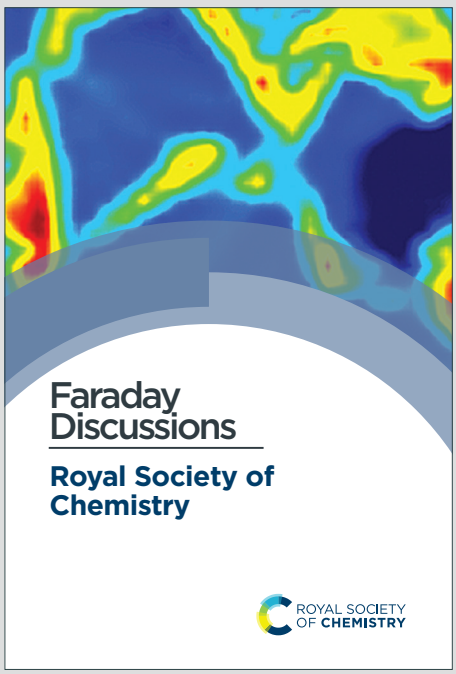
Faraday Discussions

Accepted Manuscript



This article can be cited before page numbers have been issued, to do this please use: C. Taylor, P. Butler and G. M. Day, *Faraday Discuss.*, 2024, DOI: 10.1039/D4FD00105B.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/faraday_d

View Article Online
DOI: 10.1039/D4FD00105B

Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Predictive crystallography at scale: mapping, validating, and learning from 1,000 crystal energy landscapes

Christopher R. Taylor[a], Patrick W. V. Butler[a], and Graeme M. Day[a*†]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Computational crystal structure prediction (CSP) is an increasingly powerful technique in materials discovery, due to its ability to reveal trends and permit insight across the possibility space of crystal structures of a candidate molecule, beyond simply the observed structure(s). In this work, we demonstrate the reliability and scalability of CSP methods for small, rigid organic molecules by performing in-depth CSP investigations for over 1000 such compounds, the largest survey of its kind to-date. We show that this highly-efficient force-field-based CSP approach is superbly predictive, locating 99.4% of observed experimental structures, and ranking a large majority of these (74%) as among the most stable possible structures (to within uncertainty due to thermal effects). We present two examples of insights such large predicted datasets can permit, examining the space group preferences of organic molecular crystals and rationalising empirical rules concerning the spontaneous resolution of chiral molecules. Finally, we exploit this large and diverse dataset for developing transferable machine-learned energy potentials for the organic solid state, training a neural network lattice energy correction to force field energies that offers substantial improvements to the already impressive energy rankings, and a MACE equivariant message-passing neural network for crystal structure reoptimisation. We conclude that the excellent performance and reliability of the CSP workflow enables the creation of very large datasets of broad utility and explanatory power in materials design.

1 Introduction

The discovery of new materials is important for addressing many critical societal needs, including energy production and storage, pollution remediation and healthcare. Research endeavours aimed at improving the success and efficiency of functional materials discovery, based on traditional efforts developing our understanding of the rules of crystal packing and, more recently, applications of machine learning, have benefited greatly from the availability of databases of stable crystal structures.

A major resource for such efforts has been the growth of accessible, curated databases of crystal structures and (some of) their properties. The most general and widely-used include the Cambridge Structural Database (CSD)[1] for organic and organometallic systems, and the Inorganic Crystal Structure Database[2]. These resources are unparalleled in their volume of experimental crystal structure information, but do not currently offer information about calculated or hypothetical structures. Historically, such data was limited to specialised areas (such as the Atlas of Prospective Zeolite Structures[3]), though recent developments have made remarkable progress in generalising this concept, including extensions to the Crystallography Open Database (COD)[4–7] and the Materials Project[8]. In the field of organic molecular crystals, however, much of our understanding of the rules of crystal packing derive from databases of experimentally observed structures.

Modern computational chemistry, employing both molecular and solid-state simulation techniques, can add significantly to the information that is available from experimentally determined crystal structures, and identify previously unverifiable trends. Computational studies of polymorphism[9–13] have evaluated the typical lattice energy differences between crystalline polymorphs of organic molecules, studies of conformations of flexible molecules[11,14] in their observed crystal structures have improved understanding of the limits of molecular strain in stable structures, and studies of the thermodynamics of co-crystallisation[15–18] have aided in rationalising a complex phenomenon with ramifications for experimental design.

A more complete and, crucially, predictive view of organic crystal packing can be obtained from crystal structure prediction (CSP)[19,20]. A core concept in CSP is the crystal energy landscape (or CSP landscape) – the set of plausible crystal packings for a chemical species (or combination thereof), representing an exploration of the crystalline configuration space to identify candidate

a School of Chemistry, University of Southampton, Southampton SO17 1BJ, United Kingdom.

† Electronic Supplementary Information (ESI) available alongside publication: computational details of CSP and ML calculations, summary of molecule set including CSD refcodes, and summary of CSP matches to experiment. Additionally, complete CSP landscapes and ML models available from DOI: 10.5258/SOTON/D3094.

* Author to whom correspondence should be addressed.

View Article Online
DOI: 10.1039/D4FD00105B

structures predicted to be stable (which ideally includes any observed structures), ranked in terms of this stability. These CSP landscapes are of great utility in understanding and rationalising the thermodynamic and kinetic behaviour of crystal systems; multiple low-energy minima that are close-lying on a CSP landscape may be indicative of a significant risk of polymorphism, while dynamical simulations exploring these landscapes give insight into the kinetic trapping of metastable forms[21] and the observed absence of other predicted forms[22,23].

Moreover, the energy landscape forms the foundation of the energy-structure-function maps that have in recent years demonstrated great power in materials discovery[24–27]. By associating computed properties (e.g. gas uptake, charge carrier mobility) with hypothetical crystal structures from the CSP landscape, it becomes possible to predict whether a molecule is a promising candidate for creating new functional materials, i.e. if it has one (or more) favourably-ranked crystal structures which are predicted to achieve the desired property.

The techniques and challenges in the field of organic CSP have been reviewed[19,20]; we provide a brief overview for the sake of context. CSP is typically considered a combination of two broad challenges: efficient and thorough sampling of the configuration space of crystal packing, and accurate, cost-effective structural optimisation, ranking, and property calculation.

The sampling of hypothetical crystal packing arrangements is made extremely difficult due to the "curse of dimensionality"—the number of independent degrees of freedom to sample creates a vast configurational space. As a result, simple grid-based sampling approaches must be eschewed in favour of more sophisticated techniques, such as low-discrepancy quasi-random sampling[28,29] and genetic algorithm approaches.[30]

Meanwhile, the optimisation and ranking methods must be accurate enough to describe the fine balance of different intermolecular interactions (electrostatics, dispersion, hydrogen-bonding, etc.), resolving lattice energy differences often smaller than a kJ mol^{-1}, while sufficiently cost-effective to be applied to very large numbers (»10^5) of trial crystal structures. Historically, this has entailed the use of simple empirical force fields, but modern developments often employ tailor-made force fields[18,31] or machine-learned potentials derived from *ab initio* calculations. Still, empirical force fields retain their power even today due to their efficiency and broad transferability, often being the initial step of a hierarchy of increasingly accurate (and expensive) energy models employed in one CSP workflow.

Despite these two broad and ongoing challenges, organic CSP has demonstrated enormous success in diverse applications, crucially proving itself to be a truly predictive technique, guiding synthesis and discovery of novel forms of porous materials[24], highly-flexible pharmaceutical molecules[31], co-crystals[32], simple molecules previously thought to be monomorphic[33] and templating of predicted metastable polymorphs.[34] A recurrent landmark in the field is the series of Cambridge Blind Tests of CSP, showcasing the diversity of methods (and success rates thereof) employed within different CSP techniques to predict experimental crystal structures without any knowledge beyond the molecular chemical diagram[35]. Recent iterations of the Blind Test have demonstrated that CSP method development is successfully keeping apace with the complexity of molecules and crystal structures specifically selected to stress-test it.

Some of the most consequential recent developments in CSP have employed machine learning (ML) techniques in one form or another[36]. Among the most intuitive applications is the use of ML to learn relationships between the structure and lattice energy, either by learning the difference (i.e. Δ-ML) between lattice energies computed at a lower level of theory to those at a higher level,[37–40] or to learn the relationship between the lattice energy and the ML descriptors directly through the training of ML potentials[41,42]. These approaches have achieved high accuracy predictions at a fraction of the cost of the full periodic density functional theory (DFT) reference calculations – particularly significant given the latter's ubiquity in recent Blind Test entries.

That said, ML techniques have further applications in CSP beyond improved optimisations and energy rankings. In particular, ML and related approaches applied to databases of experimental crystal structures and properties have demonstrated success in predicting NMR chemical shifts[43] and in formulating models of molecular hydrogen bond propensities within crystals[44]. Recent work by Cersonsky[45] has demonstrated machine-learning of the relationship between crystal lattice energies and the relative contributions to these by different chemical functional groups, paving the way for data-driven insights into new crystal engineering techniques. ML also been shown to have potential to enhance the analysis of crystal energy landscapes, by identifying structure-function relationships that might evade simple inspection but nevertheless offer explanatory and predictive power[46].

Our aims in this work are threefold. Firstly, we seek to demonstrate the capability of our method for efficient, large-scale rigid-molecule organic CSP by presenting the results of the largest to-date CSP study, applying the methods to over 1000 molecules with observed crystal structures in the CSD. Secondly, we in turn assess the quality and reliability of our method by evaluating how often the experimental crystal structures of these molecules are reproduced in our CSPs, and how well they are ranked energetically compared to hypothetical structures on their CSP landscapes. Finally, we demonstrate example applications of this dataset, including assessing the distribution of CSP-derived space group preferences as a function of predicted lattice energy, spontaneous resolution of chiral molecules and the training of transferable machine-learned energetic models from a very large set of CSP landscapes.

2 Computational Methods

2.1 Molecule Selection

We used the CSD's ConQuest software and Python API[1] to search the CSD for crystal structures of rigid molecules to which our CSP methods could be applied on a very large scale. Restricting our search to solved crystal structures (i.e. with coordinates for all atoms in the asymmetric unit) of single chemical species (no co-crystals, solvates, inclusion compounds), we additionally filtered structures based on the following criteria: containing only elements from C, H, N, O, F; $Z' \leq 1$; molecular weight less than 230;

View Article Online
DOI: 10.1039/D4FD00105B

and importantly containing no rotatable bonds (as defined by the CSD's internal criteria).

2.2 Molecular Geometry Optimisation

For each molecule, we began by extracting its in-crystal conformation from the corresponding CSD entry (where there is more than one entry, we select only the first listed in the database). This molecular conformation was optimised in DFT (as implemented in Gaussian[47]), using the PBE0[48,49] exchange-correlation functional, a Pople-type 6-311G** basis[50], and Grimme's D3 dispersion correction[51] with Becke-Johnson damping[52].

Distributed multipole analysis (DMA,[53,54] using the GDMA package) was performed on the resulting molecular conformers' charge densities to obtain atom-centred multipoles, as part of the model potential applied during lattice energy minimisation of trial crystal structures; multipoles up to hexadecapole were calculated for all atoms. The MULFIT[55,56] software was used to fit atomic point charges for each molecule to best describe the molecular electrostatic potential. The resulting molecular conformers and their sets of multipoles are then used as the inputs to CSP, with each unique molecule represented by a single conformer and corresponding set of atomic multipoles and charges.

2.3 Crystal Structure Generation and Optimisation

CSP was performed using the Global Lattice Energy Explorer (GLEE) package,[28] which uses quasi-random sampling of crystal packing variables to generate trial crystal structures uniformly distributed across the lattice energy landscape, followed by rigid-molecule lattice energy minimisation using an anisotropic atom-atom intermolecular force field. All resulting local energy minima are treated as possible crystal structures of the molecule.

Space group symmetry is used to reduce the dimensionality of the search space, so that only the position and orientation of molecules in the asymmetric unit are sampled, with all other molecules in the unit cell generated by symmetry. In this study, we restrict ourselves to generating crystal structures with one independent molecule in the asymmetric unit ($Z' = 1$). We sampled the 26 most commonly observed space groups for organic molecular crystals (listed in Supporting Information); these space groups cover over 99.4% of $Z' \leq 1$ structures in the CSD. These space groups were sampled equally, irrespective of their observed frequency in the CSD: quasi-random structures are generated and lattice energy minimised until 10,000 successfully energy minimised crystal structures were found in each space group (260,000 structures per molecule). The CSP process is highly parallelisable, as each crystal structure structure is independent.

Trial crystal that passed geometric checks were lattice energy-minimised in three stages. Non-electrostatic interactions (principally intermolecular dispersion and exchange-repulsion) were described by the FIT $exp-6$ force field[57,58], supplemented by fluorine parameters from Williams and Houpt.[59] At the final stage of optimisation, performed using DMACRYS,[58] the FIT potential was applied with atomic multipole electrostatics. Full details are provided in Supporting Information.

It is commonplace that multiple unique initial configurations optimise to the same local energy minimum. We remove these duplicates by fast comparison of simulated powder X-ray diffraction patterns, followed by structural comparisons using the COMPACK[60] algorithm as implemented in the CSD Python API.

Locating Experimental Structures on the Landscapes

To assess our CSP workflow's performance, comparison of the known experimental crystal structures to the sets of predictions was automated using the COMPACK algorithm between the experimental crystal structures of these molecules and every unique crystal structure of that molecule in the CSP set.

2.4 Machine Learned Interatomic Potentials

To investigate the potential of the CSP dataset to train data-derived models, a subset of the predicted crystal structures with energies within 8.0 kJ mol^{-1} of the global energy minimum on their CSP landscape was selected for training a lattice energy correction to the FIT+DMA force field. The subset was determined by active learning via query-by-committee using a committee of eight high-dimensional neural network potentials (NNPs), with selected structures evaluated by DFT+D single points at the B86bPBE+XDM level. From this, a dataset of the crystal structures and the corresponding energy correction between the FIT+DMA and B86bPBE+XDM lattice energies (ΔE) was created. B86bPBE+XDM lattice energies were calculated as the total energy less the energies of the isolated molecules from the unit cell calculated with the same basis set and tolerances.

The initial dataset (before the active learning iterations) was generated by randomly selecting up to 10 low energy predicted crystal structures for each compound, resulting in a dataset of 10,249 structures approximately evenly distributed across the compounds. To evaluate transferability, the total dataset was partitioned into a training dataset consisting of the CSP structures for a randomly selected ca. 85% of the compounds and an extrapolation test set consisting of CSP structures for the remaining compounds. A further in-domain test set was formed by randomly extracting one structure per compound from the training set. The NNPs were then trained on the remaining training dataset to yield a Δ-ML model capable of predicting the lattice energy correction. The standard deviation between the ensemble of NNPs was used to estimate the uncertainty of predictions, and was exploited in the active learning iterations to add high uncertainty candidates from the remaining low energy predicted structures of the training compounds, which overall added a further 1000 structures to the training set. Corrected CSP landscapes were calculated using the final model by adding predicted lattice energy corrections to the FIT+DMA energies (FIT+DMA+Δ-ML). Additionally, for performing unconstrained geometry optimisations MACE equivariant message-passing neural network (MPNN) models were trained using a dataset derived from that of the NNP correction model. Further details of the datasets and machine learning models are provided in the supporting information.

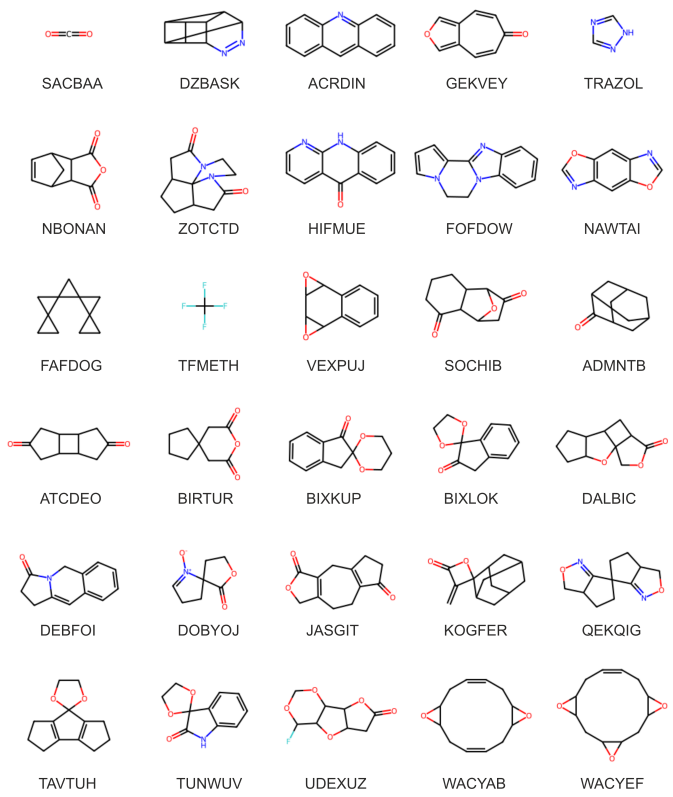


Fig. 1 Molecular diagrams and crystal structure CSD reference codes for (top three rows) a random selection of the 1007 molecules included in the large-scale CSP study. The bottom three rows show molecules in the set with the largest differences between in-crystal and optimised molecular geometries (as measured by all-atom RMSD); the CSP landscapes for these molecules were re-optimised using the transferrable MACE model (final section).

3 Results

3.1 Diversity of Survey Set

Our aforementioned search criteria yielded 1007 distinct molecules crystallising in 1040 crystal structures observed in the CSD. The constraint of no rotatable bonds necessarily limits chemical diversity, but despite this our candidate set still displays a variety of chemical functionalities and molecular structures, as seen in Figure 1 and Table 1. A complete list of molecules, including formulae, CSD refcode identifiers, SMILES strings, and systematic and common names is available in the Supporting Information.

functional group	count	functional group	count	functional group	count
benzene	418	ether	572	ester	134
pyridine	76	amide	297	carbonyl	702
imidazole	35	imide	43	imine	13
furan	22	lactone	91	secondary amine	405
piperdine	51	epoxide	73	tertiary amine	537
urea	22	ketone	269	halogen (F)	240

Table 1 Total counts of selected functional groups across the full set of molecules, as assessed by RDKit[61] from SMILES strings.

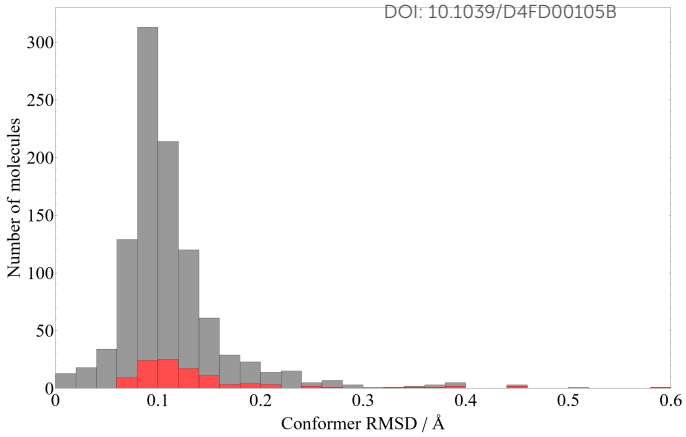


Fig. 2 A histogram of the RMSD between all atomic positions in the gas-phase optimised CSP candidate molecules, relative to their initial in-crystal conformations. The red portions of each bar indicate the molecules in that bin with spiro carbon centres.

Deviation between experimental and gas-phase conformations

Despite restricting our set to molecules containing no rotatable bonds, this does not preclude molecular flexibility entirely. More complex collective motions cannot be described in terms of a single torsional angle about a covalent bond, and so molecules displaying such conformational flexibility are present in this candidate set – the prototypical example is a ring "flip" or buckling, such as the boat-chair interconversion of cyclohexane rings.

As a measure of the typical deviation in molecular geometry between the observed crystal structures and the DFT optimised molecules used in CSP, we present in Figure 2 the histogram of all-atom root-mean-squared deviations (RMSD) in atomic positions between the gas-phase conformers used in CSP and their in-crystal initial conformations as extracted from the CSD.

As might be expected from such rigid molecules, the average RMSD after gas-phase optimisation is very small – approximately 0.11 Å, which corresponds to e.g. adjusting the C-H bond lengths in fluorobenzene by 0.16 Å. While the distribution is skewed towards small conformational changes, the outliers with larger RMSD values demonstrate the limitations of defining molecular flexibility in terms of rotatable bonds alone. The largest RMSD values correspond to systems where changes in ring conformation cause large overall molecular changes – the largest observed RMSD of 0.60 Å occurs in 7-oxa-1-azaspiro(4.4)non-1-en-6-one 1-oxide (CSD reference code: DOBYOJ), in which a 5-membered saturated ring can twist about a spiro carbon centre. The highest molecular RMSD values are largely associated with molecules containing spiro carbon atoms (denoted by the red portions of bars in Figure 2).

Regardless, the overall inflexibility displayed by this set serves as a strong indicator that our assumption of a single, near-crystalline conformer of each molecule is reasonable, and unlikely to be a systematic source of error in predicting many known structures of molecules in this set.

View Article Online
DOI: 10.1039/D4FD00105B

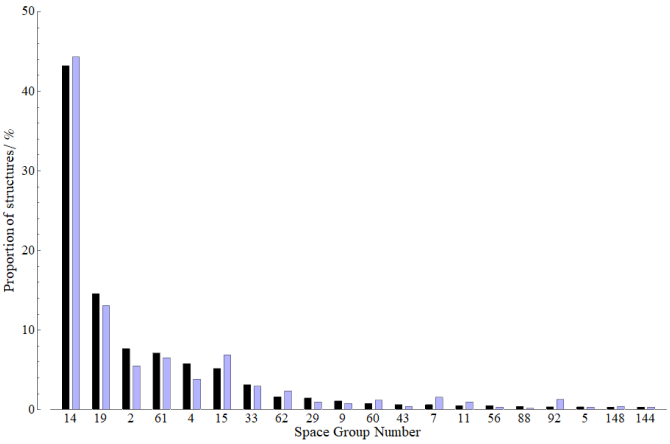


Fig. 3 The relative frequency of space groups observed for crystal structures (where $Z' \leq 2$) of molecules in the CSD with a molecular weight under 230 (black) and the subset (blue) of these molecules with no rotatable bonds selected for our CSP surveys. The distributions are presented for only the 20 most common space groups (of the general molecule case) for clarity.

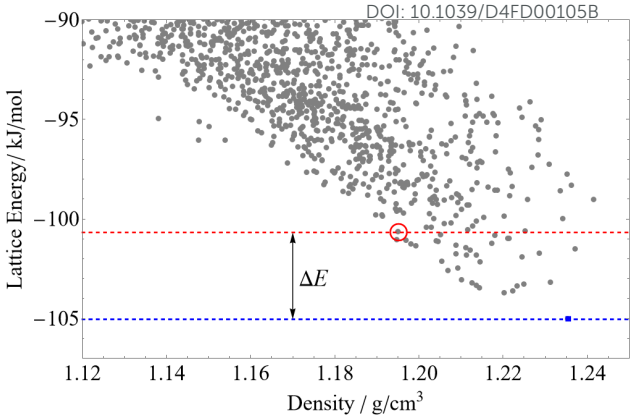


Fig. 4 An example CSP landscape for a single molecule – the set of predicted lattice energy minima from our CSP workflow. Each point represents a local energy minimum, and thus a stable hypothetical crystal packing. The blue square point is the global energy minimum, which in the absence of experimental information is taken to be the most likely crystal structure. If the red circled point is a match to the experimentally-observed crystal structure, ΔE is the difference in energy between it and the global minimum, the energy rank used in this work as a measure of the quality of the calculated energies.

Crystalline diversity

Figure 3 shows a comparison of the distribution of crystallographic space groups for most small molecules in the CSD to the rigid-molecule subset selected for our CSP survey. There is no significant difference in the relative occurrence of different crystalline symmetries between the two sets, though the relative ordering of space groups by frequency varies slightly. For example, space group 15 ($C2/c$) is observed slightly more frequently in our rigid set and space group 2 ($P\bar{1}$) slightly less so. Regardless, our rigid molecule set is reasonably representative of the range of crystal packing symmetries observed in the CSD.

In contrast, our rigid molecule set displays a diminished frequency of hydrogen bonding (H-bonding) in the observed crystal structures compared to the CSD more generally, as might be expected by excluding rotatable bonds. 26.7% (277) of crystal structures of molecules in our subset contain at least one intermolecular H-bond. This compares to 62.3% of structures of molecules of similar size without rotatable bond restrictions, demonstrating the bias introduced by the omission of common H-bonding groups due to their flexibility, including alcohols ($-$OH), carboxylic acids ($-$COOH), and primary amines ($-NH_2$). However, the proportion of H-bonded systems is still significant enough that meaningful H-bond chemistry is incorporated in our survey, albeit underrepresented. Consequently, our set in turn overrrepresents chemistry such as π-π stacking and "weak" (i.e. more isotropic, less localised) interactions, and assessments of our CSP energy model's performance must be made with these biases in mind.

3.2 Quality Assessment of the CSP Results

We propose that the dataset of predicted crystal structure landscapes across a large, diverse set of molecules is valuable for the development of future predictive models. To evaluate the quality of the dataset, we assess three aspects: completeness of the landscapes; how well, geometrically, the CSP calculations reproduce the known crystal structures of these molecules and the quality of the relative energies of the predicted structures.

The 26 space groups included in our standard search include those with an observed frequency above 0.05 % in the CSD. Naturally, in a very large survey of molecules, we include some that crystallise in less frequent space groups; for these 8 cases, we added the space group of the observed crystal structure to the search. In 3 additional cases, the symmetry of the observed structure means that it could not have been sampled without performing CSP with multiple independent molecules (Z' > 1); in these cases, we include the datasets in our study, knowing that the observed crystal structures could not have been located. Of those that could have been located, the searches find matches for 1034 of the 1040 observed crystal structures: reasons for the 6 missed matches are discussed below.

A main assumption in CSP is that the observed crystal structures of a molecule correspond to the lowest energy possible structures. We use this assumption to assess the quality of the calculated energies. Figure 5 summarises the energetic ranking of experimentally observed crystal structures within the CSP landscapes via the distribution of ΔE, the energy difference between the prediction matching the experimental crystal structure and the global lattice energy minimum crystal structure on their parent molecule's CSP landscape; an example landscape is shown in Figure 4. In the case of molecules with chiral centers and only one stereoisomer present, ΔE is calculated only among Sohncke space groups (those whose symmetry elements contain only translations, rotations and rototranslations).

Of the 1034 experimentally determined crystal structures where a matching structure was identified on the CSP landscape, 424 (41%) correspond to the global lattice energy minimum from CSP. Those that do not correspond to global energy minima could

View Article Online
DOI: 10.1039/D4FD00105B

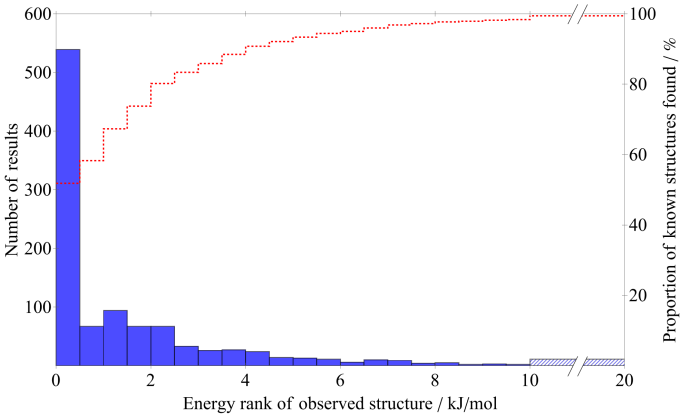


Fig. 5 A histogram of the frequency (blue bars) with which our CSP workflow achieves a match to the experimented-known structures of our molecule set, grouped by the relative energy of that match compared to the CSP global minimum (0.5 kJ/mol bins). The red dashed line relates to the secondary (right) y-axis, the proportion of known structures located successfully as a function of the relative energy at that bin and below. Note the broken x-axis; the highest-energy bin (blue hatching) encompasses all matches with relative energy greater than 10 kJ/mol.

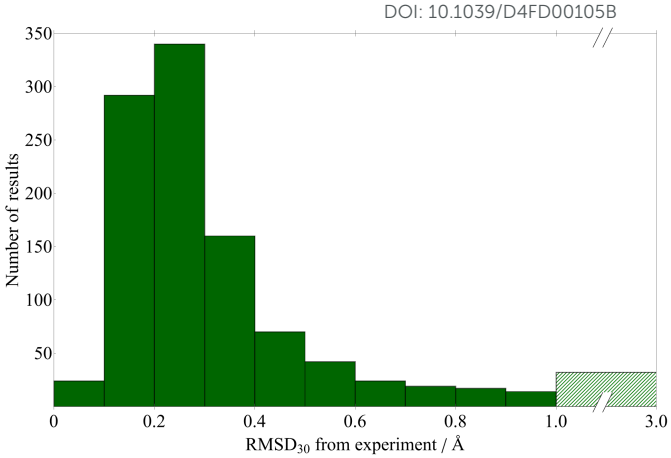


Fig. 6 A histogram of the geometric deviation between experimentally-determined crystal structures and the corresponding matching structures from CSP. The deviation is measured as the RMSD in atomic positions within 30-molecule clusters from experimental and CSP structures. Note the broken x-axis – the largest-deviation bin (green hatching) includes all matches with $RMSD_{30}$ greater than 1.0 Å.

be due to inaccuracy of the model potential (FIT+DMA), neglect of other contributions to the lattice free energy or where the kinetics of crystallisation favour a metastable structure. Nyman and Day found that, for crystals of rigid molecules, lattice vibrational contributions to room temperature free energy differences between polymorphs rarely exceed 2 kJ/mol[9,10]; 767 (74%) of observed crystal structures are found within 2 kJ/mol of the global energy minimum – the estimated error introduced by neglecting lattice vibrations and thermal expansion in the CSP calculations.

Furthermore, the known structure(s) almost always lie within 8 kJ/mol of the global lattice energy minimum (1011, 97.8%, of the observed crystal structures), consistent with the observation[10] that known polymorphic pairs of small, rigid molecules are rarely separated by more than this energy difference. Thus, the energy model used here, combining empirically parameterised repulsion-dispersion with atomic multipole electrostatics, provides energy ranking of crystal structures that is consistent with observed polymorphism, within the limits of temperature-free lattice energy-based predictions.

We quantify the geometric quality of the predictions using an all-atom RMSD within 30-molecule clusters ($RMSD_{30}$) from experimentally-determined crystal structures and their corresponding match within the CSP sets. A histogram of $RMSD_{30}$ (Figure 6) shows that geometric agreement is generally very good: $RMSD_{30}$ is below 0.4 Å in 78.9% (816) of matches. As a visualisation of this level of agreement, Figure 7 shows an overlay of the X-ray determined crystal structure of (1aR,2aS,5aS,5bS)-perhydro-4H-oxireno(3,4)cyclopenta(1,2-b)furan-4-one[62] (CSD reference code SIBJIX) and the predicted global lattice energy minimum, with $RMSD_{30} = 0.393$ Å. As a reference for these values, consider the $RMSD_{30}$ between structural determinations of the same crystal structure at ambient and low temperature: $RMSD_{30} = 0.204$ Å between neutron diffraction crystal structures of naphthalene at 5 K and 295 K, and $RMSD_{30} = 0.160$ Å between 20 K and 330 K crystal structures of form I paracetamol[63]. 327 matches have $RMSD_{30}$ below 0.204 Å, i.e. have geometric deviations that are of a magnitude that can be explained by the temperature-free nature of structural optimisations used in CSP. While known crystal structures are reproduced very well by CSP in most cases, there are a small number where agreement is less satisfactory: in 32 cases (3 % of structures), $RMSD_{30} > 1$ Å. Despite what might be assumed, we find no significant correlation between the $RMSD_{30}$ of the experimental match and the molecular RMSD of the parent gas-phase conformer used for CSP. Assuming an experimental match is found, the geometric quality of the match is only weakly sensitive to the difference in the molecular conformation used for CSP; even the most extreme outlier in molecular conformational change (the aforementioned DOBYOJ) achieves a reasonable geometric match to the experimental crystal structure, with $RMSD_{30} = 0.654$ Å.

It is evident that an overwhelming proportion of such rigid molecules can successfully be treated with the CSP workflow implemented here. Our sampling procedure followed by a cost-effective, approximate minimisation method successfully locates the vast majority of observed crystal structures with excellent geometric agreement to the experimental structure, and routinely ranks them as among the most stable structres on the landscape.

Analysis of individual outliers

The cases where no match whatsoever is found to the experimental structure are limited – approximately 0.5% of the experimental structures considered.

One missed match was for the molecule 4,8b-dihydropyrrolo[3,4-b]indole-1,3(2H,3aH)-dione (CSD reference code BUVGAC), which displays a large deviation in molecular geometry between the observed crystal and the gas phase optimised geometry used in CSP. Flexibility in the fused ring system allows the molecule to "fold up" in the gas-phase optimisation, sufficiently to change its packing behaviour. As a result, the

View Article Online
DOI: 10.1039/D4FD00105B

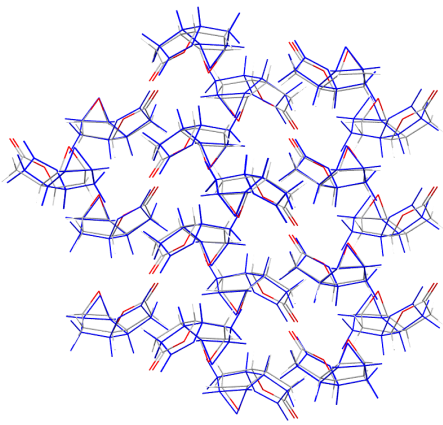


Fig. 7 Overlay of 30-molecule clusters from the X-ray determined crystal structure (atoms coloured by element, CSD reference code SIBJIX) with the matching prediction (blue) – the global energy minimum structure for (1aR,2aS,5aS,5bS)-perhydro-4H-oxireno(3,4)cyclopenta(1,2-b)furan-4-one. For this match, $RMSD_{30} = 0.393$ Å.

experimental crystal structure is not a minimum on a CSP landscape derived from the gas-phase conformer.

In the case of DNNEPH10 (1,8-dinitroso-naphthalene), we fail to find the experimental structure despite an apparently rigid, planar molecule that changes very little in the gas-phase optimisation. However, even optimising the known crystal structure using our FIT+DMA energy model results in a final structure that does not match, i.e. the experimental structure appears to be unstable at our level of theory. This may indicate a failure of our energy model for a case of somewhat unusual chemistry.

Two of the missed matches are unusual cases in which a refcode "stem" (the initial six letters, typically shared by a "family" of multiple CSD entries of the same species) is used by crystal structures containing distinct chemical bonding arrangements. Our procedure takes a single representative of a CSD refcode family as the source for the molecular connectivity, which was assumed to stay fixed. In the case of refcodes XUGHUD/XUHGUD01, there is a tautomeric difference, which led to no match with XUHGUD01. For IHEPUG/IHEPG02, the subject molecule is a diastereomeric fused ring system, which exists in an *anti* configuration in IHEPUG, but a *syn* configuration in IHEPUG02. It is arguable that these crystal structures should not be considered part of the same "family", as these distinct bonding arrangements are not interconvertible. CSP was performed with the isomer found in IHEPUG, so no crystal structure matching IHEPUG02 was located. These two systems demonstrate a shortcoming of our approach, in that we assumed that a given CSD refcode "stem" always denotes the same molecular connectivity, including protonation states. Fortunately, these are the only instances of this assumption failing in the entire set.

Our final missed match occurs for QIBCEK, benzo(f)phthalazin-4(3H)-one, another rigid planar molecule. However, upon inspection, we posit that there are flaws in the experimental determination of this structure as held in the CSD – there are extremely close hydrogen contacts (<1.4 Å) and an unsatisfied potential hydrogen-bonding arrangement despite a 1:1 donor-acceptor availability. While we retain it as a missed match for the purposes of conservatively asssessing our CSP method's performance, we also emphasise that such a large-scale, unbiased workflow has potentially identified an incorrect experimental structure simply through its absence from the CSP landscape.

3.3 Revisiting Empirical Rules

Large databases of experimentally determined crystal structures have been analysed to uncover trends in the packing preferences of organic molecules. The availability of high quality crystal energy landscapes should allow organic solid state researchers to gain deeper insight into these preferences and, we hope, to discover new rules that will benefit the field of crystal engineering. We give two examples here: the unequal frequency with which molecules occupy the possible space groups and the spontaneous resolution of chiral molecules.

3.3.1 Space Group Preferences

There are strong space group preferences for experimentally observed crystal structures: over 80% of molecular crystals occupy 5 of the 230 three-dimensional space groups. Having applied an approach in generating trial crystal structures that is unbiased in the how the 26 space groups included in our searches are sampled, we analyse the results of CSP to investigate space group preferences within the low energy structures on the set of crystal energy landscapes.

We preface our analysis by emphasising that the space group frequencies presented in Figure 8 are those for the CSP landscapes, i.e. only those of structures generated using an asymmetric unit containing a complete molecule (in these single-species systems, $Z'=1$) and without detecting and assigning additional symmetry after the minimisation. In contrast, those of the observed crystal structures in the CSD are the full, maximal-symmetry space group (allowing fractions of molecules in the asymmetric unit, i.e . $Z' \leq 1$). Hence. the space groups enumerated for the CSP set can be thought of as "lower bounds" on the symmetry – higher-symmetry space groups may be assignable if the molecules present internal symmetry.

The global lattice energy minimum for each molecule is the energetically preferred packing; the distribution of space groups among the global minima structures are highly consistent with the observed statistics – perhaps as expected, as CSP has been demonstrated in this work to often identify observed structures of these molecules as global minima.

Space group frequencies are often explained by close packing arguments: the commonly observed space groups have combinations of symmetry elements that facilitate close packing of irregular shapes.[64] Examining the space group distribution amongst the densest predicted crystal structure for each molecule gives a similar distribution to that from global energy minima; minimising energy and maximising density lead to similar space group preferences. However, there are differences after the three most popular space groups: the next few have almost equal frequencies among high density structures, suggesting that they are equally good at promoting close packing and that observed differences between these space groups relate to subtler influences of sym-

View Article Online
DOI: 10.1039/D4FD00105B

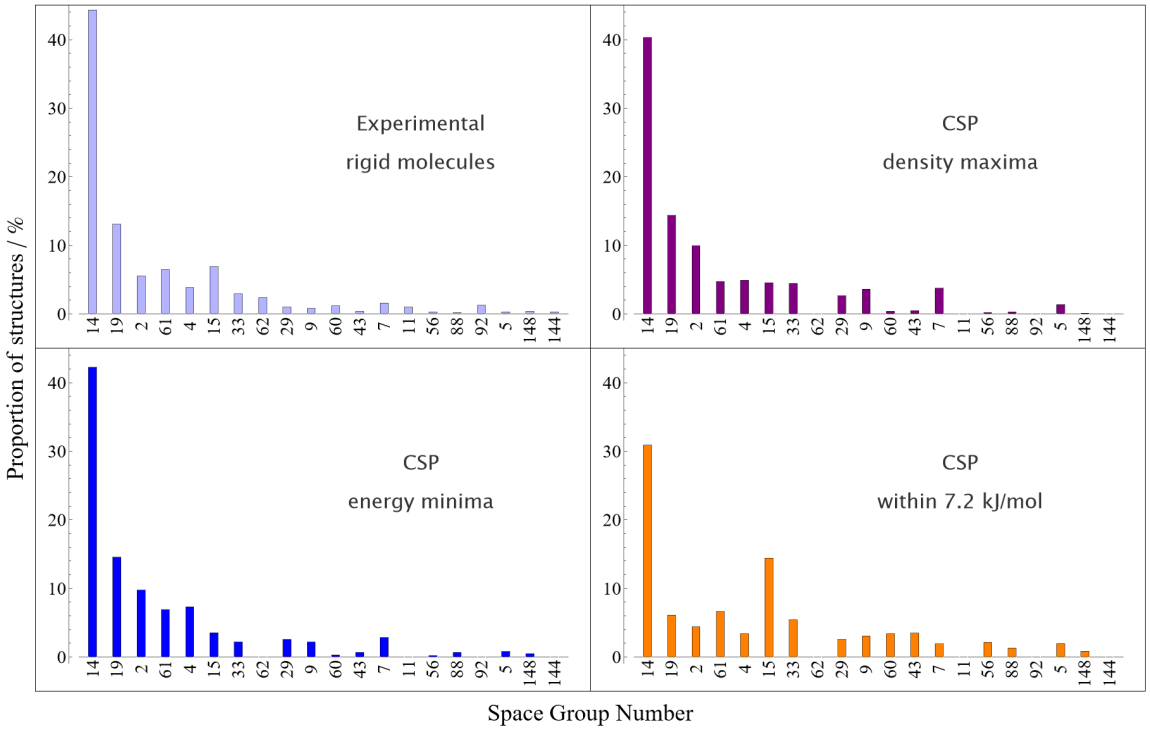


Fig. 8 The relative frequencies of space groups of crystal structures. Light blue (top left) are those of rigid molecules in the CSD (as in Figure 3), while the rest are obtained from CSP landscapes at the global density maximum (purple, top right), at the global energy minimum (blue, bottom left), and within 7.2 kJ/mol of the global energy minimum (orange, bottom right). The ordering of space groups on the x-axis is chosen to match that in Figure 3. Only the 20 most common space groups for rigid molecules are presented for clarity.

metry on lattice energy.

Considering hypothetical structures higher on the CSP energy landscape (up to the usual energy limit of polymorphism, 7.2 kJ mol^{-1}, in the bottom right panel of Figure 8), we see a further flattening of the distribution, and an overrepresentation of space group 15 ($C\,2/c$) compared to global energy minima structures. These changes in distribution with energy, which we do not examine in deeper detail here, are only available from access to complete energy landscapes and are relevant in the discovery of high energy, metastable materials, which have sometimes been observed to have attractive properties.[24,65] Thus, we feel that large datasets of CSP landscapes hold potential for generalising our understanding of symmetry preferences in molecular crystals.

3.3.2 Spontaneous Resolution of Chiral Molecules

As a second example of insight that can be gained from large numbers of crystal energy landscapes, we examine the tendency for spontaneous resolution of chiral molecules. It is generally accepted that crystallisation from a racemic solution of a chiral molecule more frequently yields racemic crystals rather than undergoing spontaneous resolution into a mixture of crystals, each containing a single stereoisomer.[66] However, information in structural databases alone is limited: knowing whether crystals were grown from racemic or enantiomerically pure solution is necessary to interpret the incidence rate of spontaneous resolution and the molecular characteristics that influence this behaviour. Furthermore, where enantiomers separate upon crystallisation, it is not possible to grow racemic crystals, so that comparison of racemic vs enantiomerically pure crystal structures is not possible. Computed crystal energy landscapes make it possible to compare the structures and relative energies of the alternative crystallisation outcomes.

In Figure 9, we show the difference in stability and density between the energy minimum across all Sohncke (i.e. enantiopure) space groups and the minimum across enantiogenic (i.e. racemic) space groups for the 356 molecules in our set containing at least one chiral centre. This energy difference represents the propensity for spontaneous resolution of a racemic solution compared to a racemic crystal of both enantiomers.

In general, there is a slight but consistent lattice energy penalty to enantiopurity – the average difference in lattice energy between the enantiopure global minimum and the racemic global minimum is 2.7 kJ/mol, favouring the racemate. Of the 356 chiral molecules, racemic crystallisation is preferred in 86% (305) of molecules and spontaneous resolution is predicted to occur for approximately 14% (51) of molecules, although those with small lattice energy differences could be influenced by thermal contributions that are not included here. Moreover, this is accompanied (or perhaps driven) by improved packing in the racemate – on average, optimal enantiopure structures are 2.3% less dense than optimal racemic structures, consistent with the empirical Wallach's rule[67,68].

View Article Online
DOI: 10.1039/D4FD00105B

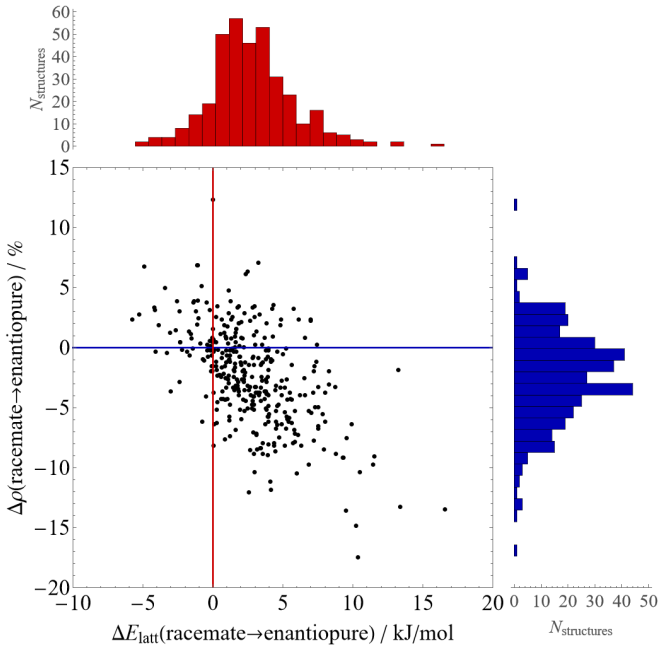


Fig. 9 The difference in lattice energy (ΔE_{latt}, top red histogram) and relative density ($\Delta\rho$, right blue histogram) between the energy minimum in Sohncke (enantiopure) space groups and that in racemic space groups, for all molecules in our set containing at least one chiral centre. The scatter plot (center) displays the relationship between these values for each comparison (red and blue lines indicate the origin, i.e. no change in either quantity).

3.4 Machine Learned Interatomic Potentials

One of the more straightforward applications of machine learning in CSP is for the prediction of high quality lattice energies, to reduce the cost of geometry optimisations or of the final energy ranking of structures.[69] Such approaches have been demonstrated in molecular organic CSP by training models on the landscapes of individual molecules.[37,39,40] The CSP dataset developed in this work has significant potential for training data-derived models for organic crystals that could be applied more broadly.

3.4.1 Transferable Δ-ML lattice energy corrections

To illustrate the use of CSP to train transferable machine-learned energy models, we trained a committee neural network potential using atom-centred symmetry functions[70,71] for lattice energy corrections to the force field used in this work (FIT+DMA), correcting the lattice energies to the B86bPBE+XDM level. An initial model was trained on 7950 selected crystal structures from ca. 85% of the CSP landscapes (up to 9 crystal structures per landscape), randomly selected from within 8 kJ mol^{-1} of the global energy minimum of each landscape. This corresponds to just under 5% of the crystal structures (166,395) within this energy range for these landscapes. One crystal structure per landscape was withheld as an in-domain test set, while 10 crystal structures per landscape from the remaining CSP landscapes are used as an extrapolation test set. Following initial training, active learning was applied to identify crystal structures from the training landscapes with highest uncertainty in the lattice energy correction. After two iterations (adding 1000 training structures), a slight decrease in errors was observed in the test set, but no improvement in the extrapolation set (see Supporting Information), so training was halted. We also tried a third iteration of active learning with a wider energy window on each landscape, potentially including more diverse structures, but this did not lower the errors on the test or extrapolation sets. Consequently, we decided to proceed with the model after two iterations of active learning.

The performance of the resulting model on the held out test set shows remarkably low errors (Figure 10), returning an MAE of just 0.93 kJ mol^{-1}. Moreover, a similarly low MAE of 1.57 kJ mol^{-1} is achieved on the extrapolation test set, which contained crystal structures of compounds not included in the training of the correction. Compared to the errors for the baseline FIT+DMA, which returned MAEs of 7.80 and 7.95 kJ mol^{-1} on the test set and extrapolation set respectively, the correction offers a marked improvement in accuracy. The fact that the errors for the correction are slightly higher on the extrapolation set shows that there are limitations to the transferability of the correction. We expect that better transferability can be achieved as we increase the number of CSP landscapes available for training; the generation of additional landscapes can be targeted to weaknesses in the underlying force field and molecular types where the current machine learned correction has large errors.

Although the performance of the correction on the test sets is encouraging, an important question is whether the improved energies are significant enough to produce improved stability rankings of organic crystals. To evaluate the influence on ranking, we applied the correction to all of the CSP landscapes, re-ranking all the predicted structures based on the corrected energies. Thereafter, we compared the ranking of the matches to the experimental structures in terms of both their ranking and their energy above the global minimum with that found using FIT+DMA. Structures with predicted uncertainties greater than 25 kJ mol^{-1} were omitted from this analysis. This amounted to 257 structures (out of 3.9 million total) with examination indicating the structures were predominately high energy structures containing voids.

The resulting distributions (Figure 11) over both the training and extrapolation compounds illustrate clearly that the corrected energies in general improve the rankings of the experimental structures. For instance, the correction results in an increase from 424 (FIT+DMA) to 501 (FIT+DMA+Δ-ML) of observed structures ranked as global energy minima, and an increase from 767 (FIT+DMA) to 839 (FIT+DMA+Δ-ML) within 2 kJ mol^{-1} of the global energy minimum (the approximate limit of vibrational contributions to free energy differences). These improvements are observed in the CSP landscapes of the extrapolation and training molecules (see Supporting Information). Importantly, the correction also greatly improves the worst ranked experimental matches with the proportion ranked above 35 in energy ranking decreasing by over a third, from 69 (FIT+DMA) to 45 (FIT+DMA+Δ-ML) (see Supporting Information).

View Article Online
DOI: 10.1039/D4FD00105B

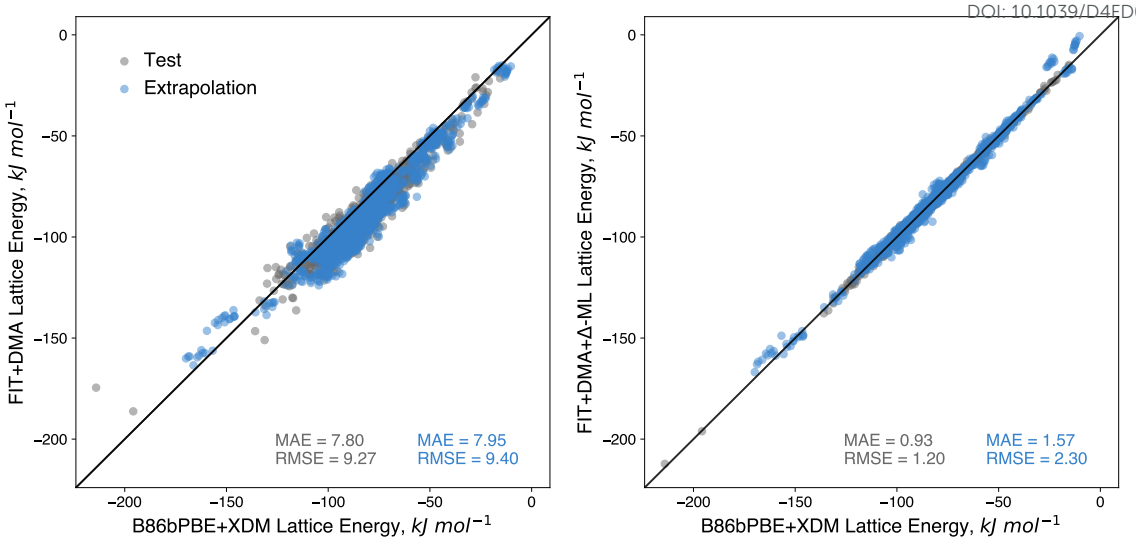


Fig. 10 Correlation of (left) force field (FIT+DMA) lattice energies and (right) force field with machine learned correction vs DFT (B86bPBE+XDM) lattice energies. The test set (grey data points) consists of crystal structures from CSP landscapes of molecules that were seen during training. The extrapolation set consists of predicted crystal structures from molecules withheld from training.

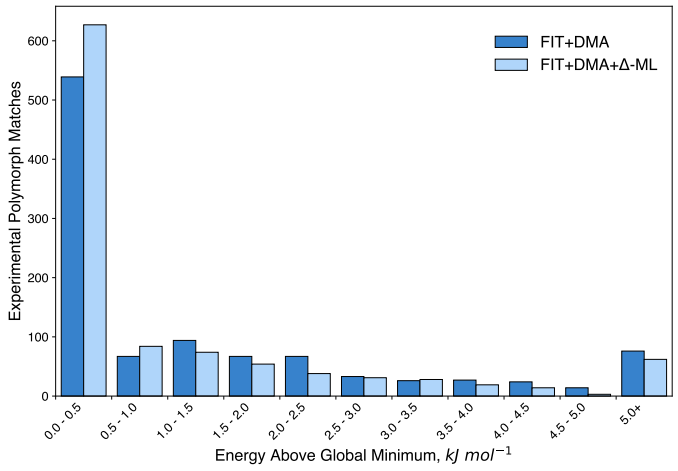


Fig. 11 Histogram of the relative energies of matches to the experimentally-determined crystal structures over all CSPs before (FIT+DMA) and after (FIT+DMA+Δ-ML) applying the data-derived B86bPBE+XDM lattice energy correction. Separate disctributions for molecules from the training and extrapolation sets are shown in the Supporting Information.

3.4.2 Transferable MACE total energy model

While the Δ-ML approach successfully improves the quality of the CSP energy rankings, some of the CSP matches to experimental crystal structures are still relatively high on their energy landscape, even after applying the lattice energy correction. Considering the demonstrated accuracy of the correction, the high relative energies could be the result of limitations with the rigid-body lattice energy minimised geometries, which an energy correction is unable to remedy. Indeed, many of the structures with high relative energies are for compounds which had large geometric deviations between the experimental in-crystal molecular conformations and the gas-phase optimised molecular geometries used (and kept rigid) during CSP. Improving the performance of CSP for these more flexible molecules in our study likely requires re-optimising the predicted crystal structures and relaxing the rigid-molecule constraint.

The FIT+DMA+Δ-ML model is not suitable for this task because it is only a correction to the intermolecular contribution to the lattice energy. Therefore, using a dataset derived from the CSP structures selected for the energy correction model, with perturbed atomic coordinates to sample conformational degrees of freedom, we trained a total energy MACE (higher order equivariant message passing neural network) model. Full details are provided in the Supporting Information. The trained MACE model was then applied to geometry optimise the predicted structures of 15 compounds with large differences in molecular geometry between the observed crystal structure and the DFT-optimised molecule (see Figure 1, bottom three rows).

Re-optimisation with the trained MACE model yielded considerable improvements in the geometric agreement of predicted structures with experiment and of their energy ranking on the CSP landscapes (Table 2). $RMSD_{30}$ between experimental and predicted structures decreased upon re-optimisation for 14 of the 15 compounds, by up to 1.4 Å, and moved the match to the ob-

View Article Online
DOI: 10.1039/D4FD00105B

crystal structure	FIT+DMA RMSD$_{30}$ (Å)	FIT+DMA ΔE (kJ mol^{-1})	MACE RMSD$_{30}$ (Å)	MACE ΔE (kJ mol^{-1})
ATCDEO	0.341	7.26	**0.125 (-0.216)**	**6.67 (-0.59)**
BIRTUR	0.396	6.06	**0.391 (-0.005)**	**0.02 (-6.03)**
BIXKUP	0.813	16.32	**0.208 (-0.605)**	**0.00 (-16.32)**
BIXLOK	1.048	4.95	**0.265 (-0.783)**	**0.00 (-4.95)**
DALBIC	0.343	1.45	**0.206 (-0.137)**	3.70 (+2.25)
DEBFOI	1.208	4.18	**0.110 (-1.098)**	**0.00 (-4.18)**
DOBYOJ	0.654	1.27	**0.640 (-0.014)**	3.55 (+2.28)
JASGIT	1.837	15.77	**1.776 (-0.061)**	**9.72 (-6.05)**
KOGFER	0.247	0.37	0.248 (+0.001)	**0.00 (-0.37)**
QEKQIG	0.824	4.84	**0.139 (-0.684)**	**0.00 (-4.84)**
TAVTUH	1.751	12.03	**0.325 (-1.426)**	**1.87 (-10.16)**
TUNWUV	1.103	9.55	**0.399 (-0.704)**	**0.00 (-9.55)**
UDEXUZ	0.725	7.41	**0.216 (-0.509)**	**0.75 (-6.66)**
WACYAB	0.721	9.81	**0.216 (-0.505)**	**0.92 (-8.89)**
WACYEF	0.975	14.02	**0.200 (-0.775)**	**0.00 (-14.02)**

Table 2 Crystal structure RMSD$_{30}$ and energetic ranking of matches to the experimentally-determined crystal structures within CSP landscapes for 15 molecules re-optimised with the transferable MACE model. Changes between the rigid-molecule force field reslts (FIT+DMA) and MACE are shown in parentheses. Bold entries highlight where MACE re-optimisation leads to improvement.

served structure closer to the global energy minimum in all but 2 cases, with 7 becoming the global energy minimum structure. Figure 12 illustrates the improved geometric agreement for one of these molecules.

Re-optimisation of CSP structures was also run for 4,8b-dihydropyrrolo[3,4-b]indole-1,3(2H,3aH)-dione, where no match to the experimental crystal structure (CSD reference code BUVGAC) was identified in the CSP. However, no match was identified after re-optimisation with the MACE model; in this case, molecular flexibility would have been required during structure generation, rather than post-CSP re-optimisation. Now that a transferable MACE model has been trained, it could potentially be implemented earlier in the CSP workflow.

It should be noted though that the MACE model is a semi-local model and so neglects long-range interactions that can be important for highly accurate relative energies. Overall, the results of the unconstrained optimisations confirm that the high relative energies of these crystal structure matches estimated by FIT+DMA+Δ-ML are largely a result of limitations with the rigid-body lattice energy minimised geometries. It is worth reiterating that the limitations of the CSP geometries only affected a small number of structures; in the vast majority of cases, as shown earlier, the predicted structures at the FIT+DMA level achieved high quality matches to the experimentally-determined crystal structures. Undoubtedly, the success of the FIT+DMA+Δ-ML energy correction is a reflection of the performance of the baseline empirical force field.

4 Conclusions

We have presented (to our knowledge) the largest CSP dataset produced to-date, serving as a computational survey of crystal packing in the organic solid state for small, rigid molecules. Using established, well-characterised CSP methods, we produced crystal structure energy landscapes for over 1000 such molecules, all of which have at least one known, solved crystal structure available in the CSD. In total, our CSP landscapes contain over 4 million unique crystal structures, each with an associated lattice energy at a consistent level of theory.

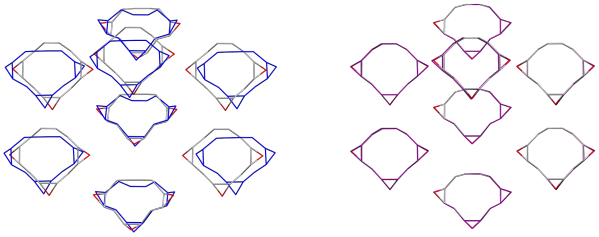


Fig. 12 Overlay of the experimentally determined crystal structure (atoms coloured by element) of (exo,exo,exo)-1,2:4,5:7,8-triepoxycyclododec-10-ene (CSD reference code WACYEF) with (left) the matching structure from the force field (FIT+DMA) CSP (blue) and (right) the matching CSP structure after re-optimisation with the transferable MACE model (purple). Hydrogen atoms are hidden for clarity. The large structural deviation in the FIT+DMA structre is driven by deviation in the molecular geometry.

We have assessed the quality of the dataset by evaluating its reliability at predicting the known crystal structures of these molecules, both in terms of the quality of the geometric match of the crystal structures and the resulting energy ranking of the experimental forms on their respective landscapes compared to other hypothetical structures. Our CSP approach is overwhelmingly successful at predicting crystal structures of these simple molecules – over 99% of all experimental structures have a match located in our CSP searches. 41% of experimental structures are predicted to be the global energy minimum on their landscapes, and 74% are found within 2 kJ/mol of this minimum, a margin equivalent to the estimated error introduced by ignoring thermal effects and ranking based solely on static, 0 K lattice energies. Geometrically, the typical discrepancy between experimental structures and their closest predicted matches is comparable to the thermal fluctuations in experimental crystal structure solutions of the same solid form obtained at low temperatures versus ambient conditions. Such remarkable performance demonstrates the consistency and accuracy of our chosen methods for optimising and ranking these structures.

Such a large dataset of many possible crystal packings should prove a valuable resource for identifying a variety of crystal packing trends, and we make this data available to the community as part of this work. Herein, we studied space group distributions among low-energy hypothetical structures compared to those observed in the CSD, and find substantial overlap, particularly among the most common space groups. We also demonstrated the potential of this varied dataset to explore chirality in the organic solid state, finding very good agreement with established empirical rules concerning the propensity of racemic mixtures to crystallise in racemic crystal structures as opposed to separate enantiopure crystals.

Additionally, we have shown the power of such large-scale CSP to train transferable machine learned potentials for organic solid-state systems. A committee neural network potential trained on single-point periodic DFT lattice energies achieved excellent ac-

View Article Online
DOI: 10.1039/D4FD00105B

curacy in correcting our force field energy landscapes to the DFT level, reducing energy errors by approximately 8 fold. While the NNP performance is reduced on molecules reserved as an extrapolation set compared to those seen in training, the potential still demonstrates improvements to the quality of the resulting CSP rankings increasing the number of experimental structures ranked as the global energy minimum by 18% overall. We further demonstrated the development of a transferable MACE potential using structures derived from the CSP landscapes to allow re-optimisation of crystal structures, testing it successfully on those molecules in our set where the molecular geometry distortion between the conformation in the known crystal structure and the gas-phase-optimised one was largest. The results showed improved structural agreement with experimental structures in almost all cases and much improved energy rankings, moving several poorly-ranked observed structures to the global energy minimum.

Using these mature, well-tested CSP methods alongside modern machine learning approaches, we have demonstrated the ability of CSP to create very large, diverse datasets of hypothetical crystal structures, and the utility of this information in both understanding broad trends in organic crystal structures and in training more advanced energetic models for refinement and transferability. It is our hope that the variety and quantity of CSP data presented here, alongside our demonstrations of possible applications, enables the greater organic solid-state computational community to develop even more sophisticated models and techniques in pursuit of truly predictive, computational data-driven discovery of novel materials.

Author Contributions

Christopher R. Taylor: Conceptualisation, Methodology, Software, Investigation, Validation, Formal analysis, Data curation, Writing - Original Draft, Writing - Review & Editing. Led on production, curation and analysis of large-scale CSP datasets. **Patrick W. V. Butler**: Methodology, Software, Investigation, Validation, Formal analysis, Writing - Original Draft, Writing - Review & Editing. Training of ML energy models and their evaluation. **Graeme M. Day**: Conceptualisation, Validation, Formal analysis, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Funding acquisition.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This project has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement no. 856405, ERC-ADAM). P.W.V.B. is funded by a University of Southampton Presidential Scholarship. The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work. Via our membership of the UK's HEC Materials Chemistry Consortium, which is funded by EPSRC (EP/X035859), this work used the ARCHER2 UK National Supercomputing Service (http://www.archer2.ac.uk).

Notes and references

1 C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, 2016, **72**, 171–179.
2 D. Zagorac, H. Müller, S. Ruehl, J. Zagorac and S. Rehme, *Journal of Applied Crystallography*, 2019, **52**, 918–925.
3 M. M. J. Treacy, I. Rivin, E. Balkovsky, K. H. Randall and M. D. Foster, *Microporous and Mesoporous Materials*, 2004, **74**, 121–132.
4 A. Le Bail, *Journal of Applied Crystallography*, 2005, **38**, 389–395.
5 S. Gražulis, D. Chateigner, R. T. Downs, A. F. T. Yokochi, M. Quirós, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck and A. Le Bail, *Journal of Applied Crystallography*, 2009, **42**, 726–729.
6 A. Vaitkus, A. Merkys and S. Gražulis, *Journal of Applied Crystallography*, 2021, **54**, 661–672.
7 A. Vaitkus, A. Merkys, T. Sander, M. Quirós, P. A. Thiessen, E. E. Bolton and S. Gražulis, *Journal of Cheminformatics*, 2023, **15**,.
8 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Materials*, 2013, **1**, 011002.
9 J. Nyman and G. M. Day, *Phys. Chem. Chem. Phys.*, 2016, **18**, 31132–31143.
10 J. Nyman and G. M. Day, *CrystEngComm*, 2015, **17**, 5154–5165.
11 A. J. Cruz-Cabeza and J. Bernstein, *Chemical Reviews*, 2014, **114**, 2170–2191.
12 A. J. Cruz-Cabeza, S. M. Reutzel-Edens and J. Bernstein, *Chemical Society Reviews*, 2015, **44**, 8619–8635.
13 P. Sacchi, P. Neoptolemou, R. J. Davey, S. M. Reutzel-Edens and A. J. Cruz-Cabeza, *Chemical Science*, 2023, **14**, 11775–11789.
14 H. P. G. Thompson and G. M. Day, *Chemical Science*, 2014, **5**, 3173–3182.
15 H. C. S. Chan, J. Kendrick, M. A. Neumann and F. J. J. Leusen, *CrystEngComm*, 2013, **15**, 3799–3807.
16 C. R. Taylor and G. M. Day, *Crystal Growth & Design*, 2018, **18**, 892–904.
17 L. M. LeBlanc, S. G. Dale, C. R. Taylor, A. D. Becke, G. M. Day and E. R. Johnson, *Angewandte Chemie International Edition*, 2018, **57**, 14906–14910.
18 J. Hoja and A. Tkatchenko, *Faraday Discussions*, 2018, **211**, 253–274.
19 G. M. Day, *Crystallography Reviews*, 2011, **17**, 3–52.
20 G. J. O. Beran, *Chemical Science*, 2023, **14**, 13290–13312.
21 S. Yang and G. M. Day, *Commun Chem*, 2022, **5**, 1–13.
22 N. F. Francia, L. S. Price, J. Nyman, S. L. Price and M. Salvalaglio, *Crystal Growth & Design*, 2020, **20**, 6847–6862.
23 P. W. V. Butler and G. M. Day, *Proceedings of the National Academy of Sciences*, 2023, **120**, e2300516120.

View Article Online
DOI: 10.1039/D4FD00105B

24 A. Pulido, L. Chen, T. Kaczorowski, D. Holden, M. A. Little, S. Y. Chong, B. J. Slater, D. P. McMahon, B. Bonillo, C. J. Stackhouse, A. Stephenson, C. M. Kane, R. Clowes, T. Hasell, A. I. Cooper and G. M. Day, *Nature*, 2017, **543**, 657–664.

25 J. E. Campbell, J. Yang and G. M. Day, *J. Mater. Chem. C*, 2017, **5**, 7574–7584.

26 G. M. Day and A. I. Cooper, *Advanced Materials*, 2018, **30**, 1704944.

27 C. Zhao, L. Chen, Y. Che, Z. Pang, X. Wu, Y. Lu, H. Liu, G. M. Day and A. I. Cooper, *Nature Communications*, 2021, **12**, 817.

28 D. H. Case, J. E. Campbell, P. J. Bygrave and G. M. Day, *Journal of chemical theory and computation*, 2016, **12**, 910–924.

29 P. G. Karamertzanis and C. C. Pantelides, *Journal of Computational Chemistry*, 2005, **26**, 304–324.

30 F. Curtis, X. Li, T. Rose, A. Vásquez-Mayagoitia, S. Bhattacharya, L. M. Ghiringhelli and N. Marom, *Journal of Chemical Theory and Computation*, 2018, **14**, 2246–2264.

31 M. A. Neumann, J. v. d. Streek, F. P. A. Fabbiani, P. Hidber and O. Grassmann, *Nature Communications*, 2015, **6**, 1–7.

32 D.-K. Bučar, G. M. Day, I. Halasz, G. G. Z. Zhang, J. R. G. Sander, D. G. Reid, L. R. MacGillivray, M. J. Duer and W. Jones, *Chemical Science*, 2013, **4**, 4417–4425.

33 C. R. Taylor, M. T. Mulvee, D. S. Perenyi, M. R. Probert, G. M. Day and J. W. Steed, *Journal of the American Chemical Society*, 2020, **142**, 16668–16680.

34 D. H. Case, V. K. Srirambhatla, R. Guo, R. E. Watson, L. S. Price, H. Polyzois, J. K. Cockcroft, A. J. Florence, D. A. Tocher and S. L. Price, *Crystal Growth & Design*, 2018, **18**, 5322–5331.

35 A. M. Reilly, R. I. Cooper, C. S. Adjiman, S. Bhattacharya, A. D. Boese, J. G. Brandenburg, P. J. Bygrave, R. Bylsma, J. E. Campbell, R. Car, D. H. Case, R. Chadha, J. C. Cole, K. Cosburn, H. M. Cuppen, F. Curtis, G. M. Day, R. A. DiStasio Jr, A. Dzyabchenko, B. P. van Eijck, D. M. Elking, J. A. van den Ende, J. C. Facelli, M. B. Ferraro, L. Fusti-Molnar, C.-A. Gatsiou, T. S. Gee, R. de Gelder, L. M. Ghiringhelli, H. Goto, S. Grimme, R. Guo, D. W. M. Hofmann, J. Hoja, R. K. Hylton, L. Iuzzolino, W. Jankiewicz, D. T. de Jong, J. Kendrick, N. J. J. de Klerk, H.-Y. Ko, L. N. Kuleshova, X. Li, S. Lohani, F. J. J. Leusen, A. M. Lund, J. Lv, Y. Ma, N. Marom, A. E. Masunov, P. McCabe, D. P. McMahon, H. Meekes, M. P. Metz, A. J. Misquitta, S. Mohamed, B. Monserrat, R. J. Needs, M. A. Neumann, J. Nyman, S. Obata, H. Oberhofer, A. R. Oganov, A. M. Orendt, G. I. Pagola, C. C. Pantelides, C. J. Pickard, R. Podeszwa, L. S. Price, S. L. Price, A. Pulido, M. G. Read, K. Reuter, E. Schneider, C. Schober, G. P. Shields, P. Singh, I. J. Sugden, K. Szalewicz, C. R. Taylor, A. Tkatchenko, M. E. Tuckerman, F. Vacarro, M. Vasileiadis, A. Vazquez-Mayagoitia, L. Vogt, Y. Wang, R. E. Watson, G. A. de Wijs, J. Yang, Q. Zhu and C. R. Groom, *Acta Crystallographica Section B Structural Science, Crystal Engineering and Materials*, 2016, **72**, 439–459.

36 J. Behler and G. Csányi, *The European Physical Journal B*, 2021, **94**, 142.

37 F. Musil, S. De, J. Yang, J. E. Campbell, G. M. Day and M. Ceriotti, *Chem. Sci.*, 2018, **9**, 1289–1300.

38 D. McDonagh, C.-K. Skylaris and G. M. Day, *Journal of Chemical Theory and Computation*, 2019, **15**, 2743–2758.

39 O. Egorova, R. Hafizi, D. C. Woods and G. M. Day, *J. Phys. Chem. A*, 2020, **124**, 8065–8078.

40 P. W. V. Butler, R. Hafizi and G. M. Day, *The Journal of Physical Chemistry A*, 2024, **128**, 945–957.

41 V. Kapil and E. A. Engel, *Proceedings of the National Academy of Sciences*, 2022, **119**, e2111769119.

42 D. P. Kovács, J. H. Moore, N. J. Browning, I. Batatia, J. T. Horton, V. Kapil, W. C. Witt, I.-B. Magdău, D. J. Cole and G. Csányi, *MACE-OFF23: Transferable Machine Learning Force Fields for Organic Molecules*, 2023, `http://arxiv.org/abs/2312.15211`, arXiv:2312.15211 [physics].

43 F. M. Paruzzo, A. Hofstetter, F. Musil, S. De, M. Ceriotti and L. Emsley, *Nature Communications*, 2018, **9**, 4501.

44 P. T. A. Galek, L. Fábián, W. D. S. Motherwell, F. H. Allen and N. Feeder, *Acta Crystallographica Section B: Structural Science*, 2007, **63**, 768–782.

45 R. K. Cersonsky, M. Pakhnova, E. A. Engel and M. Ceriotti, *Chemical Science*, 2023, **14**, 1272–1285.

46 J. Yang, S. De, J. E. Campbell, S. Li, M. Ceriotti and G. M. Day, *Chemistry of Materials*, 2018, **30**, 4361–4371.

47 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, *Gaussian09 Revision D.01*.

48 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.

49 C. Adamo and V. Barone, *The Journal of Chemical Physics*, 1999, **110**, 6158–6170.

50 K. Raghavachari, J. S. Binkley, R. Seeger and J. A. Pople, *The Journal of Chemical Physics*, 1980, **72**, 650–654.

51 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *The Journal of Chemical Physics*, 2010, **132**, 154104.

52 A. D. Becke and E. R. Johnson, *The Journal of Chemical Physics*, 2005, **123**, 154101.

53 A. J. Stone and M. Alderton, *Molecular Physics*, 2002, **100**, 221–233.

54 A. J. Stone, *Journal of Chemical Theory and Computation*, 2005, **1**, 1128–1132.

View Article Online
DOI: 10.1039/D4FD00105B

55 P. J. Winn, G. G. Ferenczy and C. A. Reynolds, *The Journal of Physical Chemistry A*, 1997, **101**, 5437–5445.

56 G. G. Ferenczy, P. J. Winn and C. A. Reynolds, *The Journal of Physical Chemistry A*, 1997, **101**, 5446–5455.

57 D. S. Coombes, S. L. Price, D. J. Willock and M. Leslie, *The Journal of Physical Chemistry*, 1996, **100**, 7352–7360.

58 S. L. Price, M. Leslie, G. W. A. Welch, M. Habgood, L. S. Price, P. G. Karamertzanis and G. M. Day, *Phys. Chem. Chem. Phys.*, 2010, **12**, 8478–8490.

59 D. E. Williams and D. J. Houpt, *Acta Crystallographica Section B*, 1986, **42**, 286–295.

60 J. A. Chisholm and S. Motherwell, *J. Appl. Crystallogr.*, 2005, **38**, 228–231.

61 *RDKit: Open-source cheminformatics.*, `https://www.rdkit.org`.

62 C. Peifer, M. Tschertsche, D. Schollmeyer and S. Laufer, *Acta Crystallographica Section E*, 2007, **63**, o1359–o1360.

63 C. Wilson, *Zeitschrift für Kristallographie - Crystalline Materials*, 2000, **215**, 693–701.

64 M. J. Buerger, *Science*, 1962, **135**, 912–912.

65 C. M. Aitchison, C. M. Kane, D. P. McMahon, P. R. Spackman, A. Pulido, X. Wang, L. Wilbraham, L. Chen, R. Clowes, M. A. Zwijnenburg, R. S. Sprick, M. A. Little, G. M. Day and A. I. Cooper, *J. Mater. Chem. A*, 2020, **8**, 7158–7170.

66 C. P. Brock and J. D. Dunitz, *Chemistry of Materials*, 1994, **6**, 1118–1127.

67 O. Wallach, *Justus Liebigs Annalen der Chemie*, 1895, **286**, 90–118.

68 C. P. Brock, W. B. Schweizer and J. D. Dunitz, *Journal of the American Chemical Society*, 1991, **113**, 9811–9820.

69 R. J. Clements, J. Dickman, J. Johal, J. Martin, J. Glover and G. M. Day, *MRS Bulletin*, 2022, **47**, 1054–1062.

70 J. Behler, *The Journal of Chemical Physics*, 2011, **134**, 074106.

71 M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi and P. Marquetand, *The Journal of Chemical Physics*, 2018, **148**, 241709.