Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue



Cite this: *Chem. Sci.*, 2023, **14**, 1272

All publication charges for this article have been paid for by the Royal Society of Chemistry

# A data-driven interpretation of the stability of organic molecular crystals†

Rose K. Cersonsky, *[a] Maria Pakhnova, [a] Edgar A. Engel [b] and Michele Ceriotti [a]

Due to the subtle balance of intermolecular interactions that govern structure–property relations, predicting the stability of crystal structures formed from molecular building blocks is a highly non-trivial scientific problem. A particularly active and fruitful approach involves classifying the different combinations of interacting chemical moieties, as understanding the relative energetics of different interactions enables the design of molecular crystals and fine-tuning of their stabilities. While this is usually performed based on the empirical observation of the most commonly encountered motifs in known crystal structures, we propose to apply a combination of supervised and unsupervised machine-learning techniques to automate the construction of an extensive library of molecular building blocks. We introduce a structural descriptor tailored to the prediction of the binding (lattice) energy and apply it to a curated dataset of organic crystals, exploiting its atom-centered nature to obtain a data-driven assessment of the contribution of different chemical groups to the lattice energy of the crystal. We then interpret this library using a low-dimensional representation of the structure–energy landscape and discuss selected examples of the insights into crystal engineering that can be extracted from this analysis, providing a complete database to guide the design of molecular materials.

Received 9th November 2022
Accepted 6th December 2022

DOI: 10.1039/d2sc06198h

rsc.li/chemical-science



## 1 Introduction

Understanding molecular crystallization is critical to many fields of chemical sciences – from anticipating pharmaceutical stability and solubility,[1–5] to preventing[6] or fostering[7] aggregation in organic electronics, to understanding complex formation in biological macromolecules.[8,9]

Yet, molecular crystallization is a complex process that involves multiple cooperative and competing forces. Initial nucleation is typically motivated by strong interactions between functional groups.[10,11] The structural patterns associated with these guiding interactions (deemed "supramolecular synthons") and their hierarchies are often the focus of experimental and computational studies in crystal structure prediction.[12,13] Nevertheless, once molecules have moved within closer range, many factors, including weaker interactions, the expulsion of solvent molecules, and geometric packing, will then determine the short- and potentially long-range order, leading to many potentially-stable polymorphs for a given stoichiometry. In the past decades, there has been a growing push to develop a "holistic" view of molecular crystallization,[14,15] not only taking into account the nearest-neighbor contacts but also the interplay of these interactions with other components of the molecular assembly.

While it is simpler to rationalize single-site interactions, the interplay of many competing interactions necessitates diverse, high-throughput studies.[14] Thus, molecular crystallization has emerged as a hotbed for computational inquiry. This focus has led to considerable theoretical and software developments for qualitative and quantitative analyses, including those tailored to crystal structure prediction (CSP)[16–19] and the representation of electrostatic surfaces and molecular geometry.[20,21] Even more recently, machine learning has been used to understand the individual configurational and energy landscapes of molecules;[22–28] however, such techniques have yet to be applied in the general, holistic vein required to extract the qualitative insights that can be used to support crystal design efforts.

To study molecular crystallization in this broad lens, we have curated a dataset of roughly 3260 C + H + N + O + S-containing molecular crystals from those reported in Cordova *et al.*[29] In Cordova *et al.*,[29] these crystals were initially selected by querying the Cambridge Structural Database (CSD) to identify a diverse set of synthesizable molecular assemblies, including some that had been experimentally stabilized at extreme conditions. The experimental properties of the full dataset are summarized in ESI Appendix A3.†

The stability of molecular crystals is traditionally studied through the binding (lattice) energy, which is computationally

[a]*Laboratory of Computational Science and Modeling (COSMO), École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. E-mail: Rose.Cersonsky@wisc.edu*

[b]*TCM Group, Trinity College, Cambridge University, Cambridge, UK*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d2sc06198h

determined by evaluating the ground-state energies for both the crystal and its molecular components in the dilute gas limit, here computed using DFT-PBE-D2 calculations of each crystal and its relaxed molecular components at ambient conditions.

From here, we build an atom-centered regression model for this lattice energy, demonstrating the improvements in accuracy and reduction in model complexity from using a physics-informed approach. This atom-centered approach, wherein we represent each molecular crystal using the average ML descriptor for each of its atomic constituents, facilitates estimating the contribution of each atom, or group of atoms, to the binding energy. Then, employing a combination of supervised and unsupervised machine learning models, we can determine and interpret each molecular moiety's intermolecular interactions. Using these approaches, we show how physically-motivated machine learning models can not only "rediscover" the known maxims of crystal engineering, but provide insight and guidance for crystal design. We have made our datasets, and analyses openly-available through the Materials Cloud,[31] with interactive components aimed to guide future molecular design and narrower or targeted studies.

# 2 Notation

In this study, we employ atom-centered descriptors[32] to identify the contributions of specific collections of atoms to the binding of a crystal. Given the many atomic and energetic entities (atoms, molecules, crystals, total energy *versus* lattice energy), we rely on many numerical representations and equations; hence we start by establishing a consistent notation we will use throughout the text.

## 2.1 Descriptors

To reflect the physics of atomic interactions, we use symmetry-adapted descriptors to encode/describe the geometric arrangement of atoms in their atomistic configurations, specifically the 3-body SOAP descriptors outlined in ESI Appendix B1.† Each of these input descriptors is written as $x_\sigma^{(\mathrm{i})}$, where the subscript σ signifies the collection of atoms being described, including the entire crystal (c), a molecule (m), or an atom (a). $n_\sigma$ is the number of atoms in the given collection. Thus it follows that $n_\mathrm{c}$ is the number of atoms in a given crystal, and $n_\mathrm{m}$ is the number of atoms in a given molecule. Because we discuss analogous atoms or molecules in both the solid and gas phases, we use the superscript (i) to denote the phase (crystalline solid (s) or dilute gas (g)).

The descriptor for a given collection should be assumed as the average of the descriptors for the constituent atoms:

$$x_\sigma^{(\mathrm{i})} = \frac{1}{n_\sigma}\sum_{\mathrm{a}\in\sigma} x_\mathrm{a}^{(\mathrm{i})} \tag{1}$$

For example, the descriptor for the atoms in a molecule in a dilute gas is $x_\mathrm{m}^{(\mathrm{g})} = \frac{1}{n_\mathrm{m}}\sum_{\mathrm{a}\in\mathrm{m}} x_\mathrm{a}^{(\mathrm{g})}$. If we were to look at the same molecule in the crystalline solid we would get $x_\mathrm{m}^{(\mathrm{s})} = \frac{1}{n_\mathrm{m}}\sum_{\mathrm{a}\in\mathrm{m}} x_\mathrm{a}^{(\mathrm{s})}$. A schematic of these concepts is shown in Fig. 1, using the co-crystal 5-aminotetrazole monohydrate (CSD ref. AMTETZ[30]) as an example.

## 2.2 Energies and regressions

We use $E_\sigma$ to denote the total energy of a collection of atoms. In this study, the total energies of the crystals $E_\mathrm{c}$ are taken from those reported in Cordova *et al.*[29] and the total energies of the molecules $E_\mathrm{m}$ are determined by DFT-PBE-D2 calculations, as described in ESI Appendix A2.† We use $e_\sigma \equiv E_\sigma/n_\sigma$ to indicate the per-atom energy. Note that we express all energies in kJ mol$^{-1}$, where $e_\sigma$ is to be interpreted as having the units of kJ per mol of atoms. Constructing a linear regression amounts to the ansatz.

$$e_\sigma = x_\sigma w_\sigma + \varepsilon_\sigma \tag{2}$$

where $w_\sigma$ is the regression weights and $\varepsilon_\sigma$ the residual errors. The lattice energy (also referred to as the binding or cohesive energy in literature) of a molecular crystal is given by $\Delta_\mathrm{c}$, where

$$\Delta_\mathrm{c} = E_\mathrm{c} - \sum_{\mathrm{m}\in\mathrm{c}} E_\mathrm{m} \tag{3}$$

With the average lattice energy per atom given by,

$$\delta_\mathrm{c} \equiv \Delta_\mathrm{c}/n_\mathrm{c} = e_\mathrm{c} - \sum_{\mathrm{m}\in\mathrm{c}} \frac{n_\mathrm{m}}{n_\mathrm{c}} e_\mathrm{m} \tag{4}$$

Later, we will use our regression model to determine the atomic contributions to the lattice energy, which we will denote $\delta_\mathrm{a}$, where $\delta_\mathrm{c} = \frac{1}{n_\mathrm{c}}\sum_{\mathrm{a}\in\mathrm{c}}\delta_\mathrm{a}$. We will also consider the contributions for different collection of atoms, and will denote the average lattice energy contribution as $\delta_\sigma = \frac{1}{n_\sigma}\sum_{\mathrm{a}\in\sigma}\delta_\mathrm{a}$. When we regularize these contributions using a Gaussian filter (discussed in Section 3.2 and Appendix B2†), we will use a tilde to give $\tilde{\delta}$.

# 3 Results and discussion

In the following, we consider crystals and gas-phase molecules, both of which have been geometry-optimized by minimizing their configurational energies with respect to the atomic positions, as described in Appendix A.† Unless stated otherwise, we use as our featurization the 3-body SOAP vectors (as described in Appendix B1†) and build a regularized ridge regression models using scikit-learn.[33] All models were trained on the same training set of 2707 crystals (or the corresponding 3242 molecules). We report errors on a mutually-exclusive set of 551 crystals (or the corresponding 628 molecules). When interpreting the results, it is important to consider that the test set has been selected at random, and is therefore representative of the makeup of the CSD, while the training structures were selected with farthest point sampling[34,35] to maximize the diversity, and therefore contain a large fraction of unstable, "extreme" cases (Table S1†).

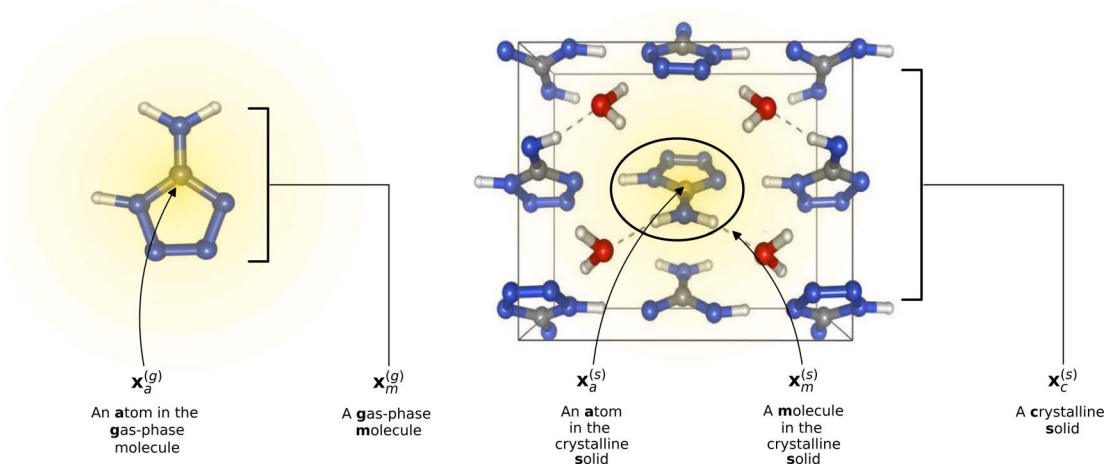


Fig. 1 Visualization of descriptor notation, as described in Section 2.1, visualized for 5-aminotetrazole monohydrate (CSD ref. AMTETZ[30]). Each descriptor contains the information of an atom and its neighborhood (shown in yellow shading).

### 3.1 Building a model for the lattice energy

One can estimate the atomic contributions to a target property (and thereby assess the contributions of specific molecular motifs) by building a robust machine learning model on an atom-centered descriptor.[36] Suppose we have a descriptor

$$x_\sigma = \frac{1}{n_\sigma}\sum_{a\in\sigma} x_a \tag{5}$$

and train a regression model on some target $y$ such that $y = x_\sigma w + \varepsilon$, where $w$ is the regression weight and $\varepsilon$ is the residual error from the regression. We can then estimate the approximate contribution of each atom by computing $y_a = x_a w$.

**3.1.1 Combining models of $e_c$ and $e_m$.** Given eqn (4), it is possible to build a model for the lattice energy from two separate models for crystal and molecular energy, replacing each energy $e$ with its approximation *via* linear regression (eqn (2))

$$\delta_c = x_c^{(s)} w_c + \varepsilon_c - \sum_{m\in c}\frac{n_m}{n_c}\left(x_m^{(g)} w_m + \varepsilon_m\right). \tag{6}$$

Eqn (6) may then be rewritten as:

$$\delta_c = x_c^{(s)} w_c - x_c^{(g)} w_m + \varepsilon \tag{7}$$

where we have defined $\varepsilon \equiv \varepsilon_c - \sum_{m\in c}\frac{n_m}{n_c}\varepsilon_m$. In this scheme, the regression of the lattice energy is implicitly limited by the errors of the independent regressions; therefore, if we obtained a good fit for $e_c$ and $e_m$, this should be a fairly robust way to predict the lattice energy.

When we predict the crystal and molecular atomic energies $e_c$ and $e_m$, we obtain RMSEs of 1.15 kJ mol$^{-1}$ and 0.727 kJ mol$^{-1}$, respectively, which are acceptably small compared to the intrinsic variance of the baselined‡ target energies of the test set, which have standard deviations of 4.402 kJ mol$^{-1}$ and 4.251 kJ mol$^{-1}$, respectively. However, the intrinsic variance of the lattice energies is smaller (1.965 kJ mol$^{-1}$); therefore, the resulting RMSE of 0.916 kJ mol$^{-1}$ from eqn (6) is unsatisfactory and suggests that the errors in the independent regressions generally overlap with the lattice energy contributions.

**3.1.2 Building a model directly on $\delta_c$.** With the reduced variance of the target (lattice energy), it thus makes sense to construct the regression model directly on our target. Building a regression on the gas-phase descriptors $x_c^{(g)}$, while conceptually nonsensical (the gas-phase descriptors of the molecules contain no information on the intermolecular interactions), yields an RMSE of 1.101 kJ mol$^{-1}$. Regressing on the solid-phase descriptors $x_c^{(s)}$ improves the regression substantially, achieving an RMSE of 0.778 kJ mol$^{-1}$.

Yet, conceptually, neither of these two representations ($x_c^{(s)}$ and $x_c^{(g)}$) contain the full set of relevant information – the molecular descriptor $x_c^{(g)}$ is missing information on intermolecular interactions, and the crystal descriptor $x_c^{(s)}$ is unaware of the conformational changes that the molecules undergo upon crystallization. The necessity of this missing information is confirmed when we regress on concatenated descriptors $\{x_c^{(s)}, x_c^{(g)}\}$ and our RMSE drops to 0.671 kJ mol$^{-1}$.

Furthermore, eqn (7) provides another way to similarly (and more explicitly) encode the nature of the problem into our choice of representation. Given a descriptor that appropriately distinguishes between periodic crystals and molecules, a regression model can predict their energies using the same regression weights, $w = w_c = w_m$. Substituting this into eqn (7),

$$\delta_c = x_c^{(s)} w - x_c^{(g)} w + \varepsilon \tag{8}$$

$$\delta_c = x_c^{(s-g)} w + \varepsilon \tag{9}$$

where we define $x_c^{(s-g)} \equiv (x_c^{(s)} - x_c^{(g)})$ as the so-called "remnant" descriptor and $\varepsilon$ again denotes the residual errors. Explicitly adapting our representation $x_c^{(s-g)}$ to the nature of the lattice energy results in a yet better result to learning on $\{x_c^{(s)}, x_c^{(g)}\}$: 0.571 kJ mol$^{-1}$, despite being in a smaller feature space. Conceptually, this descriptor still encodes the 3-body correlation between an environment and its neighbors but

explicitly incorporates the change in molecular geometry upon crystallization and reduces the weights of atomic triplets whose interactions are primarily intramolecular and/or the same in gas and solid phase.

**3.1.3 Extension to non-linear models.** This result is mirrored in non-linear regression models, where again, a superior result is obtained by either constructing a kernel on $x_c^{(s-g)}$ or taking the difference of non-linear feature vectors (see ESI Appendix C2†). An optimized RBF kernel on the remnant descriptors yields a similar RMSE to the linear model, likely due to the restricted dataset size and its diversity. We get some improvement (by ~0.06 kJ mol$^{-1}$ compared to the best linear model) by taking the difference of the non-linear features defined by the kernels of the crystalline and molecular descriptors. This result further emphasizes the rationale behind the remnant approach, and suggests that one can, more generally, improve accuracy by combining non-linear feature constructions to mimic the mathematical formulation of target properties. Kernel hyperparameter optimization has little impact on this conclusion, as we demonstrate with corroborating results using a parameter-free kernel, also in ESI Appendix C2.†

When the molecular geometry is known *a priori*, these results suggest that linear (summarized in Table 1) and non-linear regressions (ESI Appendix C2†) for the lattice energy should be built on descriptors conceptually akin to $x_c^{(s-g)}$, rather than $x_c^{(s)}$, as has been common practice in the literature.[23,25,27,28] Thus, in the remainder of the text, we will employ ML fingerprints and models based on the remnant descriptor.

### 3.2 Estimating the contributions of molecular motifs

**3.2.1 Regularizing the atomic contributions.** With our target-adapted regression model, we can assign effective contributions to each atomic environment, where we take the remnant descriptor of each atomic environment and compute

$$\delta_a = x_a^{(s-g)} w \tag{10}$$

Despite the mathematical logic behind this step, the lack of physical underpinnings for this decomposition may result in energy being arbitrarily partitioned between neighboring atoms. This leads to disproportionately large contributions of opposite size being assigned, not dissimilar to how a regression may overfit by assigning large regression weights. To ease this effect, we can apply a Gaussian filter to each $\delta_a$. For the $i^{th}$ atom, this results in

$$\tilde{\delta}_i = \sum_j \delta_j \frac{f(i,j)}{f(j,k)} \tag{11}$$

where $\sum_j$ runs over all neighbors of $i$ and $\sum_k$ runs over all neighbors of $j$ (defined by a cutoff of 2 Å). For neighbors $a$ and $b$ and interatomic distance $d_{ab}$, $f(a, b) = \exp[d_{ab}^2/2\varsigma^2]$. This procedure, introduced for the electronic density of states in Ben Mahmoud *et al.*,[37] has the effect of regularizing the decomposition without changing the regression results, *i.e.*, $\delta_c = \frac{1}{n_c}\sum_{a\in c}\delta_a = \frac{1}{n_c}\sum_{a\in c}\tilde{\delta}_a$. We show this effect of the filter on the distribution of atomic contributions in ESI Appendix B2.† It is worthwhile to conceptually compare our data-driven decomposition with one based on an empirical model of interactions, or with one of the many atoms-in-molecules decompositions of the energy computed by quantum-chemical calculations. On one hand, our approach makes it harder to explicitly interpret the stabilizing power of a motif in terms of physical terms (electrostatics, dispersion…). On the other, in many cases force fields and energy decompositions have a high degree of arbitrariness, and the accurate prediction of the total binding energy comes from a cancellation of errors in the individual components. The atomic contributions eqn (11) are obtained with the only requirement of being smooth, and (since they are built using a remnant descriptor) to correlate with the structural features associated with the crystal-forming process. As we shall see, their nature allows one to recognize the role played by collective effects – such as steric hindrance, or molecular distortions – contributing to our goal of a holistic view of lattice stability.

**3.2.2 Visualizing the contributions of different motifs.** Taking the 3242 molecules from our training set, we use SMARTS descriptors[38] and RDKit Substructure Matching[39] to identify the atoms belonging to common molecular motifs, finding 46 010 motifs. Details of this procedure and our table of SMARTS strings are given in ESI Appendix B3 and Table S3,† respectively. For each motif, we determine the effective cohesive interaction $\tilde{\delta}_\sigma$ as§

$$\tilde{\delta}_\sigma = \frac{1}{n_\sigma}\sum_{a\in\sigma}\tilde{\delta}_a \tag{12}$$

We plot the span of lattice energy contributions for motifs with greater than 200 instances in the dataset in Fig. 2. The functional groups are arranged in order of increasing average cohesive interactions. Nearly all functional groups, on average, are stabilizing, although we see a clear trend in the nature of the functional groups from left to right. On the left (the motifs leading to the strongest intermolecular interactions), there are groups typically associated with hydrogen bonding (*e.g.*, carboxyls and waters). As we move to the right, the molecular motifs are, on average, weakly binding, with the largest range of

Table 1 Results of linear regression exercises. In each linear regression, an independent, 5-fold cross-validated model was build on 2707 crystals (or the 3242 coinciding molecules)[a]

| Regression equation | Eqn | RMSE | MAE |
|---|---|---|---|
| $e_c = x_c^{(s)} w_c$ | (2) | 1.15 | 0.863 |
| $e_m = x_m^{(g)} w_m$ | (2) | 0.727 | 0.563 |
| $\delta_c = x_c^{(s)} w_c - \sum_{m\in c}\frac{n_m}{n_c}(x_m^{(g)} w_m)$ | (6) | 0.916 | 0.652 |
| $\delta_c = x_c^{(s)} w$ | | 0.778 | 0.552 |
| $\delta_c = x_c^{(g)} w$ | | 1.101 | 0.723 |
| $\delta_c = x_c^{(s-g)} w$ | (9) | 0.571 | 0.404 |
| $\delta_c = \{x_c^{(s)}, x_c^{(g)}\} w$ | | 0.671 | 0.461 |

[a] Here we report the errors (in kJ mol$^{-1}$) on a separate set of 551 crystals (or the coinciding 628 molecules). In each regression equation $w$ is unique to that regression.

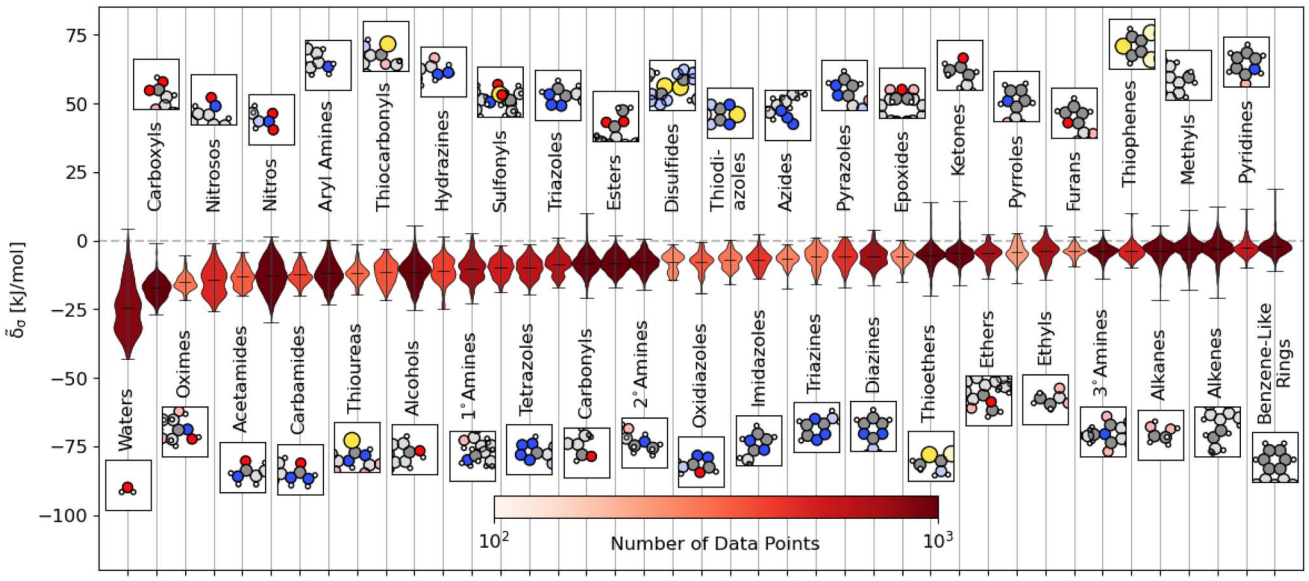


Fig. 2 Distribution of energetic contributions for different functional groups. For each functional group, we have taken the averaged remnant descriptor $x_\sigma^{(s-g)}$ and computed the estimated contribution to the binding energy $\tilde{\delta}_\sigma$ using the regressions detailed in eqn (10) and the filtering procedure in eqn (11). We have arranged the functional groups in order of average contribution, with a representative example is shown above or below the violin plot with the functional group highlighted. We have limited this figure to those functional groups with more than 200 instances in the dataset (see Fig. S4† for all groups). The lines on each plot denote each group's extreme and mean contributions. The plots are colored by the number of examples within the dataset, ranging from 4 (pentazole) to 5313 (methyl groups). Wider sections of the violin plot represent a higher probability that members of the population will take on the given value; the skinnier sections represent a lower probability.

interactions coming from the most broadly-defined groups, including the alkanes, alkenes, and benzene-like rings.

This trend is further demonstrated by plotting the structure–property map of all motifs using Principal Covariates Regression (PCovR), a hybrid supervised-unsupervised dimensionality reduction technique first introduced in De Jong and Kiers[40] and adapted to chemical systems in Helfrecht *et al.*[41] This technique produces a latent-space mapping that arranges different motif classes based on their structural similarity and correlation to a set of target properties. In Fig. 3, we show a map using the average remnant descriptor for each motif and their average energy contribution, using contour lines to show where 90% of such motifs fall on the PCovR map. One sees that, in this case, the first axis of this plot ($PCov_1$) correlates strongly with the (learnable) cohesive interactions. The second axis ($PCov_2$) allows us to resolve structural differences between motifs with similar energetic contributions. In this mapping, we can learn from the spread of each group. For example, the 868 water molecules (light blue in Fig. 3) span the greater portion of the left-hand side of the figure, highlighting the chemical diversity of intermolecular water interactions. Juxtapose this with the 2627 nitro and nitroso groups (pink in Fig. 3) that span a smaller region in PCovR space, implying a narrower range of intermolecular interactions. Here we have combined several groups for visual simplicity; however, we have included plots highlighting each functional group in Fig. S6–S9,† including the sample sizes and range of contributions.

The PCovR framework also provides a blueprint for analyzing the interactions of different structural motifs – given a single motif type, what characteristics of a molecular environment lead to a more stabilizing interaction? In the following sections, we will take a look at the stabilizing environments for a few classes of functional groups, starting with

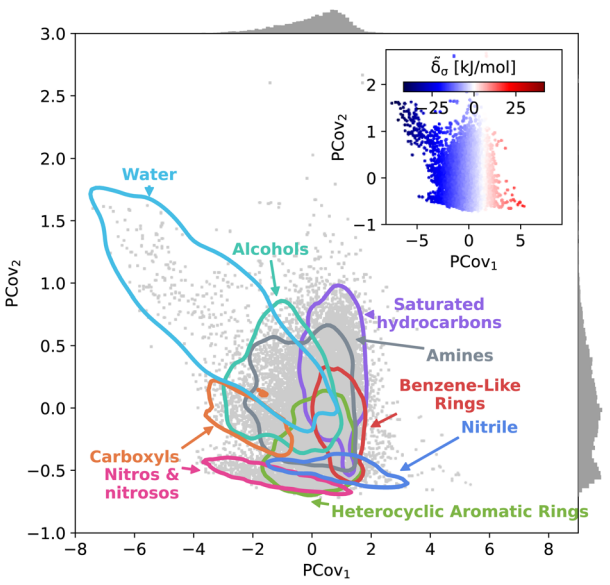


Fig. 3 Principal Covariates Regression (PCovR) map of the interactions of molecular motifs. A structure–property map of molecular motifs, denoting major classes of motifs and outlining the regions where the 90$^{th}$ percentile of these motifs occur. (Inset) the same map colored by the cohesive interactions, ranging from blue (strongly attractive) to white (neutral) to red (strongly resistant). Histograms on the upper and right borders show the distribution of motifs along the covariates.

the well-known stabilizing interactions of water and carboxylic groups, then moving onto two groups with a wide range of intermolecular interactions, 6-membered aromatic carbon rings and nitro groups. With each functional group, we generate a new PCovR using only the averaged remnant descriptors and effective interactions for the instances of that group, such that the structural diversity embedded in the map reflects the diversity of interactions, rather than the diversity of the molecules. We have included similar maps for all other molecular motifs in an online data repository.[31,48]

**3.2.3 Waters.** We begin with a ubiquitous molecular crystal stabilizer: water. The estimated contributions of the 868 water molecules in this dataset span a range of −42.92 to 4.16 kJ mol$^{-1}$ (average $e$ of atoms, multiply by 3 to obtain the contribution per water molecule), with the majority of interaction strengths occurring at around $-24.65 \pm 9.06$ kJ mol$^{-1}$. We generate a new PCovR shown in the left panel of Fig. 4. On the bottom of Fig. 4, we show the crystalline conformation and the molecules recolored by $\tilde{\delta}_a$.

First, we look at a common parameter for measuring the stabilizing effect of water: hydrogen bonding (H-bonding). Here, we have calculated H-bonds based on when the O⋯H or H⋯X distance is less than 2.5 Å and the dihedral angle of O⋯H–X or OH⋯X is greater than 150°. From the right side of Fig. 4, we see that the number of H-bonds donated to the water molecule (O⋯H) does not correlate with the cohesive interaction of the water molecules. There is some qualitative correlation/anti-correlation between the nature of these donated H-bonds and the second principal covariate (Pearson Correlation Coefficient, or PCC, = 0.49, −0.59 for the number of O⋯H–N and O⋯H–O, respectively). There is a mild anti-correlation between the number of H-bonds the water itself donates (OH⋯X) and the first covariate, with a PCC of −0.33. The second principal covariate is strongly correlated and anti-correlated with the number of OH⋯N and OH⋯O interactions, achieving a PCC of 0.69 and −0.73, respectively. Waters with primarily OH⋯N-type hydrogen bonds are at the top of the map (*e.g.*, Fig. 4(e), CSD ref. VOHBUR[46]), with OH⋯O-type at the bottom of the map (*e.g.*, Fig. 4(c) and (d), CSD ref. LEBJUX[44] and LACTOS12 [45]).

This analysis emphasizes that the number of hydrogen bonds does not fully capture all of the nuances of water

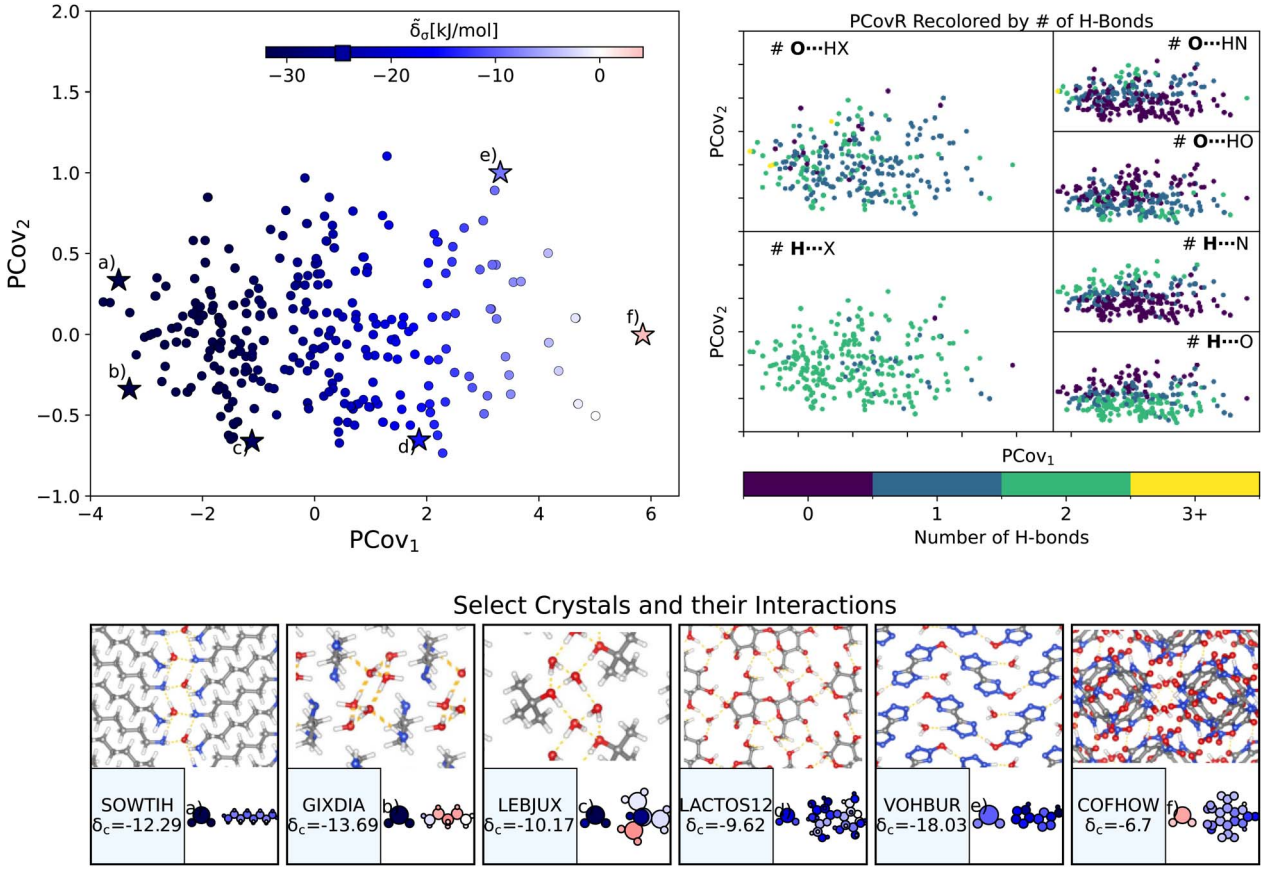


Fig. 4 The interactions of water molecules. (Left) Principal Covariates Regression (PCovR) map, where the color of each point denotes the estimated cohesive interaction of that motif and a marker on the color bar denotes the average value for all waters. (Right) The PCovR map recolored by the number of hydrogen bonds (H-bonds), separating those donated to the oxygen atom (top) from those donated by the hydrogen atoms (bottom). The insets on the bottom visualize several extremal or interesting environments. Select crystalline configurations and energy assignments: CSD ref. (a) SOWTIH,42 (b) GIXDIA,[43] (c) LEBJUX,[44] (d) LACTOS12,[45] (e) VOHBUR,[46] and (f) COFHOW.[47] In each panel, the bottom row shows the total lattice energy of the crystal (in kJ mol$^{-1}$) and the corresponding molecules where the atoms have been recolored by their estimated lattice energy contribution (on the same scale as on the PCovR map).

stabilization – the majority of water molecules participate in 2–3 such interactions, and the energy of these bonds can span a wide range. In O···H–X interactions, there is little energetic difference based on whether the acceptor is a nitrogen or oxygen atom – both types of hydrogen bonds span the full range of energies. The nature of the acceptor is encoded in the covariate orthogonal to the chemical features most correlated with interaction strength (*i.e.*, the nature of the acceptor is primarily correlated with the second covariate).

We see that the strongest water interactions in 1,6-diaminohexane monohydrate (CSD ref. SOWTIH,[42] Fig. 4(a)) and 1,3-diaminopropane trihydrate (CSD ref. GIXDIA,[43] Fig. 4(b)), where the water molecules associate with other water molecules and the amine group of their co-crystalline molecule. Our weakest contribution, by far, occurs in 4,5,6,7-tetranitro-1,3-dihydro-2*H*-benzimidazol-2-one hemihydrate (CSD ref. COFHOW,[47] Fig. 4(f)), where the water molecules sit interstitial to the imidazole molecules, prohibited from forming hydrogen bonds and potentially interfering with the stabilization of the imidazole clusters.

**3.2.4 Carboxylic acid groups.** As a strong electron donor, carboxylic acids are considered a key motif in molecular crystallization,[21,54,55] which is supported by their strong negative lattice energy contribution, here ranging from $-26.72$ kJ mol$^{-1}$ to $-1.11$ kJ mol$^{-1}$, with the majority of interaction strengths occurring in the $-17.17 \pm 3.82$ kJ mol$^{-1}$ range. Taking the 1023 carboxylic acid groups, we generate a new PCovR shown in the left panel of Fig. 5. On the right and bottom of Fig. 5, we have included panels showing, for select motifs, the crystalline conformation and molecules recolored by $\tilde{\delta}_a$.

The strongest contributions are found in 1,2-di(2-pyridyl) ethylene (CSD ref. FIBHOP[49]) in a succinic acid molecule (Fig. 5(a)) that forms two sets of supramolecular synthons: one homosynthon with the other succinic acid (Fig. 5(b)), and one heterosynthon with the pyridine group (consistent with the literature on the strength of carboxylic–pyridine interactions[56–58]). Interestingly, this crystal also contains one of the most weakly interacting groups (Fig. 5(b)), in the second succinic acid molecule that only participates in the single homosynthon.

Carboxylic acids form the strongest cohesive interactions when participating in multiple synthons, particularly heterosynthons (typified by Fig. 5(a) and (c), and noted in earlier literature[64]). Moving to the right, we see the contribution decrease commensurate to the number of interactions. For example, in 3,5-pyrazoledicarboxylic acid (CSD ref. SEPNUX,[51] Fig. 5(d)), there are two carboxylic acid groups that have drastically different energy contributions – one that forms a doublet homosynthon and the other is without close contacts. In an extreme case (CSD ref. CARCAZ,[53] Fig. 5(f)), the carboxylic acid group is prevented from interacting due to the bulkiness of the overall molecule, leading to a neutral contribution.

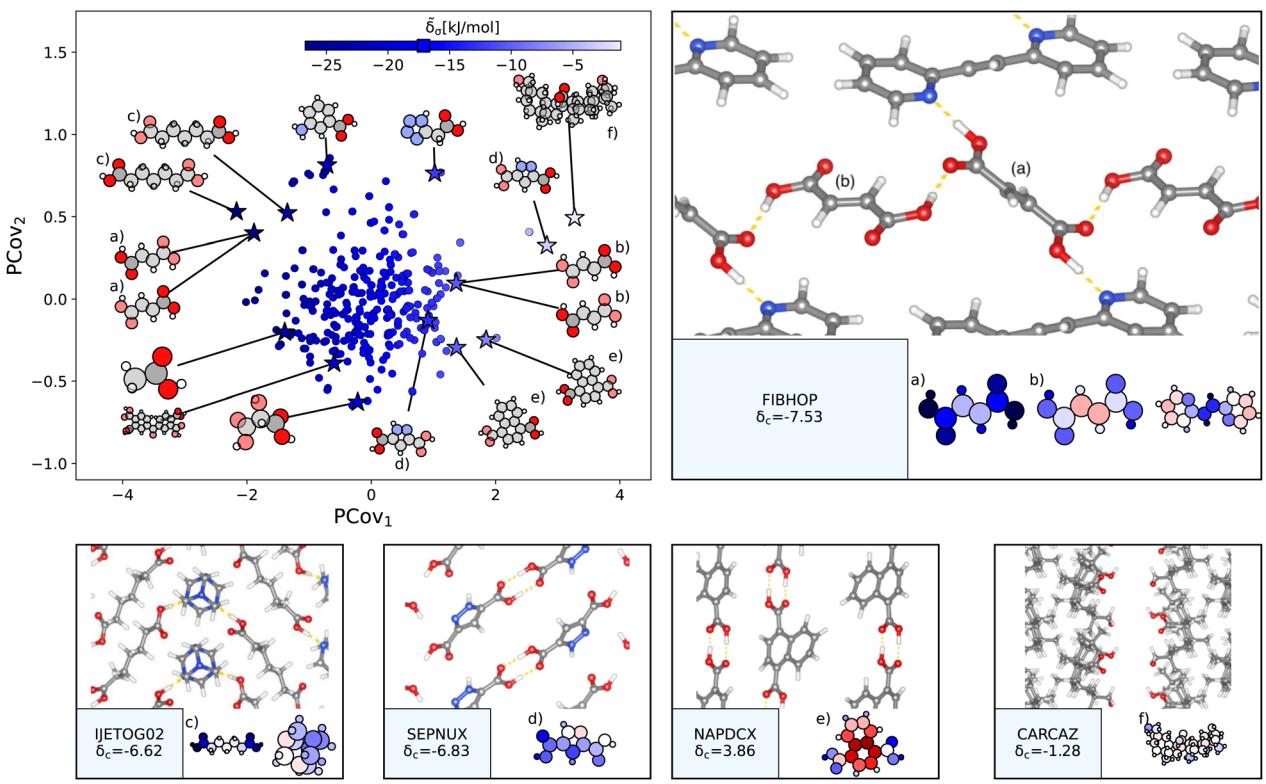


Fig. 5 The interactions of carboxylic acid groups. (Left) Principal Covariates Regression (PCovR) map, where the color of each point denotes the estimated cohesive interaction of that motif, and a marker on the color bar denotes the average value for all carboxylic acid groups. The insets visualize several extremal or interesting motifs. (Right) Select crystalline configurations and energy assignments: CSD ref. (a and b) FIBHOP,49 (c) IJETOG02,[50] (d) SEPNUX,[51] (e) NAPDCX,[52] and (f) CARCAZ.[53] In each panel on the right, the bottom row shows the total lattice energy of the crystal (in kJ mol$^{-1}$) and the corresponding molecules where the atoms have been recolored by their estimated lattice energy contribution.

An interesting success of this energy assignment is the ability to identify stabilizing motifs in otherwise unstable or metastable crystals. This is the case for CSD ref. NAPDCX[52] (Fig. 5(e)), an unstable 1,4-naphthalene-dicarboxylic acid that has an overall positive lattice energy at ambient pressure and temperature.¶ Despite this instability, we can clearly identify a binding interaction between carboxylic acid groups.

**3.2.5 6-Membered unsaturated carbon rings.** 6-Member unsaturated carbon rings (consistent with benzene molecules but more broadly-defined to include branched rings) show weak intermolecular interactions ranging from $-11.27$ kJ mol$^{-1}$ to 18.75 kJ mol$^{-1}$, with the majority of interactions occurring in the $-2.19 \pm 3.0$ kJ mol$^{-1}$ range. Similar to Section 3.2.4, we generate a new PCovR map using the averaged remnant descriptors and effective interactions using the 3280 benzene-like motifs, as shown in the left panel of Fig. 6. Again, we have included a panel on the right showing the crystalline conformation and molecules colored by $\tilde{\delta}_a$ for select configurations.

The most strongly-binding benzene-like motifs occur in molecules where (1) the ring is functionalized by strongly interacting groups, (2) the interactions of these groups facilitate planar molecular geometry, and (3) stacking occurs between the benzene-like rings with these auxiliary groups. We see this in 2,4,6-trinitrobenzene-1,3,5-triamine (CSD ref. TATNBZ03,[59] Fig. 6(b)), where the aromatic carbon ring stacks above the primarily intramolecular nitro–amine interaction and in Fig. 6(a) (CSD ref. BENZAC19[60]), where they stack above the carboxylic acid homosynthon.

There are various reasons for weakly-binding benzene-like motifs, including weak stacking and steric hindrance. As is evident from Fig. 6(c) and (d), rings will resist crystallization when the interactions of the end groups lead to deformation of the ring geometry. Take for example phenanthrene (CSD ref. PHENAN14,[61] Fig. 6(c)), a high-pressure polymorph that is unstable at ambient conditions (therefore has an overall positive lattice energy for the DFT reference used). Interestingly, we can pinpoint the localization of this deformation by looking at the atoms with the strongest positive contribution. While the keen reader may infer that this is solely due to the remnant descriptor reflecting the difference in strained and relaxed molecular geometry, we will note that a large difference in these representations can also coincide with a wealth of stabilizing intermolecular interactions, demonstrating that this simple linear model can differentiate molecular deformation from the introduction of new interactions.

This is further supported by comparing the motifs of this polymorph with its ambient-pressure, stable counterpart (CSD ref. PHENAN08[63]) to see how the nature of the same molecule changes based upon the interactions in the crystal. Both polymorphs adopt a similar herringbone crystal structure; however, the decreased molecular distortion and increased interactions between the auxiliary hydrogens and neighboring aromatic rings in PHENAN08[63] result in a significantly lower lattice energy of $\delta_c = -4.58$. In Fig. 7, we project the motifs of PHENAN08[63] and PHENAN14[61] onto our PCovR map from Fig. 6, we see this reflected by a left-shift of the motifs on the map, where the center ring moves from strongly resisting crystallization (Fig. 7(c)) to weakly interacting (Fig. 7(f)) and the

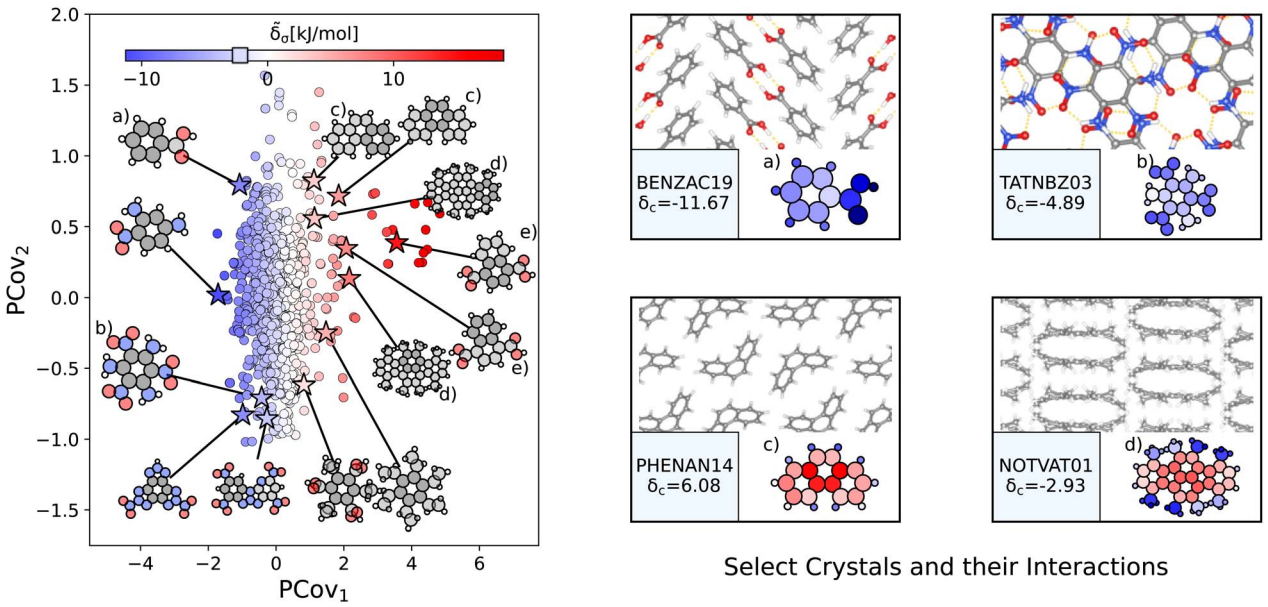


Fig. 6 The interactions of benzene-like rings. (Left) Principal Covariates Regression (PCovR) map, where the color of each point denotes the estimated cohesive interaction of that motif and a marker on the color bar denotes the average value of benzene-like rings. The insets visualize several extremal or interesting motifs. (Right) Select crystalline configurations and energy assignments: CSD ref. (a) TATNBZ03,59 (b) BENZAC19,[60] (c) PHENAN14,[61] and (d) NOTVAT01.[62] We also highlight the benzene-like motif from Fig. 5(f) in (e). In each panel on the right, the bottom row shows the total lattice energy of the crystal (in kJ mol$^{-1}$) and the corresponding molecules where the atoms have been recolored by their estimated lattice energy contribution (on the same scale as on the left panel).

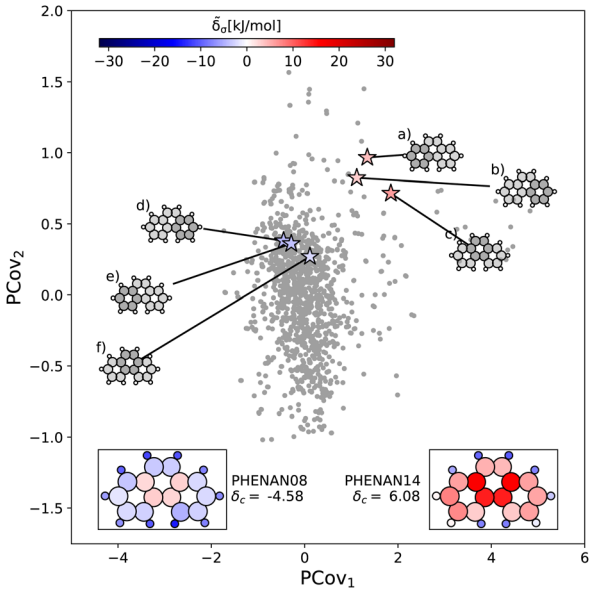


Fig. 7 Comparing the motifs in polymorphs of phenanthrene. Here we project two distinct polymorphs of phenanthrene onto the PCovR map shown in Fig. 6. At ambient conditions, one polymorph is stable (PHENAN08 [63]), while the other is unstable (PHENAN14,[61] also shown in Fig. 6(c)). Looking at the same motifs in the unstable and stable phases, we see a shift leftwards as the motifs go from resisting crystallization to weakly binding. In the lower insets, we have recolored the atoms of the phenanthrene molecule based upon their contribution to the lattice energy in the different polymorphs. Note that the bicyclic carbon atoms, while no longer distorted, weakly resist crystallization, as they prevent the auxiliary hydrogens from more closely interacting with neighboring $\pi$ bonds by distorting the molecule.

periphery rings move from weakly resisting crystallization (Fig. 7(a) and (b)) to weakly binding (Fig. 7(e) and (f)). It is worth noting that PHENAN08 [63] is an out-of-sample data point ($\varepsilon$ = 0.2 kJ mol$^{-1}$), demonstrating that the analysis in Fig. 6 is applicable beyond the initial reference set. We have included images of the PHENAN08 [63] crystal configuration in Fig. S5.†

**3.2.6 Nitro groups.** Nitro groups, defined as a nitrogen atom bonded to two terminal oxygen atoms, range in cohesive contributions from −29.9 kJ mol$^{-1}$ to 1.56 kJ mol$^{-1}$, with most interaction strengths being −12.76 ± 5.66 kJ mol$^{-1}$. Similar to our previous examples, we generate a new PCovR using the averaged remnant descriptors and effective interactions of the 2129 nitro groups, as shown in the left panel of Fig. 8. Again, we have included a panel on the right showing the crystalline conformation and constituent molecules colored by $\tilde{\delta}_a$. Unlike carboxyl and benzene-like groups, the chemical diversity of nitro interactions is limited – this is either due to the chemical nature of nitro interactions or the availability of nitro-containing crystals in CSD.

The resonant or partial charge of the oxygen atoms leads to strong binding in hydrogen-rich environments, supported by the results in Fig. 8. This is best typified by *trans-N,N*-dimethyl-2-nitrovinylamine (CSD ref. MNETAM01 [70]), a molecule where the nitro group is strongly interacting with the $CH_3$ end groups with some potential $\pi$-hole stacking[71] between the nitrogen moieties, as shown in Fig. 8(a). The strength of these binding interactions lessens with the strength of the electron donors, with smaller contributions in crystals where the primary O···H interaction is with amine donors (*e.g.*, Fig. 8(c) and (d), CSD ref. CUPYUJ,[67] KEDJUB[68]). In some of these cases, the binding is likely weakened by intramolecular interactions, similar to the contributions of the nitro groups in 2,4,6-trinitrobenzene-1,3,5-triamine (CSD ref. TATNBZ03,[59] Fig. 8(f), seen earlier in Fig. 6(b)). Finally, to the right of the map, we see the strongest repulsive interactions from nitro groups in proximity to other nitro or aromatic nitrogen groups, such as the nitro–oxidiazole interaction in 3-(3,5-dinitro-1*H*-pyrazol-4-yl)-4-nitro-1,2,5-oxadiazole (CSD ref. LAYSOV,[69] Fig. 8(f)).

### 3.3 A case study: ethenzamide co-crystals

We conclude by demonstrating how these models and methods can be used in the more practical context of crystal design. Ethenzamide is a common analgesic that has been the subject of numerous co-crystallization studies[72–81] due to the poor solubility of its homocrystalline form.[82] On the Cambridge Structure Database, there are currently 47 reported co-crystals of ethenzamide, of which there are 29 crystals that fit within the scope of this study and contain complete crystallographic information. The co-forming molecules in these co-crystals are primarily hydrobenzoic acids, nitrobenzoic acids, and dicarboxylic acids, as well as a 3-toluic acid co-crystal[77] and two saccharin co-crystals.[81] A list of these crystals with their experimental and computed properties is given in ESI Appendix A4.†

We first compute the relaxed energies of the co-crystals and their molecular components, following the procedures outlined in ESI Appendix A† to obtain the reference geometries and binding energies of each crystal. For reference, our previous model built using eqn (9) achieves an RMSE of 0.45 kJ mol$^{-1}$ and an MAE of 0.35 kJ mol$^{-1}$ more than sufficient to distinguish between the different categories of co-forming molecules, yet unable to provide any guidance in isomeric contexts (we have included a labeled parity plot in Fig. S10†). Following the procedure outlined in ESI Appendix B,† we identify the functional groups within the ethenzamide and estimate the contribution of their interactions to the molecular binding.

As shown in Fig. 9, unsurprisingly, most of the binding interactions occur due to the acetamide group in the ethenzamide, with a 3.5 kJ mol$^{-1}$ difference between the weakest contributing acetamide motif and the most strongly contributing ethyl group, which is beyond the error in the overall model. From Fig. 9, we can also see that, while the ethyl and benzene-like rings behave similarly to other similar motifs across the entire dataset, the acetamide and ether groups are generally more and less stabilizing, respectively, than their counterparts at large. With the ether groups, this is reasonable – the geometry of the ether prevents much intermolecular interaction. With the acetamide group, this demonstrates that there is a large range of engineering that can happen to affect crystalline stability, which might be beneficial when considering molecular solubility.

From here, we use the PCovR of acetamide groups to identify other acetamide motifs that behave similarly or dissimilarly to

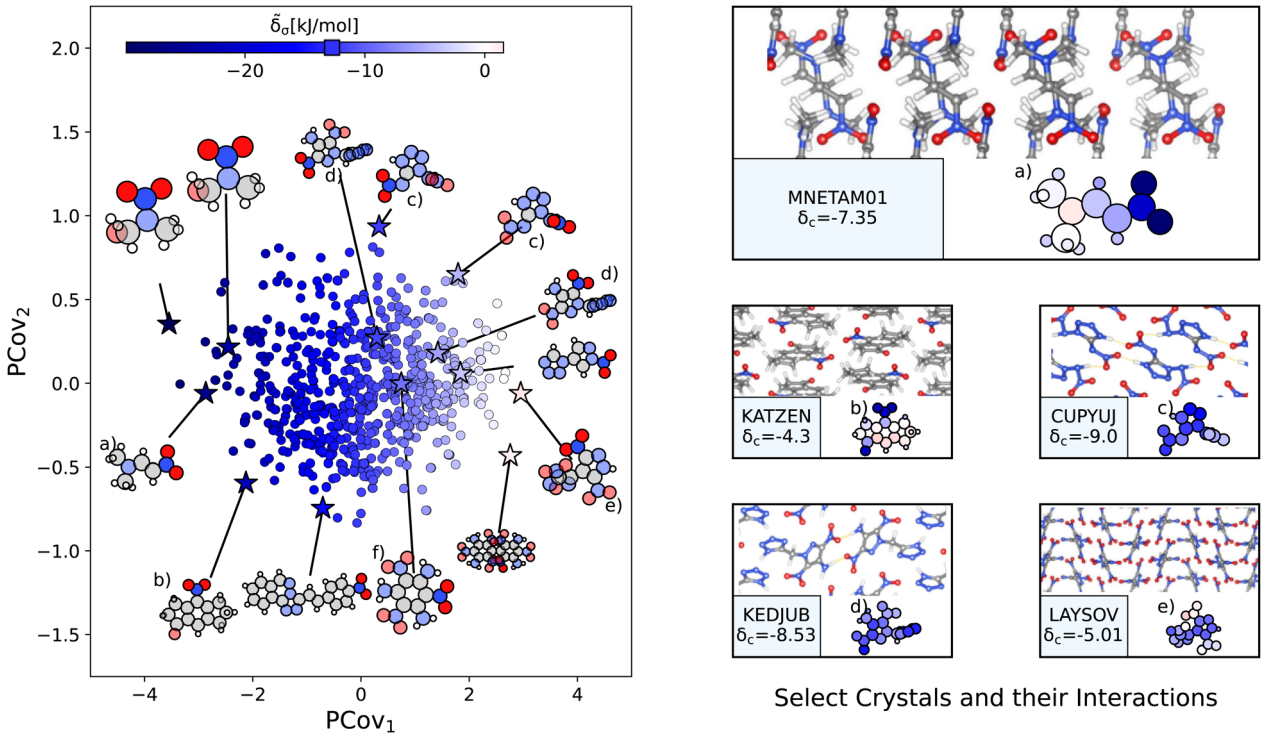


Fig. 8 The interactions of nitro groups. (Left) Principal Covariates Regression (PCovR) map, where the color of each point denotes the estimated cohesive interaction of that motif. (Right) Select crystalline configurations and energy assignments: CSD ref. (a) TIJKEC,65 (b) KATZEN,[66] (c) CUPYUJ,[67] (d) KEDJUB,[68] and (e) LAYSOV.[69] We also highlight the nitro group of TATNBZ03 [59] from Fig. 6(b) in (f). In each panel on the right, the bottom row shows the total lattice energy of the crystal (in kJ mol$^{-1}$) and the corresponding molecules where the atoms have been recolored by their estimated lattice energy contribution (on the same scale as on the left panel).

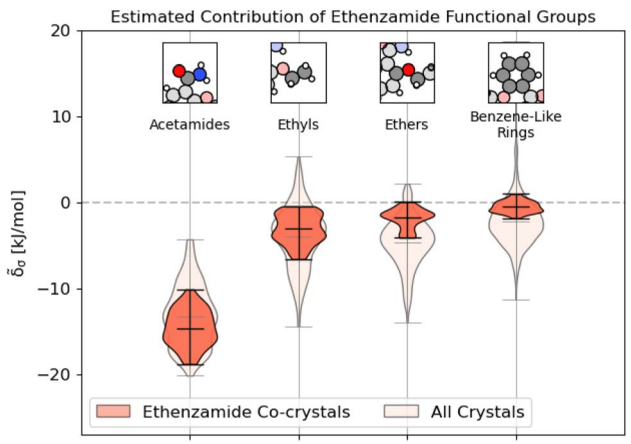


Fig. 9 Distribution of energetic contributions for the functional groups of ethenzamide. Following the procedure outlined in 3.2, we have computed the estimated contribution to the binding energy $\tilde{\delta}_\sigma$. Similar to Fig. 2, we have arranged the functional groups in order of average contribution, and the lines on each plot denote each group's extreme and mean contributions. Wider sections of the violin plot represent a higher probability that members of the population will take on the given value; the skinnier sections represent a lower probability. Here, darker sections refer to the distribution in the functional groups of the ethenzamide molecules, with lighter sections showing the distribution for the same functional group across the entire training set.

those we see in the known ethenzamide co-crystals, as highlighted in Fig. 10. We first train our PCovR model on the acetamide groups in the training set, and project those from the ethenzamide dataset into the corresponding latent space. Because the interactions across the training set are much more diverse than within the ethenzamide set, we plot along the first and third covariate to show the best distinction between the two datasets.‖ We define chemical similarity based upon the Euclidean distance in PCovR space – that is, we identify acetamide groups that appear at a similar place to those in ethenzamide co-crystals on the map in Fig. 10. Note that because we compute the distance using all covariates, some points that seem extremal in Fig. 10 are not, as they are closer or further from the ethenzamide acetamide groups in other dimensions.

We highlight the molecules that form the most similar acetamide networks to those in the ethenzamide dataset in Fig. 10 using an (o) marker and showing the molecule below. Those closest in PCovR space are molecules that form single acetamide homosynthons (*e.g.*, Fig. 10(b) and (c), CSD ref. MEGDOS[84] and LORMOV[85]) or heterosynthons with a carboxylic acid group (*e.g.*, Fig. 10(a), CSD ref. HXBNZM[83]).

Perhaps more interesting are the acetamide groups that form different interactions, highlighted in the bottom panel in Fig. 10. These groups give insight into the other supramolecular synthons that form with ethenzamide across a range of stabilizing and destabilizing contributions. We see strong

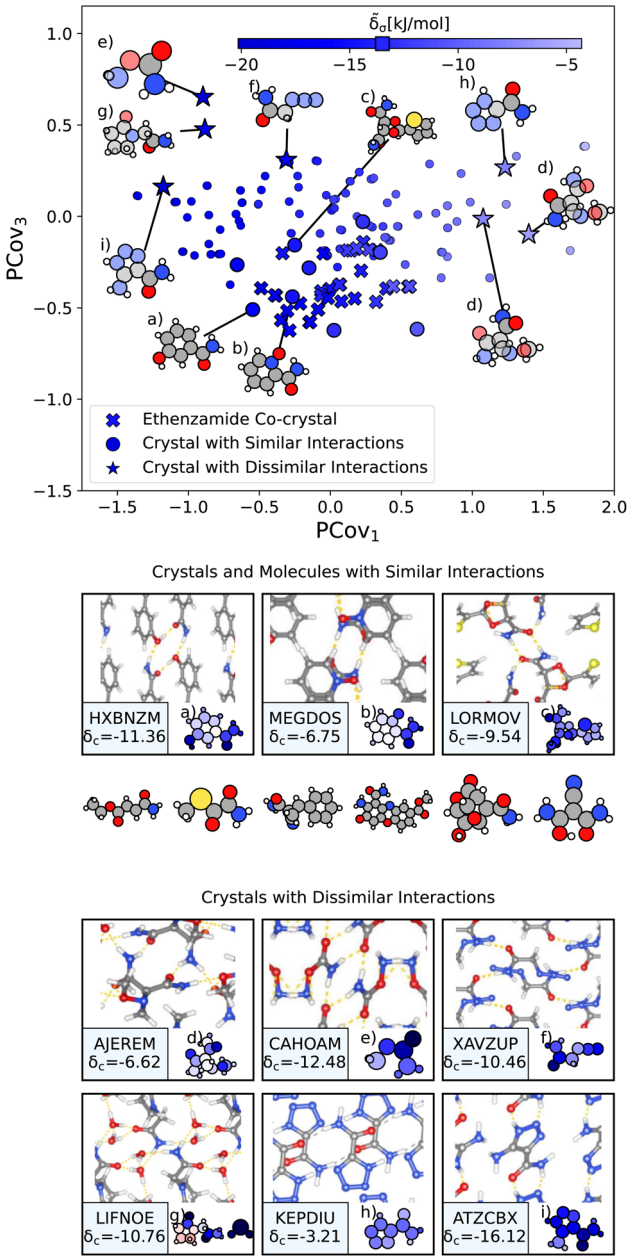


Fig. 10 The interactions of acetamide groups. (Top) Principal Covariates Regression (PCovR) map, where the color of each point denotes the estimated cohesive interaction of that motif. Here we have plotted along the first and third covariates to best distinguish the acetamide groups of the known ethenzamide co-crystals from other groups. The markers correspond to: (x) known ethenzamide co-crystals, (o) crystals with similar acetamide interactions, and (*) crystals with dissimilar acetamide interactions. (middle) Crystals and molecules with similar acetamide interactions as those in the known ethenzamide co-crystals, including CSD ref. (a) HXBNZM,83 (b) MEGDOS,[84] and (c) LORMOV.[85] (bottom) Crystals dissimilar acetamide interactions to the known ethenzamide co-crystals: CSD ref. (d) AJEREM,86 (e) CAHOAM,[87] (f) XAVZUP,[88] (g) LIFNOE,[89] (h) KEPDIU[90] and (i) ATZCBX.[91] Insets have been ordered from highest to lowest similarity to the interactions in the known ethenzamide co-crystals.

interactions in triazole-5-carboxaldehyde (Fig. 10(i), CSD ref. ATZCBX[91]), where the acetamide group forms a heterosynthon with the triazole group, and in *O*-carbamoylhydroxylamine (Fig. 10(e), CSD ref. CAHOAM[87]), where the small size of the molecule facilitates both multiple homosynthons between acetamide groups, as well as heterosynthons with the oxygen of the hydroxylamine groups. In 2-oxopyrrolidineacetamide dihydrate (Fig. 10(g), CSD ref. LIFNOE[89]), a network of hydrogen bonds is formed between acetamide groups and water molecules. In azidoacetamide (Fig. 10(f), CSD ref. XAVZUP[88]), we see an acetamide homosynthon formed at an offset so that the azide group can stack directly above the NH···O interaction. We see weaker interactions in tetrazole-5-carboxamide (Fig. 10(h), CSD ref. KEPDIU[90]), where the acetamide group is interacting with the azole group, which, when compared with triazole-5-carboxaldehyde (Fig. 10(i)), demonstrates a large range for acetamide-azole synthon binding. Finally, in 1-methoxyaziridine-2,2-dicarboxamide (Fig. 10(d), CSD ref. AJEREM[86]), despite multiple acetamide interactions, there is a weaker acetamide network, likely due to the geometry of the molecule itself.

We do not suggest that these molecules could be used directly as co-formers; the training set was obtained with diversity as the primary goal, with no regard for availability, toxicity, ease of synthesis, or stability. Instead, each of these related and unrelated crystals gives insight into the types of interactions that may beget new ethenzamide co-crystals. The molecules shown in Fig. 10 can be used as inspiration to identify co-former candidates from libraries of biocompatible compounds and to guide future crystallization studies.

## 4 Conclusions

Molecular crystallization is a complex, multi-faceted process, that poses tremendous challenges to both quantitative modeling, and to the derivation of qualitative design principles. In this work, we propose a data-driven strategy to build a database of the interaction motifs that are found in a diverse set of molecular crystals, to determine semi-quantitatively their contribution to the lattice energy, and to generate a library of molecular motifs that can be used to interpret the stability of known crystals and to assist the design of new ones.

In doing so, we have to strike a balance between several conflicting goals. By selecting structures from the CSD with maximal structural diversity, we ensure that we cover a broad range of chemical and packing motifs, while remaining focused on structures that are known to be experimentally realizable. By using a general-purpose, atom-centered structural representation that is capable of describing arbitrary structural correlations, we ensure that our data analysis is flexible and that it does not incorporate pre-conceived notions about molecular bonding. At the same time, we ensure that the model focuses on the features that are most relevant to determine crystal stability by building a remnant descriptor that mimics the definition of the lattice energy as a difference between the total energies of the crystal and its constituents.

The resulting models achieve a respectable mean absolute error of about 0.4 kJ mol$^{-1}$ in predicting the atomic contributions to crystal stability using these descriptors that gives us a semi-quantitative estimate of the contribution of each atomic

environment to the lattice energy and to compare between different co-crystals or between polymorphs that are stable at very different conditions. In order to translate these atomic contributions in a language that can be useful to crystal chemistry, we then assemble them to estimate the stabilizing power of traditional chemical groups (carboxylic acids, amines, …) and build data-driven maps that facilitate the comparison of different chemical environments by expressing simultaneously the structural variability and correlation with the lattice energy contribution. For each chemical moiety we provide an interactive map (on Materials Cloud[48]) that allows to juxtapose different types of crystal environments, to identify structural patterns that are either stabilizing or destabilizing, and to contrast them with conventional motifs (*e.g.* hydrogen-bonding), demonstrated here for a few selected cases. As we demonstrate for phenanthrene, it is also possible to use these maps to compare polymorphs of the same molecule, and to analyze molecular motifs for a structure that is not part of our original reference set. With these tools, we aim to guide those designing molecular co-crystals in identifying suitable co-formers, as demonstrated for the analgesic ethenzamide.

We hope that this library of molecular motifs will prove useful to applications to specific crystal-design problems. More broadly, we believe that the general ML protocol that we follow, combining regression of the ultimate target property with unsupervised analysis of molecular motifs, can inspire similar applications to the study of other classes of materials, ranging from metal and covalent organic frameworks to self-assembled monolayers and biological systems.

## Data availability

Data for this paper, including interactive visualizations and labelled molecular motifs, are available at MaterialsCloud at **https://molmotifs.matcloud.xyz**. The raw data and visualization files are available at **https://doi.org/10.24435/materialscloud:71-21**.

## Author contributions

RKC and MC designed the study and wrote the manuscript. RKC computed the molecular energies and geometries, built the machine learning models, and designed the figures. MP separated the crystals into molecular components, screened the dataset before relaxation calculations, started the molecular energy calculations, and edited the manuscript. EAE advised on the dataset provenance and curation and edited the manuscript.

## Conflicts of interest

There are no conflicts to declare.

## Acknowledgements

This project was funded by NCCR Marvel Inspire Fellowship (MP), NCCR Marvel (RKC & MC), Trinity College (EAE), and ERC Grant 677013-HBMAP (RKC & MC). The authors would like to acknowledge Federico Giberti, Andrea Anelli, and Guillaume Fraux for fruitful conversations at the study's start and culmination.

## Notes and references

‡ To improve the regressions of crystal and molecular energies, we subtract a baseline determined by linear regression of the atomic composition on the total energies.

§ The lattice energy of the crystal is not the sum of these motif contributions, as (1) both are averaged quantities, and (2) a single crystal may have overlapping motifs.

¶ Of the approximately 3200 crystals studied in this article, only 23 have a positive lattice energy. This is covered in more detail in ESI Appendix A3.

‖ There is an interactive map of the first four covariates within our online repository at Cersonsky *et al.*[48]

1 M. A. E. Yousef and V. R. Vangala, *Cryst. Growth Des.*, 2019, **19**, 7420–7438.
2 M. K. Dudek and K. Druzbicki, *CrystEngComm*, 2022, **24**, 1665–1678.
3 L. Iuzzolino, A. M. Reilly, P. McCabe and S. L. Price, *J. Chem. Theory Comput.*, 2017, **13**, 5163–5171.
4 S. Datta and D. J. W. Grant, *Nat. Rev. Drug Discovery*, 2004, **3**, 42–57.
5 T. Beyer, G. M. Day and S. L. Price, *J. Am. Chem. Soc.*, 2001, **123**, 5086–5094.
6 M. M. Azrain, M. R. Mansor, S. H. S. M. Fadzullah, G. Omar, D. Sivakumar, L. M. Lim and M. N. A. Nordin, *Synth. Met.*, 2018, **235**, 160–175.
7 J. Mei, N. L. C. Leung, R. T. K. Kwok, J. W. Y. Lam and B. Z. Tang, *Chem. Rev.*, 2015, **115**, 11718–11940.
8 E. Krissinel and K. Henrick, *J. Mol. Biol.*, 2007, **372**, 774–797.
9 Q. C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter, T. Maniatis, A. Califano and B. Honig, *Nature*, 2012, **490**, 556+.
10 P. Ganguly and G. R. Desiraju, *CrystEngComm*, 2010, **12**, 817–833.
11 R. Davey, G. Dent, R. Mughal and S. Parveen, *Cryst. Growth Des.*, 2006, **6**, 1788–1796.
12 A. Dey, M. T. Kirchner, V. R. Vangala, G. R. Desiraju, R. Mondal and J. A. Howard, *J. Am. Chem. Soc.*, 2005, **127**, 10545–10559.
13 J. Sarma and G. R. Desiraju, *Cryst. Growth Des.*, 2002, **2**, 93–100.
14 G. R. Desiraju, *Angew. Chem., Int. Ed.*, 2007, **46**, 8342–8356.
15 M. K. Corpinot and D.-K. Bucar, *Cryst. Growth Des.*, 2019, **19**, 1426–1453.
16 I. J. Bruno, J. C. Cole, P. R. Edgington, M. Kessler, C. F. Macrae, P. McCabe, J. Pearson and R. Taylor, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2002, **58**, 389–397.
17 I. J. Bruno, J. C. Cole, M. Kessler, J. Luo, W. S. Motherwell, L. H. Purkis, B. R. Smith, R. Taylor, R. I. Cooper, S. E. Harris, *et al.*, *J. Chem. Inf. Model.*, 2004, **44**, 2133–2144.
18 M. A. Neumann, F. J. Leusen and J. Kendrick, *Angew. Chem.*, 2008, **120**, 2461–2464.
19 J. Hoja, H.-Y. Ko, M. A. Neumann, R. Car, R. A. DiStasio and A. Tkatchenko, *Sci. Adv.*, 2019, **5**, eaau3338.

20 M. A. Spackman and D. Jayatilaka, *CrystEngComm*, 2009, **11**, 19–32.
21 M. A. Spackman and J. J. McKinnon, *CrystEngComm*, 2002, **4**, 378–392.
22 O. Egorova, R. Hafizi, D. C. Woods and G. M. Day, *J. Phys. Chem. A*, 2020, **124**, 8065–8078.
23 F. Musil, S. De, J. Yang, J. E. Campbell, G. M. Day and M. Ceriotti, *Chem. Sci.*, 2018, **9**, 1289–1300.
24 S. Wengert, G. Csányi, K. Reuter and J. T. Margraf, *J. Chem. Theory Comput.*, 2022, **18**, 4586–4593.
25 S. Wengert, G. Csányi, K. Reuter and J. T. Margraf, *Chem. Sci.*, 2021, **12**, 4536–4546.
26 V. Kapil and E. A. Engel, *Proc. Natl. Acad. Sci. U.S.A.*, 2022, **119**, e2111769119.
27 A. Seko, H. Hayashi, K. Nakayama, A. Takahashi and I. Tanaka, *Phys. Rev. B*, 2017, **95**, 144110.
28 T. Bereau, D. Andrienko and O. A. von Lilienfeld, *J. Chem. Theory Comput.*, 2015, **11**, 3225–3233.
29 M. Cordova, E. A. Engel, A. Stefaniuk, F. Paruzzo, A. Hofstetter, M. Ceriotti and L. Emsley, *J. Phys. Chem. C*, 2022, **126**, 16710–16720.
30 K. Britts and I. L. Karle, *Acta Crystallogr.*, 1967, **22**, 308.
31 L. Talirz, S. Kumbhar, E. Passaro, A. V. Yakutovich, V. Granata, F. Gargiulo, M. Borelli, M. Uhrin, S. P. Huber, S. Zoupanos, C. S. Adorf, C. W. Andersen, O. Schütt, C. A. Pignedoli, D. Passerone, J. VandeVondele, T. C. Schulthess, B. Smit, G. Pizzi and N. Marzari, *Sci. Data*, 2020, 7(1), 299.
32 M. J. Willatt, F. Musil and M. Ceriotti, *J. Chem. Phys.*, 2019, **150**, 154110.
33 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
34 G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler and M. Ceriotti, *J. Chem. Phys.*, 2018, **24**, 241730.
35 R. K. Cersonsky, B. A. Helfrecht, E. A. Engel, S. Kliavinek and M. Ceriotti, *Mach. Learn.: Sci. Technol.*, 2021, **2**, 035038.
36 B. A. Helfrecht, R. Semino, G. Pireddu, S. M. Auerbach and M. Ceriotti, *J. Chem. Phys.*, 2019, **151**, 154112.
37 C. Ben Mahmoud, A. Anelli, G. Csányi and M. Ceriotti, *Phys. Rev. B*, 2020, **102**, 235130.
38 *SMARTS: A Language for Describing Molecular Patterns*, 1997, **http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html**.
39 *RDKit: Open-source cheminformatics*, **https://www.rdkit.org**.
40 S. De Jong and H. A. Kiers, *Chemom. Intell. Lab. Syst.*, 1992, **14**, 155–164.
41 B. A. Helfrecht, R. K. Cersonsky, G. Fraux and M. Ceriotti, *Mach. learn.: sci. technol.*, 2020, **1**, 045021.
42 S. Janeda and D. Mootz, *Z. Naturforsch., B: J. Chem. Sci.*, 1998, **53**, 1197.
43 S. Janeda and D. Mootz, *Z. Naturforsch., B: J. Chem. Sci.*, 1999, **54**, 103.
44 D. Mootz and D. Staben, *Z. Naturforsch., B: J. Chem. Sci.*, 1993, **48**, 1325.
45 M. Schreyer, L. Guo, S. Thirunahari, F. Gao and M. Garland, *J. Appl. Crystallogr.*, 2014, **47**, 659.
46 T. M. Klapotke, M. Q. Kurz, R. Scharf, P. C. Schmid, J. Stierstorfer and M. Suceska, *ChemPlusChem*, 2015, **80**, 97.
47 J. Sarlauskas, *New Trends Res. Energ. Mater., Proc. Semin., 17th*, 2014, **17**, 1005.
48 R. K. Cersonsky, M. Pakhnova, E. A. Engel and M. Ceriotti, *Lattice Energies for a Diverse Set of Molecular Crystals*, 2022.
49 A. Briceno, D. Leal, G. Ortega, G. D. de Delgado, E. Ocando and L. Cubillan, *CrystEngComm*, 2013, **15**, 2795.
50 M. Gardon, C. B. Pinheiro and G. Chapuis, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2003, **59**, 527.
51 P. Roussel, F. Bentiss, M. Drache, P. Conflant, M. Lagrenee and J.-P. Wignacourt, *J. Mol. Struct.*, 2006, **798**, 134.
52 J. L. Derissen, C. Timmermans and J. C. Schoone, *Cryst. Struct. Commun.*, 1979, **8**, 533.
53 T. K. Chen, D. C. Ales, N. C. Baenziger and D. F. Wiemer, *J. Org. Chem.*, 1983, **48**, 3525.
54 M. C. Etter, *Acc. Chem. Res.*, 1990, **23**, 120–126.
55 M. C. Etter, *J. Phys. Chem.*, 1991, **95**, 4601–4610.
56 P. Vishweshwar, A. Nangia and V. M. Lynch, *J. Org. Chem.*, 2002, **67**, 556–565.
57 T. R. Shattock, K. K. Arora, P. Vishweshwar and M. J. Zaworotko, *Cryst. Growth Des.*, 2008, **8**, 4533–4545.
58 P. Chen, Z. Zhang, S. Parkin, P. Zhou, K. Cheng, C. Li, F. Yu and S. Long, *RSC Adv.*, 2016, **6**, 81101–81109.
59 Z. Chua, C. G. Gianopoulos, B. Zarychta, E. A. Zhurova, V. V. Zhurov and A. Alan Pinkerton, *Cryst. Growth Des.*, 2017, **17**, 5200.
60 W. Cai and A. Katrusiak, *CrystEngComm*, 2012, **14**, 4420.
61 F. P. A. Fabbiani, D. R. Allan, W. I. F. David, S. A. Moggach, S. Parsons and C. R. Pulham, *CrystEngComm*, 2004, **6**, 504.
62 A. Sygula, F. R. Fronczek and P. W. Rabideau, *Tetrahedron Lett.*, 1997, **38**, 5095.
63 V. Petricek, I. Cisarova, L. Hummel, J. Kroupa and B. Brezina, *Acta Crystallogr., Sect. B: Struct. Sci.*, 1990, **46**, 830.
64 L. Leiserowitz, *Acta Cryst. B*, 1976, **32**, 775–802.
65 A. D. Vasiliev, A. M. Astachov, Yu. V. Kekin, L. A. Kruglyakova and R. S. Stepanov, *Acta Crystallogr., Sect. C: Cryst. Struct. Commun.*, 2001, **57**, 1192.
66 M. Salla, M. S. Butler, R. Pelingon, G. Kaeslin, D. E. Croker, J. C. Reid, J. M. Baek, P. V. Bernhardt, E. M. J. Gillam, M. A. Cooper and A. A. B. Robertson, *ACS Med. Chem. Lett.*, 2016, **7**, 1034.
67 D. Fischer, T. M. Klapotke and J. Stierstorfer, *Angew. Chem., Int. Ed.*, 2015, **54**, 10299.
68 D. Kumar, G. H. Imler, D. A. Parrish and J. M. Shreeve, *Chem.–Eur. J.*, 2017, **23**, 7876.
69 A. B. Sheremetev, I. L. Yudin, N. V. Palysaeva and K. Yu. Suponitsky, *J. Heterocycl. Chem.*, 2012, **49**, 394.
70 M. J. Dianez, A. Lopez-Castro and R. Marquez, *Acta Crystallogr., Sect. C: Cryst. Struct. Commun.*, 1985, **41**, 981.
71 A. Bauzá, T. J. Mooibroek and A. Frontera, *Chem. Comm.*, 2015, **51**, 1491–1493.
72 R. Khatioda, P. Bora and B. Sarma, *Cryst. Growth Des.*, 2018, **18**, 4637.

73 K. K. Sarmah, K. Boro, M. Arhangelskis and R. Thakuria, *CrystEngComm*, 2017, **19**, 826.
74 R. Khatioda, B. Saikia, P. J. Das and B. Sarma, *CrystEngComm*, 2017, **19**, 6992.
75 S. Aitipamula, P. Shan Chow and R. B. H. Tan, *CrystEngComm*, 2009, **11**, 1823.
76 S. Aitipamula, A. B. H. Wong, P. Shan Chow and R. B. H. Tan, *CrystEngComm*, 2012, **14**, 8515.
77 V. M. Hariprasad, S. K. Nechipadappu and D. R. Trivedi, *Cryst. Growth Des.*, 2016, **16**, 4473.
78 S. Aitipamula, P. Shan Chow and R. B. H. Tan, *Cryst. Growth Des.*, 2010, **10**, 2229.
79 A. Kozak, P. H. Marek and E. Pindelska, *J. Pharm. Sci.*, 2018, **108**, 1476.
80 S. Aitipamula, P. Shan Chow and R. B. H. Tan, *CrystEngComm*, 2010, **12**, 3691.
81 S. Aitipamula, P. Shan Chow and R. B. H. Tan, *CrystEngComm*, 2009, **11**, 889.
82 S. Pagola and P. W. Stephens, *Acta Crystallogr., Sect. C: Cryst. Struct. Commun.*, 2009, **65**, o583.
83 Y. Katsube, Y. Sasada and M. Kakudo, *Bull. Chem. Soc. Jpn.*, 1966, **39**, 2576.
84 L. S. Reddy, N. J. Babu and A. Nangia, *Chem. Commun.*, 2006, 1369.
85 W. Xu, Z. Yang, X.-H. Li, Bo-N. Liu and De-C. Wang, *Acta Crystallogr., Sect. E: Struct. Rep. Online*, 2009, **65**, o764.
86 R. G. Kostyanovsky, V. R. Kostyanovsky, G. K. Kadorkina and K. A. Lyssenko, *Mendeleev Commun.*, 2003, 111.
87 I. K. Larsen, *Acta Chem. Scand.*, 1968, **22**, 843.
88 M. Kumasaki, K. Kinbara, Y. Wada, M. Arai and M. Tamura, *Acta Crystallogr., Sect. E: Struct. Rep. Online*, 2001, **57**, o6.
89 F. P. A. Fabbiani, D. R. Allan, W. I. F. David, A. J. Davidson, A. R. Lennie, S. Parsons, C. R. Pulham and J. E. Warren, *Cryst. Growth Des.*, 2007, **7**, 1115.
90 N. Fischer, T. M. Klapotke, S. Rappengluck and J. Stierstorfer, *ChemPlusChem*, 2012, **77**, 877.
91 A. Kalman, K. Simon, J. Schawartz and G. Horvath, *J. Chem. Soc., Perkin Trans. 2*, 1974, 1849.