JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

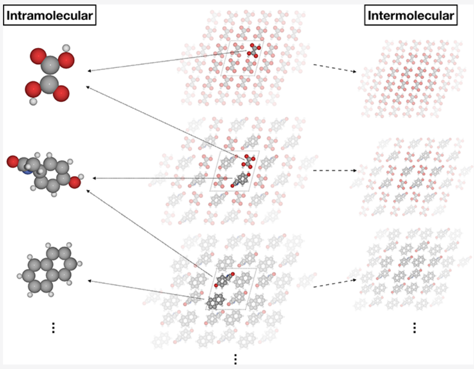
A Hybrid Machine Learning Approach for Structure Stability Prediction in Molecular Co-crystal Screenings

Simon Wengert, Gábor Csányi, Karsten Reuter, and Johannes T. Margraf*

Cite This: *J. Chem. Theory Comput.* 2022, 18, 4586–4593 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Co-crystals are a highly interesting material class as varying their components and stoichiometry in principle allows tuning supramolecular assemblies toward desired physical properties. The *in silico* prediction of co-crystal structures represents a daunting task, however, as they span a vast search space and usually feature large unit cells. This requires theoretical models that are accurate and fast to evaluate, a combination that can in principle be accomplished by modern machine-learned (ML) potentials trained on first-principles data. Crucially, these ML potentials need to account for the description of long-range interactions, which are essential for the stability and structure of molecular crystals. In this contribution, we present a strategy for developing Δ-ML potentials for co-crystals, which use a physical baseline model to describe long-range interactions. The applicability of this approach is demonstrated for co-crystals of variable composition consisting of an active pharmaceutical ingredient and various co-formers. We find that the Δ-ML approach offers a strong and consistent improvement over the density functional tight binding baseline. Importantly, this even holds true when extrapolating beyond the scope of the training set, for instance in molecular dynamics simulations under ambient conditions.



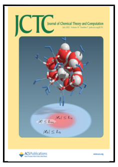
1. INTRODUCTION

The physical properties of a molecular crystal are strongly dependent on the arrangement of its building blocks in the solid state.[1] In aggregate-induced emission, for instance, interactions in the crystalline phase (or even in concentrated solution) cause otherwise non-luminescent molecules to become emissive.[2] Similarly, piezochromic luminescent materials change the color of their emission when intermolecular arrangements in the solid state are altered by external mechanical stimuli.[3] Beyond these specific examples, the large variety of crystal forms detected and characterized for certain molecules reveals that the crystal structure impacts many other properties as well, such as aqueous solubility,[4] charge transport,[5] or plastic deformation[6] to name but a few.

Being able to control molecular arrangements in the solid state, consequently, enables tuning materials toward desired properties.[7] The design of multi-component molecular crystals, so-called co-crystals, is promising in this respect as it provides a versatile route to this goal.[8] Here, the molecule of interest crystallizes in the presence of another compound, a so-called co-former. Co-crystallization has garnered interest in both academia and industry as a strategy for the design of materials with improved performance. Applications include non-linear optics,[9] energetic materials,[10] and, most notably, pharmaceuticals.[11] Here, active pharmaceutical ingredients are often combined with co-formers to improve their bioavailabilty (e.g., by tuning the dissolution rate, solubility, compressibility, and thermal stability of the co-crystal).[12,13]

The space of possible co-formers is generally quite large. For pharmaceuticals, the "generally regarded as safe" (GRAS) list is often used, which contains hundreds of molecules considered as safe for human consumption. The synthesis of multi-component crystals thus provides a large design space. Unfortunately, the successful formation of a co-crystal from its compounds is by no means trivial.[14] Indeed, recrystallization is actually a common technique for purifying compounds, i.e., to separate them from one another. Moreover, the stability and structure of a potential co-crystal are hard to predict as they result from a delicate balance between relatively weak interactions.[15] Unlike conventional covalent chemistry, the synthesis of co-crystals is thus much more difficult to plan and often a game of trial and error. A more targeted approach would therefore be highly desirable. Here, computational methods could play an important role, e.g., by predicting whether a given co-former will lead to stable co-crystals and which structural motifs are likely to be formed for a given combination. This would allow narrowing the list of potential co-formers down to a few promising candidates and thus

Received: April 8, 2022
Published: June 16, 2022



dramatically reduce the number of necessary experiments and associated costs.

The *in silico* search for molecular crystal structures faces some major challenges, however.[16] On one hand, the large search space of potential structures requires evaluating the stability of a large number of trial crystals. On the other hand, highly accurate (and thus computationally expensive) levels of theory need to be applied for a reliable prediction of crystal lattice energies.[17] Even for single-component crystals, this leads to a difficult trade-off between adequately exploring the space of possible structures and using sufficiently accurate methods to evaluate their stability. This situation is exacerbated on several fronts when screening for appropriate co-formers. First, a separate crystal structure search needs to be performed for each potential co-former. Second, the unit cells of co-crystals are typically significantly larger than those of single-component crystals as quantified by the number of molecules in the unit cell (Z) and the number of symmetry independent molecules (Z'). This means that there are more degrees of freedom to optimize ($Z' > 1$), while each energy evaluation is also more expensive (large Z). Finally, the stoichiometry of the stable co-crystal is typically unknown, which adds an additional dimension to the search space. As a consequence, computationally efficient and accurate potentials for crystal structure search and co-crystals in particular are highly desirable.

Owing to their outstanding accuracy-to-cost ratio, modern machine-learned (ML) potentials are in principle highly promising in this context. Challenges arise, however, from the importance of long-range contributions due to electrostatics or dispersion. Although recent advances in long-range ML potentials[18−22] bear good prospect for modeling condensed molecular systems, short-ranged ML potentials are still prevalent and, thus, generally less frequently applied in this context than for gas-phase molecules or ionic solids. As a notable exception, Montes-Campos et al. have nonetheless developed accurate ML potentials for molecular multi-component systems and applied them to the related field of ionic liquids.[23] In this case, they benefited from the fact that the dynamics of liquids are only weakly influenced by long-range interactions, as is also the case for ion mobilities in solid electrolytes.[24] The importance of long-range interactions for the relative stabilities of molecular crystal polymorphs is well established, however.[25]

Kapil and Engel overcame this issue by using short-ranged ML potentials for sampling, in combination with additional *ab initio* calculations for stability ranking.[26] This allowed them to obtain highly accurate thermodynamic stabilities incorporating the combined effects from the electronic structure, quantum nuclear effects, and thermal contributions. In contrast, a Δ-ML[27] ansatz bypasses the need for subsequent *ab initio* calculations by combining local ML models with appropriate (long-ranged) baselines. This has proven to be highly useful for molecular crystal structure prediction (CSP).[28,29]

In a previous study, we presented a framework for the data-efficient generation of Δ-ML models for single-component molecular crystals, which benefits from a separate treatment of inter- and intramolecular interactions.[29] In this contribution, we present recent advances in extending this approach to co-crystals. Our approach is designed with the co-former screening setting in mind.[30] Consequently, we will consider a single active pharmaceutical ingredient (paracetamol) combined with four different co-formers, as shown in Figure 1. These systems have been proposed and extensively characterized by Karki et al.[13] Being one of the most common pharmaceuticals worldwide, paracetamol is a prototypical active pharmaceutical ingredient, while the co-formers oxalic acid (Oxa), naphthalene (Nap), phenazine (Phe), and theophylline (Thp) cover a wide range in terms of polarity, functional groups, and molecular shapes, inducing various types of intermolecular interactions and arrangements in the solid state.

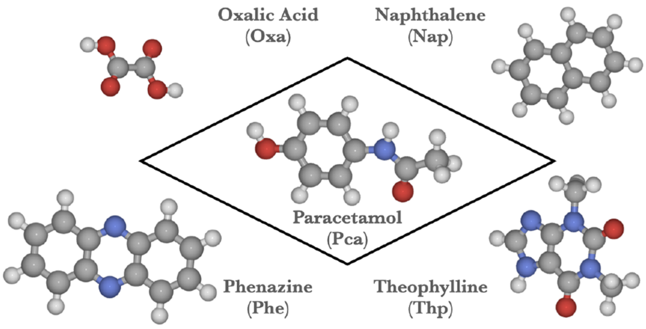


Figure 1. Central active pharmaceutical ingredient paracetamol (Pca) and the co-formers oxalic acid (Oxa), naphthalene (Nap), phenazine (Phe), and theophylline (Thp). Gray spheres: C, blue spheres: N, red spheres: O, white spheres: H.

2. METHODS

2.1. General Approach. The approach we previously developed[29] for single-component crystals has two main features. First, it combines a short-ranged ML potential with a long-ranged physical baseline (Δ-ML). Second, the ML potential is split into an intramolecular and intermolecular correction. The same idea was also used in local approximate models[31] for lattice energy minimizations of molecular crystals. We found this splitting to be advantageous because these interactions occur on different length scales. Additionally, reference data for the intramolecular correction can be generated cheaply from gas-phase calculations. It is even possible to use a different level of theory for this purpose. Below, we briefly summarize the main points of the method, highlighting the extensions that were developed for co-crystals.

2.2. Baseline Method. The dispersion-corrected density functional tight binding (DFTB) method represents an ideal baseline for CSP. First, it is efficient enough to be applied in a setting where several thousands of organic crystal structures need to be optimized.[32] In addition, the modern third-order variant of DFTB[33] combined with the 3ob[34] parameterization provides an accurate description of electrostatics, charge transfer, and polarization. Finally, the missing dispersion contributions can be corrected efficiently, e.g., via the D4 method.[35,36] The baseline method in this work is thus defined as DFTB3(3ob)+D4 (DFTB+D4 in the following).

2.3. Machine Learning Method. The intra- and intermolecular corrections to the baseline will be defined as Gaussian approximation potentials (GAP)[37,38] using the smooth overlap of atomic position (SOAP)[39] representation. These GAP models are fitted to both energies and forces. To account for the presence of different molecular building blocks in co-crystals, a separate intramolecular correction is fitted for each. In contrast, a single intermolecular correction is used to describe the interactions among paracetamol and the four co-formers. The energy expression of the combined DFTB+D4 and GAP model (termed Δ-GAP in the following) thus reads

$$E^{\Delta\text{-GAP}} = E^{\text{DFTB+D4}}_{\text{crystal}} + \Delta E^{\text{inter}}_{\text{crystal}} + \sum_{t}^{N_{\text{types}}} \sum_{i}^{N_t} \Delta E_i^{\text{intra},t} \tag{1}$$

where for each of the N_{types} possible components, the corresponding intramolecular GAP correction is applied to each molecule i (in which N_t is the number of molecules of type t present in the given unit cell). Note that intra- and intermolecular corrections are applied to energies, forces, and stresses. The models can thus be used for full unit cell relaxations and constant pressure molecular dynamics.

2.4. Target Method. The high-level target method to which the correction is fitted will be hybrid DFT (using the PBE0[40] functional) with a many-body dispersion[25,41] (MBD) correction. PBE0+MBD provides a sophisticated description of the interactions relevant to organic solids. The importance of MBD contributions and hybrid functionals for the stability assessment of molecular crystals has been highlighted by Hoja and Tkatchenko.[42] For the X23 database, containing van der Waals (vdW)-bonded, hydrogen-bonded, and mixed molecular crystals, this combination has been shown to yield lattice energies within chemical accuracy (43 meV) when compared to (back-corrected) experimental enthalpies of sublimation.[43] Moreover, LeBlanc et al. found in their studies on multicomponent acid−base crystals that the exact-exchange mixing employed in hybrid DFT is essential to cure significant geometry errors introduced by the delocalization error of semi-local functionals.[44] Due to the prohibitive computational and memory requirements of PBE0+MBD with large basis sets, we define the target method—called PBE(0)+MBD hereafter—as a composite scheme: The intramolecular part is fully described by PBE0+MBD with a tightly converged basis of numerical atomic orbitals (NAO). The intermolecular part is described by PBE+MBD[45] with the same basis, plus the difference PBE+MBD to PBE0+MBD in a smaller NAO basis. A similar scheme was used by Hoja et al.,[17] who found it to yield lattice energies in excellent agreement with converged PBE0+MBD calculations.

2.5. Training Data. The structures entering the training set ultimately define the information that is available about the target function. In the context of co-crystal screening studies, the training set should thus include combinations of the molecule of interest with all co-formers. To train the intermolecular model, we selected samples from a pool of ca. 10,000 trial structures created with the PyXtal package.[46] In this initial pool, a wide range of compositions was considered for each combination to span all possible stoichiometries. These trial candidates were locally relaxed at the DFTB+D4 level of theory. To obtain a diverse set of training structures from this pool, we then employed the farthest point sampling (FPS)[47] heuristic. Here, the SOAP kernel was used as a similarity measure between atomic environments and structures were sequentially added to the training set by selecting the most dissimilar structures to the current training set at each iteration. Note that there are several possibilities to define global similarity metrics between structures, given a local similarity metric like SOAP.[48] Herein, we simply used the maximal dissimilarity between any two atomic environments.[49] From this process, 1000 training structures were obtained, 250 for each crystal/co-former pair (including the corresponding single-component crystals).

We further included 77 structures corresponding to the experimentally known single-component crystals and randomly perturbed structures derived from them. The rationale behind this is that the experimental information about the single-component crystals is usually available in co-crystal studies. This allows us to include some additional information on highly stable interactions, though not for the important paracetamol/co-former contacts. The consequences of this bias in the training set will be discussed in detail below.

In contrast to the intermolecular correction, the training data for the intramolecular model is computationally cheap to generate as it only requires single-point calculations on monomer configurations in the gas phase. To obtain these configurations, monomer geometries were extracted from the training crystals. These were further supplemented, with configurations from gas-phase molecular dynamics simulations and local relaxations, to extensively cover the configurational space of each building block. Further details on the training sets and all training data are provided in the Supporting Information.

3. RESULTS AND DISCUSSION

To validate the presented approach, we will first test its performance on a diverse set of crystal structures as one would encounter in a CSP workflow. To this end, a test set of 1000 structures was generated in an analogous procedure to the training set generation. Here, the FPS selection included the training set to maximize the distance between test and training structures (see the Supporting Information for details). All test structures were subsequently relaxed at the Δ-GAP level. Lattice energies and force errors for this test set are summarized in Figure 2. For lattice energy calculation, we used

$$E^{\text{latt}}_{\text{crystal}} = (E_{\text{crystal}} - n_{\text{A}}E_{\text{gas,A}} - n_{\text{B}}E_{\text{gas,B}})/(n_{\text{A}} + n_{\text{B}}) \tag{2}$$

where the difference between the energy of the crystal, E_{crystal}, and the energies, E_{gas}, of its optimized molecular compounds is computed first and then normalized by the total number of compounds in the crystal unit cell. Note that lattice energies of single-component crystals have been calculated in the same way using $n_{\text{B}} = 0$.

In Figure 2 (top), Δ-GAP and DFTB+D4 predicted lattice energies are shown in comparison with the PBE(0)+MBD target values. The reference energies cover a broad range of ca. 1 eV per molecule and are mostly negative. This indicates that the random search in general leads to reasonable candidate structures, which are stable with respect to sublimation. The DFTB+D4 lattice energies are reasonably well correlated with this reference but display significant scatter. Furthermore, the lattice energies are systematically underestimated, leading to a mean absolute error (MAE) of 183 meV. Applying intra- and intermolecular corrections to this baseline in the Δ-GAP scheme strongly improves the agreement with the target, resulting in an overall MAE of only 34 meV. This is achieved both by eliminating the systematic underestimation of the lattice energies and by reducing the scatter in the predictions, as indicated by the significantly smaller standard deviation (STD) of the Δ-GAP errors (32 meV vs 83 meV). Indeed, the Δ-GAP energies actually show a slight offset toward more negative values due to the fact that the structures are minima on the Δ-GAP potential energy surface.

An even more substantial improvement is observed for force predictions (see Figure 2, bottom). Here, DFTB+D4 displays a broad error distribution and a correspondingly large MAE of 324 meV/Å. In contrast, the error distribution of predicted Δ-

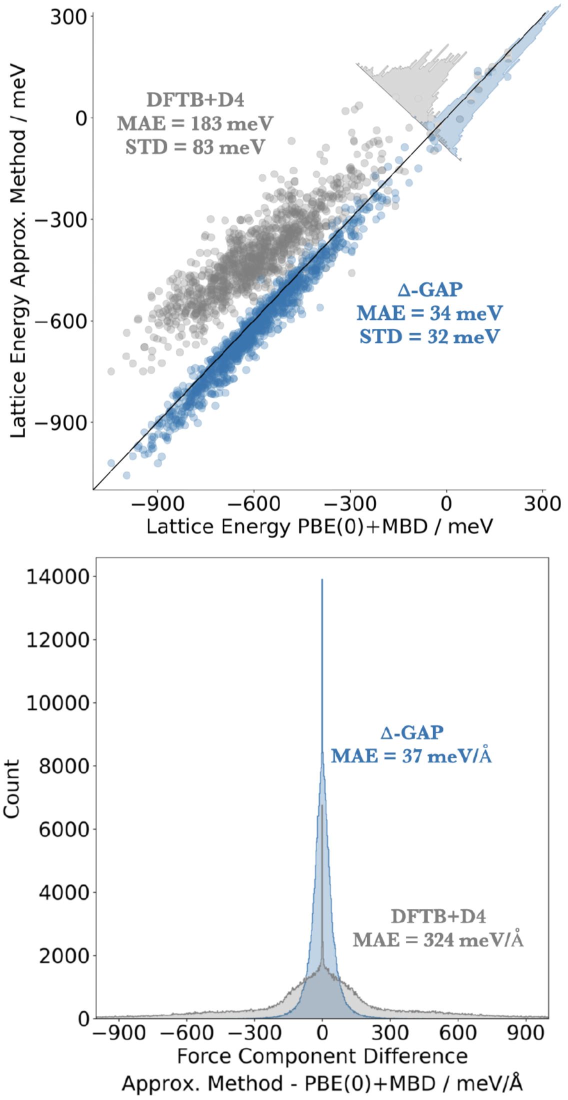


Figure 2. Correlation plot for the DFTB+D4 baseline and Δ-GAP lattice energies per molecule of PcaOxa, PcaNap, PcaPhe, and PcaThp test crystals (both single-component and co-crystals) against the PBE(0)+MBD target level of theory (top) and the corresponding differences in force components (bottom). Note that the slight shift of the Δ-GAP lattice energy distribution toward lower values compared to PBE(0)+MBD is due to the fact that the test set structures are minima on the Δ-GAP potential energy surface, while the training structures are minima on the DFTB+D4 surface (see text). The spike in the distributions of force component differences results from certain force components being zero by symmetry at all levels of theory.

GAP force components is much narrower and the MAE almost an order of magnitude lower. Importantly, while the lattice energy error of DFTB+D4 is fairly systematic, the force error cannot be corrected in a simple way and will lead to substantial deviations in the predicted structures. This is of particular relevance in the context of CSP, where accurate structure relaxations are often by far the most expensive component. Due to their small force errors, Δ-GAP relaxations should provide near PBE(0)+MBD quality structures at a fraction of the computational costs.

While the above results are promising, it should be emphasized that the training and test structures used herein are merely local minima. In particular, they are somewhat less dense and less stable than the known experimental structures for these co-crystals (see the Supporting Information). In future applications, this should be mitigated by using a more advanced CSP search algorithm (ideally together with an accurate ML potential as proposed herein) to generate more realistic structures. From the perspective of this paper, there is also a positive aspect to this discrepancy between training and experimental structures though, as it creates an opportunity to test the extrapolative capabilities of the presented approach. To this end, we test the accuracy of our method on the known experimental structures of each co-crystal.

For all experimental co-crystal structures, atomic positions and unit cell parameters were fully relaxed using the DFTB+D4 baseline, Δ-GAP model, and the PBE(0)+MBD target. For comparison, we also performed calculations at the PBE+MBD level, which is often used for relaxations instead of the more expensive hybrid PBE0 functional. These results are summarized in Figure 3.

Relative density deviations with respect to the PBE(0)+MBD geometry are shown in Figure 3a. We find that the DFTB+D4 structures are significantly contracted, in agreement with previous studies where this was attributed to insufficient Pauli-repulsion at longer distances.[32,50] In contrast, the Δ-GAP structures are in much better agreement, with only slightly higher densities. For comparison, PBE+MBD shows slightly larger but more systematic density deviations of around 3%. In contrast to Δ-GAP and DFTB+D4, this is due to systematically lower densities, which are likely a consequence of differences in the molecular electrostatic potentials predicted by semi-local and hybrid functionals.

On an atomistic level, crystal structures are typically compared with the RMSD_{15} metric,[51] as shown in Figure 3b. To this end, the root mean square deviation of the positions of non-hydrogen atoms in 15-molecule clusters extracted from the relaxed crystal structures is calculated. We again use the PBE(0)+MBD structures as the reference. As for the densities, the DFTB+D4 baseline displays the most significant structural discrepancies with the target. These are mostly due to reduced intermolecular distances, such as the spacings in the layered structures PcaOxa, PcaNap, and PcaThp and variations in molecular orientation (see Figure 4 and the Supporting Information for further examples). For PcaNap, additional discrepancy is caused by the intramolecular adjustment of paracetamol to the crystal environment. Here, the DFTB+D4 baseline predicts a weaker out-of-plane rotation of the C=O group, as highlighted in the inset. In all cases, these deviations are mitigated by the ML correction, though the effects are less distinct for PcaThp, which is already reasonably well described by the baseline. Finally, PBE+MBD is slightly more accurate and systematic than Δ-GAP, albeit at a much higher computational cost (by roughly 3 orders of magnitude, see the Supporting Information). Indeed, the structural discrepancies are in this case entirely due to the aforementioned density deviations, whereas the relative positions and orientations of the molecules are in good agreement with the PBE(0)+MBD relaxed structures.

In addition to these geometric comparisons, the relaxed structures were also evaluated from an energetic perspective. This is relevant when structures from the approximate method are used as inputs for single-point calculations or relaxations with higher level methods. Here, small structural deviations—bond distances for instance—can significantly impact predicted energies and energy differences. To evaluate the

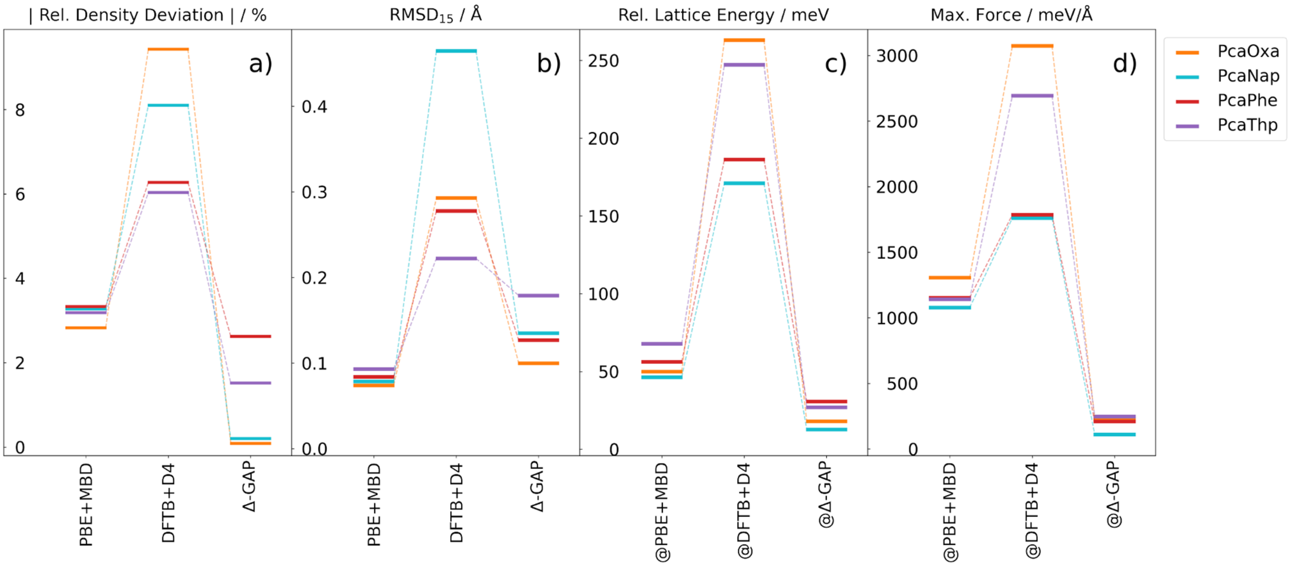


Figure 3. Comparison between PBE+MBD, the DFTB+D4 baseline, and Δ-GAP results on experimental co-crystals for PcaOxa, PcaNap, PcaPhe, and PceThp against the PBE(0)+MBD target level of theory in terms of the absolute values for percentage density deviations (a), the RMSDs between overlaying 15-mers sliced from crystal structures (b), lattice energies per molecule relative to PBE(0)+MBD optimized structures obtained from single-point calculations on structures optimized on the approximate levels of theory specified in the figure (c), and the corresponding maximum remaining PBE(0)+MBD forces (d).

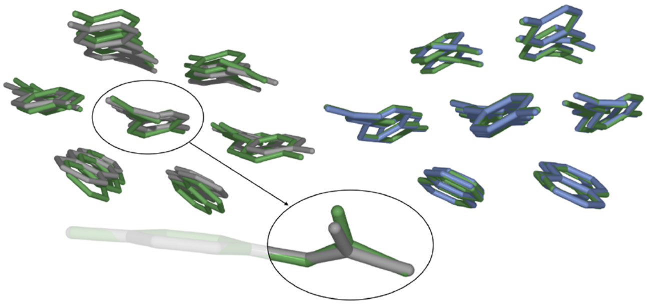


Figure 4. Overlay of the PBE(0)+MBD (green) optimized experimental PcaNap co-crystal with DFTB+D4 (gray) and Δ-GAP (blue). For DFTB+D4, a separate overlay is shown for paracetamol conformers extracted from the crystal environment.

quality of the structures in this context, single-point PBE(0)+MBD calculations were performed on the geometries predicted by the approximate levels of theory. Figure 3c illustrates the errors in lattice energies obtained from these calculations, while Figure 3d shows the corresponding maximum force. Here, the Δ-GAP values are lowest in all cases, indicating that they are closest to the PBE(0)+MBD minimum from an energetic perspective. The deviations of PBE+MBD are similarly systematic but significantly higher. Finally, the DFTB+D4 results are more scattered and generally poorer with maximum forces of up to 3 eV/Å for the putative minima and lattice energy errors of up to 250 meV.

Overall, the Δ-GAP model is thus a robust and significant improvement on DFTB+D4, even when applied outside the range of the training set. Perhaps surprisingly, it is even an improvement over the much more expensive PBE+MBD method in many respects, when comparing with the PBE(0)+MBD target. Of course, the ultimate test is comparison with experimental structures, however. Here, we somewhat unexpectedly found that the PBE+MBD densities are actually closer to the experimental values than the ones predicted by PBE(0)+MBD (and consequently also by Δ-GAP, see Figure 5).

These apparent deviations can be resolved by considering thermal effects, however. Computationally relaxed crystal structures correspond to the 0 K limit, whereas crystallographic experiments are usually performed at finite temperature and pressure. The over-contraction of PBE(0)+MBD will thus be counteracted by thermal expansion. An advantage of computa-

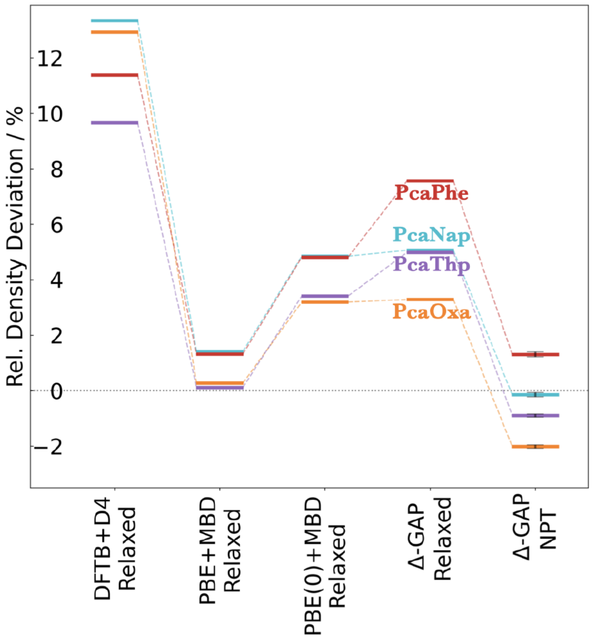


Figure 5. Percentage deviations from experimental measured densities for PcaOxa, PcaNap, PcaPhe, and PceThp co-crystals optimized with the DFTB+D4 baseline, PBE+MBD, the PBE(0)+MBD target level of theory, and Δ-GAP, as well as for densities obtained from Δ-GAP NPT simulations (298 K and 1 bar). For NPT, results corresponding to standard errors of the deviations are illustrated.

tionally efficient approaches like Δ-GAP is that they allow for including such effects in a straightforward manner by performing molecular dynamics in the NPT ensemble (at 298 K and ambient pressure). As shown in Figure 5, the average densities across these trajectories are indeed in very good agreement with the experiment. This also indicates that the PBE+MBD (0 K) densities are in fact fortuitously close to the experiment as the inclusion of thermal expansion effects would likely also cause them to decrease by ca. 5%.

Importantly, such finite temperature simulations would be computationally prohibitive on the hybrid DFT level. Being an efficient surrogate for PBE(0)+MBD, Δ-GAP thus allows performing simulations that would otherwise be impossible. These results also further underscore the robustness of our ML approach, given that the experimental structures are outside the scope of the training set and no crystal MD data was used for training at all. This is thanks to the strong physical prior that the DFTB+D4 baseline provides and the smoothness of the GAP correction. Additional improvements could be obtained by combining the current approach with more advanced structure search algorithms[52−54] and by iteratively refining the GAP correction in an active learning workflow.

4. CONCLUSIONS

We have presented an approach for Δ-ML potentials applicable to both pure crystals and co-crystals of variable composition. This Δ-GAP approach enables efficient global crystal structure searches with near hybrid DFT accuracy, at a much reduced cost. Building on a previous approach for single-component crystals, we fit separate intramolecular corrections for each component and a single intermolecular correction for all active molecule/co-former pairs. Our approach strongly reduces energy and force errors with respect to the baseline model.

Notably, the training structures used herein were generated with a simple random search procedure and consequently display markedly lower densities and stabilities than the known experimental co-crystals. Nevertheless, the Δ-GAP potentials are able to predict the structures of experimental polymorphs with high accuracy, outperforming PBE+MBD at a much lower computational cost. This shows that this approach is highly robust in an extrapolative regime. In future work, we aim to combine these potentials with more advanced CSP search algorithms.[52−54]

Finally, it should be noted that many-body dispersion can be rather long-ranged in some cases,[55] while our baseline method relies on the D4 correction, which lacks these effects. Since the intermolecular ML contributions are by construction short-ranged due to the use of a local representation, long-range many-body dispersion effects are thus currently neglected in our approach. This could be mitigated by including a physical many-body dispersion model in the baseline. An efficient ML-based MBD implementation that makes this computationally feasible has recently been reported.[56,57]

5. COMPUTATIONAL DETAILS

DFT calculations were performed with the all-electron code FHI-aims,[58] using the PBE[45] and PBE0[40] functionals. A post-SCF dispersion correction was applied using the MBD[25,41] method. Two accuracy levels with a large or small basis set have been used (compare Section 2). Large basis set calculations correspond to tier2 settings and tight integration grids, while small basis set calculations correspond to tier1 settings and light integration grids. DFTB3[33] calculations were performed using DFTB+[59] together with the 3ob[34] parametrization and the D4[35,36] dispersion correction without non-additive effects. For periodic calculations, the number of **k** points (n) in each direction is chosen as the smallest integer satisfying the relation $n{\cdot}a \geq x$, where a is the unit cell length along that direction and $x = 30$. GAP potentials were trained and evaluated using the QUIP[60] package. Candidate crystal structures were created with the PyXtal[46] package.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.2c00343.

Method and additional details on co-crystal stabilities, density dependence of the lattice energies for experimental co-crystals, structural overlay of optimized experimental co-crystals, density and lattice energy comparison between training and experimental co-crystals, molecular dynamics simulations, and comparison of computational costs (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Johannes T. Margraf − *Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany;* orcid.org/0000-0002-0862-5289; Email: johannes.margraf@ch.tum.de

Authors

Simon Wengert − *Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany; Chair of Theoretical Chemistry, Technische Universität München, 85747 Garching, Germany*

Gábor Csányi − *Engineering Laboratory, University of Cambridge, Cambridge CB2 1PZ, United Kingdom*

Karsten Reuter − *Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.2c00343

Funding

Open access funded by Max Planck Society.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

S.W. and K.R. acknowledge funding from Deutsche Forschungsgemeinschaft (DFG) under Project RE1509/18-2.

■ REFERENCES

(1) Aitipamula, S.; Chow, P. S.; Tan, R. B. H. Polymorphism in cocrystals: a review and assessment of its significance. *CrystEngComm* **2014**, *16*, 3451−3465.

(2) Hong, Y.; Lam, J. W. Y.; Tang, B. Z. Aggregation-induced emission. *Chem. Soc. Rev.* **2011**, *40*, 5361−5388.

(3) Sagara, Y.; Kato, T. Mechanically induced luminescence changes in molecular assemblies. *Nat. Chem.* **2009**, *1*, 605−610.

(4) Bernstein, J. *Polymorphism in Molecular Crystals*; International Union of Crystallography Monographs on Crystallography; Oxford University Press: Oxford, U.K., 2007.

(5) Jurchescu, O. D.; Mourey, D. A.; Subramanian, S.; Parkin, S. R.; Vogel, B. M.; Anthony, J. E.; Jackson, T. N.; Gundlach, D. J. Effects of

polymorphism on charge transport in organic semiconductors. *Phys. Rev. B* **2009**, *80*, No. 085201.

(6) Nichols, G.; Frampton, C. S. Physicochemical characterization of the orthorhombic polymorph of paracetamol crystallized from solution. *J. Pharm. Sci.* **1998**, *87*, 684−693.

(7) Aakeröy, C. B.; Sandhu, B. Solid Form Landscape and Design of Physical Properties. In *Engineering Crystallography: From Molecule to Crystal to Functional Form*; Roberts, K. J., Docherty, R., Tamura, R., Eds.; Springer: Dordrecht, The Netherlands, 2017; pp 45−56.

(8) Gunawardana, C. A.; Aakeröy, C. B. Co-crystal synthesis: fact, fancy, and great expectations. *Chem. Commun.* **2018**, *54*, 14047−14060.

(9) Choi, E.-Y.; Jazbinsek, M.; Lee, S.-H.; Günter, P.; Yun, H.; Lee, S. W.; Kwon, O.-P. Co-crystal structure selection of nonlinear optical analogue polyenes. *CrystEngComm* **2012**, *14*, 4306−4311.

(10) Aakeröy, C. B.; Wijethunga, T. K.; Desper, J. Crystal Engineering of Energetic Materials: Co-crystals of Ethylenedinitramine (EDNA) with Modified Performance and Improved Chemical Stability. *Chem. − Eur. J.* **2015**, *21*, 11029−11037.

(11) Steed, J. W. The role of co-crystals in pharmaceutical design. *Trends Pharmacol. Sci.* **2013**, *34*, 185−193.

(12) Schultheiss, N.; Newman, A. Pharmaceutical Cocrystals and Their Physicochemical Properties. *Cryst. Growth Des.* **2009**, *9*, 2950−2967.

(13) Karki, S.; Friščić, T.; Fábián, L.; Laity, P. R.; Day, G. M.; Jones, W. Improving Mechanical Properties of Crystalline Solids by Cocrystal Formation: New Compressible Forms of Paracetamol. *Adv. Mater.* **2009**, *21*, 3905−3909.

(14) Springuel, G.; Norberg, B.; Robeyns, K.; Wouters, J.; Leyssens, T. Advances in Pharmaceutical Co-crystal Screening: Effective Co-crystal Screening through Structural Resemblance. *Cryst. Growth Des.* **2012**, *12*, 475−484.

(15) Aakeröy, C. Is there any point in making co-crystals? *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2015**, *71*, 387−391.

(16) Reilly, A. M.; et al. Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2016**, *72*, 439−459.

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

(18) Grisafi, A.; Ceriotti, M. Incorporating long-range physics in atomic-scale machine learning. *J. Chem. Phys.* **2019**, *151*, 204105.

(19) Xie, X.; Persson, K. A.; Small, D. W. Incorporating Electronic Information into Machine Learning Potential Energy Surfaces via Approaching the Ground-State Electronic Energy as a Function of Atom-Based Electronic Populations. *J. Chem. Theory Comput.* **2020**, *16*, 4256−4270.

(20) Ko, T. W.; Finkler, J. A.; Goedecker, S.; Behler, J. A fourthgeneration high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nat. Commun.* **2021**, *12*, 398.

(21) Veit, M.; Wilkins, D. M.; Yang, Y.; DiStasio, R. A.; Ceriotti, M. Predicting molecular dipole moments by combining atomic partial charges and atomic dipoles. *J. Chem. Phys.* **2020**, *153*, No. 024113.

(22) Staacke, C. G.; Wengert, S.; Kunkel, C.; Csányi, G.; Reuter, K.; Margraf, J. T. Kernel charge equilibration: efficient and accurate prediction of molecular dipole moments with a machinelearning enhanced electron density model. *Mach. Learn.: Sci. Technol.* **2022**, *3*, 015032.

(23) Montes-Campos, H.; Carrete, J.; Bichelmaier, S.; Varela, L. M.; Madsen, G. K. H. A Differentiable Neural-Network Force Field for Ionic Liquids. *J. Chem. Inf. Model.* **2022**, *62*, 88−101. 34941253

(24) Staacke, C.; Heenen, H.; Scheurer, C.; Csányi, G.; Reuter, K.; Margraf, J. On the Role of Long-Range Electrostatics in Machine-Learned Interatomic Potentials for Complex Battery Materials. *ACS Appl. Energy Mater.* **2021**, *4*, 12562−12569.

(25) Ambrosetti, A.; Reilly, A. M.; DiStasio, R. A.; Tkatchenko, A. Long-range correlation energy calculated from coupled atomic response functions. *J. Chem. Phys.* **2014**, *140*, 18A508.

(26) Kapil, V.; Engel, E. A. A complete description of thermodynamic stabilities of molecular crystals. *Proc. Natl. Acad. Sci. U. S. A.* **2022**, *119*, No. e2111769119.

(27) Bartók, A. P.; Gillan, M. J.; Manby, F. R.; Csányi, G. Machine-learning approach for one- and two-body corrections to density functional theory: Applications to molecular and condensed water. *Phys. Rev. B* **2013**, *88*, 054104.

(28) McDonagh, D.; Skylaris, C.-K.; Day, G. M. Machine-Learned Fragment-Based Energies for Crystal Structure Prediction. *J. Chem. Theory Comput.* **2019**, *15*, 2743−2758.

(29) Wengert, S.; Csányi, G.; Reuter, K.; Margraf, J. T. Dataefficient machine learning for molecular crystal structure prediction. *Chem. Sci.* **2021**, *12*, 4536−4546.

(30) Aakeröy, C. B.; Grommet, A. B.; Desper, J. Co-crystal Screening of Diclofenac. *Pharmaceutics* **2011**, *3*, 601−614.

(31) Kazantsev, A. V.; Karamertzanis, P. G.; Adjiman, C. S.; Pantelides, C. C. Efficient Handling of Molecular Flexibility in Lattice Energy Minimization of Organic Crystals. *J. Chem. Theory Comput.* **2011**, *7*, 1998−2016.

(32) Iuzzolino, L.; McCabe, P.; Price, S.; Brandenburg, J. G. Crystal structure prediction of flexible pharmaceutical-like molecules: density functional tight-binding as an intermediate optimisation method and for free energy estimation. *Faraday Discuss.* **2018**, *211*, 275−296.

(33) Gaus, M.; Cui, Q.; Elstner, M. DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB). *J. Chem. Theory Comput.* **2011**, *7*, 931−948.

(34) Gaus, M.; Goez, A.; Elstner, M. Parametrization and Benchmark of DFTB3 for Organic Molecules. *J. Chem. Theory Comput.* **2013**, *9*, 338−354.

(35) Hourahine, B.; et al. DFTB, a software package for efficient approximate density functional theory based atomistic simulations. *J. Chem. Phys.* **2020**, *152*, 124101.

(36) Caldeweyher, E.; Ehlert, S.; Hansen, A.; Neugebauer, H.; Spicher, S.; Bannwarth, C.; Grimme, S. A generally applicable atomic-charge dependent London dispersion correction. *J. Chem. Phys.* **2019**, *150*, 154122.

(37) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.

(38) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. Gaussian Process Regression for Materials and Molecules. *Chem. Rev.* **2021**, *121* (16), 10073−10141.

(39) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.

(40) Adamo, C.; Cossi, M.; Barone, V. An accurate density functional method for the study of magnetic properties: the PBE0 model. *J. Mol. Struct.: THEOCHEM* **1999**, *493*, 145−157.

(41) Tkatchenko, A.; DiStasio, R. A.; Car, R.; Scheffler, M. Accurate and Efficient Method for Many-Body van der Waals Interactions. *Phys. Rev. Lett.* **2012**, *108*, 236402.

(42) Hoja, J.; Tkatchenko, A. First-principles stability ranking of molecular crystal polymorphs with the DFT+MBD approach. *Faraday Discuss.* **2018**, *211*, 253−274.

(43) Reilly, A. M.; Tkatchenko, A. Understanding the role of vibrations, exact exchange, and many-body van der Waals interactions in the cohesive properties of molecular crystals. *J. Chem. Phys.* **2013**, *139*, No. 024705.

(44) LeBlanc, L. M.; Dale, S. G.; Taylor, C. R.; Becke, A. D.; Day, G. M.; Johnson, E. R. Pervasive Delocalisation Error Causes Spurious Proton Transfer in Organic Acid−Base Co-Crystals. *Angew. Chem., Int. Ed.* **2018**, *57*, 14906−14910.

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

(46) Fredericks, S.; Parrish, K.; Sayre, D.; Zhu, Q. PyXtal: A Python library for crystal structure generation and symmetry analysis. *Comput. Phys. Commun.* **2021**, *261*, 107810.

(47) Ceriotti, M.; Willatt, M. J.; Csányi, G. Machine Learning of Atomic-Scale Properties Based on Physical Principles. *Handb. Mater. Model.* **2018**, 1−27.

(48) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754−13769.

(49) Timmermann, J.; Lee, Y.; Staacke, C. G.; Margraf, J. T.; Scheurer, C.; Reuter, K. Data-efficient iterative training of Gaussian approximation potentials: Application to surface structure determination of rutile IrO_2 and RuO_2. *J. Chem. Phys.* **2021**, *155*, 244107.

(50) Mortazavi, M.; Brandenburg, J. G.; Maurer, R. J.; Tkatchenko, A. Structure and Stability of Molecular Crystals with Many-Body Dispersion-Inclusive Density Functional Tight Binding. *J. Phys. Chem. Lett.* **2018**, *9*, 399−405.

(51) Chisholm, J. A.; Motherwell, S. *COMPACK*: a program for identifying crystal structure similarity using distances. *J. Appl.Crystallogr.* **2005**, *38*, 228−231.

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

(53) Tom, R.; Rose, T.; Bier, I.; O'Brien, H.; Vázquez-Mayagoitia, Á.; Marom, N. Genarris 2.0: A random structure generator for molecular crystals. *Comput. Phys. Commun.* **2020**, *250*, 107170.

(54) Song, H.; Vogt-Maranto, L.; Wiscons, R.; Matzger, A. J.; Tuckerman, M. E. Generating Cocrystal Polymorphs with Information Entropy Driven by Molecular Dynamics-Based Enhanced Sampling. *J. Phys. Chem. Lett.* **2020**, *11*, 9751−9758.

(55) Hoja, J.; Reilly, A. M.; Tkatchenko, A. First-principles modeling of molecular crystals: structures and stabilities, temperature and pressure. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2017**, *7*, No. e1294.

(56) Poier, P. P.; Lagardère, L.; Piquemal, J.-P. $O(N)$ Stochastic Evaluation of Many-Body van der Waals Energies in Large Complex Systems. *J. Chem. Theory Comput.* **2022**, *18*, 1633−1645.

(57) Poier, P. P.; Jaffrelot Inizan, T.; Adjoua, O.; Lagardère, L.; Piquemal, J.-P. Accurate Deep Learning-Aided Density-Free Strategy for Many-Body Dispersion-Corrected Density Functional Theory. *J. Phys. Chem. Lett.* **2022**, *13*, 4381−4388.

(58) Blum, V.; Gehrke, R.; Hanke, F.; Havu, P.; Havu, V.; Ren, X.; Reuter, K.; Scheffler, M. Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **2009**, *180*, 2175−2196.

(59) Aradi, B.; Hourahine, B.; Frauenheim, T. DFTB, a Sparse Matrix-Based Implementation of the DFTB Method. *J. Phys. Chem. A* **2007**, *111*, 5678−5684.

(60) Bartók, A. P.; Csányi, G. Gaussian approximation potentials: A brief tutorial introduction. *Int. J. Quantum Chem.* **2015**, *115*, 1051−1057.