ANNUAL REVIEWS

Annual Review of Chemical and Biomolecular Engineering

Crystal Structure Prediction Methods for Organic Molecules: State of the Art

David H. Bowskill, Isaac J. Sugden, Stefanos Konstantinopoulos, Claire S. Adjiman, and Constantinos C. Pantelides

Molecular Systems Engineering Group, Centre for Process Systems Engineering, Department of Chemical Engineering, and Institute for Molecular Science and Engineering, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom; email: c.pantelides@imperial.ac.uk

Annu. Rev. Chem. Biomol. Eng. 2021. 12:593–623

First published as a Review in Advance on March 26, 2021

The *Annual Review of Chemical and Biomolecular Engineering* is online at chembioeng.annualreviews.org

https://doi.org/10.1146/annurev-chembioeng-060718-030256

Keywords

crystal structure prediction, polymorphs, lattice energy, free energy of crystals

Abstract

The prediction of the crystal structures that a given organic molecule is likely to form is an important theoretical problem of significant interest for the pharmaceutical and agrochemical industries, among others. As evidenced by a series of six blind tests organized over the past 2 decades, methodologies for crystal structure prediction (CSP) have witnessed substantial progress and have now reached a stage of development where they can begin to be applied to systems of practical significance. This article reviews the state of the art in general-purpose methodologies for CSP, placing them within a common framework that highlights both their similarities and their differences. The review discusses specific areas that constitute the main focus of current research efforts toward improving the reliability and widening applicability of these methodologies, and offers some perspectives for the evolution of this technology over the next decade.

1. INTRODUCTION

The ability of organic molecules to crystallize into different solid-state forms (polymorphs) is central to the discovery and manufacture of novel crystalline materials. The detailed study of this phenomenon has led to an increasing understanding of the prevalence of polymorphism and of its large impact on physicochemical properties, which, in turn, has led to the growth of an active community of researchers focused on methods for crystal structure prediction (CSP).

The objective of CSP methods is generally to produce a small, yet complete, set of crystal structures that are likely to be observed experimentally. In the context of product development and manufacture, this set of structures often needs to be obtained with little or no prior knowledge of the investigated molecule(s) beyond its molecular connectivity diagram(s).

CSP can be an invaluable complement to experimental polymorph screening (1) in the pharmaceutical and agrochemical industries, among others. It can provide useful insights for the interpretation of experimental results (2), support the resolution of structures from powder X-ray diffraction (PXRD) (3) or nuclear magnetic resonance (NMR) (4) patterns, and facilitate the development of experimental protocols for the crystallization of previously unobserved polymorphs (5, 6). An energy landscape produced using CSP can also enhance the understanding of the behavior of a target molecule. For example, Abramov (7) concluded from a CSP study that the pharmaceutical compound crizotinib is unlikely to be polymorphic due to large energy differences between candidate structures. This knowledge can inform risk assessments undertaken during the course of drug development or help in identifying a drug formulation or manufacturing conditions that prevent costly interruptions of supply, such as those that occurred with ritonavir (8). Beyond structural characterization, research is under way to enable the use of CSP to investigate structure–property relationships (9–11) or as a large-scale screening method to bypass expensive experimental efforts (12).

The development of CSP methods has been accelerated through lessons learned during the CSP blind tests organized by the Cambridge Crystallography Data Centre. The results of the most recent (sixth) blind test demonstrated the significant progress that has been achieved over the past 2 decades, with more complex crystals now well within our predictive reach (13).

The sets of candidate crystal structures produced by CSP are ranked by increasing energy. The metric most commonly used for this ranking is the lattice energy at 0 K and 0 Pa, that is, the difference between the energy of a static lattice arrangement of molecules and that of static molecules in the gas phase and at infinite separation. Typically, the crystal structures of practical interest are those whose energy is within a certain cutoff limit of the global minimum in energy. This approach is driven by the observation that the lattice energy difference between pairs of experimentally observed polymorphs rarely exceeds 10 kJ/mol (14). It has proved effective in many cases, although the number of crystal structures predicted within such a cutoff is often greater than the number of distinct structures that have been identified experimentally (15). Prediction of the likely crystalline forms of many systems of practical relevance remains challenging due to characteristics, such as molecular size and flexibility or the existence of multicomponent crystals (e.g., solvates/hydrates with unknown or variable stoichiometry). For instance, the vast majority of pharmaceutically relevant molecules exhibit significant flexibility: Of a data set of 5,941 structures extracted from the Cambridge Structural Database (CSD) (16), of which 74% are drug-like, 92% exhibited some flexibility (17).

Because of these complicating factors, effective CSP methods need to achieve a delicate balance between the dual goals of identifying all likely structures and obtaining an accurate energy ranking. This requires the development of methods that enable both an extensive search of the space of possible structures and an accurate evaluation of inter- and intramolecular interactions—two potentially conflicting objectives in the context of finite computational resources.

In this article, we discuss the generic features that are common to state-of-the-art CSP methods for organic molecules, aiming to reflect upon the key elements of successful CSP methods and to motivate further methodological advances. We focus on generally applicable algorithms and models, and direct the reader to recent reviews (18, 19) for an overview of what can be achieved with CSP methods today. The extensive literature on the prediction of the thermodynamics and properties of specific crystal structures is considered only to the extent to which it has had a more general impact on CSP methodologies.

In Section 2, we present a high-level blueprint for CSP, which is common to most of the systematic methods available today. We review the various components of this blueprint in subsequent sections. In Section 3, we summarize the methods that are used to generate candidate structures in the early stages of CSP methods. In Section 4, we describe different approaches to modeling the lattice energy within CSP approaches, as this is the most common metric for ranking candidate structures. However, in Section 5, we review recent efforts to embed free energy calculations within the CSP framework. In Section 6, we describe how the elements considered in the previous sections have been synthesized to produce some of the CSP methodologies that are most commonly used today. Finally, in Section 7, we present our perspectives on the state of the art in CSP and some of the directions for future progress in this area.

2. CRYSTAL STRUCTURE PREDICTION FRAMEWORK

Practical CSP methods typically seek to determine low-lying minima of an energy function with respect to optimization variables that define the crystal structure, namely the unit cell variables (lattice lengths and angles) and the positions of all atoms in the unit cell. Optimization of each crystal is often carried out with respect to 1 of 230 potential space group symmetries. Space group symmetry allows the positions of all atoms in the crystal to be determined by knowledge of the asymmetric unit cell, thereby drastically reducing the space of optimization variables to the number of atoms/molecules in the asymmetric unit cell. In this context, each space group symmetry must be explored independently to identify all unique structural configurations. An examination of the CSD indicates that the vast majority of crystal structures can be represented by a subset of these symmetry groups; in fact, ~60 space groups account for ~97.9% of all organic crystal structures (20). It is common for the asymmetric unit cell to contain only one molecule ($Z' = 1$). In such cases, the crystal structure depends on the position (e.g., of the molecular center of mass), the orientation, and the conformation of that molecule. Up to 20% of organic polymorphs contain more than one molecule in the asymmetric unit cell (17), while crystals with multiple species, such as cocrystals, hydrates, and solvates, inherently contain multiple independent molecules. In our experience, each additional molecule in the asymmetric unit typically increases the cost of an effective CSP search by approximately four to five times.

For flexible molecules, packing forces lead to changes in the molecular conformation. Not only is the number of variables larger than for a rigid molecule, but also one must achieve a sufficiently accurate description of the interplay between intermolecular and intramolecular interactions (i.e., the cost of deforming a molecule from its in vacuo conformation). Currently, the practical limit on the number of flexible degrees of freedom (torsional angles, bond angles, and bond lengths) that can be treated in the context of a $Z' = 1$ search is approximately 10 (6).

The computational cost of CSP would be prohibitive if all possible structural configurations were evaluated at the full accuracy that would normally be required for reliable predictions. To overcome this challenge, CSP methodologies employ a multistage approach (**Figure 1**). This approach reduces the polymorphic space to successively smaller sets of structures, so that the energy of the structures in each set can be evaluated with increasingly accurate models. Each stage thus

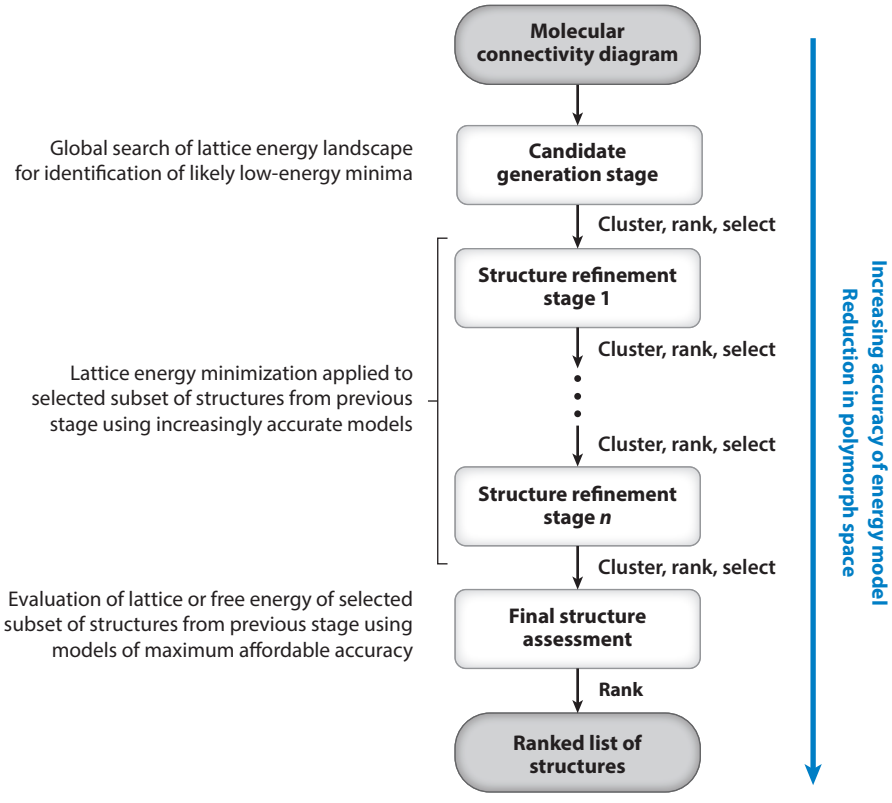

Figure 1

A flowchart of typical crystal structure prediction methodologies.

incorporates a different model of the crystal's energy, given that the evaluation of energy (and usually its partial derivatives with respect to the variables defining the crystal structure) is the largest component of the computational cost at each stage. The relative costs of these energy models can span many orders of magnitude across the stages of an investigation. For example, using simple force fields, one can evaluate the energy of a crystal structure in a matter of seconds, whereas more accurate electronic structure methods may take many thousands of CPU hours per structure. Significantly reducing the number of potential structures considered at each stage makes it feasible to apply models approaching the desired level of accuracy in later stages without expending too much computational effort on nonrelevant candidate structures during the earlier stages.

The initial candidate generation stage shown in **Figure 1** is common to all CSP workflows. Its purpose is to explore the energy landscape by generating structures that cover a broad range of crystal geometries. The number of configurations that need to be explored depends on the number of molecules in the asymmetric unit cell and the number of flexible degrees of freedom of those molecules, and it must be sufficiently large to ensure that no practically significant structures are missed. For example, the quasi-random search employed by our group requires approximately 10^5–10^6 minimizations (21) of the lattice energy function from different initial structures to generate a sufficiently complete list of candidates.

At the end of each stage, the set of structures produced is typically processed using clustering algorithms to remove crystallographically equivalent structures, thereby preventing unnecessary duplication of refinement calculations in subsequent stages. Structures are also screened to remove

those whose energy difference from the structure of lowest energy exceeds a specified cutoff; the value of the latter depends both on the expected polymorph energy differences and on the confidence in the accuracy of the model used. Several studies in this area can help suggest suitable energy cutoffs. For example, using a hybrid quantum mechanical (QM) force field–type approach, Nyman & Day (14) found that in a set of 508 polymorphic molecules the free energy differences between polymorphs did not exceed 10 kJ/mol in 99.5% of cases. In a smaller data set of 55 flexible polymorph pairs and with the use of more accurate electronic structure calculations, Cruz-Cabeza et al. (17) also reported that the energy difference never exceeded 10 kJ/mol. Thus, a cutoff of 10 kJ/mol appears reasonable for an analysis of the final landscape generated by CSP. In contrast, at earlier stages of the process, it may be necessary to apply less stringent cutoffs to compensate for the lower accuracy of the energy models used at those stages.

At the end of each stage, all clustered structures not eliminated via the energy cutoff criterion are passed to the following stage. Following the initial candidate structure generation stage, it is common for one or two refinement stages to be used. A final assessment is then typically applied to the structures resulting from the last refinement stage by using a highly accurate model of either lattice or free energy. Ideally, the model should reminimize the structure energies by using the corresponding geometries from the last refinement stage as initial guesses. However, the energy model employed may be too expensive for such optimization calculations, in which case it is used only to recalculate the structure's energy at the fixed geometry determined at the last refinement stage. The final output of the CSP investigation is a ranked list of likely polymorphs.

The general approach outlined in **Figure 1** encompasses a broad range of modern CSP methodologies. An important extension of this simple sequential workflow incorporates an element of feedback, where information from later stages can be fed back to the candidate structure generation stage to improve the search in an iterative fashion (22, 23). For example, structural and energetic data at the dispersion-corrected density functional theory (DFT-D) level can be used to improve the accuracy of the simpler energy models used at the candidate generation stage, or statistical information can be employed to adjust the energy cutoffs used to select structures for further refinement. Next, we consider the individual stages of the general framework introduced in this section in more detail.

3. CANDIDATE GENERATION

3.1. Overview

The general methodology shown in **Figure 1** starts by conducting a global search to generate an initial set of candidate structures. State-of-the-art algorithms for candidate structure generation can be grouped into two broad categories, namely unbiased and biased search methods (see Sections 3.2 and 3.3). In all cases, given the high dimensionality of the problem and the large number (hundreds of thousands to millions) of structures that must be evaluated, relatively cheap energy models with limited accuracy are often used at this stage. As a result, a relatively high refinement cutoff of 20–30 kJ/mol is typically applied (13) to the list of candidate structures obtained in order to ensure that relevant structures are not excluded from consideration at the later refinement stages. The number of structures taken from the candidate generation stage to the first refinement stage is typically of the order of hundreds to thousands.

3.2. Unbiased Search Methods

In unbiased search methods, numerous initial structures are generated with the aim to achieve broad sampling of the space of possible crystal structures. Each such structure is then used as a starting point for a local energy minimization so that the candidate generation step results in a

first ranked list of putative structures. Several methods have been proposed to achieve an effective sampling. Their common characteristic is that equal computational effort is devoted to areas of high and low energy, although in some cases starting points with a particularly high energy relative to the structure of lowest energy identified so far in the search are discarded without local minimization (24).

The simplest form of unbiased search is performed by constructing a regular grid in the space of the variables defining the crystal structure. Because the number of points in the grid increases exponentially with the number of variables, the applicability of this method is limited. Only 3 of the 25 groups taking part in the sixth blind test made use of such an approach (13).

An alternative strategy is to generate starting structures randomly as samples from a uniform probability distribution over the variable ranges. Albeit simple to implement, pseudorandom sampling may lead to nonuniform coverage of the variable space. Quasi-random, low-discrepancy sequence techniques, such as that by Sobol' (25), lead to a more even coverage of the domain of interest for a given number of generated points. Moreover, unlike in regular grid methods, termination of a Sobol' search after sampling any number of points generally results in the best possible sampling of the domain that is achievable with that number of points. As a result, it becomes possible to later continue the search to consider additional points while maintaining the quality of the sampling, implying that the number of points to be sampled does not need to be fixed in advance. These methods have been successfully implemented by several groups in investigations of organic molecules (24, 26, 27).

3.3. Biased Search Methods

Biased search methods attempt to limit the set of structures that are generated and assessed to the more promising areas of the lattice energy surface. They include several classes of algorithms, as outlined below.

The Monte Carlo simulated annealing (MCSA) approach (28, 29) involves the exploration of the energy landscape starting from a random structure and moving from one structure to a neighboring one. A new structure is accepted unconditionally if its energy is lower than that of the previous structure (30). In contrast, if its energy is higher than the previous structure, the new structure is accepted with a probability defined as an increasing function of a so-called temperature parameter. Successive search cycles performed at different temperatures are used to explore the energy landscape. Starting from an initially high value, which allows many moves to be accepted, the temperature is gradually reduced so that only moves to similar- or lower-energy configurations are accepted. The temperature is then increased again to allow the search to move into a different area of the energy landscape, and the process is repeated in a cyclic manner. Any promising (i.e., low-energy) configurations identified during the search may serve as starting points for lattice energy minimization to determine crystal geometries that correspond to local minima in the energy surface.

A methodological development of the MCSA approach is that of parallel tempering (31, 32), wherein multiple MCSA simulations are run in parallel, each at a different temperature. At each step, it is possible for the current configurations to be exchanged between two different simulations; the probability of such an exchange taking place is computed according to the Metropolis criterion. Overall, the use of parallel tempering ensures a more thorough exploration of the space, and has proved to be successful in the GRACE CSP implementation (33).

Evolutionary algorithms (34, 35), of which genetic algorithms are the most common, attempt to mimic the principles of natural selection: The global minimum energy crystal structure is analogous to the best-fit species. From a population (i.e., set) of crystal structures, those with a high

fitness (i.e., low lattice energy) are allowed to procreate (i.e., exchange structural features between the two parent structures to create a new structure) and mutate (i.e., be modified via Monte Carlo moves). In this way, the relevant low-energy structures are discovered over successive generations. There has been some success in using genetic algorithms in the prediction of crystal structures of inorganic (36, 37) and small organic (38) molecules. However, perhaps due to the large separation between minima on the organic crystal potential energy surface, genetic algorithms can get trapped at local minima, and may oversample restricted areas of configurational space. Evolutionary niching has been successfully applied to diversify the sampling of crystal structures (39), but these developments need to be tested further, particularly on large, flexible molecules, before any firm conclusions can be drawn.

Particle swarm optimization (PSO) methods (40) operate on a population of particles, each corresponding to an evolving crystal structure. At each step, the algorithm tracks the lowest-energy structure p_i encountered by each particle i so far, as well as the lowest-energy structure g identified by the entire population. Each structure i is then updated via a randomized move informed by both its own p_i and the global g. While PSO methods have been applied primarily to the prediction of inorganic crystal structures, they have recently been used in the prediction of flexible organic molecules (23), in combination with MCSA.

In general, biased search methods can save computation by focusing the search on the more promising areas of the energy landscape. However, their efficacy depends crucially on the choice of the parameters that determine the way in which the search moves within the landscape. Selecting an appropriate set of values of these parameters for any particular system under consideration in a CSP study is a nontrivial undertaking. In contrast, unbiased methods may spend significant parts of their computation on areas of the energy landscape that may ultimately be found not to contain any low-energy crystal structures. However, they may be more reliable and universally applicable as they do not involve any arbitrary parameters.

3.4. Handling Molecular Flexibility in Global Search

There are two broad approaches to dealing with molecular flexibility during the initial candidate generation stage. The first attempts to identify a number of crystallographically relevant molecule conformers before the start of the global search; it then performs a number of separate global searches, each based on a different conformer and treating it as a rigid molecule. The second approach is to treat flexible degrees of freedom as additional optimization variables within the global search.

Albeit simpler to implement, the use of multiple rigid searches suffers from some serious deficiencies. In particular, it may result in candidate structures that are more constrained and hence exhibit higher lattice energies (41), as there is no possibility for the energy to be reduced via small adjustments in the molecular confirmation. Overall, this may cause some potentially relevant structures to be eliminated at the end of the initial candidate structure generation stage as their energy is found to be above the cutoff.

Furthermore, choosing a finite set of distinct molecular conformers that will not result in any relevant crystal structures being excluded is itself a nontrivial task. In many cases, it is sufficient to include conformers with intramolecular energy up to 20 kJ/mol above that of the lowest-energy conformation. However, such a criterion may cause some relevant conformers to be overlooked (19), since intermolecular energy gains have been reported to compensate for intramolecular energy costs of up to 57 kJ/mol (42).

For these reasons, handling flexibility in terms of a set of continuous variables being optimized simultaneously with the rest of the variables determining the crystal structure is generally a more

reliable approach. However, it requires reasonably accurate and efficient methods for computing intramolecular energy during lattice energy minimization. We consider this topic in more detail in the next section.

4. LATTICE ENERGY MODELS FOR USE IN CRYSTAL STRUCTURE PREDICTION

Lattice energy is the most frequently used metric in predicting and ranking candidate crystal structures (13). This approximation is reasonable at 0 K and ambient pressure due to the small differences in relative zero-point energy (ZPE) between different crystal structures (43, 44). Due to the large number of lattice energy calculations required, the availability of reliable yet computationally tractable models becomes paramount in the early stages of any CSP study. At the later stages of the CSP, the number of candidate structures is reduced, allowing the use of more accurate but more expensive energy models. For this reason, it is desirable to develop a spectrum of energy models of increasing accuracy and cost. This has led to the use of both force fields and affordable periodic DFT-D methods across the various stages of CSP. These two types of models and their use within CSP are discussed in the following subsections.

4.1. Force Field Models of Lattice Energy

Force field methods are the most prevalent in the early stages of CSP studies because of their computational simplicity in comparison to alternatives such as electronic structure methods (Section 4.2). In particular, their relatively low computational cost enables the handling of the additional computational complexity arising from increasing molecular size and flexibility.

In force field models, the lattice energy of a crystal, ΔU^{latt}, is broken down into contributions arising from the internal structure of each molecule (i.e., intramolecular interactions) and from the interactions with other molecules in the crystal environment (i.e., intermolecular interactions):

$$\Delta U^{\text{latt}} = \Delta E^{\text{intra}} + E^{\text{inter}}. \quad 1.$$

Here, ΔE^{intra} denotes the difference between the intramolecular energy of the molecule(s) in the crystal and the minimum intramolecular energy of the molecule(s) in vacuo, and E^{inter} denotes the intermolecular energy. ΔE^{intra} depends on the molecular conformation, while E^{inter} is additionally a function of the geometry of the asymmetric unit cell and the position(s) of the molecule(s) within it.

To achieve the accuracy required for successful CSP, it is often necessary to tailor the most important interactions to the system of study. In this section, we review some common methods for deriving models for the intra- and intermolecular contributions, and for integrating them into effective lattice energy models for CSP.

4.1.1. Intramolecular energy. The intramolecular energy contribution (ΔE^{intra}) is relevant for flexible organic molecules. It can be viewed as the energetic cost of deforming a molecule from its gas-phase ground state into its geometry in the crystal. The choice of model for the intramolecular energy is critical for ensuring that the energetics are appropriately balanced with intermolecular terms so that crystal geometries are found to correspond to minima in the lattice energy landscape (45).

A classical description of intramolecular energy is provided by transferable force fields, such as DREIDING (46). In this approach, the intramolecular energy is evaluated using a simple function of atomic distances and angles. Early CSP studies (e.g., 47) made use of functions with transferable

parameters characterizing different types of atom–atom interactions (e.g., bond stretching, bending). More recently, this approach has lost ground to methodologies based on more customized approaches (13). In particular, due to the nature of valence shell electron interactions in covalent bonds and their conformation-dependent coupling, intramolecular energy can best be characterized using a QM description based on the specific system being studied.

Therefore, the derivation of force fields from a limited set of QM computations of the molecule(s) of interest has emerged as a key element of most modern CSP frameworks. Two main classes of approaches for such a parameterization have been proposed: those that rely on the so-called isolated-molecule assumption, namely that the internal interactions of a molecule can be evaluated independently of the other molecules in the crystal, and those that do not, requiring more demanding calculations of the energy of a periodic crystal.

Under the isolated-molecule assumption, QM calculations are carried out for a range of conformers and used to build a model that relates intramolecular energy to conformational degrees of freedom (26, 45, 48, 49). The influence of the crystalline environment is sometimes taken into account by imposing an external field on the molecule, specifically by using the polarizable continuum model (PCM) (50) and specifying a dielectric constant value, typically ranging from 3 to 11 (51, 52). By restricting the analysis of flexibility to the small number of unique molecules that appear in a given crystal and treating them as isolated (i.e., non-spatially-extended) systems, the use of DFT or even more accurate post-Hartree-Fock calculations to evaluate ab initio geometries and energies becomes relatively affordable and readily accessible via modern QM software packages, such as Gaussian 09 (53). Nevertheless, directly evaluating the intramolecular energy of an isolated molecule ab initio for each molecular geometry being considered during the course of a CSP study is prohibitively expensive; more sophisticated approaches are therefore required.

In the simplest practical realization of the isolated-molecule strategy, several conformers are identified at the start of the CSP study, their internal energy is computed, and a separate rigid-molecule crystal structure search is conducted for each conformer (e.g., 48). Conformational flexibility is thus reduced to a small set of discrete choices. To limit computational cost further, force fields can be used to generate conformers, while single-point DFT calculations are used to refine the energy evaluation. However, as explained in Section 3.4, focusing on a fixed set of rigid conformers may result in loss of reliability of the global search stage of the CSP.

An alternative approach is to use information derived from specific conformers to build a model that provides a locally valid approximation of the relationship between conformational flexibility and intramolecular energy. This approach was implemented in the program UPACK (54) by use of a second-order Taylor expansion of the energy as a function of all atomic positions. This requires the ab initio computation of the gradients and second derivatives of the intramolecular energy with respect to internal degrees of freedom at a given conformation and reusing this information in the minimization of the lattice energy over a neighborhood of conformations close to this point. The conformers at which the intramolecular energy and its derivatives are computed at the QM level are essentially sampling points of the intramolecular energy function. While this approach can provide a highly accurate approximation of the intramolecular energy, the need to compute first and second derivatives with respect to all conformational degrees of freedom at all sampling points renders it computationally impractical, even for moderately sized molecules (54).

The computational cost of performing ab initio isolated-molecule calculations can be drastically reduced by recognizing that only a few molecular degrees of freedom (often a subset of the molecule's torsion angles) are typically flexible enough to have an appreciable impact on a crystal's geometry and energy. It is thus possible to conduct the search for possible conformations in the reduced space of these variables (the independent conformational degrees of freedom). For any given set of values of these variables, the values of the other, dependent, variables (i.e., the

bond lengths, bond angles, and remaining torsion angles) and of the intramolecular energy at each sampling point can be obtained by minimizing the intramolecular energy with respect to the dependent variables. Overall, this minimization defines the dependent variables and the intramolecular energy as functions of the independent variables. Two strategies have been proposed for constructing suitable approximants to these functions: restricted Hermite interpolants (24) and local approximate models (LAMs) based on second-order Taylor expansions combined with optimality conditions (45).

The mapping proposed by Karamertzanis & Pantelides (24), based on restricted Hermite interpolants, avoids the need for partial derivative information above first order, so that the cost of evaluating each sampling point is quite small. However, Hermite interpolants require that the sampling points be constructed on grids, which limits the practical applicability of the approach to molecules containing up to three or four independent degrees of freedom due to the resulting computational cost. Nevertheless, larger molecules can be handled if they can be partitioned into nearly independent groups of torsion angles so that each group can have up to three or four independent variables (26).

To handle larger and more flexible molecules, LAMs have been developed to make use of second-order Taylor expansions constructed from partial derivatives, ensuring quadratic accuracy regardless of dimensionality (45, 49). The cost of constructing and using such LAMs can be further reduced by storing LAMs in databases so that they can be reused across multiple computations involving the same molecule (45) and by developing adaptive strategies for determining the points at which the LAMs are generated (55). These advances, combined with the introduction of continuous and differentiable models derived from weighted combinations of LAMs (56), have enabled accurate and efficient modeling of the intramolecular energy of flexible molecules with up to seven or eight independent degrees of freedom.

As an alternative to the isolated-molecule approach, one can derive a force field based on QM calculations of periodic crystals. Neumann (22) has developed a methodology to parameterize a DREIDING-type force field using periodic DFT-D calculations for a set of crystal structures that contain different conformers of the molecules of interest. A number of sampling points are used to fit parameters that include stretching, bending, torsional, inversion, and angle-bend-inversion coupling terms. The resulting tailor-made force field (TMFF) provides an approximate lattice energy function that takes into account the crystalline environment, but the need to sample a sufficiently broad range of conformers and corresponding crystal structures comes at a significant computational cost. The methodology recently developed at XtalPi (23) follows a similar strategy, using cloud computing to carry out a large set of monomer, dimer and periodic DFT-D calculations for force field parameterization.

4.1.2. Intermolecular energy. On the basis of the dominant interactions between molecules in the crystalline phase (57), the intermolecular energy can be partitioned into separate contributions from electrostatic (E^{elec}), inductive (E^{ind}), and repulsive/dispersive ($E^{\text{rep/disp}}$) interactions:

$$E^{\text{inter}} = E^{\text{elec}} + E^{\text{ind}} + E^{\text{rep/disp}}. \qquad 2.$$

4.1.2.1. Electrostatic interactions. Electrostatic interactions are present in all organic crystals, but they make a particularly important contribution to the lattice energy in crystals that involve polar and/or charged compounds. In classical force fields, the electrostatic term (E^{elec}) accounts for the interactions of the charged particles (protons and electrons) in different molecules in the crystalline phase. A key modeling decision is the framework to be used to approximate the electrostatic potential arising from the charge distribution within the molecule. The range of approaches

used in existing CSP methods includes atom-centered point charges, off-atom point charges, and multipoles.

Point charges are a useful first approximation of the charge distribution and hence are used in several CSP force fields. As with the intramolecular energy contribution, it is advantageous to tailor the electrostatic contributions to the molecule(s) of interest. Atom-centered point charges can be fitted to the electrostatic potential obtained via periodic DFT-D calculations (22) or, alternatively, to the electron density of an isolated molecule computed at DFT or (post-)Hartree-Fock levels of theory. The point charges derived in this manner interact through a classical Coulomb potential. Although there is a plethora of charge fitting schemes in the literature, including real-space (58) and basis-space (59) approaches, empirical fitting schemes such as ChelpG (60) or HLY (61) are often used in practice. Importantly in the context of flexible molecules, the LAMs that have been developed to model the dependence of the intramolecular energy on conformational degrees of freedom have also been extended to model the dependence of point charges on conformation (55, 62), leading to a more accurate representation of the electrostatic energy.

To achieve an even more accurate approximation of the full electrostatic potential, researchers have proposed more detailed schemes, such as the discretization of the electron density (63) and off-atom satellite charges (62), although these approaches are not currently in widespread use in CSP methods. Another approach is to include a hydrogen-bonding term, as in the DREIDING force field and, optionally, in the TMFF (22). Finally, the explicit modeling of higher-order orientation-dependent interactions, such as lone pair (dipole moment) and π–π stacking (quadrupole moment) interactions, via multipole expansions has been used extensively in the CSP context. By placing a multipole expansion of the charge density on each atom site, one can substantially improve the accuracy of the potential (64). The parameterization of multipole expansions for a given electrostatic potential relies on a partitioning scheme to isolate atomic fragments. Such partitioning schemes can be viewed as somewhat arbitrary; the effectiveness of each procedure is judged on the basis of the accuracy and the convergence speed of the expansion. For over a decade, distributed multipole analysis (DMA) (65, 66) has been the most popular method for calculating atomic multipoles within CSP methods. More recent partitioning schemes may produce expansions with faster convergence but have yet to be commonly incorporated in CSP methods. These include the basis-space implementation of the iterated stockholder atom algorithm (67) proposed by Misquitta et al. (68) and the iterative Hirshfeld partitioning scheme (Hirshfeld-I) (69, 70). Such approaches may result in a partitioning of the electron density that produces more chemically reasonable multipole distributions (68). In any case, once multipole moments have been derived, they can be used within lattice energy models to provide a good trade-off between accuracy and efficiency of calculations. As with point charges, the dependence of multipoles on conformational degrees of freedom can be represented via LAMs (45), providing an inexpensive model of the impact of flexibility on electrostatic interactions.

The so-called molecular electrostatic potentials derived from isolated-molecule QM calculations depend on whether the latter computations are carried out in vacuum or in an electric field [i.e., using a PCM to represent the effect of the crystalline environment on the electrostatic potential of individual molecules (51)]. In the latter case, the electrostatic potential attributed to each molecule is affected by induced polarization as a result of the uniform electric field imposed by the dielectric constant. Regardless of the approach used, the effect of the crystal geometry on the induction energy is neglected during the crystal structure optimization. An additional term may therefore be required to capture this effect, as described in the next section.

4.1.2.2. Induction energy. Inductive effects constitute a relatively small yet nonnegligible proportion of a crystal's cohesive energy, contributing approximately 20–40% of the electrostatic

interaction (71) in both polar and nonpolar molecules. Due to the costs and complexity associated with rigorously incorporating inherently non-pairwise-additive inductive effects in CSP, the E^{ind} term in Equation 2 is usually omitted. Although this is often a reasonable approximation, for some systems, such as those involving hydrates or salts, accounting for this term may be important for achieving a sufficiently accurate evaluation of lattice energy and, consequently, more accurate stability rankings among the low-energy forms (72). Recently, several approaches aiming to capture this contribution to the lattice energy have emerged.

One strategy is to embed an additional term in the classical force field model in order to represent polarizability. In the same vein as distributed multipoles, the induction energy in a system of molecules can thus be modeled in a classical force field by use of distributed polarizabilities. This approach has sometimes been applied in the context of CSP, where a self-consistent set of induced multipoles is derived at each crystal configuration as a function of the polarizability of each site (71, 72). Accurate atomic polarizabilities have been derived from symmetry-adapted perturbation theory based on density functional theory [SAPT(DFT)] calculations by using methods such as the Williams–Stone–Misquitta (WSM) localization scheme (73, 74). However, this approach comes at a significant computational penalty, both because of the cost of deriving atomic polarizabilities using SAPT(DFT) (72) and because of the additional computational complexity of the resulting lattice energy model.

Alternatively, the charge density can be derived from calculations in the crystal phase or an approximation thereof. This approach is exemplified by the self-consistent electronic response to point charges (SCERP) method (71) and the use of q-grids (75); the latter approach was applied in the most recent blind test (13). Such methods circumvent the need to define atomic polarizabilities, although the calculation of the electrostatic density is more demanding because it must be carried out with either molecular clusters (71) or periodic calculations (75). Furthermore, because the electrostatics thus derived depend on the specific crystalline environment, they may not be applicable if there are significant changes to the geometry of the crystal.

4.1.2.3. Repulsive/dispersive energy. The $E^{\text{rep/disp}}$ term in Equation 2 accounts for close-range repulsive and long-range dispersive interactions. The repulsive energy arises from the Pauli exclusion principle and decays exponentially at long range, while dispersion is generally an attractive force caused by correlated electron density fluctuations. With regard to first-order interactions, dispersion decays with interatomic distance r_{ij} between atom sites i and j as r_{ij}^{-6}. As a result, a popular functional form used to describe the repulsive/dispersive interactions in CSP is the Buckingham potential:

$$E^{\text{rep/disp}} = \sum_{ij} A_{ij} \exp(-B_{ij} r_{ij}) - \frac{C_{ij}}{r_{ij}^6}, \qquad 3.$$

where A_{ij}, B_{ij}, and C_{ij} are atom-specific parameters. These parameters can be estimated either from experimental data representative of organic crystals or from computational data for the specific molecule(s) of interest. In the former case, transferable parameters for pairwise interactions between atom types are estimated by fitting to crystalline geometries and sublimation enthalpies. This estimation requires the selection of a suitable training set, a task made difficult by the scarcity of data for some atom types and, sometimes, the presence of large and unknown experimental uncertainties, particularly for sublimation enthalpies (76). Furthermore, the choice of atom types requires careful consideration. On one hand, transferability is improved by treating atoms of the same chemical element as different atom types in order to account for environmental dependence (e.g., hybridization). On the other hand, it is important to keep the number of atom types small

enough to avoid overparameterization in view of the relatively limited quantity of available experimental data.

Commonly used force field parameterizations based on experimental data include the FIT (77–82) and W01 (83) potentials. These parameter sets were derived using a model of the lattice energy that does not account for intramolecular energy ($\Delta E^{\text{intra}} = 0$) and in which the only electrostatic interactions considered are Coulombic (charge–charge) in nature; the charges were derived from an electrostatic potential computed at the Hartree-Fock level of theory. Recent efforts have highlighted the importance of using parameters in $E^{\text{rep/disp}}$ that are consistent with the level of theory/basis set and specific functional form (the DMA multipole expansion) used in obtaining E^{elec} (with E^{ind} set to zero) (84, 85). Moreover, the common practice of employing combining rules to evaluate cross interactions between unlike atom types in most parameterizations of the Buckingham potential does not always have a sound physical basis; instead, the fitting of unlike parameters directly to experimental data has been found to lead to more reliable models (84, 85). A benefit of fitting to experimental data is that the resulting empirical term in the force field accounts not only for repulsive/dispersive interactions but also, to some extent, for features of the lattice energy that are poorly understood or difficult to model, such as electrostatic penetration, charge transfer, many-body contributions, and even induction (86).

The second strategy for parameterizing the Buckingham potential is to tailor the parameters to the molecule(s) under study. In this case, a set of crystal structures for the latter is generated using DFT-D calculations, and the resulting geometries and lattice energies are used to estimate the potential parameters, thereby accounting more closely for the environmental dependence of the interatomic interactions. The repulsive/dispersive potential parameters can be estimated simultaneously with other potential parameters, such as bond strengths or point charges (22), or independently in combination with electrostatic contributions derived from isolated-molecule calculations (87). In either case, a sufficiently large and diverse set of structures needs to be used in order to ensure that the derived potential is applicable to other crystal structures of the same system. In contrast, given the high computational cost of DFT-D calculations, it is important to keep the number of such structures as low as possible.

A problem with the use of the Buckingham potential in CSP is that it predicts unphysical behavior at very close ranges. This issue can be addressed by modifying the dispersive interaction using appropriate damping functions (88). Higher-order dispersion interactions of the forms r_{ij}^{-8} and r_{ij}^{-10} can also be influential, as can three-body interactions modeled through the Axilrod–Teller–Muto (ATM) equation (89). Such interactions may lead to the stabilization of certain systems, such as benzene crystals (90). The repulsive part of the interaction can also be modified to account for anisotropic interactions at short range (72), which is particularly relevant for heavier atoms. However, although it is relatively straightforward to incorporate such extended potentials within CSP algorithms, it is usually difficult to fit models that incorporate higher-order dispersion and anisotropic interactions with transferable force fields because of the significant increase in the number of parameters and their high degree of correlation (91).

Tailored potentials derived from SAPT(DFT) calculations have been proposed as an alternative that circumvents these challenges. Specifically, SAPT(DFT) can be applied to a large number (hundreds or thousands) of different configurations of pairs of molecules (dimers) at different relative distances and orientations. The resulting energies can then be used to fit an analytical force field tailored to the molecule(s) under consideration (92, 93). Alternatively, dispersion coefficients can be calculated from frequency-dependent polarizabilities obtained via methods such as the WSM localization scheme (73, 74).

SAPT(DFT)-fitted potentials offer a high level of accuracy for the computation of energy of noncovalent interactions and have already been applied to CSP for several small organic molecules

(72, 91–95), including some submissions in the most recent blind test (13). However, even for relatively small rigid molecules, deriving potentials via SAPT(DFT) methods involves the estimation of hundreds of parameters from the results of tens of thousands of expensive dimer calculations (92). Thus, the application of this approach to larger, flexible molecules may be problematic at present. Repulsive/dispersive parameters derived in a transferable way from experimental data or fitted to molecule-specific DFT-D calculations seem to be a more practical proposition for systems of the size and complexity of interest to current CSP investigations.

4.2. Electronic Structure Methods

Electronic structure methods offer a different treatment of molecular interactions based on an entirely QM approach. Instead of partitioning molecular interactions into intra- and intermolecular contributions, the total energy of the crystal is calculated from atomic positions in the periodic crystalline environment, naturally taking account of the effect of molecular conformation on the intermolecular potential and hence the energy. In general, standard electronic structure calculations are significantly more expensive than their force field counterparts. Nevertheless, some cheaper alternatives have been proposed for use in CSP.

4.2.1. Periodic density functional theory calculations. Periodic DFT methods are prevalent in modern material design approaches because they offer a good compromise between accuracy and cost compared with competing QM methods. Periodic DFT is becoming more routinely applied in CSP. The number of research groups using periodic DFT calculations in blind test CSP calculations increased from 2 in 2011 to 12 in 2016, demonstrating the growing popularity of these methods (13).

Several programs for periodic DFT calculations exist, differing mainly in the basis sets used to represent the electron density. Plane waves are the most commonly used (96), but atom-centered Gaussians (97) and natural atomic orbitals (98) have also been implemented in commercial and academic packages. The success of periodic DFT methods is contingent on an accurate approximation of the exchange-correlation energy based on the electron density of the system. A hierarchy of exchange-correlation functions (otherwise known as density functionals) was originally proposed by Perdew & Schmidt (99), and most solid-state programs embed a variety of these, including local density approximations, generalized gradient approximations (GGAs), and meta-GGAs, each of which offer various levels of accuracy and computational cost. Hybrid functionals are also commonly available in these programs, but, at least in the context of plane wave–based codes, they can typically be used only for single-point energy calculations, even with the aid of large computational resources.

Unlike post-Hartree-Fock methods, DFT methods do not offer a systematic way to improve the results to ensure convergence toward the exact solution. In addition, DFT functionals suffer from self-interaction errors (100), and as a result of their (semi)local nature, long-range dispersion interactions are not accounted for. In fact, without some means of incorporating dispersion, periodic DFT methods produce largely overestimated crystal energies (i.e., underbound crystals). Using the Perdew–Burke–Ernzerhof (PBE) functional, average errors of approximately 50% of the total interaction energy are observed if dispersion corrections are neglected (101). Overall, periodic DFT methods can be less accurate than their cheaper force field counterparts despite their much higher cost.

In view of the above deficiencies, significant efforts have been directed toward DFT-D methods. Klimeš & Michaelides (102) provide an excellent review of the hierarchy of dispersion corrections. In some of the most successful and cost-efficient dispersion corrections currently in use,

contributions due to dispersion are treated in a classical manner, invoking r_{ij}^{-6} and higher-order interaction terms between atom sites. For these interactions to be modeled effectively, the local atomic environment should be considered in the calculation. In the D3 correction (103), dispersion coefficients for the r_{ij}^{-6} and r_{ij}^{-8} terms are calculated by scaling the interaction according to the coordination number of the atom site. More recent developments include the modification of the D3 correction to take into account a scaling dependent on atomic partial charge (104). In the exchange dipole moment model (105), Fermi hole moments are used to derive r_{ij}^{-6}, r_{ij}^{-8}, and r_{ij}^{-10} terms, while in the Tkatchenko–Scheffler (TS) method, scaled r_{ij}^{-6} terms are calculated using Hirshfeld (106) or Hirshfeld-I (107) partitioned electron densities.

Beyond pairwise corrections, dispersion can be included by use of long-range density functions that include nonlocal correlation terms in the energy, in so-called van der Waals density functionals (102). Grimme et al. (103) also consider the use of three-body interactions through the ATM equation (89), while Tkatchenko et al. (106) address the treatment of many-body dispersion (MBD) interactions by extending the TS scheme with the MBD correction, which has shown promise in correctly predicting the relative stability of polymorphs (108–111).

4.2.2. Cheap electronic structure methods. The cost associated with standard periodic DFT-D calculations is often too high for all but the last stages of a CSP study, which involve significantly reduced numbers of candidate structures. Therefore, researchers have developed electronic structure methods that are orders of magnitude cheaper than most periodic DFT methods while retaining a reasonable level of accuracy. These methods notably include minimal basis set Hartree-Fock (HF-3c) (112, 113) and dispersion-corrected density functional tight binding (DFTB-D) (114, 115). Such approaches hold promise for bridging the gap between electronic structure and force field methods; with further development, they may find use in intermediate stages of CSP investigations (116).

4.3. Assessing the Accuracy of Lattice Energy Models for Crystal Structure Prediction

Several quantities can be used to assess the accuracy of lattice energy models by comparing their predictions with available experimental evidence. These include the crystalline geometry, the energy difference between two polymorphs, and the lattice energy (or, often, sublimation energy). It is generally accepted that geometry is easiest to predict, although periodic DFT-D methods can occasionally perform worse than force field methods (112, 117). Moreover, the relative energies of polymorphs are often easier to predict accurately than absolute energies (118), thanks to cancellation of errors.

In an effort to assess the impact of modeling choices on lattice energy calculations, Reilly & Tkatchenko (119) used a test set comprising 23 crystal structures to compare the computed lattice energies with corresponding values derived from experimental sublimation energies. The authors showed that PBE (120), a GGA functional; TPSS (121), a meta-GGA functional; and others perform well when combined with D3 or MBD dispersion corrections, achieving a mean average deviation (MAD) of the lattice energies in the region of 4 kJ/mol, which is close to the uncertainty of the reference data (119, 122). Improved results could be achieved by considering hybrid functionals and, within the plane wave basis set formulation, performing single-point evaluations on (meta-)GGA-optimized structures. Unfortunately, hybrid functionals cannot be routinely used for structure optimization within plane wave codes because of their high computational cost.

Brandenburg & Grimme (44) developed a different test set (POLY59) in order to investigate predictions of the relative energies of crystal structures. This test set focused on the five systems

considered in the sixth blind test. In addition to the 9 experimentally resolved polymorphs for these systems, the test set included 10 low-energy structures for each system determined computationally using a combination of force field (26, 45, 123) and periodic DFT-D calculations for refinement. The authors found that the PBE-D3 and TPSS-D3 schemes successfully predict the most stable experimentally resolved polymorph of each system as the one of lowest energy.

Cheaper electronic structure methods, such as HF-3c, result in significantly larger errors of approximately 7 kJ/mol (112), while DFTB-D is several thousand times faster than standard DFT techniques but yields a MAD of more than 10 kJ/mol, with particularly poor results obtained for crystals exhibiting hydrogen bonding. In agreement with this assessment of DFTB, Iuzzolino et al. (116) observed that the energy rankings produced by DFTB-D for a set of six flexible molecules are worse than those achieved by some force field methods. However, they demonstrated that the good reproduction of geometries with DFTB means that it may be suitable for producing reliable starting points for DFT refinement calculations. The use of these cheap DFT methods within a CSP workflow needs to be investigated further.

While periodic DFT-D methods perform excellently in many CSP applications, some molecules have proven challenging as a result of the underlying limitations of DFT functionals. For instance, for the highly polymorphic ROY molecule, one experimental form appears unstable when modeled using the PBE (120) functional (117), and computed polymorph energies indicate a ranking of polymorphs that significantly deviates from experimental observations (124).

The X23 test set has also been used (125) to evaluate the performance of a force field model based on isolated-molecule QM calculations/LAMs, multipole electrostatics, and experimentally derived repulsive/dispersive potentials. Several competing parameter sets were used for the repulsive/dispersive term. The best parameter set was found to be FIT, with a MAD of less than 10 kJ/mol, similar to that observed with cheap DFT methods (112). Newer parameter sets, such as that proposed by Gatsiou et al. (85), have yet to be evaluated against the X23 test set. Nevertheless it is possible that, with further development of force field methods, the gap between these and DFT calculations can be bridged, perhaps by utilizing concepts from the recent success of classical dispersion corrections in DFT (125).

5. FREE ENERGY EVALUATION FOR CRYSTAL STRUCTURE PREDICTION

As mentioned above, lattice energy, evaluated at 0 K, is the most commonly used metric for ranking structures generated in CSP studies. However, at finite temperature and pressure, stability is determined by the Gibbs free energy, which provides a more direct link to practically important and experimentally measurable properties such as solubility and specific heat capacity.

The Gibbs free energy of a perfect crystal (ΔG) is defined in relation to the ground-state energy of an isolated molecule with no thermal or zero-point motion, given by

$$\Delta G(T,P) = \Delta U(T,P) + PV(T,P) - TS(T,P). \quad 4.$$

Here, ΔU is the difference in internal energy with respect to the isolated-molecule reference state at given temperature T and pressure P, V is the molar volume, and S is the specific entropy of the bulk crystal at T and P. The lattice energy, ΔU^{latt}, is equal to ΔU (0 K, 0 Pa) minus the ZPE. For the given temperature and pressure, the most stable crystal form corresponds to the global minimum in the Gibbs free energy, while local minima correspond to metastable structures. The PV term in Equation 4 can usually be neglected under ambient conditions because of its small contribution to the overall free energy. In such cases, the Helmholtz free energy is approximately equal to the Gibbs free energy and can used as the stability ranking criterion.

Consideration of free energy in CSP studies is not yet widespread because of the often limited accuracy of the computed free energy values and the high computational cost of obtaining them. In the following subsections, we review the free energy models that have been used as part of multistage CSP methodologies (**Figure 1**).

5.1. Free Energy Calculations with Lattice Dynamics

Lattice dynamics (LD) theory (126, 127) provides a formalism for the description of lattice vibrations within a crystal. Estimates of both the ZPE and thermal contributions to vibrational free energy, $F^{\text{vib}}(T)$, can be obtained from phonon frequencies, including the effects of both entropy and thermal energy. Within the LD framework, the Gibbs free energy of a crystal is given by

$$\Delta G(T,P) = \Delta U^{\text{latt}} + F^{\text{vib}}(T) + PV, \qquad 5.$$

where $F^{\text{vib}}(T)$ quantifies the ZPE and thermal contributions to free energy.

5.1.1. Lattice dynamics. The simplest form of LD is the one based on the harmonic approximation (HA) (126), which approximates the potential energy of the crystal by using a second-order series expansion around an equilibrium position (local minimum) of the lattice energy hypersurface. Under the assumption that the atoms/molecules in the crystal oscillate harmonically around their equilibrium positions, vibrational (phonon) frequencies can be calculated and used in an analytical expression for the free energy of the crystal. The equilibrium geometry and volume are obtained from a lattice energy minimization and are fixed at all temperatures. Because of this restriction, HA is limited to the calculation of properties at constant volume, such as the isochoric heat capacity (C_V). Moreover, HA performs well at low temperatures but becomes less accurate at elevated temperatures as thermal expansion and anharmonic vibrational contributions become more significant.

The quasi-harmonic approximation (QHA) (128–131) is an extension of HA that can account for either isotropic or anisotropic thermal expansion (132) and facilitates the evaluation of properties such as isobaric heat capacity (C_P). The free energy of each structure is obtained by minimizing $\Delta G(T, P)$, as defined in Equation 5, with respect to the crystal geometry. QHA can be more reliable than HA for determining relative stability and evaluating thermodynamic properties, but it comes with a significant increase in cost, as multiple HA calculations have to be performed for each unit cell volume considered.

Another extension of LD is the anharmonic approximation (AA), the aim of which is to model anharmonic contributions to the free energy (110, 111, 133). Most commonly, contributions from higher-order terms in the expansion series are used to capture the nonparabolic shape of the potential well.

5.1.2. Application of lattice dynamics in crystal structure prediction. LD is by far the most commonly used approach for taking account of free energy in CSP studies. Although AA can provide more accurate results, HA and QHA are usually preferred because of their relative simplicity and lower computational cost (21, 111, 134).

In the most recent blind test (13), 6 of 25 participating groups employed LD as part of establishing the stability ranking of low-energy crystal structures. The number of structures examined in this context depended strongly on the approximations and cost of the underlying lattice energy model and free energy method used. For example, using rigid-body force fields, Nyman & Day (14, 131) applied HA to a total of 992 structures by using the coprime linear splitting method. In contrast, groups relying on DFT or HF-3c were typically able to study only a few structures.

The Pickard group (13) accounted for the effects of anharmonicity by utilizing vibrational self-consistent field theory (133) for the evaluation of ZPE; they reported the application of this approach to only a single structure, possibly because of the very high computational cost.

In general, the use of the Helmholtz free energy in the sixth blind test led to relative rankings of the experimental polymorphs that were more accurate than rankings based on lattice energy. For molecules XXII, XXV, and XXVI, the experimental structures were predicted as the global minima in free energy by three of the groups using HA. Form B of molecule XXIII was predicted as the global minimum by Brandenburg & Grimme using the HA free energy at 0 K (13).

Beyond the results reported in the context of the sixth blind test, Nyman (134) refined 100 structures of chloridazon by using the coprime linear splitting method and rigid-body force fields. Vasileiadis (135) used an atomistic force field to refine 200 structures of tetracyanoethylene and imidazole. In a recent study (111), the target molecules of the sixth blind test were revisited and vibrational free energies were calculated using PBE+TS phonon frequencies. In all cases other than the polymorphic target XXIII, the abovementioned free energies were employed for four to eight structures per molecule, and the experimental forms were found to be the ones with the lowest free energy. For target XXIII, HA was applied to 46 structures, while QHA and AA were employed for 9. Interestingly, QHA and AA predicted that an as-yet-undiscovered form is the most stable. Throughout these studies, free energy refinement generally led to an improved ranking of the experimental form(s).

5.1.3. The importance of vibrational free energy in crystal structure prediction. Researchers have attempted to quantify the contributions of ZPE and temperature to the enthalpy and entropy of polymorphs. Nyman & Day (14) have conducted the largest study to date of rigid-body vibrational free energies. They found that in 9% of the 601 polymorphic pairs examined, the inclusion of these terms caused a reranking of polymorphic stability between 0 K and room temperature, indicating an enantiotropic phase transition. The study also confirmed that the differences in ZPE and isochoric heat capacities between polymorphs are small, and highlighted the role of entropy as the main contribution to free energy. In a later study (131), the Gibbs free energies of 864 crystal structures (475 polymorphic pairs) were evaluated between 0 K and the melting point using both HA and QHA. Approximately 20% of the polymorphic pairs were reranked, again demonstrating the importance of including temperature effects on polymorphic stability.

Overall, the inclusion of ZPE and thermal contributions is an important consideration if we are to bridge the gap between computed and experimental crystal structure landscapes. To this end, free energy contributions have been incorporated within benchmark tests, such as POLY59 (44) and X23 (119).

5.2. Free Energy Calculations with Molecular Simulations

Molecular dynamics (MD) allows for ergodic sampling of various thermodynamic ensembles (136); it provides a natural method for capturing the effects of anharmonicity, such as thermal expansion or temperature-dependent frequencies, on the free energy at finite temperature and pressure (137). The reliability of MD simulations is, however, inherently restricted by the accuracy of the force field (109). Furthermore, the presence of free energy barriers and the necessarily short simulation timescales inhibit the study of rare events such as solid–solid transitions (136). To overcome these limitations, several advanced MD methods have been applied to polymorphic stability studies, including adiabatic free energy dynamics (138, 139), path-integral MD (140), orthogonal space random walk (141), and multistate Bennett acceptance ratio (MBAR) (137, 142).

To date, MD simulations have rarely been employed as part of a CSP study. In the sixth blind test (13), only the Tuckerman group carried out MD calculations. These authors used a

SAPT(DFT)-generated force field to perform a free energy ranking of 30 structures for molecule XXII by using crystal–adiabatic free energy dynamics (138). They found the experimental structure to be the fourth-lowest-energy structure among the set of 30.

QHA LD calculations were found to be in good agreement with free energies obtained by analyzing the velocity autocorrelation for imidazole and 5-azauracil (143). A comparison of free energies obtained with QHA LD and MBAR has also been conducted for a set of polymorphic molecules (137) using point charges. For rigid molecules up to room temperature, these two methods yield free energy values that are in reasonable agreement; however, the differences in predicted values are larger for flexible and/or disordered molecules at elevated temperatures, especially close to the melting point.

Metadynamics (144, 145) is a powerful enhanced sampling method based on collective variables that facilitates the study of rare events, which may be useful to identify transitions between forms. The key challenge in the case of CSP is the identification of appropriate collective variables. Metadynamics has been employed to refine the candidate structures generated by CSP searches for two molecules, 5-fluorouracil (146) and an industrial pigment (PR179) (147). In the case of 5-fluorouracil, 60 free energy minima derived from lattice energy minimization at 0 K and 0 Pa were investigated using metadynamics. The thermal fluctuations at ambient conditions led to 25% fewer free energy minima, indicating the presence of shallow minima in the lattice energy landscape. Form II was predicted as the most stable structure at 0 K and 0 Pa and did not undergo a phase transition during the metadynamics simulation. In contrast, the most stable experimental structure, form I, transformed into a disordered structure after a few metadynamics steps. In the case of PR179, 18 candidate structures generated from CSP at 0 K and 0 Pa were used as initial configurations for metadynamics simulations. The most stable form (phase I) was predicted as the global minimum in both the lattice energy and free energy landscapes. The study (147) concluded that another predicted form lies in a deep free energy minimum, making it a plausible metastable polymorph, although no experimental evidence for this polymorph has yet been found.

5.3. Disorder in Molecular Crystals

In addition to the effects of vibrational motion on free energy, a key consideration is disorder in crystals, which occurs in approximately a quarter of all structures in the CSD (148). Disorder can add a considerable configurational entropy contribution to crystalline stability as a result of an exponential increase in the number of accessible microstates in the ensemble of configurations. Distinct static structures that are generated as part of a CSP study can belong to the same ensemble of configurations, often leading to significant stabilization of seemingly metastable structures. However, modeling and predicting disorder are challenging, even with experimental guidance; therefore, it has been difficult to incorporate this phenomenon within CSP methods.

Both dynamic and static disorder were investigated for dimethylsulfoxide:carbamazepine solvates (149) using static models of disorder and MD simulations in a combined experimental and CSP study. This study found that static disorder is dominant at lower temperatures whereas dynamic disorder becomes more dominant above room temperature.

A suitable technique for modeling substitutional and orientational static disorder is the symmetry-adapted ensemble technique (SAET) (143). The importance of configurational free energy in disordered systems has been demonstrated through studies of several organic molecular crystals, such as caffeine (150), eniluracil, and dichloro/dibromobenzene (151). In CSP studies, SAET has also been applied alongside experiments for loratadine (152) and gandotinib (153). For loratadine (152), a CSP study using the GRACE program resulted in the prediction of ordered form II as the global lattice energy minimum, in contrast with experimental measurements that

indicated form I to be the most stable structure. Consideration of isolated-site disorder analysis and SAET led to the prediction of disordered form I as the energetically most favorable. The appearance of the disordered form II was attributed to nonequilibrium effects (frozen-in crystallization). For gandotinib (153), the experimental study revealed a polymorph-rich landscape. The CSP study on neat polymorphs predicted an undiscovered form to be the lowest lattice energy structure. Form I matched a highly populated family of similar structures, whereas form II was not found in the CSP search. Consideration of vibrational free energies, at the PBE(0)-MBD+F_{vib} level, and SAET demonstrated that the structure predicted to be the most stable is not considerably more stable than form I. The case of gandotinib highlights the importance of considering configurational free energy in disordered systems, and illustrates some of the continuing computational and experimental challenges in the exploration of solid-form landscapes.

6. BUILDING AN EFFECTIVE CRYSTAL STRUCTURE PREDICTION METHODOLOGY

Most current systematic approaches to CSP for organic molecules are built around the overall scheme presented in **Figure 1** by combining selected methods among those discussed in Sections 3–5. **Table 1** lists several leading programs designed to generate exhaustive lists of candidate structures at the first stage of the CSP methodology. With the exception of XtalPi's recently developed approach (23), all of these programs were applied with various degrees of success in the most recent blind test (13). As indicated in the table, the distinguishing characteristics of these programs are primarily (*a*) the search method used to find low-energy structures, (*b*) the type of energy model used to construct the energy landscape, and (*c*) the way in which molecular flexibility is incorporated into the search.

As shown in **Table 1**, each of the search algorithms discussed in Section 3 is employed in at least one code. The computational cost of these methods tends to be very system dependent (e.g.,

Table 1 Existing programs used in the candidate generation stage of crystal structure prediction for organic molecules in recent studies

Candidate generation program	Search method	Energy model	Handling of molecular flexibility
CrystalPredictor I (24, 26)	QR	Isolated-molecule QM Restricted Hermite interpolants Atomic/off-atom charges Buckingham potential	Selected torsions (potentially partitioned in semi-independent torsion groups)
CrystalPredictor II (49, 55, 56)	QR	Isolated-molecule QM LAMs Atomic charges Buckingham potential	Selected torsions (all other molecular conformation variables adjusted via LAMs)
GLEE (27)/DMACRYS (123)	QR	Isolated-molecule QM Distributed multipoles Buckingham potential	Rigid searches for selected conformers
GRACE (22)	MC	TMFF (22)	All variables
Genarris (154)/GAtor (155)	EA	Periodic DFT-D [FHI-aims (98)]	All variables
XtalPi (23)	PSO/MC	Tailored FF (23)	All variables

Abbreviations: CSP, crystal structure prediction; DFT-D, dispersion-corrected density functional theory; EA, evolutionary algorithm; FF, force field; LAM, local approximate model; MC, Monte Carlo; QM, quantum mechanical; QR, quasi-random (Sobol') search; PSO, particle swarm optimization; TMFF, tailor-made force field.

the degree of flexibility or the value of Z'); the main distinguishing characteristics are the speed of convergence and the reliability with which all relevant structures are sampled. Through the use of the CrystalPredictor I and II programs, which utilize quasi-random search, 10^5–10^7 local minimizations were performed for each system in the sixth blind test. For the GRACE program, which makes use of Monte Carlo parallel tempering, of the order of $>10^7$ trials or energy evaluations were carried out for flexible molecules in the sixth blind test, as discussed in the supplementary information provided in Reference 13. The quasi-random search used in GLEE has been found to reach convergence with respect to the number of unique structures generated within a few thousand minimizations per space group (27) for a rigid search. The XtalPi methodology employs two complementary search techniques—a Monte Carlo method that aims to provide a reliable coverage of the landscape and a particle swarm algorithm with fast, but potentially premature, convergence—in an attempt to ensure that polymorphs are not missed due to the biases of one particular method (23). Publications available in the open literature on more recent approaches, such as Genarris/GAtor and XtalPi, do not yet provide detailed information on how many structure calculations are required for convergence of the candidate generation.

The energy models used in candidate generation are predominantly force field methods, as described in Section 4. This is because, once constructed, energy evaluations with these force fields are often cheap enough to enable extensive exploration of the lattice energy landscape. A notable exception is GAtor (155), which relies on DFT-D calculations with low numerical tolerances to evaluate and minimize the lattice energy. Notwithstanding the use of loose numerical tolerances, the cost of these calculations typically far exceeds that of their force field counterparts, which may cause difficulties when attempting to undertake a global search.

The global search algorithms also differ in their treatment of conformational flexibility. Most modern codes handle flexibility to some extent by including flexible torsions as optimization variables. This is the case for CrystalPredictor, GRACE, and XtalPi. In GLEE/DMACRYS, however, all conformational variables are fixed during the local search, and instead one relies on multiple fixed-conformer searches to explore the extent of flexibility. This is less computationally demanding and often effective, but it can sometimes cause some conformers to be erroneously dismissed as a result of their large intramolecular energy, particularly for systems containing intramolecular hydrogen bonds (19). In the case of GAtor, flexibility is included by default as a result of the use of a DFT energy model. Nevertheless, the extent of the exploration of the conformational space is dependent on the pool of structures used to initialize the evolutionary algorithm; the pool is generated using Genarris (154), which relies on rigid conformer selection. Furthermore, in its current state, GAtor is designed primarily to investigate molecules that either are rigid or exhibit a mild degree of flexibility (e.g., target XXII of the sixth blind test, which may incur some bending in its central six-membered ring) (119, 155).

Once a list of initial candidate structures has been generated, the multistage CSP methodology proceeds to one or more refinement stages. **Table 2** lists some of the currently available refinement programs. DMACRYS (123) uses a force field with multipole analysis to minimize the intermolecular component of the lattice energy with respect to the unit cell variables. The molecular conformation is kept fixed at this stage, allowing many structures to be rapidly evaluated. CrystalOptimizer (45) is a force field–based program that relies on a combination of accurate intramolecular energy and semiempirical intermolecular interactions. The use of LAMs in CrystalOptimizer, and the reduction of the number of calculations through the storage of LAMs in database, means that the program can be used to refine on the order of 1,000–2,000 structures within much more modest computational times than periodic DFT.

Most other refinement codes are general periodic DFT packages, such as VASP (96), CRYSTAL (97), and FHI-aims (98). These packages can both optimize crystal geometries and

Table 2 Existing programs used in refinement stages of crystal structure prediction for organic molecules

Refinement stages	Energy model type	Handling of molecular flexibility
DMACRYS (123)	Distributed multipoles Buckingham potential	Rigid molecules only
CrystalOptimizer (45)	Isolated-molecule QM LAMs Distributed multipoles Buckingham potential	Selected torsions and bond angles and lengths (all other molecular conformation variables adjusted via LAMs)
VASP (96)	Periodic DFT-D	All variables
FHI-aims (98)	Periodic DFT-D	All variables
CRYSTAL (97)	Periodic DFT-D/HF-3c	All variables
DFTB+ (156)	Periodic DFTB3-D	All variables

Abbreviations: DFT-D, dispersion-corrected density functional theory; DFTB, dispersion-corrected density functional tight binding; HF, Hartree-Fock; LAMs, local approximate models; QM, quantum mechanical.

perform free energy calculations. The use of a DFT package can impose quite strict limits on the number of structures that can be refined at this level; the balance between cost and reliability is often adjusted through the selection of numerical tolerances and the choice of functional. For example, in the sixth blind test (13), the Neumann group considered ~500 structures per system using loose numerical tolerances, but only 25 structures with tighter tolerances. More recently, larger numbers of DFT optimizations (up to ~3,000) have been undertaken in CSP investigations through the large-scale cloud-based implementation of XtalPi (23). Cheaper QM approaches such as HF-3c (112, 113) and DFTB-D (114, 115) are available in CRYSTAL and DFTB+ (156), respectively, and were used by the Grimme group in the sixth blind test (13).

Overall, CSP remains in a state of significant evolution. A variety of methodologies are still being explored in the literature, differing, for example, in the energy models used at each stage of the workflow and sometimes in the workflow structure itself. For instance, single-point energy calculations have been employed as an additional intermediate refinement step (135) or at the final stage in a CSP investigation (157). Also, sophisticated DFT hierarchies have been used to calculate both lattice and free energy contributions (13, 110). In contrast, the recent emergence of commercial software codes such as GRACE and XtalPi may be an indication that, notwithstanding the continuing developments in research, the technology may soon be able to yield commercially valuable results in the hands of sufficiently experienced industrial practitioners.

7. PERSPECTIVES

The emergence of systematic CSP methods has been driven by a combination of scientific curiosity, industrial need and the building of a community around the series of blind tests. Significant progress has been achieved in recent years (18, 19), with a notable increase in the size and complexity of molecules that can be studied. With rigid molecules containing more than 100 atoms (158) and flexible molecules with 10 flexible torsions (6) now within the scope of the available techniques, we have now reached the point where many pharmaceutically relevant molecules can be studied with a reasonable degree of confidence, opening the door to increasing use in industry.

At the heart of this progress has been the emergence of general CSP methodologies that can be applied to a wide range of systems without relying on substantial expert insight or intuition regarding the particular system under investigation. Most of the approaches that have been successful

in practice share a common multistage workflow consisting of an initial generation of candidate structures followed by successive refinements of an energy landscape. In this review, we have attempted to identify, compare, and contrast the different methods that have been used at each stage of this workflow. We have also considered the ways in which these individual methods have been combined to produce software codes for practical use. We hope that this analysis will stimulate further investigation into more efficient and effective codes.

On a more detailed level, it is clear that the initial candidate generation step is fundamental in the success of any CSP methodology: If a practically relevant crystal structure is missed at this stage, it will not be identified at any later one. Missing a structure may be the result of either failing to identify it, due to an incomplete exploration of the energy landscape, or identifying it with an energy that is so high that it cannot proceed to the subsequent refinement stages. Thus, advances in search methodologies and the incorporation of increasingly accurate energy models at the global search stage have been, and will continue to be, particularly important in this context. As more powerful computational resources become available, we should be able to perform more extensive searches while simultaneously increasing the sophistication of the energy models used at this initial stage (e.g., via improved force fields or cheap DFT methods). A systematic evaluation of the efficacy of different search methods toward the complete exploration of the energy landscape could also be very useful in this context.

At the other end of the multistage CSP methodologies considered in this review, the final structure assessment stage is equally important. Practical experience with currently available methodologies indicates that, even when they succeed in identifying all low-energy structures with a reasonable prediction of their geometries, the relative stability rankings of these structures are often incorrect. Therefore, it is important to perform a final assessment of these structures by using the most accurate energy models that are practically affordable. Such an assessment may take the form of a single-point evaluation of the energy of each structure followed by a final reranking. If computational cost permits, a better option would be a final refinement of each structure using the accurate energy model in the context of either an energy reminimization calculation or an MD simulation (93). Overall, this final step will determine a corrected structure geometry as well as the corresponding energy. Moreover, it is possible that refinements of two or more structures will produce the same final structure, thereby resulting in a simplification of the polymorphic landscape.

For this reason, it may be worth highlighting some promising advances on the horizon in the area of highly accurate energy models (159), even if their computational cost is currently prohibitive for use in CSP. Post-Hartree-Fock incremental fragmentation methods have been applied to systems such as benzene (160, 161), *para*-diiodobenzene (162), urea, and hexamine (163). Other approaches, such as the hybrid many-body interaction fragmentation model, have shown promise with estimated accuracies of approximately 1–2 kJ/mol (164, 165) and have also been used in the prediction of the solid-state phase diagram of methanol (166). Another class of highly accurate models that are also applicable to solid-state systems consists of those based on quantum Monte Carlo approaches (167), in particular fixed-node diffusive Monte Carlo (168). Due to recent reductions in computational cost, diffusive Monte Carlo has been applied to systems of the size of naphthalene and anthracene by use of single-point calculations, but it has the potential to tackle much larger systems (168). Uncertainties are estimated to be approximately 1 kJ/mol (169).

Beyond accurate lattice energy computations, entropic and ZPE contributions are important in determining polymorph stability (14, 131). In addition, free energies of the crystalline phase are a prerequisite for establishing a link between crystal structure and practically important properties including phase diagrams of pure substances (170), solid–solid transitions, solubility (141), and specific heat capacities (14). Free energy considerations, as well as the characterization of the effects of disorder, will be the focus of significant attention in CSP over the next decade.

Overall, it is encouraging that much greater accuracy can now be achieved in the context of single-structure evaluations of lattice or free energy. However, continued creativity in models, algorithms, and the use of high-performance computational resources will be required to translate this progress into significant advances in CSP, which requires such accurate energy predictions for very large numbers of crystal structures. A key consideration in this context is the ability to derive surrogate computationally efficient models that can match the predictions of much more expensive ones for the system of interest over a restricted range of their inputs. Additionally, the deployment of such models within mathematical optimization frameworks may require that they possess properties such as continuity and differentiability, and that their computation be free of numerical noise. The development of surrogate models, such as LAMs and TMFFs derived from isolated-molecule QM models or periodic crystal DFT-D models, has been a major enabling factor extending CSP to systems of practical significance over the past decade. The application of systematic machine learning techniques may also provide new impetus in this area (171).

Finally, we note that the recent theoretical developments in CSP have been complemented by an increasing quantity of accessible experimental data, with the CSD now containing more than one million crystal structures (16, 148). As illustrated by the blind tests, the ability to compare the final results of CSP studies with accurate experimental information is essential for assessing the current state of CSP technology and motivating further research. However, to date there has been relatively little emphasis on systematic techniques for integrating experimental information within the CSP workflow itself. As discussed in Section 4.1.2.3, one area that has received some attention in this context is that of the estimation of parameters for repulsive/dispersive potentials for atom–atom interactions from experimental data, typically taken from the CSD. However, even in that case, the emphasis has been on the generation of tables of transferable parameter values that can then be used universally across all CSP studies with no further modification.

Arguably, what is needed are more nuanced techniques which are capable of taking account of experimental information that is more directly relevant to the specific system under consideration, for instance, relating to already resolved solid forms of the same or similar molecules. Because such system-specific information is unlikely to be available in sufficient amounts to allow the reliable estimation of all relevant parameters, it may need to be combined within a Bayesian estimation framework with prior information in the form of transferable parameter values. The development of reliable and efficient algorithms and codes that could be routinely used for the estimation of energy model parameters from large numbers of experimental crystal structures would be an important advance in this context.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

1. Price SL, Braun DE, Reutzel-Edens SM. 2016. Can computed crystal energy landscapes help understand pharmaceutical solids? *Chem. Commun.* 52:7065–77
2. Braun DE, Bhardwaj RM, Florence AJ, Tocher DA, Price SL. 2012. Complex polymorphic system of gallic acid—five monohydrates, three anhydrates, and over 20 solvates. *Cryst. Growth Des.* 13:19–23
3. Habermehl S, Mörschel P, Eisenbrandt P, Hammer SM, Schmidt MU. 2014. Structure determination from powder data without prior indexing, using a similarity measure based on cross-correlation functions. *Acta Crystallogr. B* 70:347–59

4. Baias M, Dumez JN, Svensson PH, Schantz S, Day GM, Emsley L. 2013. *De novo* determination of the crystal structure of a large drug molecule by crystal structure prediction–based powder NMR crystallography. *J. Am. Chem. Soc.* 135:17501–7
5. Arlin JB, Price LS, Price SL, Florence AJ. 2011. A strategy for producing predicted polymorphs: catemeric carbamazepine form V. *Chem. Commun.* 47:7074–76
6. Neumann M, van de Streek J, Fabbiani F, Hidber P, Grassmann O. 2015. Combined crystal structure prediction and high-pressure crystallization in rational pharmaceutical polymorph screening. *Nat. Commun.* 6:7793
7. Abramov YA. 2012. Current computational approaches to support pharmaceutical solid form selection. *Org. Process Res. Dev.* 17:472–85
8. Chemburkar SR, Bauer J, Deming K, Spiwek H, Patel K, et al. 2000. Dealing with the impact of ritonavir polymorphs on the late stages of bulk drug process development. *Org. Process Res. Dev.* 4:413–17
9. Slater AG, Reiss PS, Pulido A, Little MA, Holden DL, et al. 2017. Computationally-guided synthetic control over pore size in isostructural porous organic cages. *ACS Cent. Sci.* 3:734–42
10. Pulido A, Chen L, Kaczorowski T, Holden D, Little MA, et al. 2017. Functional materials discovery using energy–structure–function maps. *Nature* 543:657–64
11. Rice B, LeBlanc LM, Otero-de-la Roza A, Fuchter MJ, Johnson ER, et al. 2018. A computational exploration of the crystal energy and charge-carrier mobility landscapes of the chiral [6]helicene molecule. *Nanoscale* 10:1865–76
12. Yang J, De S, Campbell JE, Li S, Ceriotti M, Day GM. 2018. Large-scale computational screening of molecular organic semiconductors using crystal structure prediction. *Chem. Mater.* 30:4361–71
13. Reilly AM, Cooper RI, Adjiman CS, Bhattacharya S, Boese AD, et al. 2016. Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallogr. B* 72:439–59
14. Nyman J, Day GM. 2015. Static and lattice vibrational energy differences between polymorphs. *CrystEngComm* 17:5154–65
15. Price SL. 2013. Why don't we find more polymorphs? *Acta Crystallogr. B* 69:313–28
16. Groom CR, Bruno IJ, Lightfoot MP, Ward SC. 2016. The Cambridge Structural Database. *Acta Crystallogr. B* 72:171–79
17. Cruz-Cabeza AJ, Reutzel-Edens SM, Bernstein J. 2015. Facts and fictions about polymorphism. *Chem. Soc. Rev.* 44:8619–35
18. Price SL. 2018. Is zeroth order crystal structure prediction (CSP_0) coming to maturity? What should we aim for in an ideal crystal structure prediction code? *Faraday Discuss.* 211:9–30
19. Nyman J, Reutzel-Edens SM. 2018. Crystal structure prediction is changing from basic science to applied technology. *Faraday Discuss.* 211:459–76
20. Cambridge Crystallogr. Data Cent. 2020. *CSD space group statistics: space group frequency ordering*. Data Table, Cambridge Crystallogr. Data Cent., Cambridge, UK. **https://www.ccdc.cam.ac.uk/support-and-resources/ccdcresources/cb57878568114279913f49c22d5958fc.pdf**
21. Vasileiadis M, Pantelides CC, Adjiman CS. 2015. Prediction of the crystal structures of axitinib, a polymorphic pharmaceutical molecule. *Chem. Eng. Sci.* 121:60–76
22. Neumann MA. 2008. Tailor-made force fields for crystal-structure prediction. *J. Phys. Chem. B* 112:9810–29
23. Zhang P, Wood GP, Ma J, Yang M, Liu Y, et al. 2018. Harnessing cloud architecture for crystal structure prediction calculations. *Cryst. Growth Des.* 18:6891–900
24. Karamertzanis PG, Pantelides CC. 2005. *Ab initio* crystal structure prediction. I. Rigid molecules. *J. Comput. Chem.* 26:304–24
25. Sobol' IM. 1967. On the distribution of points in a cube and the approximate evaluation of integrals. *Z. Vychisl. Mat. Mat. Fiz.* 7:784–802
26. Karamertzanis P, Pantelides C. 2007. *Ab initio* crystal structure prediction. II. Flexible molecules. *Mol. Phys.* 105:273–91
27. Case DH, Campbell JE, Bygrave PJ, Day GM. 2016. Convergence properties of crystal structure prediction by quasi-random sampling. *J. Chem. Theory Comput.* 12:910–24
28. Kirkpatrick S, Gelatt CD, Vecchi MP. 1983. Optimization by simulated annealing. *Science* 220:671–80

29. van Laarhoven PJ, Aarts EH. 1987. *Simulated Annealing: Theory and Applications*. Dordrecht, Neth.: Reidel
30. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21:1087–92
31. Hukushima K, Nemoto K. 1996. Exchange Monte Carlo method and application to spin glass simulations. *J. Phys. Soc. Jpn.* 65:1604–8
32. Swendsen RH, Wang JS. 1986. Replica Monte Carlo simulation of spin-glasses. *Phys. Rev. Lett.* 57:2607–9
33. Kendrick J, Leusen FJ, Neumann MA, van de Streek J. 2011. Progress in crystal structure prediction. *Chem. A Eur. J.* 17:10736–44
34. Holland JH. 1992. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Cambridge, MA: MIT Press
35. Paszkowicz W. 2009. Genetic algorithms, a nature-inspired tool: survey of applications in materials science and related fields. *Mater. Manuf. Process.* 24:174–97
36. Woodley SM. 2004. Prediction of crystal structures using evolutionary algorithms and related techniques. In *Applications of Evolutionary Computation in Chemistry*, ed. RL Johnston, pp. 95–132. Berlin: Springer
37. Johnston RL. 2003. Evolving better nanoparticles: genetic algorithms for optimising cluster geometries. *Dalton Trans.* 22:4193–207
38. Turner GW, Tedesco E, Harris KD, Johnston RL, Kariuki BM. 2000. Implementation of Lamarckian concepts in a genetic algorithm for structure solution from powder diffraction data. *Chem. Phys. Lett.* 321:183–90
39. Curtis F, Ros T, Marom N. 2018. Evolutionary niching in the GAtor genetic algorithm for molecular crystal structure prediction. *Faraday Discuss.* 211:61–77
40. Kennedy J, Eberhart RC. 1997. A discrete binary version of the particle swarm algorithm. In *IEEE International Conference on Systems, Man, and Cybernetics: Computational Cybernetics and Simulation*, Vol. 5, pp. 4104–8. Piscataway, NJ: IEEE
41. Day GM, Motherwell WS, Jones W. 2007. A strategy for predicting the crystal structures of flexible molecules: the polymorphism of phenobarbital. *Phys. Chem. Chem. Phys.* 9:1693–704
42. Cruz-Cabeza AJ, Bernstein J. 2013. Conformational polymorphism. *Chem. Rev.* 114:2170–91
43. Reilly AM, Tkatchenko A. 2014. Role of dispersion interactions in the polymorphism and entropic stabilization of the aspirin crystal. *Phys. Rev. Lett.* 113:055701
44. Brandenburg JG, Grimme S. 2016. Organic crystal polymorphism: a benchmark for dispersion-corrected mean-field electronic structure methods. *Acta Crystallogr. B* 72:502–13
45. Kazantsev AV, Karamertzanis PG, Adjiman CS, Pantelides CC. 2011. Efficient handling of molecular flexibility in lattice energy minimization of organic crystals. *J. Chem. Theory Comput.* 7:1998–2016
46. Mayo SL, Olafson BD, Goddard WA. 1990. DREIDING: a generic force field for molecular simulations. *J. Phys. Chem.* 94:8897–909
47. Motherwell WS, Ammon HL, Dunitz JD, Dzyabchenko A, Erk P, et al. 2002. Crystal structure prediction of small organic molecules: a second blind test. *Acta Crystallogr. B* 58:647–61
48. Kazantsev AV, Karamertzanis PG, Adjiman CS, Pantelides CC, Price SL, et al. 2011. Successful prediction of a model pharmaceutical in the fifth blind test of crystal structure prediction. *Int. J. Pharm.* 418:168–78
49. Habgood M, Sugden IJ, Kazantsev AV, Adjiman CS, Pantelides CC. 2015. Efficient handling of molecular flexibility in ab initio generation of crystal structures. *J. Chem. Theory Comput.* 11:1957–69
50. Mennucci B, Tomasi J. 1997. Continuum solvation models: a new approach to the problem of solute's charge distribution and cavity boundaries. *J. Chem. Phys.* 106:5151–58
51. Cooper TG, Hejczyk KE, Jones W, Day GM. 2008. Molecular polarization effects on the relative energies of the real and putative crystal structures of valine. *J. Chem. Theory Comput.* 4:1795–805
52. Gelbrich T, Braun DE, Ellern A, Griesser UJ. 2013. Four polymorphs of methyl paraben: structural relationships and relative energy differences. *Cryst. Growth Des.* 13:1206–17
53. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, et al. 2010. Gaussian 09, Revis. C.01. *Software Package*. **http://gaussian.com/glossary/g09/**

54. van Eijck BP, Mooij W, Kroon J. 2001. *Ab initio* crystal structure predictions for flexible hydrogen-bonded molecules. Part II. Accurate energy minimization. *J. Comput. Chem.* 22:805–15
55. Sugden I, Adjiman CS, Pantelides CC. 2016. Accurate and efficient representation of intramolecular energy in ab initio generation of crystal structures. I. Adaptive local approximate models. *Acta Crystallogr. B* 72:864–74
56. Sugden IJ, Adjiman CS, Pantelides CC. 2019. Accurate and efficient representation of intramolecular energy in ab initio generation of crystal structures. II. Smoothed intramolecular potentials. *Acta Crystallogr. B* 75:423–33
57. Stone A. 2013. *The Theory of Intermolecular Forces*. Oxford, UK: Oxford Univ. Press
58. Bader RF. 1985. Atoms in molecules. *Acc. Chem. Res.* 18:9–15
59. Mulliken RS. 1955. Electronic population analysis on LCAO–MO molecular wave functions. I. *J. Chem. Phys.* 23:1833–40
60. Breneman CM, Wiberg KB. 1990. Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *J. Comput. Chem.* 11:361–73
61. Hu H, Lu Z, Yang W. 2007. Fitting molecular electrostatic potentials from quantum mechanical calculations. *J. Chem. Theory Comput.* 3:1004–13
62. Karamertzanis PG, Pantelides CC. 2004. Optimal site charge models for molecular electrostatic potentials. *Mol. Simul.* 30:413–36
63. Gavezzotti A. 2002. Calculation of intermolecular interaction energies by direct numerical integration over electron densities. I. Electrostatic and polarization energies in molecular crystals. *J. Phys. Chem. B* 106:4145–54
64. Day GM, Motherwell WDS, Jones W. 2005. Beyond the isotropic atom model in crystal structure prediction of rigid molecules: atomic multipoles versus point charges. *Cryst. Growth Des.* 5:1023–33
65. Stone AJ, Alderton M. 1985. Distributed multipole analysis: methods and applications. *Mol. Phys.* 56:1047–64
66. Stone AJ. 2005. Distributed multipole analysis: stability for large basis sets. *J. Chem. Theory Comput.* 1:1128–32
67. Lillestolen TC, Wheatley RJ. 2008. Redefining the atom: atomic charge densities produced by an iterative stockholder approach. *Chem. Commun.* 2008:5909–11
68. Misquitta AJ, Stone AJ, Fazeli F. 2014. Distributed multipoles from a robust basis-space implementation of the iterated stockholder atoms procedure. *J. Chem. Theory Comput.* 10:5405–18
69. Bultinck P, Cooper DL, Van Neck D. 2009. Comparison of the Hirshfeld-I and iterated stockholder atoms in molecules schemes. *Phys. Chem. Chem. Phys.* 11:3424–29
70. Elking DM, Perera L, Pedersen LG. 2012. HPAM: Hirshfeld partitioned atomic multipoles. *Comput. Phys. Commun.* 183:390–97
71. Welch GWA, Karamertzanis PG, Misquitta AJ, Stone AJ, Price SL. 2008. Is the induction energy important for modeling organic crystals? *J. Chem. Theory Comput.* 4:522–32
72. Aina AA, Misquitta AJ, Price SL. 2017. From dimers to the solid state: distributed intermolecular forcefields for pyridine. *J. Chem. Phys.* 147:161722
73. Misquitta AJ, Stone AJ. 2008. Accurate induction energies for small organic molecules. 1. Theory. *J. Chem. Theory Comput.* 4:7–18
74. Misquitta AJ, Stone AJ, Price SL. 2008. Accurate induction energies for small organic molecules. 2. Development and testing of distributed polarizability models against SAPT(DFT) energies. *J. Chem. Theory Comput.* 4:19–32
75. de Klerk NJJ, van den Ende JA, Bylsma R, Grančič P, de Wijs GA, et al. 2016. q-GRID: a new method to calculate lattice and interaction energies for molecular crystals from electron densities. *Cryst. Growth Des.* 16:662–71
76. Chickos JS. 2003. Enthalpies of sublimation after a century of measurement. *Netsu Sokutei* 30:116–24
77. Williams DE, Cox SR. 1984. Nonbonded potentials for azahydrocarbons: the importance of the Coulombic interaction. *Acta Crystallogr. B* 40:404–17
78. Coombes DS, Price SL, Willock DJ, Leslie M. 1996. Role of electrostatic interactions in determining the crystal structures of polar organic molecules. A distributed multipole study. *J. Phys. Chem.* 100:7352–60

79. Beyer T, Price SL. 2000. Dimer or catemer? Low-energy crystal packings for small carboxylic acids. *J. Phys. Chem. B* 104:2647–55
80. Cox SR, Hsu LY, Williams DE. 1981. Nonbonded potential function models for crystalline oxohydrocarbons. *Acta Crystallogr. A* 37:293–301
81. Williams DE, Houpt DJ. 1986. Fluorine nonbonded potential parameters derived from crystalline perfluorocarbons. *Acta Crystallogr. B* 42:286–95
82. Hsu LY, Williams DE. 1980. Intermolecular potential-function models for crystalline perchlorohydrocarbons. *Acta Crystallogr. A* 36:277–81
83. Williams DE. 2001. Improved intermolecular force field for molecules containing H, C, N, and O atoms, with application to nucleoside and peptide crystals. *J. Comput. Chem.* 22:1154–66
84. Pyzer-Knapp EO, Thompson HPG, Day GM. 2016. An optimized intermolecular force field for hydrogen-bonded organic molecular crystals using atomic multipole electrostatics. *Acta Crystallogr. B* 72:477–87
85. Gatsiou CA, Adjiman CS, Pantelides CC. 2018. Repulsion–dispersion parameters for the modelling of organic molecular crystals containing N, O, S and Cl. *Faraday Discuss.* 211:297–323
86. Pantelides CC, Adjiman CS, Kazantsev AV. 2014. General computational algorithms for ab initio crystal structure prediction for organic molecules. In *Prediction and Calculation of Crystal Structures*, ed. S Atahan-Evrenk, A Aspuru-Guzik, pp. 22–58. Berlin: Springer
87. Bowskill DH, Sugden IJ, George N, Keates A, Webb J, et al. 2020. Efficient parameterization of a surrogate model of molecular interactions in crystals. *Comput. Aided Process Eng.* 48:493–98
88. Tang KT, Toennies JP. 1984. An improved simple model for the van der Waals potential based on universal damping functions for the dispersion coefficients. *J. Chem. Phys.* 80:3726–41
89. Axilrod BM, Teller E. 1943. Interaction of the van der Waals type between three atoms. *J. Chem. Phys.* 11:299–300
90. von Lilienfeld OA, Tkatchenko A. 2010. Two- and three-body interatomic dispersion energy contributions to binding in molecules and solids. *J. Chem. Phys.* 132:234109
91. Misquitta AJ, Stone AJ. 2016. *Ab initio* atom–atom potentials using CamCASP: theory and application to many-body models for the pyridine dimer. *J. Chem. Theory Comput.* 12:4184–208
92. Podeszwa R, Bukowski R, Rice BM, Szalewicz K. 2007. Potential energy surface for cyclotrimethylene trinitramine dimer from symmetry-adapted perturbation theory. *Phys. Chem. Chem. Phys.* 9:5561–69
93. Podeszwa R, Rice BM, Szalewicz K. 2009. Crystal structure prediction for cyclotrimethylene trinitramine (RDX) from first principles. *Phys. Chem. Chem. Phys.* 11:5512–18
94. Misquitta AJ, Welch GWA, Stone AJ, Price SL. 2008. A first principles prediction of the crystal structure of $C_6Br_2ClFH_2$. *Chem. Phys. Lett.* 456:105–9
95. Day GM, Cooper TG, Cruz-Cabeza AJ, Hejczyk KE, Ammon HL, et al. 2009. Significant progress in predicting the crystal structures of small organic molecules—a report on the fourth blind test. *Acta Crystallogr. B* 65:107–25
96. Hafner J. 2008. *Ab-initio* simulations of materials using VASP: density-functional theory and beyond. *J. Comput. Chem.* 29:2044–78
97. Dovesi R, Orlando R, Erba A, Zicovich-Wilson CM, Civalleri B, et al. 2014. CRYSTAL14: a program for the ab initio investigation of crystalline solids. *Int. J. Quantum Chem.* 114:1287–317
98. Blum V, Gehrke R, Hanke F, Havu P, Havu V, et al. 2009. *Ab initio* molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* 180:2175–96
99. Perdew JP, Schmidt K. 2001. Jacob's ladder of density functional approximations for the exchange-correlation energy. *AIP Conf. Proc.* 577:1–20
100. LeBlanc LM, Dale SG, Taylor CR, Becke AD, Day GM, Johnson ER. 2018. Pervasive delocalisation error causes spurious proton transfer in organic acid–base co-crystals. *Angew. Chem.* 130:15122–26
101. Otero-De-La-Roza A, Johnson ER. 2012. A benchmark for non-covalent interactions in solids. *J. Chem. Phys.* 137:054103
102. Klimeš J, Michaelides A. 2012. Perspective: advances and challenges in treating van der Waals dispersion forces in density functional theory. *J. Chem. Phys.* 137:120901
103. Grimme S, Antony J, Ehrlich S, Krieg H. 2010. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H–Pu. *J. Chem. Phys.* 132:154104

104. Caldeweyher E, Bannwarth C, Grimme S. 2017. Extension of the D3 dispersion coefficient model. *J. Chem. Phys.* 147:034112
105. Becke AD, Johnson ER. 2007. Exchange-hole dipole moment and the dispersion interaction revisited. *J. Chem. Phys.* 127:154108
106. Tkatchenko A, DiStasio RA Jr., Car R, Scheffler M. 2012. Accurate and efficient method for many-body van der Waals interactions. *Phys. Rev. Lett.* 108:236402
107. Bučko T, Lebègue S, Ángyán JG, Hafner J. 2014. Extending the applicability of the Tkatchenko-Scheffler dispersion correction via iterative Hirshfeld partitioning. *J. Chem. Phys.* 141:034114
108. Marom N, DiStasio RA Jr., Atalla V, Levchenko S, Reilly AM, et al. 2013. Many-body dispersion interactions in molecular crystal polymorphism. *Angew. Chem.* 52:6629–32
109. Hoja J, Reilly AM, Tkatchenko A. 2017. First-principles modeling of molecular crystals: structures and stabilities, temperature and pressure. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 7:e1294
110. Hoja J, Tkatchenko A. 2018. First-principles stability ranking of molecular crystal polymorphs with the DFT+MBD approach. *Faraday Discuss.* 211:253–74
111. Hoja J, Ko HY, Neumann MA, Car R, DiStasio RA Jr., Tkatchenko A. 2019. Reliable and practical computational description of molecular crystal polymorphs. *Sci. Adv.* 5:eaau3338
112. Brandenburg JG, Grimme S. 2013. Dispersion corrected Hartree–Fock and density functional theory for organic crystal structure prediction. In *Prediction and Calculation of Crystal Structures*, ed. S Atahan-Evrenk, A Aspuru-Guzik, pp. 1–23. Berlin: Springer
113. Sure R, Grimme S. 2013. Corrected small basis set Hartree-Fock method for large systems. *J. Comput. Chem.* 34:1672–85
114. Brandenburg JG, Grimme S. 2014. Accurate modeling of organic molecular crystals by dispersion-corrected density functional tight binding (DFTB). *J. Phys. Chem. Lett.* 5:1785–89
115. Mortazavi M, Brandenburg JG, Maurer RJ, Tkatchenko A. 2018. Structure and stability of molecular crystals with many-body dispersion-inclusive density functional tight binding. *J. Phys. Chem. Lett.* 9:399–405
116. Iuzzolino L, McCabe P, Price SL, Brandenburg JG. 2018. Crystal structure prediction of flexible pharmaceutical-like molecules: density functional tight binding as an intermediate optimisation method and for free energy estimation. *Faraday Discuss.* 211:275–96
117. Nyman J, Yu L, Reutzel-Edens SM. 2019. Accuracy and reproducibility in crystal structure prediction: the curious case of ROY. *CrystEngComm* 21:2080–88
118. Wen S, Beran GJO. 2012. Accidental degeneracy in crystalline aspirin: new insights from high-level ab initio calculations. *Cryst. Growth Des.* 12:2169–72
119. Reilly AM, Tkatchenko A. 2013. 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
120. Perdew JP, Burke K, Ernzerhof M. 1996. Generalized gradient approximation made simple. *Phys. Rev. Lett.* 77:3865–68
121. Tao J, Perdew JP, Staroverov VN, Scuseria GE. 2003. Climbing the density functional ladder: nonempirical meta–generalized gradient approximation designed for molecules and solids. *Phys. Rev. Lett.* 91:146401
122. Moellmann J, Grimme S. 2014. DFT-D3 study of some molecular crystals. *J. Phys. Chem. C* 118:7615–21
123. Price SL, Leslie M, Welch GWA, Habgood M, Price LS, et al. 2010. Modelling organic crystal structures using distributed multipole and polarizability-based model intermolecular potentials. *Phys. Chem. Chem. Phys.* 12:8478–90
124. Tan M, Shtukenberg AG, Zhu S, Xu W, Dooryhee E, et al. 2018. ROY revisited, again: the eighth solved structure. *Faraday Discuss.* 211:477–91
125. Nyman J, Pundyke OS, Day GM. 2016. Accurate force fields and methods for modelling organic molecular crystals at finite temperatures. *Phys. Chem. Chem. Phys.* 18:15828–37
126. Born M, Huang K. 1954. *Dynamical Theory of Crystal Lattices*. Oxford, UK: Clarendon
127. Dove MT. 1993. *Introduction to Lattice Dynamics*. Cambridge, UK: Cambridge Univ. Press
128. Brandenburg JG, Potticary J, Sparkes HA, Price SL, Hall SR. 2017. Thermal expansion of carbamazepine: Systematic crystallographic measurements challenge quantum chemical calculations. *J. Phys. Chem. Lett.* 8:4319–24

129. McKinley JL, Beran GJO. 2018. Identifying pragmatic quasi-harmonic electronic structure approaches for modeling molecular crystal thermal expansion. *Faraday Discuss.* 211:181–207
130. Abraham NS, Shirts MR. 2018. Thermal gradient approach for the quasi-harmonic approximation and its application to improved treatment of anisotropic expansion. *J. Chem. Theory Comput.* 14:5904–19
131. Nyman J, Day GM. 2016. Modelling temperature-dependent properties of polymorphic organic molecular crystals. *Phys. Chem. Chem. Phys.* 18:31132–43
132. Abraham NS, Shirts MR. 2019. Adding anisotropy to the standard quasi-harmonic approximation still fails in several ways to capture organic crystal thermodynamics. *Cryst. Growth Des.* 19:6911–24
133. Monserrat B, Drummond ND, Needs RJ. 2013. Anharmonic vibrational properties in periodic systems: energy, electron–phonon coupling, and stress. *Phys. Rev. B* 87:144302
134. Nyman J. 2017. *Computational predictions of structures, inclusion behaviour and properties of organic molecular crystals*. PhD Thesis, Univ. Southampton, Southampton, UK
135. Vasileiadis MM. 2013. *Calculation of the free energy of crystalline solids*. PhD Thesis, Imperial College London, London, UK
136. Frenkel D, Smit B. 2001. *Understanding Molecular Simulation: From Algorithms to Applications*. Amsterdam: Elsevier. 2nd ed.
137. Dybeck EC, Abraham NS, Schieber NP, Shirts MR. 2017. Capturing entropic contributions to temperature-mediated polymorphic transformations through molecular modeling. *Cryst. Growth Des.* 17:1775–87
138. Yu TQ, Tuckerman ME. 2011. Temperature-accelerated method for exploring polymorphism in molecular crystals based on free energy. *Phys. Rev. Lett.* 107:015701
139. Schneider E, Vogt L, Tuckerman ME. 2016. Exploring polymorphism of benzene and naphthalene with free energy based enhanced molecular dynamics. *Acta Crystallogr. B* 72:542–50
140. Reilly AM, Habershon S, Morrison CA, Rankin DWH. 2010. Simulating thermal motion in crystalline phase-I ammonia. *J. Chem. Phys.* 132:134511
141. Schnieders MJ, Baltrusaitis J, Shi Y, Chattree G, Zheng L, et al. 2012. The structure, thermodynamics, and solubility of organic crystals from simulation with a polarizable force field. *J. Chem. Theory Comput.* 8:1721–36
142. Dybeck EC, Schieber NP, Shirts MR. 2016. Effects of a more accurate polarizable Hamiltonian on polymorph free energies computed efficiently by reweighting point-charge potentials. *J. Chem. Theory Comput.* 12:3491–505
143. Gray AE, Day GM, Leslie M, Price SL. 2004. Dynamics in crystals of rigid organic molecules: Contrasting the phonon frequencies calculated by molecular dynamics with harmonic lattice dynamics for imidazole and 5-azauracil. *Mol. Phys.* 102:1067–83
144. Laio A, Parrinello M. 2002. Escaping free-energy minima. *PNAS* 99:12562–66
145. Martoňák R, Laio A, Parrinello M. 2003. Predicting crystal structures: the Parrinello–Rahman method revisited. *Phys. Rev. Lett.* 90:075503
146. Karamertzanis PG, Raiteri P, Parrinello M, Leslie M, Price SL. 2008. The thermal stability of lattice-energy minima of 5-fluorouracil: metadynamics as an aid to polymorph prediction. *J. Phys. Chem. B* 112:4298–308
147. Zykova-Timan T, Raiteri P, Parrinello M. 2008. Investigating the polymorphism in PR179: a combined crystal structure prediction and metadynamics study. *J. Phys. Chem. B* 112:13231–37
148. Taylor R, Wood PA. 2019. A million crystal structures: The whole is greater than the sum of its parts. *Chem. Rev.* 119:9427–77
149. Cruz-Cabeza AJ, Day GM, Jones W. 2011. Structure prediction, disorder and dynamics in a DMSO solvate of carbamazepine. *Phys. Chem. Chem. Phys.* 13:12808–16
150. Habgood M. 2011. Form II caffeine: a case study for confirming and predicting disorder in organic crystals. *Cryst. Growth Des.* 11:3600–8
151. Habgood M, Grau-Crespo R, Price SL. 2011. Substitutional and orientational disorder in organic crystals: a symmetry-adapted ensemble model. *Phys. Chem. Chem. Phys.* 13:9590–600
152. Woollam GR, Neumann MA, Wagner T, Davey RJ. 2018. The importance of configurational disorder in crystal structure prediction: the case of loratadine. *Faraday Discuss.* 211:209–34

153. Braun DE, McMahon JA, Bhardwaj RM, Nyman J, Neumann MA, et al. 2019. Inconvenient truths about solid form landscapes revealed in the polymorphs and hydrates of gandotinib. *Cryst. Growth Des.* 19:2947–62
154. Li X, Curtis FS, Rose T, Schober C, Vazquez-Mayagoitia A, et al. 2018. Genarris: random generation of molecular crystal structures and fast screening with a Harris approximation. *J. Chem. Phys.* 148:241701
155. Curtis F, Li X, Rose T, Vázquez-Mayagoitia Á, Bhattacharya S, et al. 2018. GAtor: a first-principles genetic algorithm for molecular crystal structure prediction. *J. Chem. Theory Comput.* 14:2246–64
156. Aradi B, Hourahine B, Frauenheim T. 2007. DFTB+, a sparse matrix-based implementation of the DFTB method. *J. Phys. Chem. A* 111:5678–84
157. Bhardwaj RM, Price LS, Price SL, Reutzel-Edens SM, Miller GJ, et al. 2013. Exploring the experimental and computed crystal energy landscape of olanzapine. *Cryst. Growth Des.* 13:1602–17
158. McMahon DP, Stephenson A, Chong SY, Little MA, Jones JT, et al. 2018. Computational modelling of solvent effects in a prolific solvatomorphic porous organic cage. *Faraday Discuss.* 211:383–99
159. Beran GJO. 2016. Modeling polymorphic molecular crystals with electronic structure theory. *Chem. Rev.* 116:5567–613
160. Bludský O, Rubeš M, Soldán P. 2008. *Ab initio* investigation of intermolecular interactions in solid benzene. *Phys. Rev. B* 77:092103
161. Yang J, Hu W, Usvyat D, Matthews D, Schütz M, Chan GKL. 2014. *Ab initio* determination of the crystalline benzene lattice energy to sub-kilojoule/mole accuracy. *Science* 345:640–43
162. Taylor CR, Bygrave PJ, Hart JN, Allan NL, Manby FR. 2012. Improving density functional theory for crystal polymorph energetics. *Phys. Chem. Chem. Phys.* 14:7739–43
163. Tsuzuki S, Orita H, Honda K, Mikami M. 2010. First-principles lattice energy calculation of urea and hexamine crystals by a combination of periodic DFT and MP2 two-body interaction energy calculations. *J. Phys. Chem. B* 114:6799–805
164. Wen S, Beran GJO. 2011. Accurate molecular crystal lattice energies from a fragment QM/MM approach with on-the-fly ab initio force field parametrization. *J. Chem. Theory Comput.* 7:3733–42
165. Wen S, Nanda K, Huang Y, Beran GJO. 2012. Practical quantum mechanics–based fragment methods for predicting molecular crystal properties. *Phys. Chem. Chem. Phys.* 14:7578–90
166. Červinka C, Beran GJO. 2018. *Ab initio* prediction of the polymorph phase diagram for crystalline methanol. *Chem. Sci.* 9:4622–29
167. Foulkes WMC, Mitas L, Needs RJ, Rajagopal G. 2001. Quantum Monte Carlo simulations of solids. *Rev. Mod. Phys.* 73:33–83
168. Zen A, Brandenburg JG, Klimeš J, Tkatchenko A, Alfè D, Michaelides A. 2018. Fast and accurate quantum Monte Carlo for molecular crystals. *PNAS* 115:1724–29
169. Hongo K, Watson MA, Sánchez-Carrera RS, Iitaka T, Aspuru-Guzik A. 2010. Failure of conventional density functionals for the prediction of molecular crystal polymorphism: a quantum Monte Carlo study. *J. Phys. Chem. Lett.* 1:1789–94
170. Červinka C, Beran GJO. 2017. *Ab initio* thermodynamic properties and their uncertainties for crystalline α-methanol. *Phys. Chem. Chem. Phys.* 19:29940–53
171. Egorova O, Hafizi R, Woods DC, Day GM. 2020. Multi-fidelity statistical machine learning for molecular crystal structure prediction. chemRxiv 12407831. **https://doi.org/10.26434/chemrxiv.12407831.v1**