npj Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Efficient Gaussian process regression for prediction of molecular crystals harmonic free energies

Marcin Krynski[1,2]✉ and Mariana Rossi[1,3]

We present a method to accurately predict the Helmholtz harmonic free energies of molecular crystals in high-throughput settings. This is achieved by devising a computationally efficient framework that employs a Gaussian Process Regression model based on local atomic environments. The cost to train the model with ab initio potentials is reduced by starting the optimization of the framework parameters, as well as the training and validation sets, with an empirical potential. This is then transferred to train the model based on density-functional theory potentials, including dispersion-corrections. We benchmarked our framework on a set of 444 hydrocarbon crystal structures, comprising 38 polymorphs and 406 crystal structures either measured in different conditions or derived from these polymorphs. Superior performance and high prediction accuracy, with mean absolute deviation below 0.04 kJ mol^{-1} per atom at 300 K is achieved by training on as little as 60 crystal structures. Furthermore, we demonstrate the predictive efficiency and accuracy of the developed framework by successfully calculating the thermal lattice expansion of aromatic hydrocarbon crystals within the quasi-harmonic approximation, and predict how lattice expansion affects the polymorph stability ranking.

npj Computational Materials (2021)7:169 ; https://doi.org/10.1038/s41524-021-00638-x

INTRODUCTION

Polymorphism and the prediction of the energetic stability of a crystal polymorph are a fundamental problem of condensed matter physics, especially for the research and applications of molecular crystals. Polymorphism is the capability of solid materials to form more than one distinct crystal structure[1,2]. It is particularly pronounced when multiple atomic or molecular packing arrangements are characterized by a similar free energy. The physicochemical properties of these systems, such as mechanical and optical characteristics, melting point, chemical reactivity, solubility, or stability are tied strongly to the crystal morphology, therefore increasing the relevance of a comprehensive structure screening and the prediction of the relative stability of polymorphs for a broad range of industries[3].

High-throughput computational screening of crystal structures based on free energies is rarely performed due to its high complexity as well as large computational effort, in particular if a first-principles potential energy surface is required[4]. It is more common to evaluate the relative stability of crystal polymorphs by calculating the lattice energy taking into account only potential energy contributions[5–9], effectively disregarding enthalpic and entropic contributions at finite temperature[2,10]. Finite pressure contributions when comparing different phases at different pressures is typically of a lower magnitude, reaching only about 1 kJ mol^{-1} per molecule for pressure difference of several gigapascals. It was shown[11] that even if the vibrational free energy difference between two given polymporphs lies typically around 2 kJ mol^{-1} per molecule, it is sufficient to cause a rearrangement of the polymporph relative stability ranking. Furthermore, even when the vibrational contribution to the relative stability is taken into account in a number of cases, the effect of the thermal expansion of the crystal unit-cell on the free energy is most frequently omitted. This is due to the typically low impact of the thermal expansion on the free energy (around 1–2 kJ mol^{-1} per molecule[12]), which is, nevertheless, also sufficient to affect the polymporph stability ranking.

The vibrational part of the free energy can be accessed by, among others, two straightforward types of calculation: within the harmonic approximation given by lattice dynamics calculations[13,14] and with statistical sampling methods that accounts for all anharmonic contributions, for example via thermodynamic integration (TI)[15–19]. Even though methods like TI are more accurate, they are also extremely computationally demanding, requiring a large amount of statistical sampling in order to achieve the necessary accuracy. This renders this technique often impossible to carry out within a high-throughput setting. Approximations to the contribution of anharmonic terms to the free energy can be accessed by a number of other methods that are less computationally demanding. However, such approximations have been shown not to present a significant improvement over the much less computationally demanding harmonic approximation for the investigation of polymporph relative stability[20]. Still, harmonic lattice dynamics are not a viable solution for high-throughput screening if force evaluations are a bottleneck, since the calculation typically involves hundreds of force evaluations for a single structure (or costly perturbation theory techniques), considering the full unit cell.

Within the last decades, the rapid increase of computer power, allied to the rise of machine learning (ML) and big-data algorithms in the realm of material science, allowed for large-scale screening of materials properties, including those related to polymorphism[10,21–28]. There are only a handful of examples where vibrational free energies[29], or other quantities related to the vibrational density of states[30–33], were successfully predicted with the assistance of ML methods. Those methods, however, do not focus on high transferability, or, if they do, rarely achieve the necessary accuracy to differentiate between polymorphs. Clearly, if one could train a very accurate ML interatomic potential for a large

[1]Fritz Haber Institute of the Max Planck Society, Faradayweg 4-6, 14195 Berlin, Germany. [2]Present address: Faculty of Physics, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland. [3]Present address: MPI for the Structure and Dynamics of Matter, Luruper Chaussee 149, 22765 Hamburg, Germany. ✉email: marcin.krynski@pw.edu.pl

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

class of systems, it would represent the best solution for the evaluation of lattice energies and free energies at the same time. However, despite the exceptional performance of many such potentials, typical root-mean-square errors on the forces lie around 20 meV $Å^{-1}$ per atom[34–39]. With such errors, the expected prediction accuracy of phonon modes is ±0.15 THz for the best performing potentials[34]. If the resulting phonon accuracy, as in ref. [34], is assumed to be constant along the entire frequency range, the harmonic free energy calculation error amounts to 0.38 kJ mol^{-1} per atom.

In this study, we target high accuracy and low computational cost for harmonic free energy predictions. We build a model for the prediction of Helmholtz harmonic free energies of molecular crystals based on Gaussian process regression (GPR) and smooth overlap of atomic positions (SOAP)[40] descriptors for representing the local atomic environments. We optimize the training and validation set selection with a computationally cheap empirical potential, confirm its transferability to a first-principles potential, and proceed to achieve a model with first-principles accuracy with a very low cost of training. For a set of hydrocarbon crystals, we are able to achieve a mean absolute error on the free energies of 0.04 kJ mol^{-1} per atom. We analyzed the stability ranking for a few families of hydrocarbon crystal polymorphs up to 300 K, highlighting the power and accuracy of the model. Furthermore, this method can predict the anisotropic lattice expansion of these crystals, allowing a cheap evaluation of volume expansion and free energies in the quasi-harmonic approximation.

RESULTS AND DISCUSSION

Because it was shown[20] that the harmonic approximation to the free energy can be a suitable estimate for the computation of the relative stability between different structures of molecular crystals, this project focuses on predicting the harmonic Helmholtz free energies F. Contributions from pressure that would be described instead by the Gibbs free energy are not considered, because the structures regarded in this study are typically observed much below 1 GPa of pressure, making this contribution to the free energy negligible. Throughout this paper, for the sake of simplicity, F is evaluated at the Γ point of the Brillouin zone of a given unit cell. We consider unit cells larger than the primitive cell where needed (see "Methods" section). The harmonic free energies are thus calculated as

$$F(V,T) = \sum_{i=1}^{3N-3}\left(\frac{\hbar\omega_i}{2} + k_B T \ln\left(1 - e^{-\frac{\hbar\omega_i}{k_B T}}\right)\right), \tag{1}$$

where ω_i is the frequency of a given phonon mode at the Γ point. When taking lattice expansion into account, the vibrational frequencies depend indirectly on the temperature such that $\omega_i = \omega_i(V(T))$.

Definition of the GPR model

The key assumption of the free energy prediction approach explored in this project is that even if free energies are defined only for the entire collection of atoms of the crystal structure, they can be decomposed into local contributions of atomic environments. The approach of casting a global property on local environments was explored previously[41,42] for the generation of an interatomic potential from quantum mechanical data. The problem of the harmonic Helmholtz free energy prediction is approached by connecting the atomic-wise free energy to the full free energy by

$$\mathbf{F} = \mathbf{M}^T\mathbf{f}, \tag{2}$$

where $\mathbf{F}$ is the vector with all measured free energies for a given crystal set of dimension N_s (number of crystal structures in the training set), $\mathbf{M}$ is an incidence matrix of dimension $N_s \times N_{ae}$ (number of atom environments in the given set) and $\mathbf{f}$ is the vector of all, unobserved, atom-wise free energies in the chosen ensemble. Then, the prediction of $\mathbf{f}$ in the training set is modeled as

$$\mathbf{f}' = \mathbf{C}\boldsymbol{a}, \tag{3}$$

where $\mathbf{C}$ is the matrix containing the similarities between pairs of atomic environments (dimension $N_{ae} \times N_{ae}$), defined as

$$C_{ij} = \sigma e^{-\frac{\sum_{d=1}^{D}\left(q_{d,i} - q_{d,j}\right)^2}{2l^2}}, \tag{4}$$

where σ is a scaling prefactor, and $\mathbf{q}_i$ is a vector of length D describing local atomic environments. C_{ij} corresponds to the Gaussian kernel. In Eq. (3), $\boldsymbol{a}$ is a vector of N_{ae} weights for each atomic environment, such that

$$\mathbf{F}' = \mathbf{M}^T\mathbf{C}\boldsymbol{a}. \tag{5}$$

Opening up this equation element-wise, the full free energy of one sample i in the training set is given by

$$F_i' = \sum_{j=1}^{N_{ae}}\sum_{k=1}^{N_{ae}}\left(\mathbf{M}^T\right)_{ij} C_{jk} a_k. \tag{6}$$

Optimizing the weights a_k is equivalent to minimizing the loss function

$$L = \sum_{i}^{N_s}\left[F_i' - F_i\right]^2 + \sigma_\epsilon^2 \boldsymbol{a}^T\mathbf{C}\boldsymbol{a}, \tag{7}$$

where σ_ϵ^2 is a regularization parameter related to the variance of the noise of the data.

Finally, substituting Eqs. (6) into (7), the minimization is straightforward and leads to

$$\boldsymbol{a} = \mathbf{M}\left(\mathbf{M}^T\mathbf{C}\mathbf{M} + \sigma_\epsilon^2\mathbf{I}\right)^{-1}\mathbf{F}, \tag{8}$$

where $\mathbf{I}$ is the identity matrix of dimensions $N_s \times N_s$. In this way, one can obtain the optimized weights with no need to define or observe atom-wise free energies.

Finally the prediction of the free energy of a new structure that is not contained in the training set is achieved by calculating

$$F(\mathbf{q}^*) = \sum_{i=1}^{N}\sum_{j=1}^{N_{ae}}\left(\mathbf{C}^{*T}\right)_{ij} a_j \tag{9}$$

where $\mathbf{C}^*$ is the similarity matrix between the atomic environments $\mathbf{q}^*$ of the new structure to the ones in the training set, with elements

$$C_{ij}^* = \sigma e^{-\frac{\sum_{d=1}^{D}\left(q_{d,i}^* - q_{d,j}\right)^2}{2l^2}}. \tag{10}$$

All hyper-parameters for the GPR model and the representations were selected by minimizing, using the steepest descent method, the negative log marginal likelihood function[43]

$$\begin{aligned} -\ln P(\mathbf{F}|(l,\sigma_\epsilon,\boldsymbol{\theta})) &= \tfrac{1}{2}\ln\left|\mathbf{M}^T\mathbf{C}(l,\boldsymbol{\theta})\mathbf{M} + \sigma_\epsilon^2\mathbf{I}\right| \\ &\quad + \tfrac{1}{2}\left(\mathbf{M}^T\mathbf{C}(l,\boldsymbol{\theta})\mathbf{M} + \sigma_\epsilon^2\mathbf{I}\right)^{-1} + \tfrac{1}{2}\ln 2\pi, \end{aligned} \tag{11}$$

where $\boldsymbol{\theta}$ is a vector containing the hyperparameters of the representations entering $\mathbf{q}$. The application of the steepest descent method is only guaranteed to find a local minimum. A wide space of hyper-parameters was considered in order to increase the probability of finding a global minimum.

In all supervised machine learning based models, the quality of the model strongly depends on the quality of the training set. Typically, selecting the training set can be done by either a random selection of samples, given that the considered ensemble is fairly homogeneous, or by implementing methods that aim at covering the sampled domain by maximizing the resulting

prediction accuracy, such as the "correlation" clustering method[44], genetic optimization[45], or *k*-fold cross-validation[46]. Unfortunately, most of the methods from the latter group require a large pool of data for which the target property, like free energy in this case, is available. In this study, because one of the objectives is to minimize the computational cost of obtaining a good training set, the applied procedure focuses on selecting an optimal training subset based exclusively on the geometrical parameters of the crystal structure.

For this purpose, the farthest point sampling (FPS)[47] method is applied, that searches for a subset of the entire investigated crystal structure ensemble that covers evenly all structural motifs of the sampled domain with minimal information overlap. First, a similarity measures between molecular crystal structure $R_{a\to b}$ is defined according to the best-match structural kernel[48] method, as it is needed for the application of the FPS

$$R_{a\to b} = \frac{1}{n_b}\sum_{i=1}^{n_b} \max_{j\in(1,n_a)} C(\mathbf{q}_j^a, \mathbf{q}_i^b), \tag{12}$$

where $C(\mathbf{q}_j^a, \mathbf{q}_i^b)$ is the kernel matrix element defined in Eq. (4), $\mathbf{q}_j^a$ and $\mathbf{q}_i^b$ are *j*th and *i*th atomic environment representations of structure *a* and *b* respectively, and similarly n_a and n_b are the number of atoms in those structures. $R_{a\to b}$ defines how well atoms of structure *a* can represent geometrical motifs of structure *b* and $R_{a\to b} \neq R_{b\to a}$. In other words, it is possible that atomic environments of structure *a* represent well those of structure *b*, while structure *b* contains geometric features not present in *a*. This method of defining the relationship between crystal structures is very similar to others typically chosen for such tasks[40,48–51], with the difference that $R_{a\to b}$ is not invariant with respect to the crystal structure index ($R_{a\to b} \neq R_{b\to a}$) so it is not a similarity metric in a strict sense. Next, according to the FPS algorithm, the training set is created by iteratively picking structures that are least represented by those already present in the training set. Since any crystal structure can be used as the starting point for the FPS algorithm, the applied method selects N_Ω potential training sets, where N_Ω is the number of crystal structures in the considered ensemble. In order to choose out of N_Ω potential training sets, we have investigated the scaled cumulative sum $I(N_m)$ of the $R_{a\to b}$,

$$I(N_m) = \frac{\sum_a^{N_m}\sum_b^{N_\Omega} R_{a\to b}}{\sum_a^{N_\Omega}\sum_b^{N_\Omega} R_{a\to b}}, \tag{13}$$

where N_m is the total number of the molecular crystal configurations in the training set. This quantity reveals how fast a given training set candidate converges to unity, which we consider to represent a full coverage of the sampled feature space. In another sense, the $I(N_m)$ quantity can be seen as the description of the information acquisition during consecutive steps of the FPS algorithm. Finally, training set with the highest recorded value of $I(N_m)$ after all $N_m = 60$ steps of the FPS algorithm is chosen. The training set size of $N_m = 60$ was chosen because above this number, the improvement of the prediction accuracy was too small to justify a larger training set and the associated increase in computational effort.

In the same spirit of maximizing accuracy and minimizing cost, with the objective of performing free energy predictions with ab initio data, the aim was to select an efficient and reliable validation set, without using the entire ensemble. Here the goal is to create such a subset that would represent well the entire set, so as to include, for example, a proportional number of outlier structures as found in the entire set. A random selection of validation set would not fulfill this criterion due to the limited size of validation set used in this project. Additionally, this task largely differs from selecting the training set, because it typically contains a greater relative number of outliers compared to the entire set. In order to optimally select the validation set, while preserving the density of outliers, a stratified approach was used. Here each crystal structure *a* is assigned a similarity index S_a, that compares a given crystal to entire set

$$S_a = \sum_b^{N_\Omega} R_{a\to b}. \tag{14}$$

The relatively high values of S_a indicate a "typical" crystal and low values indicate "outliers". Next, the entire set is sorted with respect to S_a and the validation set is chosen by selecting every *n*th element of the sorted set, with $n = \mathrm{round}(N_\Gamma/N_\Theta)$ where N_Γ and N_Θ are the target numbers of structures in the validation and training sets. All sets sorted with respect to S_a are presented in Supplementary Fig. 2.

Within the discussed framework, and common to many ML models, the choice of method encoding the atomic environments to numerical representations has an impact on the resulting performance of the model. In this project, three well-established general-use atomic environment representations[52] were selected and tested, namely: SOAP[40] that uses spherical harmonics to locally expand atomic densities, many body tensor representation (MBTR)[53] that uses distributions of different structural motifs (like radial or angular distribution functions) and atom-centered symmetry functions (ACSFs)[54] that use two-body and three-body functions detecting specific features. The Python implementations of the mentioned representations found in the DScribe package[55] were used.

Model implementation and validation

For the purpose of this work we have chosen crystals composed of seventeen different hydrocarbons: pyrene, methylcyclopentane, styrene, naphthalene, benzene, tetracene, mesitylene, pentane, pentacene, hexane, ethylbenzene, propane, heptane, phenanthrene, butane, hexacene, and anthracene. We have included most available polymporphs that could be obtained from the Cambridge Crystallographic Data Center[56] (CCDC), leading to an ensemble of 74 structures. We noted that polymporphs of very similar lattice constant in CCDC tend to be almost identical, with close to negligible differences in atomic positions, for example, the case of ANTCEN20 and ANTCEN22. Finally, the sample domain was further expanded by introducing structures with perturbations of roughly 5% in the lattice parameters, as this can lead to up to 16% increase in unit cell volume—a typical volume expansion percentage for molecular crystals[57]. The addition of crystal structures with strongly perturbed lattice parameters was found to be crucial for the later prediction of lattice expansion coefficients. Finally, $N_\Omega = 444$ crystal structures were considered in this project.

The building and testing of the framework was initially performed using a classical force-field potential (AIREBO, as detailed in "Methods" section). In the first steps of the model verification, the training and validation set selection criterion, based on the FPS method and maximization of $I(N_m)$, was evaluated. For this purpose, based on the classical force field data with prediction performed at 300 K and with SOAP atomic-environment representations, free energy mean absolute error F_{MAE} was calculated

$$F_{MAE} = \frac{1}{N}\sum_i^N |F_i - F_i'|, \tag{15}$$

where *N* is the number of structures for which the prediction is performed. The results are presented in Fig. 1a in the form of learning curves, with increasing size of the training set *N* and with the validation set. It is visible that the learning curve obtained for the chosen training set, so with the highest $I(N_m)$, shows one of the lowest F_{MAE} at the target training set size among all potential sets obtained using FPS method.

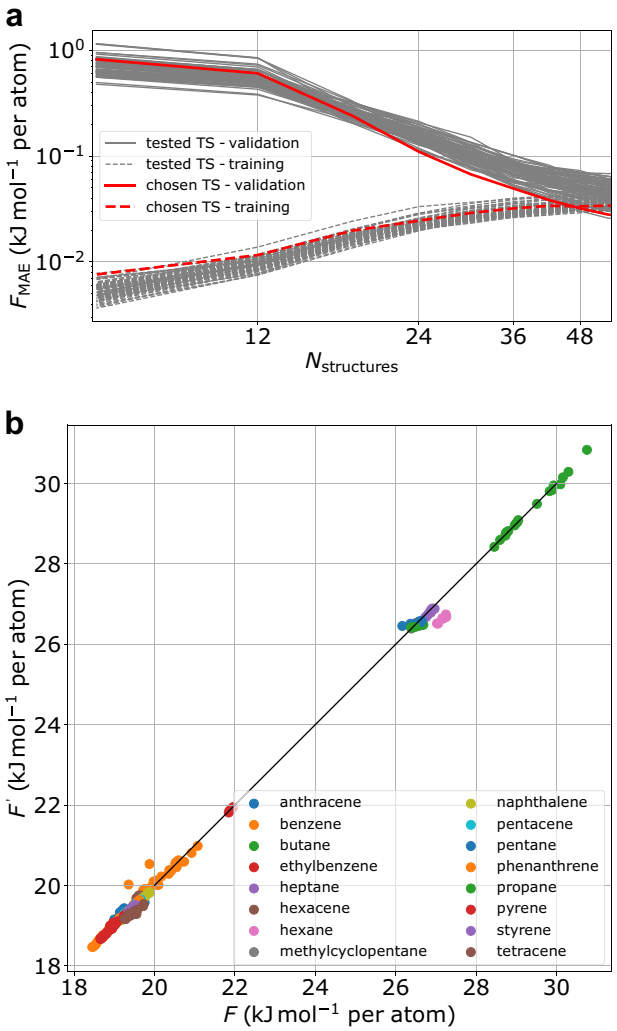

Fig. 1 Performance of the developed prediction model. a Learning curves obtained for 100 randomly chosen learning sets out of total 444 in gray, and the chosen learning set with the highest recorded value of $I(N_m)$ in red. Data was obtained using SOAP representation and the classical force field model with the free energy calculated at 300K. **b** Correlation between predicted and calculated free energies at 300 K (classical force field, SOAP representations). Different crystal families are represented by different colors.

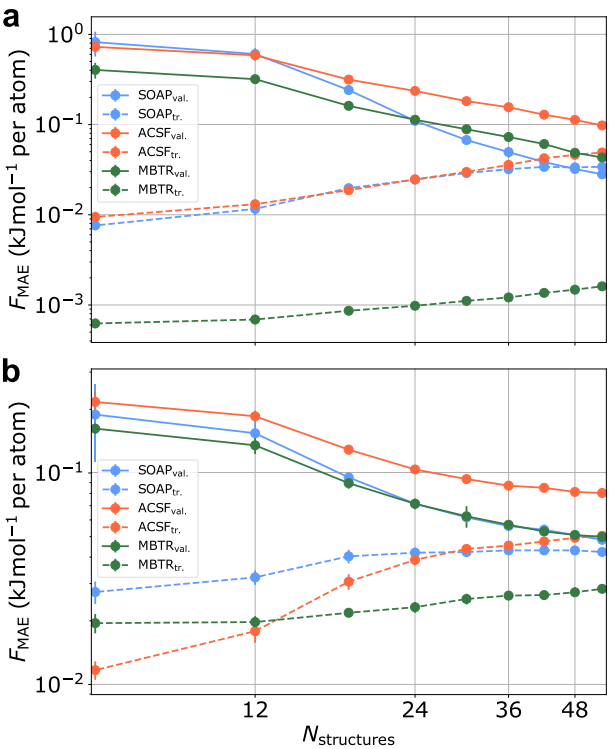

Fig. 2 Learning curves of F_{MAE} (300 K) calculated for the training (dashed line) and the validation set (solid line) obtained with SOAP, MBTR and ACSF representations. Results are presented for GPR models obtained based on: **a** the empirical AIREBO force field and **b** density-functional theory (PBE functional with pairwise van der Waals corrections) data, and are presented as a function of the number of crystal structures in the training set. Error bars are equal to the standard deviation of F_{MAE} of training sets with different structures.

Next, the linear and monotonic correlations between benchmark F and predicted F' values was assessed by calculating the Pearson and Spearman correlation coefficients. For predictions performed at 300 K with the SOAP representation they were found to be 0.9996 and 0.9894, respectively. A value so close to 1 for these coefficients indicate a good performance of the developed framework. Furthermore, due to the low cost of the lattice dynamics calculations performed using classical force field, the F_{MAE} was inspected for the entire set (300 K with the SOAP representation) and it was found to be 0.042 kJ mol^{-1} per atom. Additionally, the $F_{MAE} = 0.218$ kJ mol^{-1} per atom was obtained for 10% of the crystals with the poorest prediction and 0.023 kJ mol^{-1} per atom for the remaining 90% of samples.

Figure 1b shows the predicted free energy values F' compared with the benchmark data F for the different crystal families. The analysis gives an indication of the system-sensitive performance of the framework, revealing that crystals of pentane, pentacene, tetracene, and hexane are characterized by the poorest averaged prediction accuracy, with the F_{MAE} around 2 kJ mol^{-1} per molecule, reaching a possible free energy difference between different polymporphs[11,12]. Additionally, the predictions performed for crystal structures with strongly perturbed lattice parameters were noticeably poorer, even if the training set contained parental crystal structures. Nevertheless, the prediction accuracy overall is very high, especially considering the diversity of hydrocarbons represented.

Figure 2a shows the learning curves by monitoring F_{MAE} at 300 K with increasing training set size and a constant validation set. Additionally, the impact of the atomic environment representation on the efficiency of the method was investigated. The learning is well-behaved for all representations, as expected for properly parameterized machine learning models. The results obtained with the SOAP representation, with a 6 Å cutoff and 1 Å for the standard deviation of the Gaussian functions used to expand the atomic density, are characterized by the lowest F_{MAE}, showing that it is the best representation within the investigated set. Finally, the accuracy of the predictions are noticeably affected by the temperature at which the free energies are required, going from 0.019 kJ mol^{-1} per atom at 300 K and 0.015 kJ mol^{-1} per atom at 200 K to as low as 0.002 kJ/mol at 0 K.

Finally, it is worth noting that the atom-wise contributions to **F** are not guaranteed to have any physical meaning, but sometimes their analysis can give interesting insights. In this case, as further discussed in Supplementary Note 1, this analysis shows that there are no trends in the contribution to the free energy from different atomic species or classes of bonded atoms, which could be related to the strongly non-local character of the free energy.

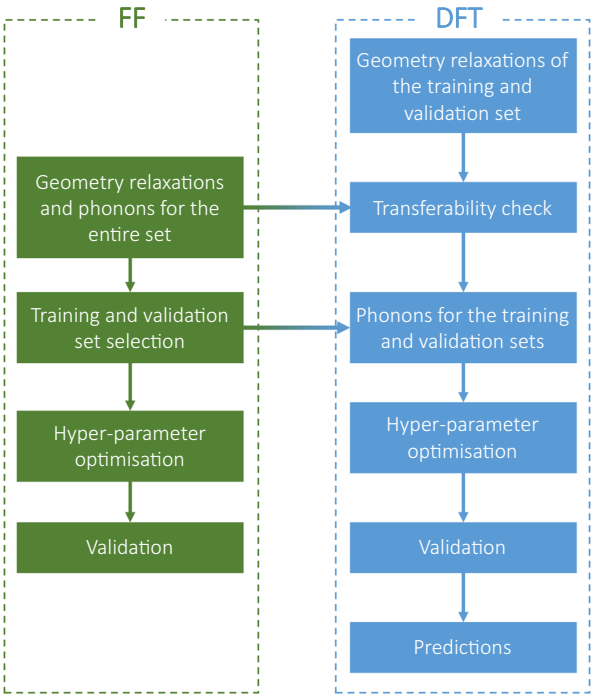

Fig. 3 Flowchart of the developed framework. An optimal and small training set is obtained by training the model on force field data (green boxes) and then used to train the model on DFT data (blue boxes).

Transferability of the prediction model

Once the framework was built and proven to deliver a satisfactory prediction of harmonic free energies based on data coming from an empirical potential, the transferability of the model when using DFT data was investigated. For that, the PBE exchange correlation functional with pairwise van der Waals interactions was employed, as detailed in "Methods" section. As a test, the similarity between relaxed structures obtained with the empirical potential and DFT was assessed by analysing the root mean square deviation (RMSD) of the atomic positions averaged over entire set. RMSD for carbon and hydrogen were 0.16 Å and 0.20 Å, respectively. Importantly, differences in the SOAP representation were also investigated by calculating the root mean square error normalized by the standard deviation ϵ_X, defined as

$$\epsilon_X = 100 \times \sqrt{\frac{\sum_d^D \frac{1}{N} \sum_i^N \left(q_{d,i}^{FF} - q_{d,i}^{DFT}\right)^2}{\sum_d^D \frac{1}{N-1} \sum_j^N \left(\overline{q}_d^{DFT} - q_{d,i}^{DFT}\right)^2}}, \quad (16)$$

where D is the number of features of the representation and N is the number of atoms for which the ϵ_X was calculated. Obtained values for both carbon and hydrogen are $\epsilon_C = 1.09$ and $\epsilon_H = 1.15$, respectively. Those results show that the overall structural features are in good agreement in these two potential energy surfaces. As a consequence of this structural similarity between two data sets, the training and validation sets obtained with the empirical potential, as explained in the previous section, can be automatically used in DFT. As a cross-check, the same training and validation set optimization procedure were independently applied on the optimized DFT structures, indeed obtaining the same results. This proved that the experience gathered from the first phase of the project, where only classical data was used, is fully transferable to the current stage, where we employ more accurate ab initio data. As a result, the more expensive ab initio lattice dynamics calculations were only performed for crystal structures included in the training and validation sets, greatly reducing the computational cost of the model generation.

Finally, the hyperparameters of the GPR model were re-optimized and were used to calculate learning curves for training and validation sets shown in Fig. 2b, with the SOAP, MBTR, and ACSF representations. All representations presented a good performance, with MBTR and SOAP yielding very similar learning curves. The obtained F_{MAE} for the SOAP representation at full training set was found to be 0.038 kJ mol^{-1} per atom. Interestingly, a fairly good prediction performance can be obtained with as little as 20 crystal structures, resulting in $F_{MAE} = 0.07$ kJ mol^{-1} per atom. Such small training sets typically do not contain all different molecular components of the crystals that are present in the entire set, but can still describe it well. The remainder of this manuscript will focus on results obtained based on the DFT data with the SOAP representation, exclusively.

The proposed framework is summarized in the flowchart in Fig. 3. In addition, as it is shown in the SI, the possibility of this model trained only on hydrocarbons to extrapolate to systems containing carbon, hydrogen, and nitrogen atoms was investigated. Although the prediction accuracy decreases as the concentration of nitrogen atoms in the samples increases, the model is not completely invalid. It shows that with a small addition of structures to the training set or building representations for new atoms that combine characteristics of the atoms that were previously trained[58,59], this framework could be easily extended to other systems.

Relative free energies of molecular crystals: stability ranking

The GPR model was employed to create a stability ranking of several families of hydrocarbon molecular crystals. Sixteen crystal families were considered, encompassing 38 polymorphs and 36 variants corresponding to different thermodynamical conditions with lattice parameters as they are given by the CCDC[56]. Additionally, 370 crystal structures with randomly distorted lattice parameters derived from the initial 74 were included. Figure 4 shows the lattice energy and the free energy obtained at various temperatures, presented as relative values to the crystal structure characterized by the lowest free energy at 300 K (full data is found in Table S2, in the SI). The identifiers of all crystal structures follow the convention used in CCDC[56]. For many crystal families, the structure with the lowest lattice energy is not the same as the one with the lowest free energy especially at the room temperature. A clear example is the pyrene crystal and its three polymorphs: Form I is represented by PYRENE02 and PYRENE03 (structures measured at 423 K and 113 K, respectively, and ambient pressure); Form II is represented by PYRENE07 and PYRENE10 (at 93 and 90 K, ambient pressure); and Form III is represented by PYRENE08 and PYRENE09 (measured at at 0.3 GPa and 298 K, and at 0.5 Gpa and 298 K, respectively). Form I is measured to be more stable than form III at all temperatures up to and beyond 430 K, at ambient pressure. Here, it is shown that the energy ranking formed based on lattice energy exclusively would place the high-pressure form III PYRENE09 (form III) structure very close to PYRENE02 (form I). An inclusion of zero-point-energy and vibrational contributions already at low temperatures irrevocably destabilizes form III.

A similar example is the benzene crystal. Here, structures of the ambient-pressure form I, represented by, for example, BENZEN15, BENZEN19, or BENZEN26, are characterized by overall lower free energy comparing to the high-pressure form II structures, like BENZEN16 and BENZEN17. Interestingly, for this crystal family, the lattice energy can provides a satisfactory relative stability ranking. However, the need for including the vibrational contributions becomes visible once a high and ambient pressure variants of one polymorph are compared, e.g., BENZEN13 and BENZEN26. It is visible in Fig. 4 that if considering only lattice energies, BENZEN13 shows the lowest energy compared to other crystal

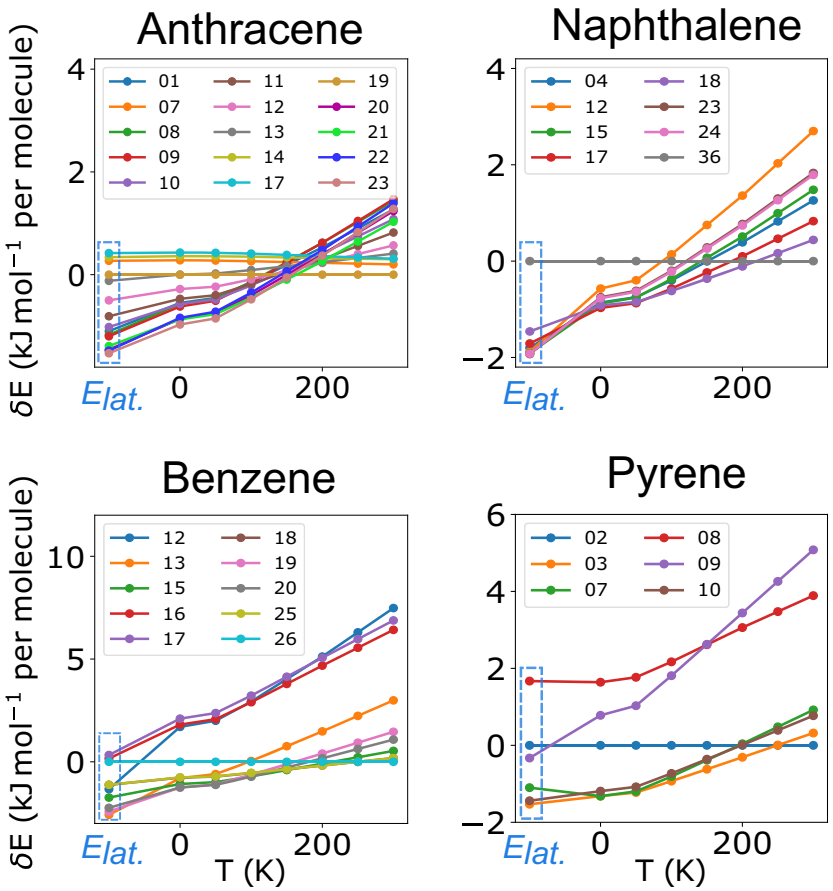

Fig. 4 Lattice energy $E_{lat.}$ and predicted free energy differences between variants of four molecular crystal families: anthracene, naphthalene, benzene, and pyrene. Full data, also for a larger number of crystals and polymporphs, is available in the Table S2 in the SI. Relative energies are calculated separately at each temperature, always with respect to the lowest free energy structure at 300 K. The numbers on the labels represent the identifiers of the crystals following the convention used in CCDC[56].

variants, with lattice energy lower than that of BENZEN26 by 2.58 kJ mol^{-1} per molecule. However, the free energy prediction shows that at 300 K, the BENZEN26 structure becomes the most stable out of all those investigated, and its relative free energy with respect to BENZEN13 is now lower by 2.92 kJ mol^{-1} per molecule, effectively swapping places in the relative ranking stability with BENZEN26. For this case, and to test the predictions of the model in practice, the free energies for both BENZEN13 and BENZEN26 structures were additionally calculated with DFT. These calculations showed that BENZEN13 is characterized by a free energy that is 3.43 (kJ mol^{-1} per molecule) higher than that of BENZEN26 at 300 K, confirming the results obtained with the GPR model.

The rearrangement of the relative stability ranking when room temperature free energy is taken into account is a very common trend among the investigated samples, and there are a number of cases, where even at 0 K the zero point energy contribution is high enough to affect the relative stability ranking. These observations are in good agreement with previous studies, where more direct methods were used[11]. In some cases, the prediction accuracy of this model is not sufficient to determine the relative stability of some structures. Nevertheless, the model is accurate enough to point towards those few that are characterized by the lowest free energies. Here, even only narrowing the pool of considered strucrures can effectively decrease the computational effort of phonon calculations required, if more accuracy is needed.

Predicting lattice expansion

Because one of the challenges in high throughput computational screening of crystal structures is accounting for thermal lattice expansion, the application of the trained free energy model was explored in this context. To illustrate the procedure, a simple case where only one of the lattice parameters is being perturbed was considered. For this purpose the BENZEN11 crystal was chosen, with the lattice parameter a being sampled within 6.52 and 7.32 Å. Next, within the quasi-harmonic approximation, the free energy was calculated and predicted as a function of a. Figure 5a shows the comparison between the GPR model and DFT calculations for the free energy at 200 K. The optimal lattice parameter a is determined by a Birch-Murnaghan[60] fit. While there are small differences between the DFT and the GPR curves, mostly consisting of a shift in energy, the resulting optimal lattice parameter a is very similar in both cases, and equal to 6.92 Å and 6.95 Å, respectively. This simple and fairly artificial example illustrates that the prediction accuracy of this framework is sufficient to be employed in the context of the lattice expansion/contraction prediction.

A more challenging task is the prediction of anisotropic lattice changes. Direct calculations of anisotropic lattice expansion requires lattice dynamic evaluations for, typically, hundreds of structures of the same crystal polymorph, making it a very costly calculation for a high-throughput setting. Although harmonic and quasi-harmonic models[61] as well as an approach based on the assumption of the linear relation between free energy and volume have been proposed to overcome this cost[62], with this framework

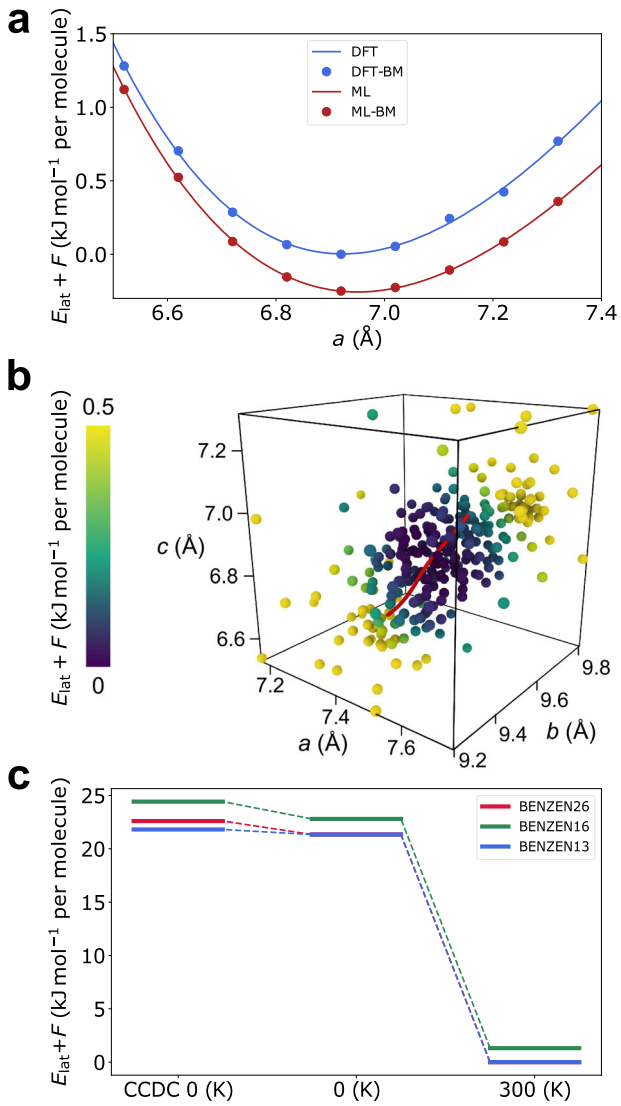

Fig. 5 Application of the developed framework to predict the thermal lattice expansion. a Predicted (red) and calculated (blue, DFT) free energies as a function of one lattice parameter of the BENZEN11 crystal at 200 K. Solid lines correspond to a Birch-Murnaghan fit. **b** 3D visualization, in lattice parameter space (*a*, *b*, *c*), of the free energy prediction including lattice expansion of the anthracene crystal at 200 K, with 300 different combinations of perturbed lattice parameters. The red line indicates the observed change in the lattice parameters when increasing the temperature from 0 to 300 K. **c** Relative predicted free energies of benzene crystal structures when considering fixed lattice constants (taken from CCDC), and when considering lattice expansion at 0 and at 300 K.

these lattice changes can be estimated without relying on any *ansatz* for the dependence of the free energy on the lattice parameters. It is worth noticing that even if the free energy predictions at various temperatures requires training the ML model multiple times, it happens with minimal computational overhead once appropriate lattice vibrations have been computed. Four molecular crystals were picked, namely $P2_1/a$ anthracene (ANTCEN), *Pbca* benzene (BENZEN), *P*1 pentacene (PENCEN01), and *Pbcn* styrene (ZZZTKA01) and hundreds of ionic relaxations with the *a*, *b*, and *c* lattice parameters perturbed by around 5% were performed. Next, for each of those perturbed structures free energy prediction at a number of temperatures from 0 to 300 K range was performed.

Figure 5b shows a 3D visualization of free energy predictions for over 300 different combinations of lattice parameters *a*, *b*, and *c* of the ANTCEN crystal. Even with such a high number of sampled lattice parameter combinations, the position of the free energy local minima might not overlap with the gathered data. In this case, in order to find the minimum in this high-dimensional space, an active learning based on the GPR algorithm is employed. Here, the GPR is used as a multi-dimensional, non-linear regressor, as implemented in the scikit-learn[63] package. In detail, the following bootstrap procedure is used:

- Identifying the position of the data point with the lowest free energy value according to the GPR 3D interpolation.
- For the chosen set of (*a*, *b*, *c*) lattice parameters perform an ionic relaxation and predict the free energy with the trained model.
- If the predicted free energy of the (*a*, *b*, *c*) sample varies from the free energy obtained by the 3D GPR regression, a new 3D GPR regression is performed, now explicitly including sample (*a*, *b*, *c*), then go back to step 1.
- If the predicted free energy of the (*a*, *b*, *c*) sample is sufficiently close to the one of the 3D GPR regression (within ±0.1%), then the scheme is stopped and the optimal lattice parameters are considered to be found.

We found that typically only around three additional relaxations and free energy predictions (per temperature) are necessary to achieve sufficient convergence of the lattice parameters. By employing this procedure to predict the anisotropic lattice changes the lattice-parameter change is calculated, as well as the full volume change of the selected crystals, as shown in Fig. S5.

The results obtained can be compared to experimental values where data is available. For anthracene the experimentally measured volume change is $V^{exp.}_{290K}/V^{exp.}_{90K} = 1.024$[64] and we obtained $V^{ML}_{290K}/V^{ML}_{90K} = 1.034$. For pentacene, the comparison is $V^{exp.}_{295K}/V^{exp.}_{90K} = 1.037$[65] and $V^{ML}_{295K}/V^{ML}_{90K} = 1.031$; for benzene $V^{exp.}_{270K}/V^{exp.}_{78K} = 1.089$[66] and $V^{ML}_{270K}/V^{ML}_{78K} = 1.068$; for styrene $V^{exp.}_{120K}/V^{exp.}_{83K} = 1.017$[67,68] and $V^{ML}_{120K}/V^{ML}_{83K} = 1.009$. The predictions are quite close to experimental data and overall a high degree of anisotropy is observed. Moreover, a deviation from a linear behavior of the free energy change with respect to volume is observed, as shown in Fig. S6.

This framework can thus be used to create the relative stability ranking including the thermal expansion effect on the free energy. Here, one example of how this can impact the relative stability and crystal form of these systems is presented. For this purpose, BENZEN13 and BENZENB26 (high and low pressure variants of the $P2_1/b2_1/c2_1/a$ benzene I polymorph[69]) are selected, as well as BENZEN16 (a high pressure $P2_1/c$ benzene II polymorph[70]). The initial lattice constants were taken from the CCDC. As shown in Fig. 5c, by simply searching for the free energy minimum at 0 K using the procedure described above, BENZEN13 and BENZEN26 were found to end up being characterized by almost identical (predicted) free energies and lattice constants. Further inspection indicated that indeed the BENZEN13 and BENZEN26 structures converged to the same structure, and the same behavior was found at all investigated temperatures. Even if somewhat expected, given that they are high and low pressure phases within the same crystal group and in the absence of any applied pressure it is natural that they both adopt the low-pressure structure, the fact that this result came from the model alone, and that the free energy predictions were able to capture this transition, shows that the method is robust. The BENZEN16 structure is stabilized by 1 kJ mol^{-1} per molecule upon increasing the temperature from 0 to 300 K, as shown in Fig. 5c. This stabilization is accompanied by an appreciable lattice expansion with a volume increase of around 6% from 0 to 300 K.

In summary, the framework proposed here provides a machine learning model with first principles accuracy for the harmonic Helmholtz free energies of molecular crystals, that is suitable for high-throughput studies. In addition, it was shown that the training and validation set of the model can be optimized using a cheaper empirical potential, and then transferred to first-principles calculations, thus substantially decreasing the cost of training, without sacrificing accuracy.

The model was tested to predict the relative energetic stability ranking of several diverse hydrocarbon polymorphs and distorted crystal structures derived from them, and the changes on this ranking with increasing temperature was studied. We observed that, in most cases, omitting thermal effects and instead using only the lattice energy produces misleading results. Furthermore, it was shown that the model can be efficiently employed to calculate the anisotropic lattice expansion—a task rarely approached due to its complexity and high computational demand when performed at the ab initio level. Unsurprisingly, taking the anisotropic lattice expansion into account leads to further changes in the stability ranking. Naturally, the same framework could be used to predict other quantities derived from vibrational properties, like the vibrational heat capacity.

The strengths of this framework lie in its low computational cost, reliability and accuracy. However, because the model is trained to directly predict free energies, one still has to deal with the computational cost of obtaining optimized structures, which we here obtained from first-principles geometry optimizations. Fitting a machine-learned interatomic potential is becoming more streamlined[71], even though these potentials rarely target the accurate description of vibrational properties due to the added complexity of including them in the learning procedure. The presented framework, on the other hand, can be easily combined with any potential that can predict structures in a reasonable manner and has the potential to be more accurate.

Extending this framework beyond hydrocarbon-based crystals could be straightforward, albeit perhaps requiring different training data. We have already observed that the framework is capable of predicting DFT free energies from FF-relaxed structures with promising accuracy (see Supplementary Note 2). Finally, targeting fully anharmonic free energies with ab initio accuracy is still a daunting task that can, nevertheless, profit from the knowledge gained in this study.

METHODS

Force field model

Geometry optimization calculations with empirical potentials were performed using LAMMPS[72] together with AIREBO[73] interatomic potentials. The conjugate gradient minimization algorithm was used with dummy parameters to ensure full convergence, namely 10^{-25} (1) and 4×10^{-25} kJ mol^{-1} Å^{-1} for energy and forces respectively and with 5×10^4 maximum iterations of the minimizer. Phonon calculations with the empirical potentials were performed using the i-PI[74] code, considering $2 \times 2 \times 2$ repetitions of the primitive cell. The phonons were calculated by finite differences with a 0.005 Å displacement in all Cartesian directions.

Density functional theory model

All ab initio simulations were performed using the FHI-aims package[75]. For this purpose, we employed *light* settings for all atomic species, together with the Perdew-Burke-Ernzerhof exchange-correlation functional[76] and many-body dispersion corrections[77]. We have used $5 \times 5 \times 5$ k-point sampling of the Brillouin zone. A self-consistency convergence criterion of 10^{-5} eV Å^{-1} was imposed on the forces, which ensured that energies were converged to 10^{-7} eV or below. The relaxation was performed using the trust radius version of the Broyden–Fletcher–Goldfarb–Shanno[78,79] optimization algorithm with the maximum residual force component threshold equal to 10^{-4} eV Å^{-1}. Lattice dynamics calculations were performed through finite differences using Phonopy[80]. The atomic displacements were of 0.002 Å in all Cartesian directions. The size of the supercell was individually chosen for the different molecular crystals, with the requirement that at least twice the distance between molecular centers of mass of adjacent molecules was comprised by the vector lengths in each direction.

Framework development

The framework for the GPR model was developed using Python3 and Fortran95 languages. The SOAP, MBTR, and ACSF representations were calculated using the DScribe[55] package.

DATA AVAILABILITY

All data necessary to replicate and interpret the free energy prediction framework discussed in this article can be accessed in the NOMAD repository with identifier https://doi.org/10.17172/NOMAD/2020.09.16-1.

CODE AVAILABILITY

A preliminary version of the core functionalities of presented framework is available in https://github.com/sabia-group/fep.git.

Received: 11 January 2021; Accepted: 21 September 2021;
Published online: 15 October 2021

REFERENCES

1. Desiraju, G. R. *Crystal Engineering: The Design of Organic Solids.* (Elsevier, 1989).
2. Cruz-Cabeza, A. J., Reutzel-Edens, S. M. & Bernstein, J. Facts and fictions about polymorphism. *Chem. Soc. Rev.* **44**, 8619–8635 (2015).
3. Davey, R. J. Polymorphism in molecular crystals by Joel Bernstein. *Crystal Growth Des.* **2**, 675–676 (2002).
4. Hoja, J. et al. Reliable and practical computational description of molecular crystal polymorphs. *Sci. Adv.* **5**, eaau3338 (2019).
5. Körbel, S., Marques, M. A. L. & Botti, S. Stability and electronic properties of new inorganic perovskites from high-throughput ab initio calculations. *J. Mater. Chem. C* **4**, 3157–3167 (2016).
6. Curtarolo, S., Kolmogorov, A. N. & Cocks, F. H. High-throughput ab initio analysis of the Bi-In, Bi-Mg, Bi-Sb, In-Mg, In-Sb, and Mg-Sb systems. *Calphad* **29**, 155–161 (2005).
7. Hart, G. L. W., Curtarolo, S., Massalski, T. B. & Levy, O. Comprehensive search for new phases and compounds in binary alloy systems based on platinum-group metals, using a computational first-principles approach. *Phys. Rev. X* **3**, 041035 (2013).
8. Price, S. L. Predicting crystal structures of organic compounds. *Chem. Soc. Rev.* **43**, 2098–2111 (2014).
9. Musil, F. et al. Machine learning for the structure-energy-property landscapes of molecular crystals. *Chem. Sci.* **9**, 1289–1300 (2018).
10. Curtarolo, S. et al. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
11. Nyman, J. & Day, G. M. Static and lattice vibrational energy differences between polymorphs. *CrystEngComm* **17**, 5154–5165 (2015).
12. Nyman, J. & Day, G. M. Modelling temperature-dependent properties of polymorphic organic molecular crystals. *Phys. Chem. Chem. Phys.* **18**, 31132–31143 (2016).
13. Born, M. & Huang, K. *Dynamical Theory of Crystal Lattices* (Clarendon Press, 1954).
14. Vasileiadis, M. *Calculation of the Free Energy of Crystalline Solids* (Imperial College, 2013).
15. Vega, C., Sanz, E., Abascal, J. L. F. & Noya, E. G. Determination of phase diagrams via computer simulation: methodology and applications to water, electrolytes and proteins. *J. Condens. Matter Phys.* **20**, 153101 (2008).
16. Ghiringhelli, L. M., Los, J. H., Meijer, E. J., Fasolino, A. & Frenkel, D. Modeling the phase diagram of carbon. *Phys. Rev. Lett.* **94**, 145701 (2005).
17. Polson, J. M. & Frenkel, D. Calculation of solid-fluid phase equilibria for systems of chain molecules. *J. Chem. Phys.* **109**, 318–328 (1998).
18. Rossi, M., Gasparotto, P. & Ceriotti, M. Anharmonic and quantum fluctuations in molecular crystals: a first-principles study of the stability of paracetamol. *Phys. Rev. Lett.* **117**, 115702 (2016).
19. Cheng, B. & Ceriotti, M. Computing the absolute Gibbs free energy in atomistic simulations: applications to defects in solids. *Phys. Rev. B* **97**, 054102 (2018).
20. Kapil, V., Engel, E., Rossi, M. & Ceriotti, M. Assessment of approximate methods for anharmonic free energies. *J. Chem. Theory Comput.* **15**, 5845–5857 (2019).

21. Bazterra, V. E., Ferraro, M. B. & Facelli, J. C. Modified genetic algorithm to model crystal structures. i. benzene, naphthalene and anthracene. *J. Chem. Phys.* **116**, 5984–5991 (2002).
22. Oganov, A. R. & Glass, C. W. Crystal structure prediction using ab initio evolutionary techniques: principles and applications. *J. Chem. Phys.* **124**, 244704 (2006).
23. Price, S. L. From crystal structure prediction to polymorph prediction: interpreting the crystal energy landscape. *Phys. Chem. Chem. Phys.* **10**, 1996–2009 (2008).
24. Pickard, C. J. & Needs, R. J. Ab initio random structure searching. *J. Condens. Matter Phys.* **23**, 053201 (2011).
25. Day, G. M. Current approaches to predicting molecular organic crystal structures. *Crystallogr. Rev.* **17**, 3–52 (2011).
26. Yu, T.-Q. & Tuckerman, M. E. Temperature-accelerated method for exploring polymorphism in molecular crystals based on free energy. *Phys. Rev. Lett.* **107**, 015701 (2011).
27. Oganov, A. R., Pickard, C. J., Zhu, Q. & Needs, R. J. Structure prediction drives materials discovery. *Nat. Rev. Mater.* **4**, 331–348 (2019).
28. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
29. Legrain, F., Carrete, J., van Roekeghem, A., Curtarolo, S. & Mingo, N. How chemical composition alone can predict vibrational free energies and entropies of solids. *Chem. Mater.* **29**, 6220–6227 (2017).
30. Legrain, F. et al. Vibrational properties of metastable polymorph structures by machine learning. *J. Chem. Inf. Model.* **58 12**, 2460–2466 (2018).
31. Carrete, J., Li, W., Mingo, N., Wang, S. & Curtarolo, S. Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling. *Phys. Rev. X* **4**, 011019 (2014).
32. van Roekeghem, A., Carrete, J., Oses, C., Curtarolo, S. & Mingo, N. High-throughput computation of thermal conductivity of high-temperature solid phases: the case of oxide and fluoride perovskites. *Phys. Rev. X* **6**, 041061 (2016).
33. Raimbault, N., Grisafi, A., Ceriotti, M. & Rossi, M. Using Gaussian process regression to simulate the vibrational Raman spectra of molecular crystals. *New J. Phys.* **21**, 105001 (2019).
34. George, J., Hautier, G., Bartók, A. P., Csányi, G. & Deringer, V. L. Combining phonon accuracy with high transferability in Gaussian approximation potential models. *J. Chem. Phys.* **153**, 044104 (2020).
35. Rowe, P., Csányi, G., Alfè, D. & Michaelides, A. Development of a machine learning potential for graphene. *Phys. Rev. B* **97**, 054303 (2018).
36. Bartók, A. P., Kermode, J., Bernstein, N. & Csányi, G. Machine learning a general-purpose interatomic potential for silicon. *Phys. Rev. X* **8**, 041048 (2018).
37. Marques, M. R. G., Wolff, J., Steigemann, C. & Marques, M. A. L. Neural network force fields for simple metals and semiconductors: construction and application to the calculation of phonons and melting temperatures. *Phys. Chem. Chem. Phys.* **21**, 6506–6516 (2019).
38. Behler, J. Representing potential energy surfaces by high-dimensional neural network potentials. *J. Condens. Matter Phys.* **26**, 183001 (2014).
39. Pukrittayakamee, A. et al. Simultaneous fitting of a potential-energy surface and its corresponding force fields using feedforward neural networks. *J. Chem. Phys.* **130**, 134101 (2009).
40. De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **18**, 13754–13769 (2016).
41. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
42. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
43. Christopher, B. *Pattern Recognition and Machine Learning*. (Springer, 2006).
44. Häse, F., Valleau, S., Pyzer-Knapp, E. & Aspuru-Guzik, A. Machine learning exciton dynamics. *Chem. Sci.* **7**, 5139–5147 (2016).
45. Browning, N. J., Ramakrishnan, R., von Lilienfeld, O. A. & Roethlisberger, U. Genetic optimization of training sets for improved machine learning models of molecular properties. *J. Phys. Chem. Lett.* **8**, 1351–1359 (2017).
46. Hansen, K. et al. Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* **9**, 3404–3419 (2013).
47. Eldar, Y., Lindenbaum, M., Porat, M. & Zeevi, Y. Y. The farthest point strategy for progressive image sampling. *IEEE Trans. Image Process* **6**, 1305–1315 (1997).
48. Rupp, M., Proschak, E. & Schneider, G. Kernel approach to molecular similarity based on iterative graph similarity. *J. Chem. Inf. Model.* **47**, 2280–2286 (2007).
49. De, S. et al. Relation between the dynamics of glassy clusters and characteristic features of their energy landscape. *Phys. Rev. Lett.* **112**, 083401 (2014).
50. De, S. et al. Energy landscape of fullerene materials: a comparison of boron to boron nitride and carbon. *Phys. Rev. Lett.* **106**, 225502 (2011).
51. Sadeghi, A. et al. Metrics for measuring distances in configuration spaces. *J. Chem. Phys.* **139**, 184118 (2013).
52. Langer, M. F., Goeßmann, A. & Rupp, M. Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning. https://arxiv.org/2003.12081 (2020).
53. del Rosario, Z., Rupp, M., Kim, Y., Antono, E. & Ling, J. Assessing the frontier: active learning, model accuracy, and multi-objective candidate discovery and optimization. *J. Chem. Phys.* **153**, 024112 (2020).
54. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).
55. Himanen, L. et al. DScribe: library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **247**, 106949 (2020).
56. CCDC. https://www.ccdc.cam.ac.uk/ Accessed 25 May 2020 (2021).
57. Beran, G. J. O., Hartman, J. D. & Heit, Y. N. Predicting molecular crystal properties from first principles: finite-temperature thermochemistry to NMR crystallography. *Acc. Chem. Res.* **49**, 2501–2508 (2016).
58. Grisafi, A. et al. Transferable machine-learning model of the electron density. *ACS Cent. Sci.* **5**, 57–64 (2019).
59. van der Giessen, E. et al. Roadmap on multiscale materials modeling. *Model. Simul. Mater. Sci. Eng.* **28**, 043001 (2020).
60. Murnaghan, F. D. The compressibility of media under extreme pressures. *Proc. Natl Acad. Sci. USA* **30**, 244–247 (1944).
61. Raimbault, N., Athavale, V. & Rossi, M. Anharmonic effects in the low-frequency vibrational modes of aspirin and paracetamol crystals. *Phys. Rev. Mater.* **3**, 053605 (2019).
62. Nyman, J., Pundyke, O. S. & Day, G. M. Accurate force fields and methods for modelling organic molecular crystals at finite temperatures. *Phys. Chem. Chem. Phys.* **18**, 15828–15837 (2016).
63. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
64. Mason, R. The crystallography of anthracene at 95°K and 290°K. *Acta Crystallogr.* **17**, 547–555 (1964).
65. Mattheus, C. C. et al. Polymorphism in pentacene. *Acta Crystallogr. C* **57**, 939–941 (2001).
66. Madelung, O., Rössler, U. & Schulz, M. (eds.) *Ternary Compounds, Organic Semiconductors* (Springer, 2000).
67. Bond, A. D. & Davies, J. E. Styrene at 120K. *Acta Crystallogr. E* **57**, o1191–o1193 (2001).
68. Yasuda, N., Uekusa, H. & Ohashi, Y. Styrene at 83K. *Acta Crystallogr. E* **57**, o1189–o1190 (2001).
69. Budzianowski, A. & Katrusiak, A. Pressure-frozen benzene I revisited. *Acta Crystallogr. B* **62**, 94–101 (2006).
70. Katrusiak, A., Podsiadło, M. & Budzianowski, A. Association ch···π and no van der Waals contacts at the lowest limits of crystalline benzene i and ii stability regions. *Cryst. Growth Des.* **10**, 3461–3465 (2010).
71. McDonagh, D., Skylaris, C.-K. & Day, G. M. Machine-learned fragment-based energies for crystal structure prediction. *J. Chem. Theory Comput.* **15**, 2743–2758 (2019).
72. Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **117**, 1–19 (1995).
73. Stuart, S. J., Tutein, A. B. & Harrison, J. A. A reactive potential for hydrocarbons with intermolecular interactions. *J. Chem. Phys.* **112**, 6472–6486 (2000).
74. Kapil, V. et al. i-pi 2.0: a universal force engine for advanced molecular simulations. *Comput. Phys. Commun.* **236**, 214–223 (2019).
75. Blum, V. et al. Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **180**, 2175–2196 (2009).
76. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
77. Tkatchenko, A., DiStasio, R. A., Car, R. & Scheffler, M. Accurate and efficient method for many-body van der Waals interactions. *Phys. Rev. Lett.* **108**, 236402 (2012).
78. Pfrommer, B. G., Côté, M., Louie, S. G. & Cohen, M. L. Relaxation of crystals with the quasi-Newton method. *J. Comput. Phys.* **131**, 233–240 (1997).
79. Nocedal, J. & Wright, S. J. *Numerical Optimization* (Springer, 2000).
80. Togo, A. & Tanaka, I. First principles phonon calculations in materials science. *Scr. Mater.* **108**, 1–5 (2015).

ACKNOWLEDGEMENTS

We acknowledge useful discussions with T. Bereau, M. Langer, L. Ghiringhelli, and M. Ceriotti. We thank M. Rupp and M. Langer for a critical read of the manuscript draft. This work has been financially supported by BiGmax, the Max Planck Society's Research Network on Big-Data-Driven Materials-Science.

AUTHOR CONTRIBUTIONS

M.K. was responsible for designing, developing and programming the GPR framework, and performing all necessary simulations. M.R. was responsible for the project planning, design, and supervision. Both authors wrote the manuscript and analyzed the data.

FUNDING

Open Access funding enabled and organized by Projekt DEAL.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-021-00638-x.

Correspondence and requests for materials should be addressed to Marcin Krynski.