A complete description of thermodynamic stabilities of molecular crystals

Venkat Kapil[a,b,1,2,3] and Edgar A. Engel[c,2,3]

[a]Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom; [b]Laboratory of Computational Science and Modeling, Institut des Matériaux, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; and [c]Theory of Condensed Matter Group, Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, United Kingdom

Predictions of relative stabilities of (competing) molecular crystals are of great technological relevance, most notably for the pharmaceutical industry. However, they present a long-standing challenge for modeling, as often minuscule free energy differences are sensitively affected by the description of electronic structure, the statistical mechanics of the nuclei and the cell, and thermal expansion. The importance of these effects has been individually established, but rigorous free energy calculations for *general* molecular compounds, which simultaneously account for all effects, have hitherto not been computationally viable. Here we present an efficient "end to end" framework that seamlessly combines state-of-the art electronic structure calculations, machine-learning potentials, and advanced free energy methods to calculate ab initio Gibbs free energies for general organic molecular materials. The facile generation of machine-learning potentials for a diverse set of polymorphic compounds—benzene, glycine, and succinic acid—and predictions of thermodynamic stabilities in qualitative and quantitative agreement with experiments highlight that predictive thermodynamic studies of industrially relevant molecular materials are no longer a daunting task.

statistical mechanics | machine learning | ab initio thermodynamics | polymorphism

Molecular crystals are ubiquitous in the pharmaceutical industry (1) and show great promise for applications in organic photovoltaics (2); gas adsorption (3); and the food, pesticide, and fertilizer industries (4). Their tendency to exhibit polymorphism, i.e., to exist in multiple crystal structures, on one hand provides a mechanism to tune properties by controlling crystal structure (5) and on the other hand introduces the challenge of synthesizing and stabilizing crystal structures with desired properties (6). While thermodynamic stability at the temperature and pressure of interest is sufficient (although not necessary) to ensure long-term stability, simply understanding thermodynamic stability already poses a formidable challenge.* This is particularly true for pharmaceuticals, where free energy differences between drug polymorphs are often smaller than 1 kJ/mol (7), leading to the risk of the drug transforming into a less soluble and consequently less effective form during manufacturing, storage, or shelf life (8, 9). Indeed, the problem of late-appearing drug polymorphs is widespread (10, 11).

The pharmaceutical industry therefore spends considerable resources on high-throughput crystallization experiments to screen for polymorphs (12), into which the target structure may decay. However, crystallization experiments do not probe thermodynamic stability, and conclusive studies of the impact of temperature changes after crystallization on the stability of polymorphs (i.e., their monotropic or enantiotropic nature) (13) are often prevented by limited sample quantities. Hence, there is the appeal of theoretical crystal structure prediction (CSP) (14) based on the thermodynamic stability, which promises to complement crystallization experiments (15) by exhaustively searching for competing polymorphs.

Despite the demonstrable value of CSP for many classes of materials (16–21), and the continuing progress evidenced by a series of blind tests (15), the success of CSP for molecular crystals has been limited by the inability to routinely predict the relative stability of competing candidate structures (22). This is largely because the methods used for stability rankings typically ignore or approximate the subtle interplay of several effects, such as intricate intermolecular interactions (23), the (quantum) statistical mechanics of the nuclei (24) and the unit cell (25), and thermal expansion (26), thereby incurring errors larger than the free energy differences of interest. The importance of each of these effects has been demonstrated in isolation, but predictive stability rankings must also comprehensively account for their interplay.

Recent implementations of advanced path-integral (PI) approaches (27, 28) allow exactly accounting for the quantum statistical mechanics of the nuclei and the unit cell (29, 30) for arbitrary potential energy surfaces (PESs). At the same time, modern machine-learning potentials (MLPs) (31) permit accurately reproducing ab initio PESs and dramatically reduce the cost of performing simulations approaching ab initio accuracy (32).

Significance

Predicting stable polymorphs of molecular crystals remains one of the grand challenges of computational science. Current methods invoke approximations to electronic structure and statistical mechanics and thus fail to consistently reproduce the delicate balance of physical effects determining thermodynamic stability. We compute the rigorous ab initio Gibbs free energies for competing polymorphs of paradigmatic compounds, using machine learning to mitigate costs. The accurate description of electronic structure and full treatment of quantum statistical mechanics allow us to predict the experimentally observed phase behavior. This constitutes a key step toward the first-principles design of functional materials for applications from photovoltaics to pharmaceuticals.

---

[1]Present address: Churchill College, University of Cambridge, Cambridge CB3 0DS, United Kingdom.

[2]V.K. and E.A.E contributed equally to this work.

[3]To whom correspondence may be addressed. Email: vk380@cam.ac.uk or eae32@cam.ac.uk.

---

*Kinetics may protect thermodynamically metastable structures from decaying almost indefinitely.

Despite these advances, calculations of rigorous thermodynamic stabilities for general molecular materials have been complicated by the absence of an integrated framework, which facilitates both the rapid development of MLPs and free energy calculations including all physically relevant effects, while ensuring universal applicability to diverse systems.

In this work we present an efficient framework for ranking candidate structures of arbitrary compounds using rigorous ab initio Gibbs free energy calculations, based on the streamlined development of MLPs and their integration with PI methods. Our approach builds upon our previous work on combining PI approaches with MLPs for ice polymorphs (29, 33), but greatly enhances its accuracy, efficiency, and robustness for out-of-the-box applications to general compounds. In particular, we simplify the development of MLPs using a straightforward and inexpensive protocol for compiling ab initio reference data, which is designed to work for general organic compounds and accounts for (the often-neglected) cell flexibility and quantum nuclear motion. Additionally, robust data-driven techniques minimize the human effort involved in training the MLPs. In contrast to previous CSP ranking methods that use MLPs (34, 35), we exactly account for the quantum statistical mechanics of the nuclei and the cell and use MLPs only as a stepping stone for computing ab initio Gibbs free energies, eliminating all dependence on the MLPs and their limitations.

The reliability and general applicability of our approach are showcased by the rapid development of MLPs and correct stability predictions for crystal polymorphs of three prototypical compounds: benzene, glycine, and succinic acid. These bear the hallmarks of more complex biomolecular systems—molecular flexibility, competing polymorphs, and intermolecular interactions ranging from weak dispersive to hydrogen bonded and ionic. Importantly, the relative stability of their polymorphs is well established (36–38). We further assess the temperature and pressure dependence of relative stabilities based on gradients of Gibbs free energies, which correspond to indicators widely used by experimentalists to predict the monotropic or enantiotropic nature of the polymorphs.

Our work complements state-of-the-art CSP methods, which efficiently survey structural space to extract small sets of promising candidate structures using ab initio calculations and/or MLPs (34, 35), but struggle to reliably resolve subtle differences in stability among them (22). Combining rigorous free energy calculations, as demonstrated here, with structure searching and inexpensive CSP ranking methods constitutes an avenue to predictive CSP for complex molecular crystals of industrial importance.

Computational Framework and Systems

To predict rigorous relative stabilities, we combine PI thermodynamic integration (29) (referred to as quantum thermodynamic integration [QTI]) in the constant pressure ensemble (thereby accounting for anharmonic quantum nuclear motion and the fluctuations and thermal expansion of the cell) with density-functional-theory (DFT) calculations with the hybrid PBE0 functional (39, 40) and the many-body dispersion (MBD) correction of Tkatchenko et al. (41) and Tkatchenko and coworkers (42) (referred to as PBE0-MBD). PBE0-MBD provides an accurate description of intermolecular interactions, as benchmarked using experimental and coupled cluster theory with singlets, doublets, and perturbative triplets [CCSD(T)] lattice energies for various molecular crystals, including form I of benzene and α-glycine (43, 44). Since direct calculation of Gibbs free energies using ab initio QTI is prevented by the cost of the required energy and force evaluations (29), ab initio Gibbs free energies are calculated in a four-step process, as depicted schematically in Fig. 1 and detailed further in *SI Appendix*.
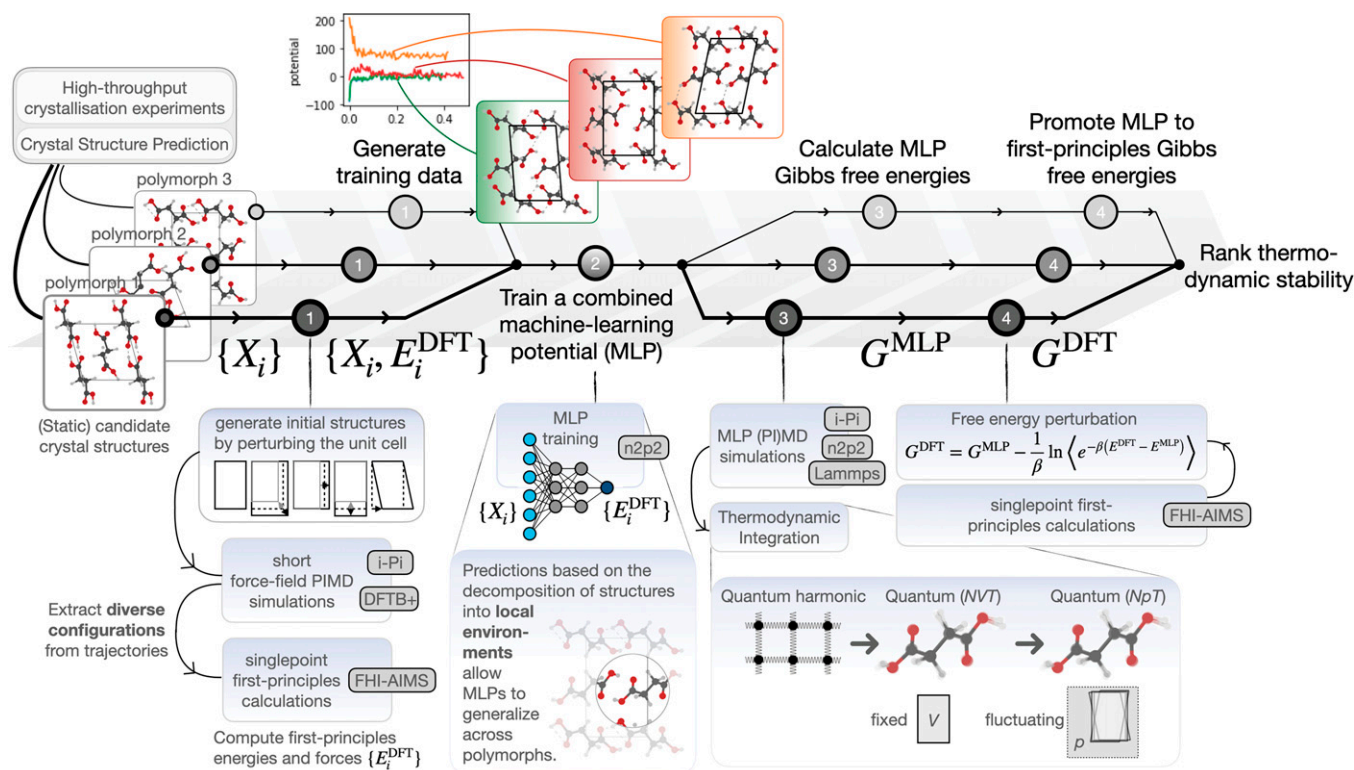

Fig. 1. Schematic representation of the workflow for computing ab initio, quantum anharmonic Gibbs free energies for candidate crystal structures. *Upper* section shows the main steps: 1) generating ab initio reference data on which to 2) train a combined MLP, which can then be used to 3) compute MLP Gibbs free energies, which one can finally 4) promote to ab initio Gibbs free energies. *Lower* section (shaded in blue) details the key aspects of how each of these steps is performed in practice.

First, we use a simple strategy to generate a minimal but exhaustive set of unit cell "training configurations," for which we then perform PBE0-MBD calculations: We perform PI simulations based on density-function tight binding (DFTB) (45) theory for unit cells with perturbed cell parameters. This allows us to gather a large number of configurations, which incorporate quantum nuclear fluctuations and cell flexibility and from which we can distill the most distinct ones using a data-driven approach (46). This strategy leverages the low cost of DFTB and its qualitative accuracy for diverse molecular crystals (47) to avoid the bottleneck that is PBE0-MBD–based configurational sampling. Due to the versatility of DFTB, it can be used to generate robust training data for almost any compound of interest.

The subsequent training of the MLPs hinges on identifying the most important "features" of the configurations, fed to the MLPs as input. These features are usually abstract functions quantifying the local density of atoms and require the careful tuning of multiple parameters (46). Here, we render training MLPs for general compounds accessible to nonexperts by automating this procedure using a "size-extensive" data-driven approach, which avoids the manual selection of features based on "prior experience." Combining these first two steps with a "tried and tested" neural network architecture (48–50) greatly simplifies and speeds up the generation of MLPs, while remaining agnostic to the system of study.

In a third step, we exploit the orders-of-magnitude lower cost of the resultant MLPs compared to the ab initio reference method, to compute Gibbs free energies for much larger simulation supercells using QTI (29). We account for anisotropic fluctuations of the simulation cell, which are important for flexible functional materials (51), and directly calculate the free energy difference between the harmonic reference systems and the physical, anharmonic system at the PI level, which substantially reduces the complexity and cost compared to the multistep integration performed in ref. 33. We note that the affordability of MLP free energies comes at the price of residual errors with respect to the ab initio reference values due to the imperfect reproduction of the reference PES. These may arise from the short-ranged nature of the MLPs (52), from information lost during the "featurization" of the configurations (53), or from insufficient training data. The typical errors in MLP predictions of configurational energies (Table 1) are small but comparable to the subtle free energy differences between polymorphs. Therefore, in a fourth and final step, we eliminate the associated errors to obtain true ab initio Gibbs free energies by computing the difference between the MLP and PBE0-MBD free energies using free energy perturbation (FEP) (33). All calculations and simulations are performed using readily available and well-documented software, and Jupyter notebooks for analysis are provided in *SI Appendix*.

As an exposé of the universal applicability of this scheme, we predict the relative stabilities of a set of prototypical systems, whose small number and size belie how representative they are of general organic molecular crystals: Benzene is the archetypal rigid, van der Waals bonded molecular crystal, while succinic acid represents general hydrogen-bonded systems, and glycine prototypes are flexible zwitter-ionic systems. This small, "irreducible" set of prototypical systems covers not only the three different types of bonding, but also the chemical space that includes pharmaceuticals such as aspirin and paracetamol. Moreover, molecular flexibility and the large-amplitude curvilinear motion of the amide group in glycine trigger the same pathologies of approximate free energy methods as more complex systems exhibiting free rotation of molecular units (24, 29) and serve as a stringent test for stability predictions.

For each compound we compute the free energy differences between the stable ambient-pressure polymorph and its closest experimentally established competitor(s): We consider forms I and II of benzene (36) and α- and β-succinic acid (37) at 100 K and α-, β-, and γ-glycine (54) at 300 K to compare with available calorimetric data (38, 55). The nearly orthorhombic simulation supercells shown in Fig. 2, which contain equivalent numbers of molecules for all polymorphs of the same compound, ensure near cancellation of center-of-mass free energies and suffice to converge stabilities with respect to finite-size effects to within 0.1 kJ/mol (*SI Appendix*).

Ab Initio Thermodynamic Stabilities

As shown in Fig. 3A, the final ab initio Gibbs free energies (shown in red) reproduce the greater stability of form I over form II of benzene and of β- over α-succinic acid, the metastability of β-glycine, and the near degeneracy of α- and γ-glycine (55). Moreover, our Gibbs free energy differences are in agreement with available calorimetry data (38, 55) to within statistical and experimental uncertainties.

The QTI approach also yields gradients of Gibbs free energies, including the molar volume, entropy, and heat capacity, which provide indication regarding pressure- and temperature-driven changes in relative stability and thus the monotropic or enantiotropic nature of compounds. For instance, since molar volumes are derivatives of the free energy with pressure, we can predict form II of benzene to become thermodynamically stable over the ambient pressure form I at 1.4 GPa (at 100 K), which is in good agreement with the experimentally determined transition pressure of 1.5 GPa (56). Similarly, we determine the entropy of β-succinic acid to be smaller than that of α-succinic acid, making the latter the preferred high-temperature polymorph, in agreement with the experimental phase behavior (57). While in the case of glycine we are able only to predict near degeneracy of α- and γ-glycine at ambient conditions, molar volumes suggest α-glycine to be the most stable phase at high pressures, which is in line with experiments showing that it remains stable up to 23 GPa (54).

By comparing rigorous free energies with estimates that exclude nuclear quantum effects (NQEs), anharmonicity, and cell expansion and flexibility, we are able to understand the extent to which these effects and their interplay contribute toward the stability of molecular crystals. Crucially, as shown in Fig. 3B, the size and sign of these effects depend entirely on the compound and the polymorphs at hand, highlighting that rigorous QTI is indispensable for predicting phase stabilities and that molecular crystals are typically stabilized by a nontrivial interplay of different physical effects, whose individual importance is belied by the subtle resultant free energy differences. For instance, the greater stability of form I of benzene hinges on an accurate description of the electronic structure, while NQEs and anharmonicity cancel out almost perfectly and thermal expansion affects both forms similarly. In contrast, in succinic acid NQEs and anharmonicity cooperatively stabilize the α form and thermal expansion differentiates the two polymorphs. In glycine NQEs and thermal expansion differently affect the stability of the α- and β-polymorphs with respect to the γ form, and neglecting any of the three effects would lead to large errors on the scale of the experimental free energy differences.

Meanwhile, the MLP-based stability predictions (shown in blue in Fig. 3A) are only limited by the accuracy, with which the
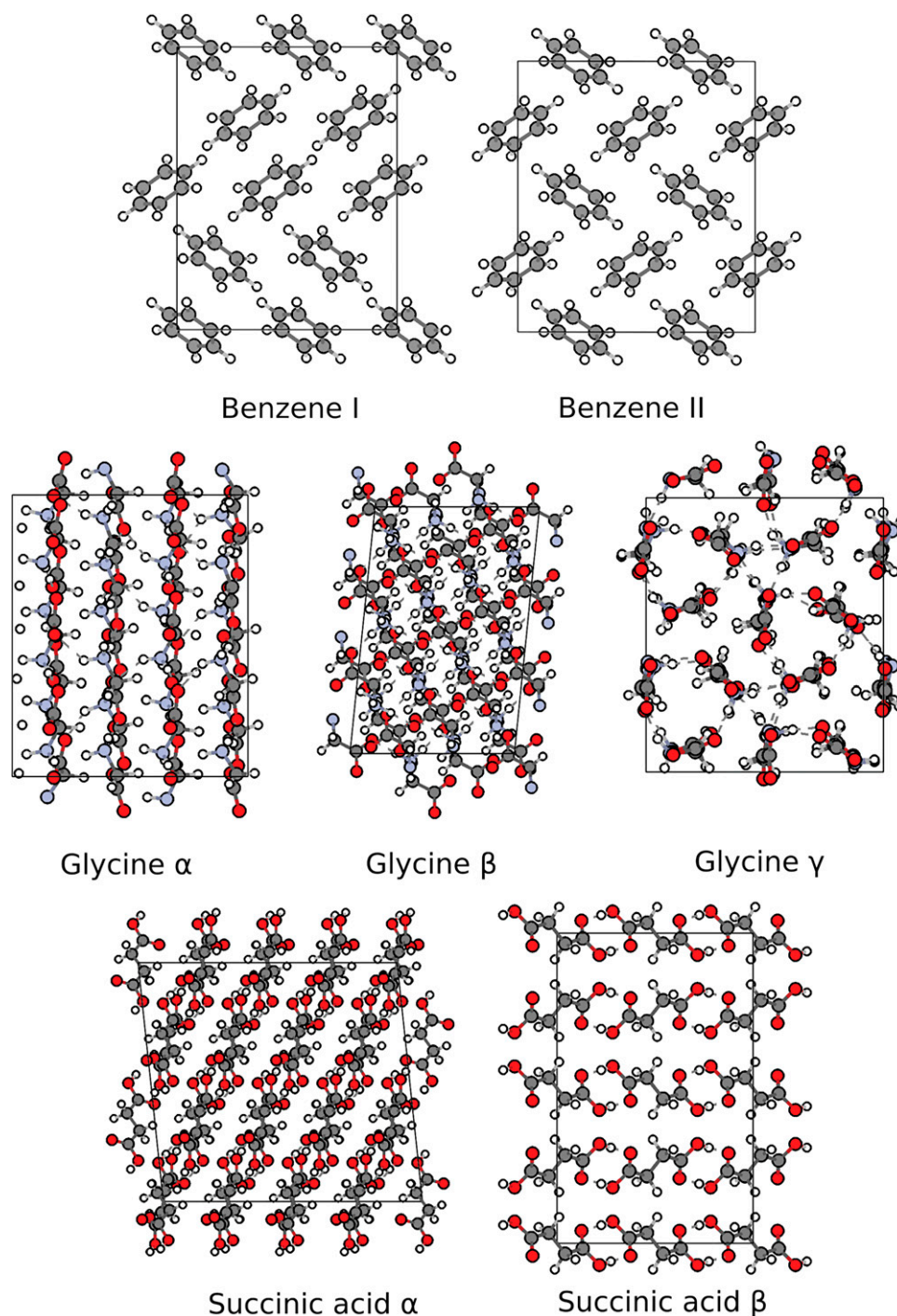
Table 1. Number of single-point PBE0-MBD calculations underlying each MLP and their respective root-mean-square errors (RMSE) in predicting energies on a separate test set of configurations from PI simulations of the experimental unit cells

System	Reference data	Energy RMSE, kJ/mol
Benzene	1,000	1.2
Glycine	4,000	1.6
Succinic acid	2,000	2.3

Fig. 2. Structures of forms I and II of benzene containing 16 molecules; forms α, β, and γ of glycine containing 24 molecules; and forms α and β of succinic acid containing 24 molecules. Hydrogen, carbon, nitrogen, and oxygen atoms are shown in white, gray, blue, and red, respectively.

MLPs reproduce the ab initio PES (Table 1), and consequently correctly reproduce the greater stability of form I of benzene compared to form II. At the same time, the incorrect MLP-based stability predictions for succinic acid and glycine highlight the critical importance of the final FEP step. Promoting MLP free energies to the ab initio level by FEP incurs only the cost of a few tens of ab initio energy and force evaluations for configurations sampled by the MLPs. We note that the cost of this step is comparable to that of common equation-of-state calculations and thus constitutes a reliable and computationally efficient means of predicting the relative stability of polymorphs.

Given that errors of 1 kJ/mol are often considered to be within "chemical accuracy," it is worth emphasizing that the compounds considered here are not hand-picked, "pathological" examples, but expected to be representative of many biomolecular compounds. The small free energy differences between polymorphs, which are smaller than $k_B T$ but can be resolved experimentally (38, 55) due to the kinetic suppression of interconversion between polymorphs, constitute a very stringent test of our framework and its ability to accurately capture phase stability. By matching the subkilojoule per mole accuracy of calorimetry experiments, it provides a robust foundation for studying
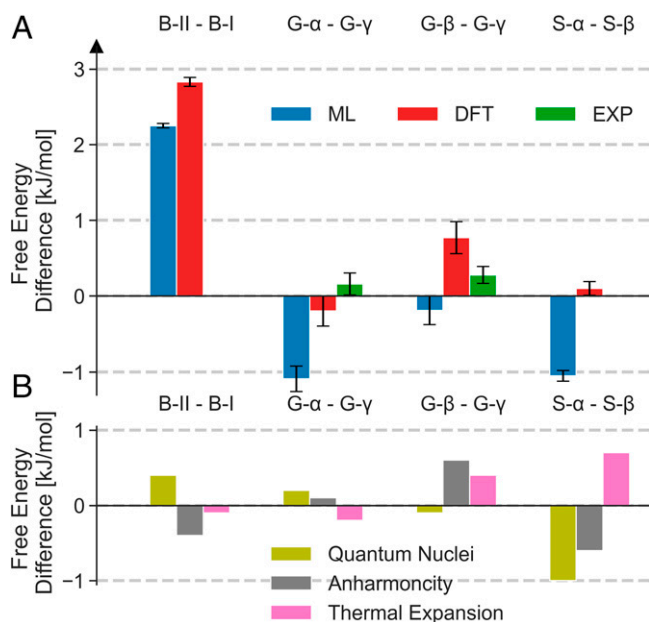
Fig. 3. (*A*) PI Gibbs free energy differences between forms II and I of benzene (B-II and B-I); α-, β-, and γ-glycine (G-α, G-β, and G-γ); and α- and β-succinic acid (S-α and S-β) calculated using PBE0-MBD–based MLPs (blue) with the QTI approach and corrected to the ab initio PBE0-MBD DFT level using free energy perturbation (red). Experimental data (38, 55) are shown in green. (*B*) Contributions of quantum nuclei (olive), anharmonicity (gray), and cell expansion and flexibility (pink) to the relative stabilities of the said polymorphs. These have been respectively obtained by comparing Gibbs free energy differences to estimates from a classical thermodynamic integration, a harmonic approximation, and a quantum thermodynamic integration using a fixed 0-K optimized cell.

transition temperatures, pressures, and rates and permits benchmarking sophisticated electronic structure theories against experiment.

Comparison with Approximate Approaches

To further highlight the advantages of the approach proposed here over established approximate methods for ranking stabilities in CSP, we assess the limitations of the most widely used approximate methods, prefaced by acknowledging their successes for a wide range of applications (58). We note that the MLPs reproduce the ab initio PESs with sufficient accuracy to assess the impact of approximations to nuclear motion and compare the respective approximate (free) energy differences between polymorphs to the corresponding exact MLP Gibbs free energies.

The current state of the art is to correct (free) energies on the basis of a single-point hybrid-functional DFT calculation for the structure relaxed using semilocal DFT (57). Thermal and quantum nuclear effects are included within a harmonic approximation (HA) (59), while thermal expansion is modeled by relaxing the cell within a quasi-harmonic approximation (QHA) (58). These corrections are generally computed at the semilocal DFT level. As shown in Fig. 4, these approaches neither universally predict the most stable form (as they exhibit errors larger than 1 kJ/mol) nor systematically converge to the full hybrid-functional QHA reference. This highlights the need to go beyond a single-point hybrid-functional DFT correction to semilocal configurational or (quasi-)harmonic free energies to consistently deliver correct stability orders and free energy differences with subkilojoule per mole accuracy. We further note that the above results benefit substantially from the fortuitous cancellation of errors (29), but the residual errors cannot be estimated and apparent physical insights may be misleading.

Since the hybrid-functional–based QHA seems to be competitive with the rigorous PI approach, is further worthwhile to put the cost of the calculations into perspective. For glycine, as the most costly example, the 4,000 PBE0-MBD calculations on unit cells constituting the reference data for the MLP, the MLP-based PI thermodynamic integration, and the 50 PBE0-MBD calculations on supercells required for the FEP contribute roughly equally to the total cost of around 148,000 core hours per polymorph. For comparison, computing PBE0-MBD HA free energies for the same simulation supercells using finite differences and nondiagonal supercells to probe individual *k*-points (60), but not leveraging the MLP, would require about three times the core hours. A PBE0-MBD QHA free energy calculation would be an order of magnitude higher in computational cost. Although (Q)HA free energies may also be computed inexpensively using MLPs, they cannot be promoted to their first-principles counterparts in a straightforward and cost-effective manner as exact MLP free energies. Despite a focus on universal applicability over efficiency, the cost of the above rigorous Gibbs free energies is thus small compared to the estimated cost of calculating free energies within the (Q)HA using hybrid-functional DFT.

Discussion

The ability of our approach to predict free energy differences with subkilojoule per mole accuracy renders it valuable in identifying "competing" polymorphs with similar lifetimes to the most stable form. It bridges the gap between theory and experiments by allowing direct comparison of free energy differences with calorimetric data—a significant improvement over current approaches, which require error-prone ad hoc extrapolations to
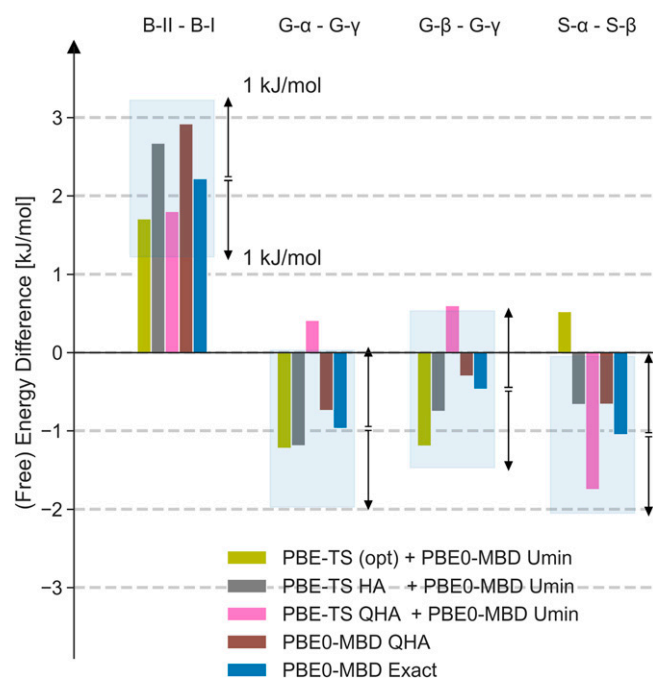

Fig. 4. MLP (free) energy differences between forms II and I of benzene (B-II and B-I); α-, β-, and γ-glycine (G-α, G-β, and G-γ); and α- and β-succinic acid (S-α and S-β) at different tiers of accuracy: fixed-cell optimization using PBE-TS with a PBE0-MBD single-point energy (green), fixed-cell optimization and harmonic free energy using PBE-TS with a PBE0-MBD single-point energy (gray), quasi-harmonic approximation (QHA) free energy using PBE-TS with a single-point PBE0-MBD correction (pink), full PBE0-MBD–based QHA (brown), and the exact PI free energy difference (blue). The shaded region indicates free energy differences within 1 kJ/mol of the respective exact PI result as a guide to the eye.

0 K (61). Moreover, 1) rigorous predictions of the entropy, molar volume, and heat capacity and 2) robust MLPs with ab initio accuracy are complements to our approach. The former are directly related to "thermodynamic rules of thumb," which are widely used by experimentalists to assess stability trends (13), while the latter enable structure determinations for experimental samples based on NMR (62) and vibrational spectra (63). Furthermore, a rigorous account of thermally induced phase transitions can be obtained without repeating the procedure at every state point. The combination of QTI with parallel tempering (64) can enhance the efficiency of performing single-temperature (or pressure) sweeps, yielding full phase diagrams, while also sampling "slow" degrees of freedom such as conformational transitions, inaccessible to approximate methods (29). All these features are highly sought after by the pharmaceutical industry, as they are made possible at a manageable computational cost.

Our protocol easily extends to stability predictions for other complex molecular crystals, as its data-driven nature accelerates MLP development, irrespective of the material under consideration. As a proof of this, we have developed MLPs for polymorphs of three complex pharmaceuticals—aspirin, paracetamol, and XXIII, the most complicated system (58) from the latest blind test of organic crystal structure prediction methods (22)—and tested them by performing PI simulations in the constant pressure ensemble, as required for QTI (*SI Appendix*). Although these MLPs have been trained on DFTB data as a proof of concept and consequently lack chemical accuracy, they remain robust and capture the molecular flexibility of these systems. Given that dynamic disorder, thermal expansion, conformational relaxation of the molecular units, and potential (dynamic) instabilities of candidate polymorphs are automatically accounted for within the QTI approach, we expect stability predictions to be very robust with respect to the nature of the candidate polymorphs and thus directly applicable to said pharmaceutical and blind-test systems.

In applications involving large numbers of polymorphs or polymorphs with large unit cells, suitable sets of reference configurations can be generated based on configurations of liquid or amorphous states at different pressures (65). This exploits that the accuracy of MLPs, which predict energies and forces on the basis of local contributions, rests on having reference data for all distinct local atomic environments (65), rather than for all polymorphs of interest. The computational cost of building the training set then remains largely independent of number and unit cell size of the polymorphs of interest. For large numbers of polymorphs the cost per polymorph thus effectively reduces to that of the MLP-based thermodynamic integration and of FEP. In practice, the computational cost of FEP can be reduced by running only as many ab initio calculations as required to reduce the statistical error to below the predicted free energy differences between polymorphs. For instance, fewer than a handful of PBE0-MBD calculations would have sufficed to conclusively establish that form I of benzene is more stable than form II. Indeed, subject to estimates of the uncertainty of the MLP predictions (66, 67), it may be possible to omit FEP altogether. Recent work on the use of higher body-order correlations in atomistic representations (68) and on including long-ranged interactions (52) promises to enable subkilojoule per mole accuracy, eliminating the need for FEP even in applications involving subtle free energy differences.

Finally, the empiricism involved in selecting the exchange-correlation functional and dispersion correction used in the DFT calculations can be removed by using PESs evaluated using beyond-DFT electronic structure theory. Crucially, our scheme extends naturally to predictions of Gibbs free energies based on quantum-chemical electronic structure methods (61, 69) such as second-order Møller-Plesset perturbation theory, random-phase approximation, coupled cluster, or quantum Monte Carlo, some of which are systematically improvable

and can thereby be rendered truly ab initio (70). While these come at an increased computational cost per calculation, recent developments in machine learning for materials science (71) promise to minimize the number of quantum-chemical calculations required to train accurate MLPs and thus to keep the overall costs in check. Indeed, recent work demonstrates the corresponding construction of robust and accurate MLPs for CCSD(T) reference data (70).

In conclusion, marrying state-of-the-art electronic structure, free energy, and machine-learning methods in a widely applicable framework enables rigorous, predictive free energy calculations for complex (organic) molecular crystals at general thermodynamic conditions. The unprecedented accuracy of our approach sets the stage for future studies of kinetic effects as well as full *p-T* phase diagrams in a reliable and computationally efficient manner, paving the way for guiding experimental synthesis of such materials. The protocol and the scripts provided in *SI Appendix* permit its application practically out of the box. Determining the relative stability of generic polymorphic compounds is a recurrent problem across different domains of science and engineering—from nucleation theory to the practical design of pharmaceuticals—and we hope that the robust and easy-to-use nature of our end-to-end protocol will facilitate reliable, accurate free energy calculations beyond those of the computational chemistry community.

Methods

Machine-Learning Potentials. We have constructed Behler–Parinello-type neural network potentials (48) for benzene, glycine, and succinic acid using the n2p2 code (72). In this framework, structures are encoded in terms of local atom-centered symmetry functions (SFs) (48). Initial sets of SFs were generated following the recipe of ref. 73. Based on the same reference structure-property data subsequently used for training, the 128 (benzene and succinic acid) and 256 (glycine) most informative SFs were extracted via principal covariates CUR selection (74).

Our data are based on Langevin-thermostated PI NVT simulations at 300 K, performed using the i-Pi force engine (28) coupled to DFTB+ (75) calculations with the 3ob parameterization (76). For each polymorph multiple cells were simulated, rescaling the experimental cell lengths and angles by up to 10 and 5%, respectively. The trajectories of PI replicas for all polymorphs of a given compound were concatenated and farthest-point sampled (77–79) to extract the most distinct configurations for feature selection and MLP training. Subsequently, ab initio reference energies and forces were evaluated for said configurations.

To minimize the computational cost of the reference calculations the MLPs are composed of a baseline potential trained to reproduce energies and forces from more affordable PBE-DFT (80) calculations with a Tkatchenko-Scheffler (TS) dispersion correction (81) (PBE-TS) and a △-learning (82) correction trained (on 10 times fewer training data) to reproduce the difference between the baseline and more expensive calculations with the hybrid PBE0 functional (39, 40) and the MBD dispersion correction (41, 42) (PBE0-MBD). For a separate test set, the MLPs reproduce PBE0-MBD energies with root-mean-square errors of 1.2 kJ/mol for benzene, 1.6 kJ/mol for glycine, and 2.3 kJ/mol for succinic acid, respectively.

Ab Initio DFT Calculations. PBE0+MBD calculations were performed using FHI-aims (83–85) with the standard FHI-aims "intermediate" basis sets and a Monkhorst–Pack k-point grid (86) with a maximum spacing of $0.06 \times 2\pi$ Å^{-1}. The PBE-TS baseline calculations for a △-learning approach were performed using Quantum Espresso v6.3, the same k-point grid, a wavefunction cutoff energy of 100 Rydberg, and the optimized, norm-conserving Vanderbilt pseudopotentials from ref. 87.

Free Energy Methods. For each polymorph the average cell was determined using MLP-based PI NST simulations (88) at the desired inverse temperature β, accounting for anharmonic quantum nuclear motion and anisotropic cell fluctuations. The difference between the Gibbs and Helmholtz free energies computed from an MLP-based PI NPT simulation based on its average cell is

$$G_{\mathrm{MLP}}(P^{\mathrm{ext}}, \beta) - A_{\mathrm{MLP}}(V, \beta) = P^{\mathrm{ext}}V + \beta^{-1} \ln \rho(V|P^{\mathrm{ext}}, \beta), \qquad [1]$$

where $\rho(V|P^{\mathrm{ext}}, \beta)$ is the probability of observing the cell volume V at external pressure P^{ext} and inverse temperature β. A standard Kirkwood construction (89) that transforms the Hamiltonian from a harmonic to an

anharmonic one provides the difference between the anharmonic and the harmonic quantum Helmholtz free energies:

$$A_{MLP}(V, \beta) - A_{MLP}^{har}(V, \beta) = \int_0^1 d\lambda \left\langle \hat{H}_{MLP} - \hat{H}_{MLP}^{har} \right\rangle_{V, \beta, \hat{H}_\lambda}, \quad [2]$$

where $\hat{H}_\lambda$ is the Hamiltonian of the MLP alchemical system with the potential $U_\lambda \equiv \lambda U_{MLP} + (1 - \lambda)U_{MLP}^{har}$, and $\langle \cdot \rangle$ is the ensemble average computed from a PI NVT simulation. The reference absolute harmonic Helmholtz free energy is obtained from a harmonic approximation using

$$A_{MLP}^{har}(V, \beta) = U_{MLP}(V) + \sum_i \left[\frac{1}{2}\hbar\omega_i + \beta^{-1} \ln \left(1 - e^{-\beta\hbar\omega_i} \right) \right], \quad [3]$$

where ω_i is the frequency of the ith phonon mode. In a final step, the ab initio Gibbs free energy is obtained from its MLP counterpart by free energy perturbation using

$$G(P^{ext}, \beta) - G_{MLP}(P^{ext}, \beta) = -\beta^{-1} \ln \left\langle e^{-\beta(U-U_{MLP})} \right\rangle_{P^{ext}, \beta, \hat{H}_{MLP}}.$$

For systems exhibiting large-amplitude curvilinear motion, the harmonic-to-anharmonic thermodynamic integration can be performed efficiently using a Padé interpolation formula (24).

Understanding the Role of Different Effects. We disentangle the role of anharmonicity directly from Eq. **2** and that of thermal expansion by comparing the Helmholtz free energies from Eq. **2** for the variable-cell geometry-optimized and mean PI NST cells. The role of the quantum nature of nuclei is quantified by comparing the classical and quantum Gibbs free energies. We calculate the former using the Helmholtz free energy of the classical harmonic oscillator as a reference and evaluating Eqs. **1** and **2** using classical molecular dynamics.

Free Energy Gradients. Volume and entropy are related to gradients of the free energy

$$V = \left(\frac{\partial G}{\partial P} \right)_{N,T}, \quad S = -\left(\frac{\partial G}{\partial T} \right)_{N,P}. \quad [4]$$

Differences between equilibrium (molar) volumes of polymorphs can directly be observed in PI NPT simulations. Meanwhile entropic differences can be computed from

$$S = -\frac{1}{T}(G - H), \quad [5]$$

with G from Eq. **1** and the enthalpy H from the associated PI NPT simulation. Linear extrapolation then permits estimating whether and at which pressures P_c and temperatures T_c the Gibbs free energy difference between polymorphs will vanish and a phase transition should be expected:

$$\Delta G = -(T_c - T)\Delta S + (P_c - P)\Delta V. \quad [6]$$

1. S. Datta, D. J. W. Grant, Crystal structures of drugs: Advances in determination, prediction and engineering. *Nat. Rev. Drug Discov.* **3**, 42–57 (2004).
2. S. R. Forrest, The path to ubiquitous and low-cost organic electronic appliances on plastic. *Nature* **428**, 911–918 (2004).
3. T. Tozawa *et al.*, Porous organic cages. *Nat. Mater.* **8**, 973–978 (2009).
4. K. Honer, E. Kalfaoglu, C. Pico, J. McCann, J. Baltrusaitis, Mechanosynthesis of magnesium and calcium salt–urea ionic cocrystal fertilizer materials for improved nitrogen management. *ACS Sustain. Chem. Eng.* **5**, 8546–8550 (2017).
5. D. Gentili, M. Gazzano, M. Melucci, D. Jones, M. Cavallini, Polymorphism as an additional functionality of materials for technological applications at surfaces and interfaces. *Chem. Soc. Rev.* **48**, 2502–2517 (2019).
6. D. Chistyakov, G. Sergeev, The polymorphism of drugs: New approaches to the synthesis of nanostructured polymorphs. *Pharmaceutics* **12**, 34 (2020).
7. A. J. Cruz-Cabeza, S. M. Reutzel-Edens, J. Bernstein, Facts and fictions about polymorphism. *Chem. Soc. Rev.* **44**, 8619–8635 (2015).
8. S. R. Chemburkar *et al.*, Dealing with the impact of ritonavir polymorphs on the late stages of bulk drug process development. *Org. Process Res. Dev.* **4**, 413–417 (2000).
9. K. Ray Chaudhuri, Crystallisation within transdermal rotigotine patch: Is there cause for concern? *Expert Opin. Drug Deliv.* **5**, 1169–1171 (2008).
10. I. B. Rietveld, R. Céolin, Rotigotine: Unexpected polymorphism with predictable overall monotropic behavior. *J. Pharm. Sci.* **104**, 4117–4122 (2015).
11. D.-K. Bučar, R. W. Lancaster, J. Bernstein, Disappearing polymorphs revisited. *Angew. Chem. Int. Ed. Engl.* **54**, 6972–6993 (2015).
12. S. L. Morissette *et al.*, High-throughput crystallization: Polymorphs, salts, co-crystals and solvates of pharmaceutical solids. *Adv. Drug Deliv. Rev.* **56**, 275–300 (2004).
13. E. H. Lee, A practical guide to pharmaceutical polymorph screening & selection. *Asian J. Pharm. Sci.* **9**, 163–175 (2014).
14. S. L. Price, Predicting crystal structures of organic compounds. *Chem. Soc. Rev.* **43**, 2098–2111 (2014).
15. J. Nyman, S. M. Reutzel-Edens, Crystal structure prediction is changing from basic science to applied technology. *Faraday Discuss.* **211**, 459–476 (2018).
16. C. J. Pickard, J. Needs, Structure of phase III of solid hydrogen. *Nat. Phys.* **3**, 473–476 (2007).
17. Y. Ma *et al.*, Transparent dense sodium. *Nature* **458**, 182–185 (2009).
18. I. Errea *et al.*, High-pressure hydrogen sulfide from first principles: A strongly anharmonic phonon-mediated superconductor. *Phys. Rev. Lett.* **114**, 157004 (2015).
19. H. L. Zhuang, R. G. Hennig, Computational discovery, characterization and design of single-layer materials. *J. Miner. Met. Mater. Soc.* **66**, 366–374 (2014).
20. A. Marrazzo, M. Gibertini, D. Campi, N. Mounet, N. Marzari, Prediction of a large-gap and switchable Kane–Mele quantum spin Hall insulator. *Phys. Rev. Lett.* **120**, 117701 (2018).
21. B. D. Conduit, N. G. Jones, H. J. Stone, G. J. Conduit, Design of a nickel-base superalloy using a neural network. *Mater. Des.* **131**, 358–365 (2017).
22. A. M. Reilly *et al.*, Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater.* **72**, 439–459 (2016).
23. N. Marom *et al.*, Many-body dispersion interactions in molecular crystal polymorphism. *Angew. Chem. Int. Ed. Engl.* **52**, 6629–6632 (2013).
24. M. Rossi, P. Gasparotto, M. Ceriotti, Anharmonic and quantum fluctuations in molecular crystals: A first-principles study of the stability of paracetamol. *Phys. Rev. Lett.* **117**, 115702 (2016).
25. E. Schneider, L. Vogt, M. E. Tuckerman, Exploring polymorphism of benzene and naphthalene with free energy based enhanced molecular dynamics. *Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater.* **72**, 542–550 (2016).
26. H.-Y. Ko, R. A. DiStasio, B. Santra, R. Car, Thermal expansion in dispersion-bound molecular crystals. *Phys. Rev. Mater.* **2**, 055603 (2018).
27. T. E. Markland, M. Ceriotti, Nuclear quantum effects enter the mainstream. *Nat. Rev. Chem.* **2**, 1–14 (2018).
28. V. Kapil *et al.*, i-PI 2.0: A universal force engine for advanced molecular simulations. *Comput. Phys. Commun.* **236**, 214–223 (2019).
29. V. Kapil, E. Engel, M. Rossi, M. Ceriotti, Assessment of approximate methods for anharmonic free energies. *J. Chem. Theory Comput.* **15**, 5845–5857 (2019).
30. V. Kapil *et al.*, Modeling the structural and thermal properties of loaded metal–organic frameworks. An interplay of quantum and anharmonic fluctuations. *J. Chem. Theory Comput.* **15**, 3237–3249 (2019).
31. V. L. Deringer, M. A. Caro, G. Csányi, Machine learning interatomic potentials as emerging tools for materials science. *Adv. Mater.* **31**, e1902765 (2019).
32. J. Lan *et al.*, Simulating the ghost: Quantum dynamics of the solvated electron. *Nat. Commun.* **12**, 766 (2021). *Nat. Commun.* **12**.
33. B. Cheng, E. A. Edgar, J. Behler, C. Dellago, M. Ceriotti, Ab initio thermodynamics of liquid and solid water. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 1110–1115 (2019).
34. S. Wengert, G. Csányi, K. Reuter, J. T. Margraf, Data-efficient machine learning for molecular crystal structure prediction. *Chem. Sci.* **12**, 4536–4546 (2021).
35. D. McDonagh, C.-K. Skylaris, G. M. Day, Machine-learned fragment-based energies for crystal structure prediction. *J. Chem. Theory Comput.* **15**, 2743–2758 (2019).
36. A. Katrusiak, M. Podsiadło, A. Budzianowski, 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).
37. J.-L. Leviel, G. Auvert, J.-M. Savariault, Succinic acid (neutron study, at 77degK) $C_4H_8O_4$. *Acta Crystallogr. B* **37**, 2185 (1981).
38. G. L. Perlovich, L. K. Hansen, A. Bauer-Brandl, The polymorphism of glycine. Thermochemical and structural aspects. *J. Therm. Anal. Calorim.* **66**, 699–715 (2001).
39. J. P. Perdew, M. Ernzerhof, K. Burke, Rationale for mixing exact exchange with density functional approximations. *J. Chem. Phys.* **105**, 9982–9985 (1996).
40. C. Adamo, V. Barone, Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **110**, 6158–6170 (1999).
41. A. Tkatchenko, R. A. DiStasio Jr., R. Car, M. Scheffler, Accurate and efficient method for many-body van der Waals interactions. *Phys. Rev. Lett.* **108**, 236402 (2012).
42. A. Ambrosetti, A. M. Reilly, R. A. Di Stasio Jr., A. Tkatchenko. Long-range correlation energy calculated from coupled atomic response functions. *J. Chem. Phys.* **140**, 018A508 (2014).
43. A. M. Reilly, A. Tkatchenko, Understanding the role of vibrations, exact exchange, and many-body van der Waals interactions in the cohesive properties of molecular crystals. *J. Chem. Phys.* **139**, 024705 (2013).
44. G. J. O. Beran, Modeling polymorphic molecular crystals with electronic structure theory. *Chem. Rev.* **116**, 5567–5613 (2016).
45. M. Elstner *et al.*, Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B Condens. Matter Mater. Phys.* **58**, 7260 (1998).
46. G. Imbalzano *et al.*, Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *J. Chem. Phys.* **148**, 241730 (2018).

CHEMISTRY

47. J. G. Brandenburg, S. Grimme, Accurate modeling of organic molecular crystals by dispersion-corrected density functional tight binding (DFTB). *J. Phys. Chem. Lett.* **5**, 1785–1789 (2014).

48. J. Behler, M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).

49. J. Behler, Neural network potential-energy surfaces in chemistry: A tool for large-scale simulations. *Phys. Chem. Chem. Phys.* **13**, 17930–17955 (2011).

50. V. Kapil, J. Behler, M. Ceriotti, High order path integrals made easy. *J. Chem. Phys.* **145**, 234103 (2016).

51. A. van der Lee, D. G. Dumitrescu, Thermal expansion properties of organic crystals: A CSD study. *Chem. Sci. (Camb.)* **12**, 8537–8547 (2021).

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

53. S. N. Pozdnyakov et al., Incompleteness of atomic structure representations. *Phys. Rev. Lett.* **125**, 166001 (2020).

54. A. Dawson et al., Effect of high pressure on the crystal structures of polymorphs of glycine. *Cryst. Growth Des.* **5**, 1415–1427 (2005).

55. V. A. Drebushchak, Y. A. Kovalevskaya, I. E. Paukov, E. V. Boldyreva, Low-temperature heat capacity of α and γ polymorphs of glycine. *J. Therm. Anal. Calorim.* **74**, 109–120 (2003).

56. K. H. Yurtseven, M. G. Senol, Temperature and pressure dependence of the linewidth for an internal mode in the solid phases of benzene. *Acta Phys. Pol. A* **124**, 698 (2013).

57. P. Lucaioli et al., Serendipitous isolation of a disappearing conformational polymorph of succinic acid challenges computational polymorph prediction. *CrystEngComm* **20**, 3971–3977 (2018).

58. J. Hoja et al., Reliable and practical computational description of molecular crystal polymorphs. *Sci. Adv.* **5**, eaau3338 (2019).

59. A. M. Reilly, A. Tkatchenko, Role of dispersion interactions in the polymorphism and entropic stabilization of the aspirin crystal. *Phys. Rev. Lett.* **113**, 055701 (2014).

60. J. H. Lloyd-Williams, B. Monserrat, Lattice dynamics and electron-phonon coupling calculations using nondiagonal supercells. *Phys. Rev. B Condens. Matter Mater. Phys.* **92**, 184301 (2015).

61. A. Zen et al., Fast and accurate quantum Monte Carlo for molecular crystals. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 1724–1729 (2018).

62. E. A. Engel, V. Kapil, M. Ceriotti, Importance of nuclear quantum effects for NMR crystallography. *J. Phys. Chem. Lett.* **12**, 7701–7707 (2021).

63. S. Shepherd, J. Lan, D. M. Wilkins, V. Kapil. Efficient quantum vibrational spectroscopy of water with high-order path integrals: From bulk to interfaces. *J. Phys. Chem. Lett.* **12**, 9108–9114 (2021).

64. Y. Sugita, Y. Okamoto, Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141–151 (1999).

65. B. Monserrat, J. G. Brandenburg, E. A. Engel, B. Cheng, Liquid water contains the building blocks of diverse ice phases. *Nat. Commun.* **11**, 5757 (2020).

66. F. Musil, M. J. Willatt, M. A. Langovoy, M. Ceriotti, Fast and accurate uncertainty estimation in chemical machine learning. *J. Chem. Theory Comput.* **15**, 906–915 (2019).

67. G. Imbalzano et al., Uncertainty estimation for molecular dynamics and sampling. *J. Chem. Phys.* **154**, 074102 (2021).

68. J. Nigam, S. Pozdnyakov, M. Ceriotti, Recursive evaluation and iterative contraction of N-body equivariant features. *J. Chem. Phys.* **153**, 121101 (2020).

69. C. Červinka, G. J. O. Beran, Ab initio prediction of the polymorph phase diagram for crystalline methanol. *Chem. Sci.* **9**, 4622–4629 (2018).

70. S. Chmiela, H. E. Sauceda, K.-R. Müller, A. Tkatchenko, Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **9**, 3887 (2018).

71. C. Schran, K. Brezina, O. Marsalek, Committee neural network potentials control generalization errors and enable active learning. *J. Chem. Phys.* **153**, 104105 (2020).

72. A. Singraber, J. Behler, C. Dellago, Library-based LAMMPS implementation of high-dimensional neural network potentials. *J. Chem. Theory Comput.* **15**, 1827–1840 (2019).

73. G. Imbalzano et al., Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *J. Chem. Phys.* **148**, 241730 (2018).

74. R. K. Cersonsky, B. Helfrecht, E. A. Engel, S. Kliavinek, M. Ceriotti, Improving sample and feature selection with principal covariates regression. *Mach. Learn. Sci. Technol.* **2**, 035038 (2021).

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

76. T. Krüger, M. Elstner, P. Schiffels, T. Frauenheim, Validation of the density-functional based tight-binding approximation method for the calculation of reaction energies and other data. *J. Chem. Phys.* **122**, 114110 (2005).

77. Y. Eldar, M. Lindenbaum, M. Porat, Y. Y. Zeevi, The farthest point strategy for progressive image sampling. *IEEE Trans. Image Process.* **6**, 1305–1315 (1997).

78. M. Ceriotti, G. A. Tribello, M. Parrinello, Demonstrating the transferability and the descriptive power of sketch-map. *J. Chem. Theory Comput.* **9**, 1521–1532 (2013).

79. R. J. G. B. Campello, D. Moulavi, A. Zimek, J. Sander, Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Trans. Knowl. Discov. Data* **10**, 1–51 (2015).

80. J. P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).

81. A. Tkatchenko, M. Scheffler, Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.* **102**, 073005 (2009).

82. R. Ramakrishnan, P. O. Dral, M. Rupp, O. Anatole von Lilienfeld, Big data meets quantum chemistry approximations: The Δ-machine learning approach. *J. Chem. Theory Comput.* **11**, 2087–2096 (2015).

83. V. Blum et al., Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **180**, 2175–2196 (2009).

84. X. Ren et al., Resolution-of-identity approach to Hartree–Fock, hybrid density functionals, RPA, MP2, and GW with numeric atom-centered orbital basis functions. *New J. Phys.* **14**, 053020 (2012).

85. S. Levchenko et al., Hybrid functionals for large periodic systems in an all-electron, numeric atom-centered basis framework. *Comput. Phys. Commun.* **192**, 60–69 (2015).

86. H. J. Monkhorst, J. D. Pack, Special points for Brillouin-zone integrations. *Phys. Rev. B* **13**, 5188 (1976).

87. M. Schlipf, F. Gygi, Optimization algorithm for the generation of ONCV pseudopotentials. *Comput. Phys. Commun.* **196**, 36–44 (2015).

88. P. Raiteri, J. D. Gale, G. Bussi, Reactive force field simulation of proton diffusion in $BaZrO_3$ using an empirical valence bond approach. *J. Phys. Condens. Matter.* **23**, 334213 (2011).

89. D. Frenkel, A. J. C. Ladd, New Monte Carlo method to compute the free energy of arbitrary solids. Application to the fcc and hcp phases of hard spheres. *J. Chem. Phys.* **81**, 3188–3193 (1984).