# Multi-fidelity Statistical Machine Learning for Molecular Crystal Structure Prediction

Olga Egorova,[a‡] Roohollah Hafizi,[b‡] David C. Woods*[a] and Graeme M. Day*[b]

The prediction of crystal structures from first principles requires highly accurate energies for large numbers of putative crystal structures. The accuracy of solid state density functional theory (DFT) calculations is often required, but hundreds or more structures can be present in the low energy region of interest, so that the associated computational costs are prohibitive. Here, we apply statistical machine learning to predict expensive hybrid functional DFT (PBE0) calculations using a multi-fidelity approach to re-evalute the energies of crystal structures predicted with an inexpensive force field. The method uses an autoregressive Gaussian process, making use of less expensive GGA DFT (PBE) calculations to bridge the gap between the force field and PBE0 energies. The method is benchmarked on the crystal structure landscapes of three small, hydrogen bonding organic molecules and shown to produce accurate predictions of energies and crystal structure ranking using small numbers of the most expensive calculations; the PBE0 energies can be predicted with errors of less than 1 kJ mol$^{-1}$ with between 4.2-6.8% of the cost of the full calculations. As the model that we have developed is probabilistic, we discuss how the uncertainties in predicted energies impact on assessment of the energetic ranking of crystal structures.

## 1 Introduction

Molecular crystal structure prediction (CSP) aims to predict the set of likely crystal structures of a molecule through computational methods alone, starting from no more than the chemical diagram. The crystal structure adopted by a molecule is important because it affects many properties of materials, including physical and chemical stability[1,2], melting points[3–5], solubility[6], morphology[7], porosity[8–10] and electronic/optoelectronic properties[11,12]. The relationship between molecular structure and crystal structure is complicated by the phenomenon of polymorphism, where different crystal structures of the same chemical composition can be accessed experimentally, either concomitantly or under different crystallization conditions. Thus, a given molecule can have very different materials properties, depending on polymorph, and control of crystal structure can be exploited in matters as simple as improving the daily experience of eating chocolate[13], or as severe as preventing drug ineffectiveness[14–16].

The role of CSP must be to predict all likely crystal structures, along with a measure of their likelihood. There have been high profile cases in the pharmaceutical industry where failing to anticipate polymorphism of a drug had serious consequences. In one case, Rotigotine, a prescription for the treatment of Parkison's disease, began crystallizing to an unknown polymorph in transdermal patches[14,15]. This new form was thermodynamically more stable than the original polymorph, less soluble, and thus less efficient. In another case, an unknown polymorph of Ritonavir, a drug for the treatment of HIV, was accidentally produced, which was considerably less soluble and thus bioavailable[16]. As well as imposing substantial costs for the drug companies, the temporary removal of these drugs from the market also interrupt the patients' treatment. CSP could form part of a strategy to minimize such risks, by guiding experiments for the preparation of undiscovered polymorphs or providing additional confidence that no such undiscovered polymorphs exist. Applications of CSP are also developing in the area of functional materials discovery by linking molecular structure to likely materials structures and, hence, to properties of interest; this process can be used to screen and prioritise potential synthetic targets.[17–20]

CSP is usually approached as a problem in global optimization. Possible crystal structures correspond to local minima on a high dimensional energy surface determined by the structural degrees of freedom defining a crystal structure (unit cell dimensions, molecular positions and orientations, and intramolecular degrees of freedom). The process of CSP can be conceptually split into exploration for putative crystal structures, followed by their ranking, usually based on calculated lattice energies. The global energy minimum is assumed to correspond to the most likely observable crystal structure.[21,22] However, the prevalence of polymorphism[23] demonstrates that other higher energy crystal structures are also important and the energetic range of observed polymorphism gives an indication of the region of the crystal energy landscape that will normally include all observable crystal structures. A study of over 500 pairs of known polymorphs[24] revealed that over half of the pairs had a lattice energy difference of less than 2 kJ mol$^{-1}$, and only 5% had an energy difference higher than 7.2 kJ mol$^{-1}$. CSP has revealed that small organic molecules often have dozens, and sometimes over 100 possible crystal structures within this small energy range.[25] A reason behind the plurality of low energy crystal structures is the weak nature of packing forces such as dispersion interactions, hydrogen bonding, and less specific polar interactions. For most molecules, there are many ways that these interactions can combine to form similarly low-energy crystal structures.

The small differences in lattice energies between competing crystal structures imply that a very good energy resolution is needed for reliable energy rankings. This requires the compu-

‡ These authors contributed equally.

[a] *Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, United Kingdom*

[b] *Computational Systems Chemistry, School of Chemistry, University of Southampton, Southampton, SO17 1BJ, United Kingdom; E-mail: G.M.Day@soton.ac.uk.*

tation of crystal energies with a high level of theory, for which solid state, periodic implementations of density functional theory (DFT) have become a popular approach.[22] The cost of solid state DFT is too high to use for the entire CSP process. This is because a thorough sampling of the lattice energy surface typically requires the generation and optimization of tens or hundreds of thousands of trial crystal structures.[26] It is, therefore, more common to take a hierarchical approach, where crystal structures are initially generated and ranked using lower cost methods (e.g. force fields), followed by re-ranking of the lowest energy predicted crystal structures using DFT. Final energy calculations using generalized gradient approximation (GGA) exchange-correlation DFT functionals usually suffice for ranking of structures in a organic molecular CSP.[27] However, there are sensitive cases that require energies at a higher level of theory, such as hybrid functional DFT.[28,29] In these cases, energy evaluation can become prohibitively expensive, even for CSP of very small molecules. The situation becomes worse as the size of the molecule increases. A fast, accurate approach to achieve energies of equivalent quality to hybrid DFT would be highly desirable.

State-of-the-art statistical machine learning methods provide promising tools to achieve this goal. They have been actively used for estimation of the potential energy surface of various classes of materials[30–34]. Gaussian Process Regression (GPR) is of particular interest here. GPR can provide a probabilistic description of a wide variety of functional behavior, interpolate computed data, and tends to outperform other prediction methods for small data problems.[35] The latter is a crucial feature when the collection of large training data sets is computationally unaffordable. Minimizing the required training set size, as well as efficient choice of such sets, especially at the hybrid level of DFT, is crucially important to obtain an effective and cost-effective approach that can be practically applied to CSP. GPR has previously been applied to directly learning the DFT relative energies of sets of predicted crystal structures,[36] as well as to learning the differences between relative energies at force field and quantum mechanical levels of theory.[36,37] These earlier studies have demonstrated promising results for the application of GPR for acheiving DFT-quality predicted rankings at substantially lowered computational cost.

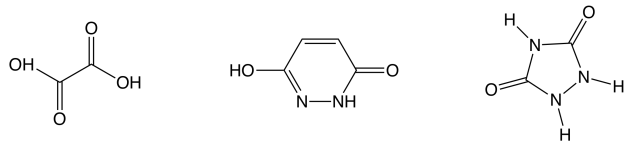
In this paper, we develop and apply multi-fidelity (multi-level) statistical machine learning methods to learn differences in lattice energies of a hierarchy of simulation methods that are commonly adopted during the structure exploration and final ranking stages of CSP. A fast, approximate force field that is used at the early stages of a global structure search is set as the baseline, and the machine learning models are developed to refine its results, first to predict GGA DFT lattice energies and then hybrid DFT. In the following, we start by optimizing data gathering costs, and we then train the machine learning model on the energy differences between models in the energy hierarchy. The quality of the constructed models is verified by application to CSP for three molecular crystals known to be challenging in terms of energy ranking: oxalic acid, maleic hydrazide, and urazole.

## 2 Methods

### 2.1 Choice of molecules and CSP datasets

The data sets used in this work consist of crystal structure-lattice energy pairs at three different levels of accuracy and computational complexity in their energy evaluation. The lowest level was evaluated using an atomic multipole-based force field (described below). The intermediate level consists of periodic GGA DFT energies using the Perdew–Burke-Ernzerhof (PBE) functional.[38] While most of the relevant physics is already accounted for in these GGA calculations, for a better treatment of the effect of electron exchange and correlation on the ranking of crystal structures, DFT calculations with PBE0 hybrid exchange-correlation functional[39] were performed as the high level.

Three small molecules were chosen to test the methods developed here: oxalic acid, maleic hydrazide and urazole (Fig. 1). All three are small enough so that it is affordable to evaluate the energies for all predicted crystal structures at the highest level of theory considered here, so that errors from the GPR models can be assessed. Two of the molecules (oxalic acid and maleic hydrazide) have known polymorphism and all three are known to be challenging for obtaining accurate energy rankings either between their known polymorphs[40] or in previous CSP studies.[41]




(a) oxalic acid (b) maleic hydrazide (c) urazole

Fig. 1 The three molecules studied here (CSD reference codes[42] for each experimentally determined form are given in parentheses): (a) oxalic acid $\alpha$ (OXALAC05[43]) and $\beta$ (OXALAC07[44]) polymorphs; (b) maleic hydrazide monoclinic (MALEHY01[45]), triclinic (MALEHY10[46]), and MH3 monoclinic (MALEHY12[47]) polymorphs and (c) urazole (KOXRIY[48]).

The data sets used here consist of low-energy crystal structures found in previous CSP studies.[37] The crystal structures were generated using quasi-random sampling followed by local energy minimization to locate the local minima on each molecule's lattice energy hypersurface using the Global Lattice Energy Explorer code.[26] Crystal structures were generated in the 11 most frequently occupied space groups ($P2_1/c$, $P2_12_12_1$, $P\bar{1}$, $P2_1$, $Pbca$, $C2/c$, $Pna2_1$, $Cc$, $Pca2_1$, $C2$, $P1$) with one independent molecule in the crystallographic asymmetric unit ($Z' = 1$). The molecular structure was held rigid throughout, at the DFT optimized geometry of the isolated molecule.

Lattice energy minimization was performed using the FIT+DMA[49] anisotropic atom-atom force field, which consists of an empirically parameterized *exp-6* intermolecular repulsion-dispersion potential and electrostatics described by atomic multipoles up to hexadecapole on all atoms. Multipoles were obtained from a distributed multipole analysis of the B3LYP/6-311G** charge density.[50] Charge–charge, charge–dipole and dipole–dipole interactions were calculated with Ewald summation. All other interactions were calculated between whole molecules with a centre-of-mass separation of less than 25 Å. Af-

ter removal of duplicates, predicted crystal structures within the lowest 25 kJ mol$^{-1}$ of the energy landscape for each molecule were kept. These datasets consist of 526, 388 and 468 crystal structures of oxalic acid, maleic hydrazide and urazole, respectively.

### 2.2 Solid state DFT

Periodic DFT single-point energy calculations were performed on the force field optimized crystal structures with both the VASP[51–54] code, using plane-wave (PW) basis sets, and CRYSTAL17[55] using Gaussian Type Orbital (GTO) basis sets. The justification of the choice of basis set is discussed in Section 3.2. Generally speaking, PWs are intrinsically periodic, fast for both energy and its gradient calculations, and the quality of a PW basis set is improvable by a single parameter, namely the electronic kinetic energy cutoff. PW calculations were used for the intermediate level (PBE) energy evaluations. However, due to the extremely high computational costs involved in exact exchange calculations with PW basis sets, the highest level, hybrid functional (PBE0) calculations were perfomed using GTO basis sets. Furthermore, we exploited symmetry in CRYSTAL17, which sped up the PBE0 energy evaluations by a factor of up to the number of equivalent molecules in the unit cell.

PW calculations in VASP used the projector-augmented wave (PAW) method and standard pseudopotentials with a plane-wave energy cutoff of 600 eV and maximum k-point spacing of 0.05 Å$^{-1}$. GTO calculations in CRYSTAL17 employed different basis sets and composite methods, details of which are described in Section 3.2. Electronic k-points were sampled uniformly to a resolution of at least 0.02 Å$^{-1}$. Truncation criteria for bielectronic integrals, TOLINTEG, of 12 12 12 12 24 was enough for good convergence in most of the structures, however, there were cases that we had to increase these values to 14 14 14 14 28.

The missing long-range correlation effects in DFT calculations are accounted for by applying the Grimme D3 dispersion correction in both PBE and PBE0 calculations.[56–58] Hereafter, we use PBE and PBE0 to refer to dispersion-corrected PBE-D3 and PBE0-D3 calculations. To correct for the basis set superposition error (BSSE)[59] present in GTO calculations, we used the geometrical counterpoise (gCP) correction with automatic parameter setup[60,61] in all GTO calculations.

Lattice energies were calculated by subtracting the intramolecular energy of the constituent molecules of a crystal in the gas-phase from the total energy of the crystal, always at the same level of theory (functional and basis set). Further details are in the supporting information.

### 2.3 Structural descriptors

Structural descriptors are required to convert the atomistic structure of each predicted crystal structure into a suitable input for the statistical machine learning method. Cartesian coordinates are not a suitable choice because they are not invariant under translation, rotation and reflection of the whole system. Among available atomic descriptors which satisfy these requirements, we use atom-centered symmetry functions[62,63] to describe the local (radial and angular) environment of each atom. We choose these descriptors of atomic local environments because of their success in the development of machine learned force fields[63] and previous machine learning applications to energy model improvement for molecular crystals.[37]

We used the recent modification to symmetry function descriptors that separates element pairs (for radial functions) and triplets (for angular functions) to provide a better resolution of the atomic environment description.[63] 32 equispaced Gaussians were used to describe the radial environment in a cut-off sphere of radius 9.3 Å around each atom. For the angular environment, a cut-off radius of 6.27 Å was used, taking 8 equispaced radial Gaussians and 8 angular ones (64 in total). Thus, for a system with $N_E$ elements, there are $N_E \times 32$ radial and $N_E(N_E+1)/2 \times 64$ angular symmetry functions for each atom.

### 2.4 Multi-fidelity Gaussian process modelling

We use Bayesian GPR[64] to model the relationship between the crystal structures and the response: lattice energies evaluated at multiple levels of accuracy and corresponding computational complexity. Each structure is uniquely identified by the vector of the structural descriptor values $x \in R^s$, concatenated symmetry functions as described in Section 2.3. In a system with $SF_r$ radial and $SF_a$ angular symmetry functions, the dimensionality of structures with $N_A$ atoms is $s = N_A \times (SF_r + SF_a)$.

GPR is an adaptive non-parametric modelling approach with prior beliefs about the behaviour of the relationship, or function, of interest described by a prior Gaussian process. Probabilistic inference and predictions are obtained by updating this prior using collected data.

For one level of computational complexity, a Gaussian process prior on $y(x)$, the response for descriptor vector $x$, is denoted $GP\{f^T(x)\beta,\, k(x,x';\phi,\sigma^2,\tau^2)\}$ and results in the responses from any collection of $n$ structures, represented by the descriptor vectors $x_i$, $i = 1\ldots n$, having the prior predictive multivariate normal distribution

$$y(x_1),\ldots,y(x_n) \sim \mathrm{N}[(f(x_1),\ldots,f(x_n))^T\beta,\, K(x_1,\ldots x_n;\phi,\sigma^2,\tau^2)].$$

The mean is a linear combination of regression functions stored in $f(x_i)$ with coefficient vector $\beta$. The $(i,j)$th element $k(x_i,x_j;\phi,\sigma^2,\tau^2)$ of the covariance matrix $K(x_1,\ldots x_n;\phi,\sigma^2,\tau^2)$ is defined through the correlation kernel $\kappa(x_i,x_j;\phi)$, variance parameter $\sigma^2 > 0$ and a regularization parameter, or nugget, $\tau^2 \geq 0$. For non-deterministic data, the nugget measures random variation around the mean response. For deterministic systems, addition of a nugget is still common and beneficial, improving the conditioning of the prior covariance matrix and providing some robustness of the assumed form of the correlation kernel.[65] Hence

$$k(x_i,x_j;\phi,\sigma^2,\tau^2) = \sigma^2[\kappa(x_i,x_j;\phi) + \tau^2\, 1(x_i = x_j)].$$

The choice of the correlation kernel should reflect prior beliefs about the degree of response smoothness and the sensitivity of the response to the difference between structures. In this work we use an isotropic squared exponential correlation kernel, with the

distance between the structures introduced through the Euclidean distance between the vectors of descriptors:

$$\kappa(x_i, x_j; \phi) = \exp\left(-\frac{1}{2}||x_i - x_j||^2/\phi\right).$$

The hyperparameter $\phi > 0$ quantifies the prior correlation between $y(x_i)$ and $y(x_j)$, with larger values resulting in higher correlation. We denote the adjusted correlation matrix by $\Sigma(\cdot, \phi)$, so that $K(\cdot; \phi, \sigma^2) = \sigma^2\Sigma(\cdot, \phi)$.

When the same response, in this case lattice energy, is measured at different levels of computational complexity and accuracy, a hierarchical Autoregressive Gaussian Process (ARGP) model[66] can be constructed, which links a higher-level complexity response, $y_t$ (e.g. PBE0 lattice energies), to a lower-complexity response, $y_{t-1}$ (PBE lattice energies), through a scaling parameter $\rho$, with the difference $\delta$ modelled as a GPR, independently of the GPR assumed for $y_{t-1}$. For our problem:

$$\begin{cases} y_0(x) = FF(x), \\ y_1(x) = \rho_0 y_0(x) + \delta_1(x), \\ y_2(x) = \rho_1 y_1(x) + \delta_2(x), \end{cases} \tag{1}$$

with

$$\delta_t(x) \sim \mathrm{GP}[\beta_t, k_t(x, x'; \phi_t, \sigma_t^2, \tau_t^2)]$$

for $t = 1, 2$ and $FF(x)$ being the force field energy. The force field provides a useful approximation of the physical contributions to the interactions between molecules and its negligible computational cost ensures $FF(x)$ is fixed and known for all possible training and test molecules.

In addition to the obvious relationships between $y_1$ and $y_2$ through $\rho_1$ and $\delta_1$, model (1) also assumes that once $y_1$ has been observed at point $x$, no other observation of $y_1$ at any $x' \neq x$ furthers our knowledge about the higher level response $y_2(x)$. That is, $\mathrm{Cov}(y_1(x'), y_2(x)|y_1(x)) = 0$, a natural Markov property for multi-level responses.

The joint prior distribution of the responses at levels 1 (PBE) and 2 (PBE0) is then

$$\begin{bmatrix} y_2(x) \\ y_1(x') \end{bmatrix} \sim \mathrm{N}\left[ \begin{array}{c} \rho_1(\rho_0 FF(x) + \beta_1) + \beta_2 \\ \rho_0 FF(x') + \beta_1 \end{array}, \right.$$

$$\left. \begin{pmatrix} \rho_1^2\sigma_1^2[1+\tau_1^2] + \sigma_2^2[1+\tau_2^2] & \rho_1 k_1(x, x'; \phi_1, \sigma_1^2, \tau_1^2) \\ \rho_1 k_1(x, x'; \phi_1, \sigma_1^2, \tau_1^2) & \sigma_1^2[1+\tau_1^2] \end{pmatrix} \right].$$

Extensive research has been conducted regarding fitting of ARGP models and their variations.[67–69] We adopt the recursive Bayesian multi-fidelity approach,[70] suitable for nested training sets. Sequential estimation of GPRs is performed for the two levels; first, a standard GPR for $y_1(x)$ is fitted on the $n_1$ structures composing the lower-level training set $D_1 = \{x_1, \ldots, x_{n_1}\}$. Secondly, the higher-level response $y_2(x)$ is modelled with a GPR on the subset $D_2 \subseteq D_1$ of size $n_2$, with the GPR prior for $y_1(x)$ in (1) being replaced by its GPR posterior obtained at the first stage.

To complete the model specification, where possible we choose conjugate normal and inverse gamma (IG) prior distributions for the model parameters, conditional on lower-level responses, that allow for straightforward updated inference and predictive distributions in closed form:

$$\rho_0, \beta_1 | \sigma_1^2, y_0 \sim \mathrm{N}(b_{10}, \sigma_1^2 R_1);\ \sigma_1^2 | y_0 \sim \mathrm{IG}(\alpha_1, \gamma_1),$$

$$\rho_1, \beta_2 | \sigma_2^2, y_1 \sim \mathrm{N}(b_{20}, \sigma_2^2 R_2);\ \sigma_2^2 | y_1 \sim \mathrm{IG}(\alpha_2, \gamma_2).$$

Here $R_t$ are appropriately sized (here – $2 \times 2$) prior variance-covariance matrices, containing the prior variance scaling between $\sigma_t^2$ and linear coefficients of the model. We chose $R_t$ to be the identity matrices, prior means for linear terms $b_{t0} = (0.5, 0)^T$, and both shape and scale parameters $\alpha_t$ and $\gamma_t$ were set to 1. Regularization parameters $\tau_t^2 = 10^{-5}$ were chosen to resolve any computational singularity issues when inverting $\Sigma_t$. After the data $y_t$ is available at level $t = 1, 2$, the posterior joint distributions of the parameters are also normal-inverse-gamma:

$$(\rho_{t-1}, \beta_t | y_t, y_{t-1}, \sigma_t^2) \sim \mathrm{N}(\tilde{\Sigma}_t \mu_t, \tilde{\Sigma}_t), \tag{2}$$

$$\tilde{\Sigma}_t = \sigma_t^2[F_t^T \Sigma_t^{-1} F_t + R_t^{-1}],^{-1}\ \mu_t = \frac{1}{\sigma_t^2}[F_t^T \Sigma_t^{-1} y_t + R_t^{-1} b_{t0}],$$

$$(\sigma_t^2 | y_t) \sim \mathrm{IG}\left(\frac{n_t}{2} + \alpha_t,\ Q_t(y_t, \gamma_t, b_{0t}, \Sigma_t, F_t)\right). \tag{3}$$

Here $\Sigma_t$ is the $n_t \times n_t$ correlation matrix at level $t$ and matrix $F_t = [y_{t-1}(D_t), 1_{n_t}]$; the detailed expression for $Q_t$ can be found in the literature.[70]

Uniform prior distributions were assumed for correlation hyperparameters $\phi_1, \phi_2$, and an empirical Bayes approach employed. For posterior inference, these parameters were set equal to their posterior mode.

Conditional on parameters $\rho_{t-1}$, $\beta_t$ and $\sigma_t^2$, the posterior predictive distribution for the highest level response (PBE0) is Gaussian. However, the marginal posterior predictive distribution, with these parameters integrated out with respect to their posterior distributions (2) and (3), is not available in closed-form. However, its mean and variance are available, and take the following form:

$$E[y_2(x_*)|y_1, y_2] = \hat{f}_2(x_*)\tilde{\Sigma}_2\mu_2 + k_2^T(x_*)\Sigma_2^{-1}(y_2 - F_2\tilde{\Sigma}_2\mu_2), \tag{4}$$

$$Var[y_2(x_*)|y_1, y_2] = (\hat{\rho}_1^2 + \tilde{\Sigma}_{2,1,1})Var[y_1(x_*)|y_1] + \frac{Q_2}{n_2 + 2(\alpha_2 - 1)}(1 - k_2^T(x_*)\Sigma_2^{-1}k_2(x_*)) + (\hat{f}_2(x_*) - k_2^T(x_*)\Sigma_2^{-1}F_2)\tilde{\Sigma}_2(\hat{f}_2(x_*) - k_2^T(x_*)\Sigma_2^{-1}F_2),^T \tag{5}$$

where $\hat{\rho}$ is the posterior mean of $\rho$, $\tilde{\Sigma}_{2,1,1}$ is the first diagonal element of $\tilde{\Sigma}_2$, $k_2(x_*) = [\kappa_2(x_*, x_{21}; \phi_2), \ldots, \kappa_2(x_*, x_{2n_2}; \phi_2)]^T$ is the vector of correlations between the unobserved structure $x_*$ and the training set for the highest level response and $\hat{f}_h(x_*) = \{E[y_l(x_*)|y_l], 1\}$. Sampling from the posterior predictive distribution can proceed via Monte Carlo methods or by using a $t$-distribution with $2\alpha_2 + n_2$ degrees of freedom, mean (4) and vari-

ance (5). After checking the adequacy of this approximation for our examples, we adopt this latter approach for computational convenience.

### 2.5 Training

For each of the three molecules, training sets for the lower fidelity response (PBE) were chosen that included between 10% to 70% of all structures (in increments of 5%). These training sets were chosen to be nested, that is, the 65% set is a subset of the 70% set and so on. For each of these training sets, nested subsets were chosen again containing between 10% to 70% of structures (in increments of 5%) for which the higher fidelity response (PBE0) was made available. The size of these high level training sets are given as percentages of the lower level training set from which they are drawn. Each subset was chosen as a maximin space-filling design[71], maximizing the minimum Euclidean distance in descriptor space between structures. Hence a collection of 169 nested training sets were constructed, containing between 1% and 49% of the whole data set. We use the nomenclature "$L\%/H\%$" training set to refer to a set containing evaluations of PBE for $L\%$ of the original structures and evaluations of PBE0 for $H\%$ of the PBE training set. The 30% of structures that were not included in any training set were reserved as a test set.

## 3 Results

### 3.1 Data Collection

Given the predicted crystal structures taken from previous work with structures optimized at the force field level, the first stage in this work was the re-evaluation of their lattice energies at higher levels of theory. Our final target in this work is the evaluation of lattice energies using the hybrid PBE0 functional with the D3 dispersion correction.

Because of the availability of fast Fourier transforms when using PWs, their use as basis sets is efficient for GGA DFT calculations. Although the exploitation of symmetry in CRYSTAL17 using GTO basis sets would be helpful for larger unit cells with more molecules involved, we decided on PBE within VASP as a fast and accurate lower level of energy re-evaluation.

The choice of the calculation method for collecting the high level PBE0 data points is more significant. This is because of the extremely high cost of exact exchange calculations when employing delocalized basis sets such as PWs, which has limited the use of such sets in hybrid functional calculations. GTOs, on the other hand, require much less (up to 2 orders of magnitude) resources to achieve equivalent accuracy of a PW method.[72] The cost of GTO calculations depends on the size of basis set and basis sets of at least triple-zeta quality including one set of polarization functions have been shown to provide good agreement with converged PW results.[72]

Thus, we have performed a comparison of GTO basis sets to PW calculations on the lattice energies of one of our sets of CSP structures, oxalic acid. We first compared the results of well-converged VASP PBE lattice energies on the full set of 526 oxalic acid crystal structures to those from CRYSTAL17 using the Ahlrich's-type split valence double-zeta (def2-SVP) and triple-zeta (def2-TZVP)[73] basis sets (Fig 2). Enlarging the basis set from def2-SVP (Fig. 2a) to def2-TZVP (Fig. 2b) reduces the mean absolute error (MAE) and maximum absolute error (MAX) in relative lattice energies (always measured relative to the global energy minimum structure, which is the same in PBE and PBE0), and improves the correlation to PW results considerably. Absolute lattice energies are also improved significantly, from an MAE of 43 kJ mol$^{-1}$ with def2-SVP to less than 1 kJ mol$^{-1}$ with def2-TZVP (Fig. S1). The PBE/def2-TZVP results show an excellent linear relationship to those from PW calculations, as well as small differences in the absolute lattice energies.

To assess whether this convergence is also observed for PBE0, calculations were performed on a subset of the oxalic acid predicted crystal structures (Fig. 2c and d). We started from 20 crystal structures in 7 space groups. However, three of the PBE0/PW single point calculation did not converge in 48 hours on 192 computing cores and were abandoned. For the remaining 17 structures, errors in relative lattice energies with the smaller def2-SVP basis set are too large (MAE $= 4.45$ and MAX $= 9.83$ kJ mol$^{-1}$), considering the close energetic spacing of predicted structures, with an $R^2$ of 0.490 to the PBE0/PW relatice lattice energies. The results using the larger def2-TZVP basis set have much better correlation to PW ($R^2 = 0.926$) while MAE and MAX errors are 1.37 and 3.41 kJ mol$^{-1}$, respectively. Given the much lower computational cost of PBE0/def2-TZVP than PW basis sets, we used this GTO basis set for PBE0 calculations of the energies for the full CSP structure sets for all three molecular systems (oxalic acid, urazole and maleic hydrazide).

We also tested two simplified DFT schemes, HSE-3c and B97-3c,[74] as possible intermediate energy models. Both methods use modified, smaller basis sets combined with semi-empirical corrections. B97-3c and HSE-3c lattice energies are both strongly correlated to the PW results ($R^2 = 0.984$ and 0.869, respectively, Fig. S2). However, when the cost of calculations is taken into account, neither method offered an advantage over PBE/PW for the small molecules studied here.

### 3.2 CSP landscapes

The energy-density distribution of predicted crystal structures of oxalic acid is presented in Fig. 3, showing all structures that lie within 25 kJ mol$^{-1}$ of the global minimum from the force field calculations. While the two known polymorphs, $\alpha$ and $\beta$, have very different packing motifs, their experimentally determined lattice energies are very close, with $\alpha$ lying slightly lower than $\beta$.[75] The force field lattice energies predict the $\beta$ polymorph to be the lowest energy structure (Fig. 3a), but incorrectly produce over 70 other crystal structures below $\alpha$, which is more than 7 kJ mol$^{-1}$ above the global minimum.

The known crystal structures of urazole and maleic hydrazide are also predicted poorly by the force field calculations; their landscapes are provided in the SI (Figs. S3 and S5). The three known polymorphs of maleic hydrazide lie between 5.74 and 7.83 kJ mol$^{-1}$ above the global minimum, and the only known crystal structure of urazole lies 3.39 kJ mol$^{-1}$ above than the global minimum. While the FIT+DMA force field has been suc-

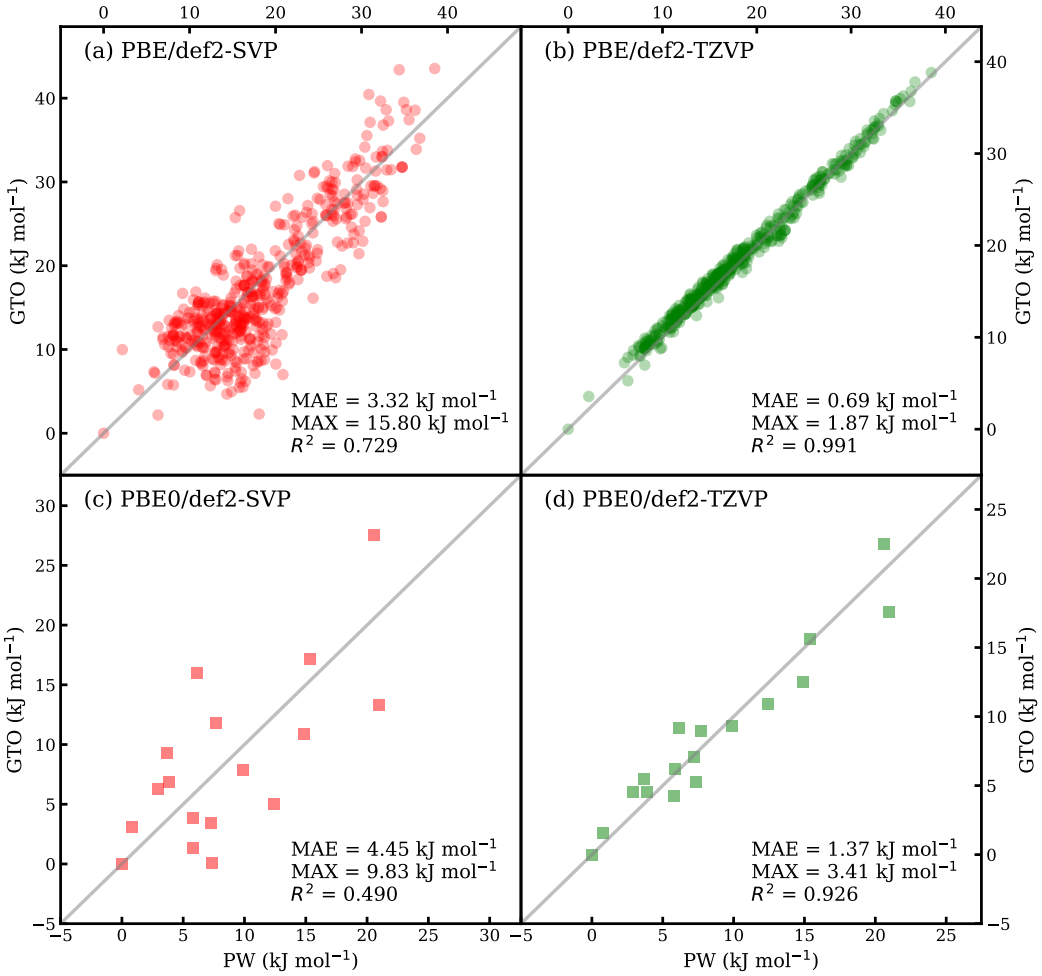



Fig. 2 Calculated relative lattice energies of the oxalic acid data set from GTO basis set calculations vs PW calculations for (a) PBE/def2-SVP, (b) PBE/def2-TZVP, (c) PBE0/def2-SVP, and (d) PBE0/def2-TZVP. The mean absolute error (MAE), maximum absolute error (MAX) and correlation to the PW values are reported in the inset of each plot. All 526 crystal structures are included in the PBE calculations, while PBE0 comparisons were preformed for 17 crystal structures.

cessful for CSP of many small molecules, these three molecules were chosen as known failures, where it is clear that a higher level of theory is necessary for successfully ranking the crystal structures.

All higher level energy evaluations in this work are performed at the force field optimized structures, with no further optimization. For all three molecules, re-evaluation of the energies at the PBE and PBE0 levels of theory improve the position of the experimentally known crystal structures on their CSP landscapes (Figure 3b,c and SI). PBE retains the low energy of $\beta$-oxalic acid and brings the $\alpha$ polymorph to the global minimum, while PBE0 further increases the relative stability of $\alpha$ to around around 5 kJ mol$^{-1}$ below that of $\beta$. These energy differences compare well with the lattice energy difference of around 3 kJ mol$^{-1}$ between $\alpha$ and $\beta$ when fully re-optimized at PBE0.[76] Similarly, the known structure of urazole is re-ranked to the global energy minimum with PBE and 2nd lowest energy structure with PBE0, 1.04 kJ mol$^{-1}$ above the global minimum. and the three polymorphs of maleic hydrazide are all brought closer to the global minimum with PBE and PBE0 (see Figs S3, S5).

Thus, we conclude that single point energy re-evaluations using solid state DFT map the force field results for all three molecules to more realistic energy rankings in which the known experimental crystal forms are either the global minimum or very close to the global minimum on their respective landscapes. Empirically, we find that PBE provides slightly better rankings of the known crystal structures for all three molecules. However, for the purposes of developing the method, we treat PBE0 as the target for learning a high level energy. The performance of PBE0 might be limited by the single point nature of the energetic re-evaluation and, so, optimization on the GPR modelled energy surface is a clear next step of development, but is reserved for future work.

Ideally, one would prefer to evaluate the predicted crystal structures by the highest accuracy possible, however, even after the speed-up gained from using GTOs and exploiting the symmetry of crystals, the cost of hybrid functional calculations is still prohibitively high. In the specific case of 526 single point energy evaluation of the oxalic acid data set, calculations at force field, PBE and PBE0 levels of theory using 40 2.0 GHz Intel Skylake processors, took under a minute, 11 hours, and 101 days, respec-

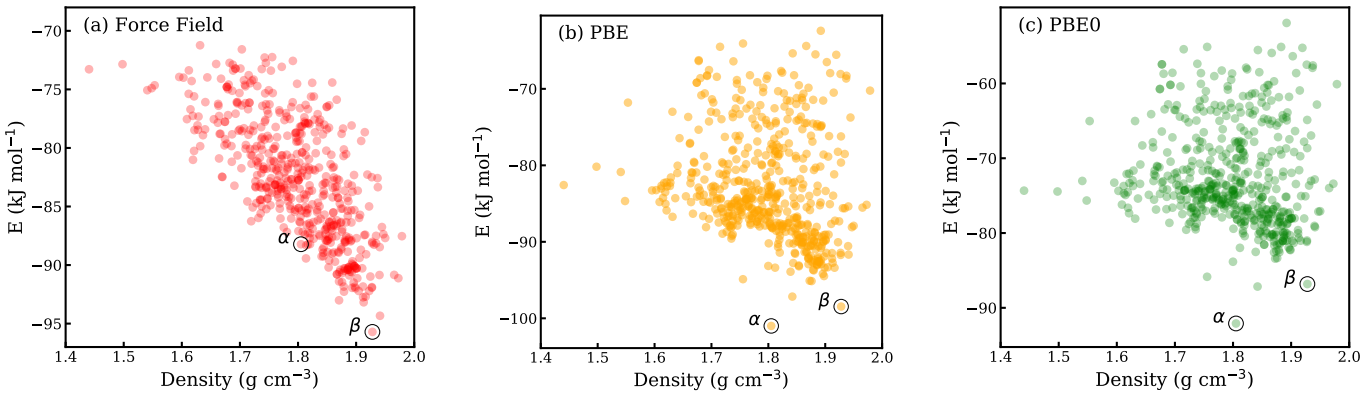



Fig. 3 Lattice energy vs density distributions of the predicted oxalic acid crystal structures calculated using (a) the FIT+DMA force field, (b) PBE, and (c) PBE0, all at the force field geometries. Each point corresponds to a distinct predicted crystal structures. The two known forms of oxalic acid, $\alpha$ and $\beta$, are marked with black circles.

tively. The corresponding cost ratio is roughly 1:1,000:220,000, and is expected to be more extreme for larger molecules. This large ratio of computational costs highlights the motivation for an efficient, reliable estimation of PBE0 lattice energies. Average timings for each molecule at each level of theory are provided in the SI (Table S1).

A multi-level approach seems promising rather than directly learning the most expensive model from the structural descriptors. This is because the force field energies are already in-hand for every structure as a result of the global structure search, and are based on FIT+DMA, a physically motivated force field that provides a good baseline. This is clear from the correlation to PBE lattice energies (Fig. 4a). In particular, FIT+DMA provides accurate electrostatic energies, whose long range might be impossible to model accurately based on descriptors of local atomic environments. Furthermore, since most physics of the problem is captured at the PBE level, the PBE/PBE0 linear correlation (Fig. 4b) is very high ($R^2 \approx 0.98$) and so much of the relationship between the expensive PBE0 energies and the structural descriptor can be learned from PBE calculations, at a much reduced computational cost. From these features, we anticipate model (1) being successful with parameter $\rho$ capturing the strong linear relationship between PBE and PBE0. Inadequacies in a simple linear correlation will be described with the GPR prior on the difference $\delta$, and learning $\delta$ allows efficient prediction of PBE0 energies using less expensive evaluations of PBE.

### 3.3 Prediction

We assess the quality of energy modelling by evaluating the prediction MAE and Continuous Ranked Probability Score (CRPS) of the posterior predictive distributions[77]. CRPS is a proper scoring rule regularly used to assess probabilistic prediction skill, which accounts for both accuracy and precision. For the probabilistic predictions from the GPR, CRPS provides a meaningful way of evaluating the distance between the posterior predictive distribution and the test observations. For a deterministic point prediction, e.g. simply using values of force field energies, CPRS corresponds with MAE. As we obtain a separate posterior predictive distribution for each structure, we will be using the mean of the scores calculated for $N_t$ structures:

$$\mathrm{CRPS}(\{G_i, y_i^E\}_{i=1}^{N_t}) = \frac{1}{N_t}\sum_{i=1}^{N_t}\int_{-\infty}^{+\infty}(G_i(y) - 1(y \geq y_i^E))^2 dy.$$

CRPS is available in closed form for the $t$-distribution we use to approximate the posterior predictive distribution, implemented in R package `scoringRules`.[78]

#### 3.3.1 Single fidelity GPR modelling

We first look at the predictions obtained from the first step of the recursive modelling: fitting a GPR model to PBE with force field measurements as a fixed predictor. As per Section 2.5, we use nested maximin designs as training sets. For comparison, we also present results for 30 randomly selected training sets of each size (boxplots in Figure 5a). The MAE of the mean of the predicted distribution in reproducing the test set PBE lattice energies decreases from just above 2 kJ mol$^{-1}$ at a 10% training fraction (53 structures for oxalic acid) to 0.5 kJ mol$^{-1}$ using 70% of structures for training (Figure 5a), with the error improvements slowing down after the training fraction reaches 40%, which is 210 structures from the oxalic acid data considered. On average, random and maximin training set selection yield similar errors at small ($< 30\%$) training fractions, but for training fractions larger than 30%, the maximin training set selection leads to noticeably smaller errors. This finding agrees with earlier observations for single fidelity GP prediction of lattice energies.[36]

The observed relationship between the error and the training fraction holds true for the crystal structure landscapes of all three molecules. The MAE and CRPS both decrease as the training proportion increases (Fig. 5b), but at a slower rate for maleic hydrazide and urazole than for oxalic acid. The different rates of decrease in the errors probably reflect more complex structural landscapes, particularly of urazole, and more difficult relationships between structural descriptors and lattice energies to model. The MAE and CRPS show similar trends and we shall hereafter evaluate the prediction quality in terms of CRPS. All corresponding MAE values can be found in SI.

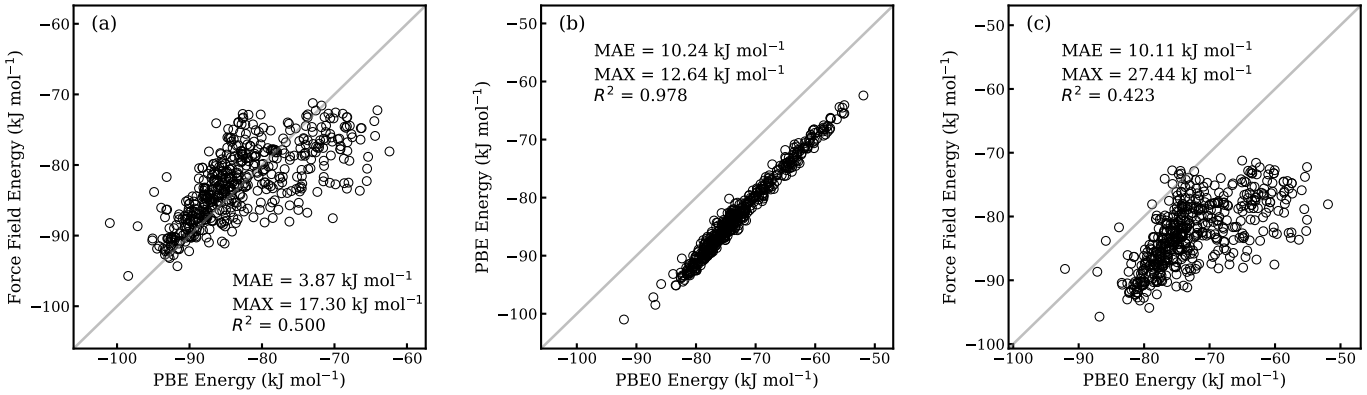



Fig. 4 Correlations between lattice energies calculated using the three energy models for the predicted crystal structures of oxalic acid. (a) Force field vs. PBE lattice energies, (b) PBE vs. PBE0, and (c) force field vs. PBE0. Mean absolute error (MAE), maximum error (MAX), and correlation ($R^2$) are reported in the inset of each panel.

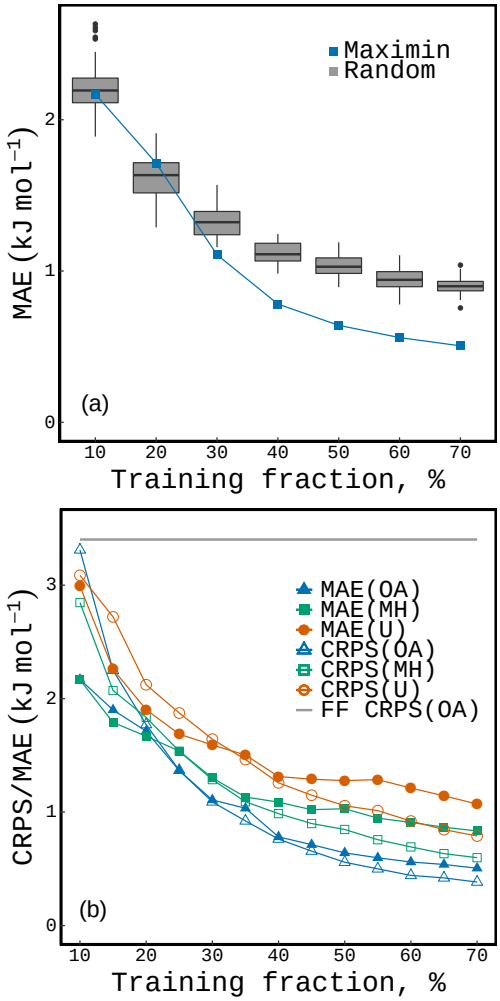



Fig. 5 Single fidelity (PBE) modelling. (a) MAE in reproducing OA lattice energies using randomly chosen and maximin training sets. The boxplots represent results from 30 randomly selected training sets for each training fraction. (b) CRPS (open symbols) and MAE (filled symbols) in reproducing lattice energies of oxalic acid (OA, blue), urazole (U, orange) and maleic hydrazide (MH, green). The force field (FF) based CRPS is 3.4 kJ/mol for OA and shown as a horizoltal grey line. FF-based CRPS of 6.4 kJ/mol for urazole and 14.1 kJ/mol for MH are not displayed.

### 3.3.2 Multi-fidelity ARGP modelling

Predictions and true values of the PBE0 lattice energies for the oxalic acid test set (30% of original structures) are presented in Figure 6 for three modelling strategies. Model A (Fig. 6a) is the ARGP model, trained on PBE evaluations for 30% of all structures (training set I) and PBE0 evaluations on 30% of this set (9% of all structures, training set II). These are compared to predictions from two single-level PBE0 GPR models: model B (Fig. 6b) is trained on PBE0 energies obtained for training set I; and model C is trained on the PBE0 energies for training set II (Fig. 6c). Hence, model B had available a greater number of expensive training points than was the case for model A, and model C had the same number of expensive training points as model A but did not make use of the larger number of cheaper points.

The CRPS from the ARGP model A is 1.114 kJ mol$^{-1}$, compared with 1.112 kJ mol$^{-1}$ for the predictions obtained from the expensive single GP model B. To obtain this very minor improvement requires an increase in computational cost of more than a factor of 3: just 9% of the expensive PBE0 data used for training model A compared to 30% for model B. The PBE calculations needed for the ARGP model have only a small influence on the total computational cost. Model C, on the other hand, has a comparable cost to the ARGP model; the difference in computational costs occurs only from using the PBE data for 30% of the structures in model A. The CRPS for this single GPR model is considerably higher, at 3.96 kJ mol$^{-1}$.

The comparisons in Figure 6 show the clear benefits of the ARGP model, whose cost and accuracy are determined by the training set sizes at the lower and higher levels. Figure 7 (and Tables S3-S5 in SI) compares the prediction errors from, and relative computation costs of, ARGP models for oxalic acid with differing sized training sets for PBE and PBE0. Analagous plots and tables are provided for maleic hydrazide and urazole in the SI. Costs are defined relative to the expense of obtaining the lattice energies of 1% of structures using PBE0. Thus, from the average timings (see Table S1), the cost of obtaining PBE evaluations for 1% of oxalic acid structures is $4.56 * 10^{-3}$. The computational costs of force field calculations and statistical modelling are negligible. As an

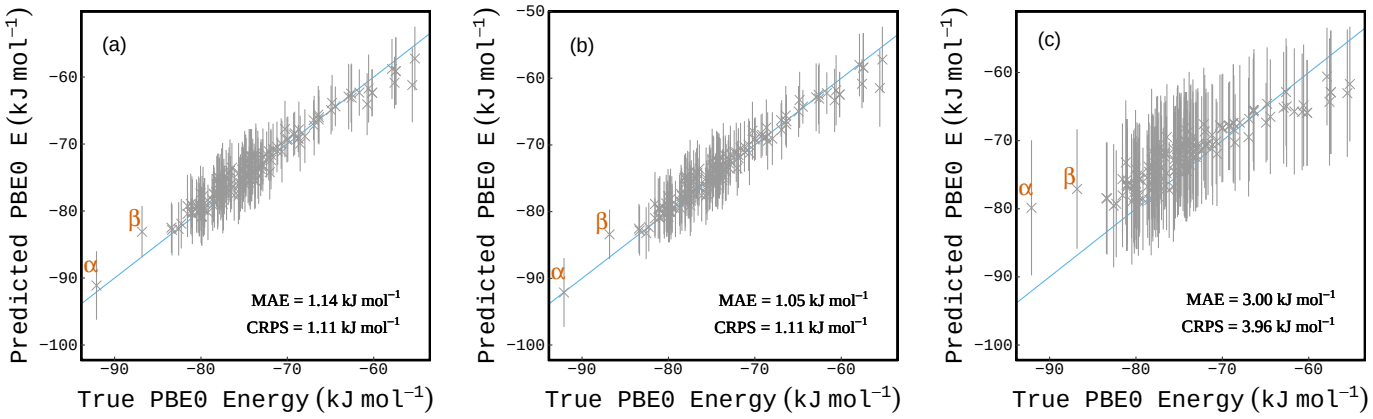



Fig. 6 Prediction of oxalic acid PBE0 lattice energies (±2 SD): (a) using the recursive approach (Model A) on 30%/30% training sets; single GP regression models: (b) on the larger, 30%, training set (Model B) and (c) on the smaller, 9%, training set (model C).

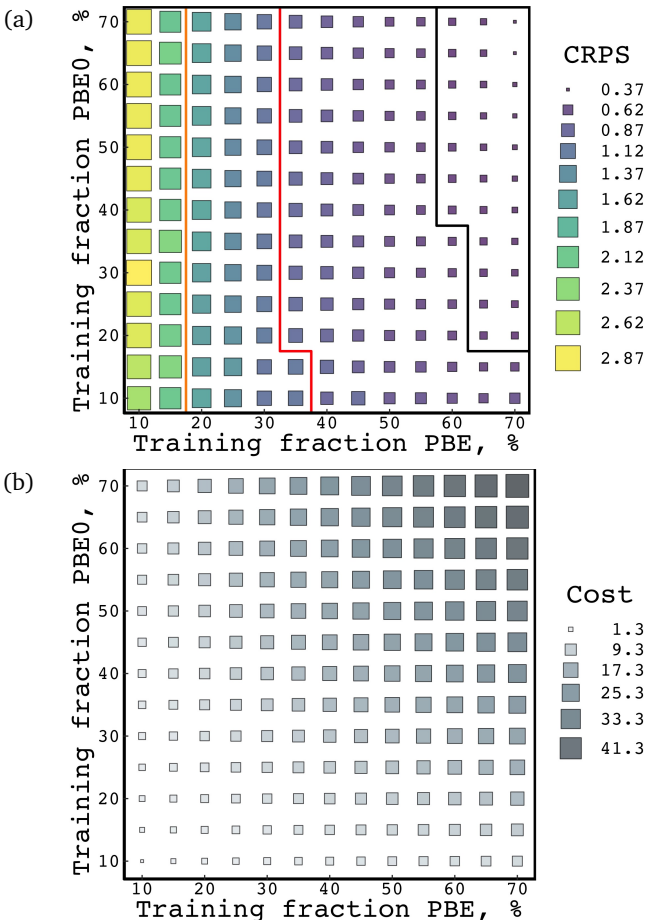



Fig. 7 Landscapes of PBE0 prediction errors and recursive modelling costs for the range of training sets for the oxalic acid data set: (a) CRPS prediction errors (kJ mol$^{-1}$); (b) Relative computational cost (where 1 corresponds to the cost of PBE0 calculations on 1% of structures). The horizontal axis gives percentage of structures for which PBE evaluations are available, and the vertical axis gives the relative percentage of the structures for which PBE0 was also available (so each cell represents the performance/cost of a single $L\%/H\%$ training set). The orange contour in (a) separates the models which give errors above and below 2 kJ mol$^{-1}$; red and black contours delineate models that give better than 1 kJ mol$^{-1}$ and 0.5 kJ mol$^{-1}$ errors, respectively.

example, fitting the ARGP model using a 30%/60% training set incurs a cost of $(4.56*10^{-3}*30)+(1*30*60/100)=18.14$. This training set is almost as expensive as a 60%/30% set (18.27), since the number of PBE0 calculations is the same, but the mean prediction CRPS decreases from 1.04 kJ mol$^{-1}$ to 0.52 kJ mol$^{-1}$, due to the increase in PBE training points in the latter set. With the high-fidelity-only (single-level) GPR modelling approach, obtaining a similar prediction error (0.53 kJ mol$^{-1}$) requires running PBE0 computations on 55% of the structures, which is 3 times more costly. Similar tables are presented in the Supplementary information for maleic hydrazide and urazole.

The typical energy differences seen between predicted crystal structures, and between observed polymorphs, set target criteria for acceptable errors, which can inform the choice the training fractions. Over 50% of observed polymorphs of organic molecules are separated in lattice energy by less than 2 kJ mol$^{-1}$,[24] making this an upper bound for acceptable errors. The orange contour in Figure 7 separates models that provide errors above and below 2 kJ mol$^{-1}$ (CRPS), with red and black contours showing stricter 1 and 0.5 kJ mol$^{-1}$ thresholds, respectively. These contours are nearly vertical, showing that the target errors can be met by increasing the lower-level, PBE, training fraction without the expense of increasing the fraction of PBE0 calculations. Due to the large cost ratio between PBE0 and PBE, the lowest cost model meeting the target error involves the smallest PBE0 training fraction: only 10% of the lower level training set. To meet a smaller target error of 0.5 kJ/mol requires 20% PBE0 for oxalic acid and is not achieved at all for maleic hydrazide and urazole. Although small numbers of PBE0 calculations are required, these are necessary for achieving the required accuracy; pure PBE-based models yield significantly larger errors (see Table S2).

The errors and costs of the lowest cost ARGP models that give errors below 2 and 1 kJ mol$^{-1}$ are listed for all three molecules in Table 1, along with the most expensive 70%/70% models. For oxalic acid, the 40%/10% model yields errors below 1 kJ mol$^{-1}$ with a cost corresponding to only 4.2% of the cost of PBE0 calculations on all crystal structures. Similar savings are possible for maleic hydrazide and urazole, yielding errors below 1 kJ mol$^{-1}$

| Molecule | PBE/PBE0 (%/%) | Number of structures | Time (CPU.hr) | CRPS (kJ/mol) | MAE |
|---|---|---|---|---|---|
| Oxalic Acid | 20/10 | 105/10 | 1,946 | 1.67 | 1.93 |
| | 40/10 | 210/21 | 4,079 | 0.84 | 1.00 |
| | 70/70 | 368/257 | 48,052 | 0.37 | 0.50 |
| Urazole | 25/10 | 117/11 | 2,971 | 1.80 | 2.00 |
| | 65/10 | 304/30 | 8,086 | 0.94 | 1.22 |
| | 70/70 | 327/228 | 59,092 | 0.84 | 1.10 |
| Maleic Hydrazide | 15/10 | 58/5 | 1,886 | 1.88 | 1.96 |
| | 50/10 | 193/19 | 7,105 | 0.94 | 1.22 |
| | 70/70 | 270/189 | 66,978 | 0.60 | 0.83 |

Table 1 Recursive modelling PBE0 energy prediction CRPS and MAE values and computational cost comparison for the molecular systems. The presented models are the least computationally expensive to provide CRPS values below 2 kJ mol,$^{-1}$ 1 kJ mol$^{-1}$ and the ones providing the best achieved score values. Training proportions are displayed together with the corresponding numbers of structures. Costs in CPU hours are calculated based on the average timings per structure (see Table S1), calculated on 2.0 GHz Intel Skylake processors.

at costs of 5.2% and 6.8% of the full PBE0 cost, respectively.

#### 3.3.3 Ranking

Another aspect of the modelling approach quality that is of a particular interest in CSP is the accuracy of the energy ranking predictions. Inference regarding the rank predictions is derived by sampling from the posterior predictive distributions for test structure lattice energies and examining the resulting rank distributions. Figure 8a displays the ranks generated from $10^5$ samples drawn from the posterior predictive distribution from training the ARGP model on the 40%/10% training sets. Figures 8b-e present a closer look at the sampled rank distributions for the first four predicted structures. We find that the known crystal structures $\alpha$ and $\beta$ are recognised with high certainty as those with the lowest energies. In general, structures in the lower energy region, which is usually of the main interest, are quite well identifiable, unlike those in the middle range, which have larger uncertainties in ranking. This is likely due to the high number of structures with very close PBE0 energy values in the energy region $> 8$ kJ mol$^{-1}$ above $\alpha$ (Fig. 6).

To assess the quality of rank predictions, we measure the Kendall rank correlation coefficient between the PBE0-based ("true") rankings and the sampled ones. This coefficient assesses the difference in the proportions of correctly and incorrectly ordered pairs among all possible pairs in two sequences of observations, thus providing a general measure of the degree of similarity between the orderings. First we consider the correlations between the sampled (as above) and the "true" ranks for the top ranked series of structures: from the first 10 to the first 100 (Fig. 9a). We find that, while among the first 10 lower energy structures, on average, more than half of the pairs are ordered concordantly, this proportion decreases as the energy range increases (i.e. in the first $20-30$ ranked structures). However, the correct ranking improves again for the test structures with higher energies. For all 158 test structures, across all $10^5$ posterior samples, the (average) prediction ordering agrees with the PBE0-based one for 86% of pairs, and does not agree for 14%, leading to a median correlation coefficient of 0.72.

Another relationship of interest is the trade-off between the amount of data used for model training (and the associated costs) and the goodness of the posterior rank predictions, which is displayed in Figure 9b, and ordered by the total number of structures used for training. We observe the expected general tendency of the quality of posterior rank predictions to increase with the training set size, and that the improvement becomes less steep as more structures are used for model training. The results clearly demonstrate that the distribution of the training structures across the two levels is important in producing higher rank correlations at low computational cost. As an example, the distribution of sampled posterior rank correlations produced by the least expensive model providing the CRPS value below 1 kJ mol$^{-1}$ (trained on 40%/10% fractions, as listed in Table 1) performs considerably better than the model trained on 30%/30% fractions, at less than half of the cost.

## 4 Conclusions

We present a statistical machine learning approach to predict high quality, hybrid functional (PBE0) DFT energies for crystal structure prediction by relating crystal structure to lattice energy, trained on a subset of predicted crystal structures through descriptors of local atomic environments. The main development that we present is a multi-fidelity model for energetic predictions. The recursive GPR modelling approach takes advantage of correlations between lattice energies calculated using different methods to increase predictive accuracy using low numbers of the highest level, most computationally demanding calculation. Thus, accurate predictions of the high-level calculated energies can be obtained at a fraction of the cost of the full calculations by making use of less computationally demanding, lower-fidelity data. The method is applied to the crystal structure landscapes of three small molecules that have proven challenging for force field-based prediction: oxalic acid, maleic hydrazide and urazole.

Crystal structure prediction is a well-suited application for a statistical modelling approach because of the large numbers of crystal structures, all with identical chemical composition and differing only in the arrangement of their constituent molecules. For the three molecules studied here, each crystal structure landscape contains between 388 and 526 distinct crystal structures in the energy range studied and we show that PBE0 lattice energies can be predicted to an accuracy of 1 kJ mol$^{-1}$ with between 4.2 and 6.8% of the computational cost of the full PBE0 calculations. This cost reduction is achieved by using a larger set of training structures calculated at a lower level of theory, here the GGA functional PBE, after which a very small number of PBE0 calculations are required for the second level of the model. As the cost of GGA and hybrid calculations scale roughly as $\mathcal{O}(n_e^3)$ and $\mathcal{O}(n_e^4)$, respectively,[79] this feature is of particular importance for larger molecules and could bring substantial cost savings.

It is also encouraging that the method has been effective de-

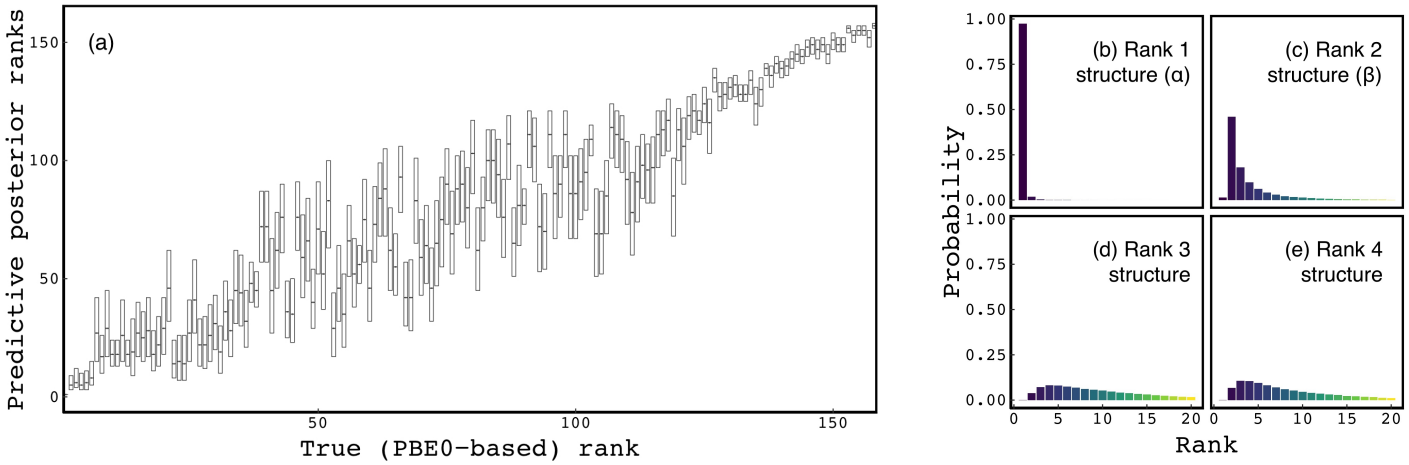



Fig. 8 (a) Posterior rank distributions for the oxalic acid test structures. $10^5$ PBE0 energy samples from the predictive posterior distribution, based on the recursive model trained on $40\%/10\%$ of the data. The test structures are ordered along the horizonal axis according to the PBE0-based ranking and edges of the boxplots outline the inter-quantile range of the sampled data. (b - e) Posterior rank probabilities for the first four oxalic acid structures.

spite differences in the types of basis set used for the GGA (plane wave basis sets) and hybrid functional (Gaussian type basis sets) calculations. That these differences have apparently not harmed the performance of the models allows us to take advantage of different solid state DFT implementations to gain maximum efficiency in the calculations. More generally, the savings that we see in high level calculations open up the use of even more accurate and demanding calculations as the highest level of the model, because of the small number of structures on which these calculations are necessary. An option here is the implementation of the multi-fidelity GP modelling approach with fragment-based lattice energy models,[37,80,81] enabling the use of wavefunction-based higher levels.

In the context of structure prediction, where the lowest energy structures are usually considered most important, identifying the energetic ordering of the structures is particularly relevant. We examined the performance of the model-based predictions in terms of the rankings and observed good rank correlations with the full PBE0 results, particularly for the lower spectrum of the energy values, corresponding to the most important structures.

As a statistical model, the predicted output for each structure is a distribution of energies, which enables an assessment of the associated uncertainty for each of the test structures. The uncertainty in final energetic predictions, as well as ranking, is an important consideration in applying a statistical machine learning model in applications of CSP. The acceptable level of uncertainty for a particular application can be used in selecting training set sizes and the prediction uncertainties are important in the interpretation of results, for example the probability that a predicted, as-yet unobserved polymorph of a pharmaceutical molecule is lower in energy than known crystal structures. The uncertainties in relative energies can also be incorporated in probabilistic interpretations of structure-energy-property maps that have been developed for materials discovery using CSP.[18–20]

## Conflicts of interest

There are no conflicts to declare.

## Acknowledgements

We thank the EPSRC for funding, via grant EP/S015418/1. We 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. We are also grateful to the UK Materials and Molecular Modelling Hub for computational resources, which is partially funded by the EPSRC (EP/P020194/1). We thank David McDonagh for sharing his experiences during the first stages of the project. RH thanks Barry Searle for his helpful comments on convergence of CRYSTAL17 calculations.

## Data access

All data supporting this study are openly available from the University of Southampton repository at https://doi.org/10.5258/SOTON/D1398

## Notes and references

1 A. Burger and R. Ramberger, *Microchimica Acta*, 1979, **72**, 273–316.
2 A. Gavezzotti, *CrystEngComm*, 2002, **4**, 343–347.
3 A. Nezzal, L. Aerts, M. Verspaille, G. Henderickx and A. Redl, *Journal of Crystal Growth*, 2009, **311**, 3863–3870.
4 A. Patterson and B. P. Groshens, *Nature*, 1954, **173**, 398–398.
5 D. Chapman, *Chemical Reviews*, 1962, **62**, 433–456.
6 L. Nicoud, F. Licordari and A. S. Myerson, *Crystal Growth & Design*, 2018, **18**, 7228–7237.
7 T. Beyer, G. M. Day and S. L. Price, *Journal of the American Chemical Society*, 2001, **123**, 5086–5094.
8 J. T. A. Jones, D. Holden, T. Mitra, T. Hasell, D. J. Adams, K. E. Jelfs, A. Trewin, D. J. Willock, G. M. Day, J. Bacsa, A. Steiner

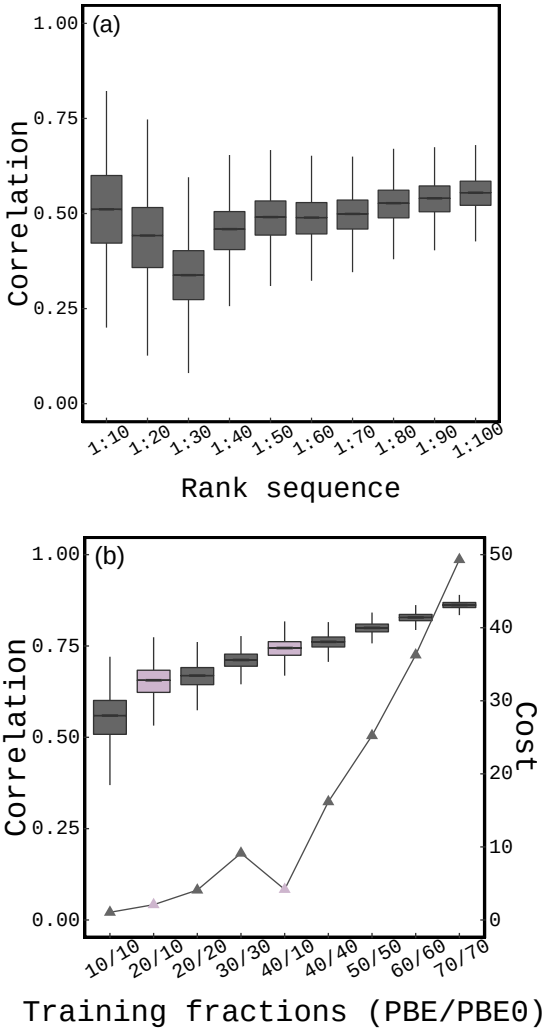



Fig. 9 Distributions of Kendall correlations for the oxalic acid test structures : (a) for the sequences of first ranked structures, based on samples from the ARGP model trained on 40%/10% of the data and (b) for different training fractions, together with the relative average computational costs.


and A. I. Cooper, *Angewandte Chemie International Edition*, 2011, **50**, 749–753.

9 E. O. Pyzer-Knapp, H. P. Thompson, F. Schiffmann, K. E. Jelfs, S. Y. Chong, M. A. Little, A. I. Cooper and G. M. Day, *Chemical Science*, 2014, **5**, 2235–2245.

10 J. T. A. Jones, T. Hasell, X. Wu, J. Bacsa, K. E. Jelfs, M. Schmidtmann, S. Y. Chong, D. J. Adams, A. Trewin, F. Schiffman, F. Cora, B. Slater, A. Steiner, G. M. Day and A. I. Cooper, *Nature*, 2011, **474**, 367–371.

11 S. R. Forrest, *Nature*, 2004, **428**, 911–918.

12 M. Muccini, *Nature materials*, 2006, **5**, 605–613.

13 S. T. Beckett, *The science of chocolate*, Royal Society of Chemistry, 2018.

14 I. B. Rietveld and R. Céolin, *Journal of pharmaceutical sciences*, 2015, **104**, 4117–4122.

15 H.-M. Wolff, L. Quere and J. Riedner, *Polymorphic form of rotigotine and process for production*, 2012, US Patent 8,232,414.

16 J. Bauer, S. Spanton, R. Henry, J. Quick, W. Dziki, W. Porter and J. Morris, *Pharmaceutical research*, 2001, **18**, 859–866.

17 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.

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

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

20 C. Y. Cheng, J. E. Campbell and G. M. Day, *Chem. Sci.*, 2020, **11**, 4922–4933.

21 G. M. Day, *Crystallography Reviews*, 2011, **17**, 3–52.

22 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 Cryst. B*, 2016, **72**, 439–459.

23 A. J. Cruz-Cabeza, S. M. Reutzel-Edens and J. Bernstein, *Chem. Soc. Rev.*, 2015, **44**, 8619–8635.

24 J. Nyman and G. M. Day, *CrystEngComm*, 2015, **17**, 5154–5165.

25 G. M. Day, J. Chisholm, N. Shan, W. S. Motherwell and W. Jones, *Crystal growth & design*, 2004, **4**, 1327–1340.

26 D. H. Case, J. E. Campbell, P. J. Bygrave and G. M. Day, *Journal of Chemical Theory and Computation*, 2016, **12**, 910–924.

27 M. Neumann, F. Leusen and J. Kendrick, *Angewandte Chemie International Edition*, 2008, **47**, 2427–2430.

28 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.

29 J. Hoja, H.-Y. Ko, M. A. Neumann, R. Car, R. A. DiStasio and A. Tkatchenko, *Science Advances*, 2019, **5**, year.

30 S. Manzhos and T. Carrington Jr, *The Journal of chemical physics*, 2006, **125**, 084109.

31 J. Behler and M. Parrinello, *Physical review letters*, 2007, **98**, 146401.

32 A. P. Bartók, M. C. Payne, R. Kondor and G. Csányi, *Physical review letters*, 2010, **104**, 136403.

33 J. Behler, *The Journal of chemical physics*, 2016, **145**, 170901.

34 P. Rowe, G. Csányi, D. Alfè and A. Michaelides, *Physical Review B*, 2018, **97**, 054303.

35 C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

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

37 D. McDonagh, C.-K. Skylaris and G. M. Day, *Journal of chemical theory and computation*, 2019, **15**, 2743–2758.

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

39 C. Adamo and V. Barone, *The Journal of chemical physics*, 1999, **110**, 6158–6170.

40 I. Nobeli and S. L. Price, *The Journal of Physical Chemistry A*, 1999, **103**, 6448–6457.

41 G. M. Day, W. D. Sam Motherwell and W. Jones, *Cryst. Growth Des.*, 2005, **5**, 1023–1033.

42 C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2016, **72**, 171–179.

43 V. R. Thalladi, M. Nüsse and R. Boese, *J. Am. Chem. Soc.*, 2000, **122**, 9227–9236.

44 S. Bhattacharya, V. G. Saraswatula and B. K. Saha, *Cryst. Growth Des.*, 2013, **13**, 3651–3656.

45 A. Katrusiak, *Acta Crystallogr., Sect. C: Cryst. Struct. Commun.*, 1993, **49**, 36–39.

46 P. D. Cradwick, *J. Chem. Soc., Perkin Trans. 2*, 1976, **0**, 1386–1389.

47 A. Katrusiak, *Acta Crystallogr., Sect. B: Struct. Sci*, 2001, **57**, 697–704.

48 F. Belaj, *Acta Crystallogr., Sect. C: Cryst. Struct. Commun.*, 1992, **48**, 1088–1090.

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

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

51 G. Kresse and J. Hafner, *Physical Review B*, 1993, **47**, 558.

52 G. Kresse and J. Hafner, *Physical Review B*, 1994, **49**, 14251.

53 G. Kresse and J. Furthmüller, *Physical review B*, 1996, **54**, 11169.

54 G. Kresse and J. Furthmüller, *Computational materials science*, 1996, **6**, 15–50.

55 R. Dovesi, A. Erba, R. Orlando, C. M. Zicovich-Wilson, B. Civalleri, L. Maschio, M. Rérat, S. Casassa, J. Baima, S. Salustro *et al.*, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2018, **8**, e1360.

56 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *The Journal of chemical physics*, 2010, **132**, 154104.

57 S. Grimme, S. Ehrlich and L. Goerigk, *Journal of computational chemistry*, 2011, **32**, 1456–1465.

58 S. Grimme, A. Hansen, J. G. Brandenburg and C. Bannwarth, *Chemical reviews*, 2016, **116**, 5105–5154.

59 F. B. Van Duijneveldt, J. G. van Duijneveldt-van de Rijdt and J. H. van Lenthe, *Chemical Reviews*, 1994, **94**, 1873–1885.

60 H. Kruse and S. Grimme, *The Journal of chemical Physics*, 2012, **136**, 04B613.

61 J. G. Brandenburg, M. Alessio, B. Civalleri, M. F. Peintinger, T. Bredow and S. Grimme, *The Journal of Physical Chemistry A*, 2013, **117**, 9282–9292.

62 J. Behler, *The Journal of chemical physics*, 2011, **134**, 074106.

63 J. S. Smith, O. Isayev and A. E. Roitberg, *Chemical science*, 2017, **8**, 3192–3203.

64 A. O'Hagan, *Journal of the Royal Statistical Society B*, 1978, **40**, 1–42.

65 R. B. Gramacy and H. K. H. Lee, *Statistics and Computing*, 2012, **22**, 713–722.

66 M. C. Kennedy and A. O'Hagan, *Biometrika*, 2000, **87**, 1–13.

67 A. I. Forrester, A. Sóbester and A. J. Keane, *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2007, **463**, 3251–3269.

68 M. Bayarri, J. Berger, R. Paulo, J. Sacks, J. Cafeo, J. Cavendish, C. Lin and J. Tu, *Technometrics*, 2007, **49**, 138–154.

69 P. Perdikaris, M. Raissi, A. Damianou, N. Lawrence and G. Karniadakis, *Technometrics*, 2017, **473**, 20160751.

70 L. Le Gratiet and J. Garnier, *International Journal for Uncertainty Quantification*, 2014, **4**, 365–386.

71 M. Johnson, L. Moore and D. Ylvisaker, *Journal of statistical planning and inference*, 1990, **26**, 131–148.

72 S. Tosoni, C. Tuma, J. Sauer, B. Civalleri and P. Ugliengo, *The Journal of chemical physics*, 2007, **127**, 154102.

73 F. Weigend and R. Ahlrichs, *Physical Chemistry Chemical Physics*, 2005, **7**, 3297–3305.

74 E. Caldeweyher and J. G. Brandenburg, *Journal of Physics: Condensed Matter*, 2018, **30**, 213001.

75 H. G. De Wit, J. A. Bouwstra, J. G. Blok and C. G. De Kruif, *J. Chem. Phys.*, 1983, **78**, 1470–1475.

76 J. Moellmann and S. Grimme, *The Journal of Physical Chemistry C*, 2014, **118**, 7615–7621.

77 T. Gneiting and A. E. Raftery, *Journal of the American Statistical Association*, 2007, **102**, 359–378.

78 A. Jordan, F. Krüger and S. Lerch, *R package version*, 2019, **1**, year.

79 A. D. Becke, *The Journal of chemical physics*, 2014, **140**, 18A301.

80 G. J. O. Beran and K. Nanda, *J. Phys. Chem. Lett.*, 2010, **1**, 3480–3487.

81 S. Wen, K. Nanda, Y. Huang and G. J. Beran, *Phys. Chem. Chem. Phys.*, 2012, **14**, 7578–7590.