli, I.; Price, S. L. A Non-Empirical Intermolecular Potential for Oxalic Acid Crystal Structures. *The Journal of Physical Chemistry A* **1999**, *103*, 6448–6457.

(93) De Wit, H. G.; Bouwstra, J. A.; Blok, J. G.; De Kruif, C. G. Vapor Pressures and Lattice Energies of Oxalic Acid, Mesotartaric Acid, Phloroglucinol, Myoinositol, and Their Hydrates. *J. Chem. Phys.* **1983**, *78*, 1470–1475.

(94) Otero-de-la Roza, A.; Johnson, E. R. A Benchmark for Non-Covalent Interactions in Solids. *The Journal of Chemical Physics* **2012**, *137*, 054103.

(95) Marom, N.; Distasio, R. A.; Atalla, V.; Levchenko, S.; Reilly, A. M.; Chelikowsky, J. R.; Leiserowitz, L.; Tkatchenko, A. Many-body Dispersion Interactions in Molecular Crystal Polymorphism. *Angew. Chem. Int. Ed.* **2013**, *52*, 6629–6632.

For Table of Contents Use Only