nature communications

Article

https://doi.org/10.1038/s41467-025-57479-1

A robust crystal structure prediction method to support small molecule drug development with large scale validation and blind study

Received: 11 January 2024

Accepted: 19 February 2025

Published online: 05 March 2025

Check for updates

Dong Zhou[1,6], Imanuel Bier[1], Biswajit Santra[1], Leif D. Jacobson[2], Chuanjie Wu[1], Adiran Garaizar Suarez[3], Barbara Ramirez Almaguer[4], Haoyu Yu[1,7], Robert Abel[1], Richard A. Friesner[5] & Lingle Wang[1]

Crystal polymorphism is an important and fascinating aspect of solid state chemistry with far reaching implications in the pharmaceuticals, agrisciences, nutraceuticals, battery and aviation industries. Late appearing more stable polymorphs have caused numerous issues in the pharmaceutical industry. Experimental polymorph screening can be very expensive and time consuming, and sometimes may miss important low energy polymorphs due to an inability to exhaust all crystallization conditions. In this paper, we report a crystal structure prediction (CSP) method with state of the art accuracy and efficiency, validated on a large and diverse dataset including 66 molecules with 137 experimentally known polymorphic forms. The method combines a novel systematic crystal packing search algorithm and the use of machine learning force fields in a hierarchical crystal energy ranking. Our method not only reproduces all the experimentally known polymorphs, but also suggests new low energy polymorphs yet to be discovered by experiment that might pose potential risks to development of the currently known forms of these compounds. In addition, we report the prediction results of a blinded study, results for Target XXXI from the seventh CSP blind test, and demonstrate how the method can be used to accelerate clinical formulation design and derisk downstream processing.

Crystal polymorphism, the existence of different crystal structures for the same chemical compound, is a common phenomenon in chemistry. Many compounds are known to form multiple polymorphs depending on the crystallization conditions. Different polymorphs can have different physical and chemical properties, such as density, melting point, hardness, color, stability, morphology, solubility, and bioavailability. Therefore, crystal polymorphism is an important and fascinating aspect of solid state chemistry with implications for various fields including pharmaceuticals, materials science (e.g., energetic materials, dyes and pigments, and organic electronics), and agriculture[1].

[1]Schrödinger Inc., New York: 1540 Broadway, 24th Floor, 10036 New York, NY, USA. [2]Schrödinger Inc., Portland: 101 SW Main Street, Suite 1300, 97204 Portland, OR, USA. [3]Bayer AG, Computational Life Science, Alfred-Nobel-Straße 50, 40789, 40789 Monheim am Rhein, Germany. [4]Bayer AG, Process Industrialization, Friedrich-Ebert-Str. 217, 42117 Wuppertal, Germany. [5]Department of Chemistry, Columbia University, New York, 10027 New York, USA. [6]Present address: Atommap Inc. 450 Lexington Avenue, 4th floor, 10017 New York, NY, USA. [7]Present address: ByteDance Inc., 151 w 42nd street, New York, NY 10036, USA. e-mail: Lingle.wang@schrodinger.com

Late-appearing polymorphs are crystalline forms that emerge unexpectedly after a long period of time or under altered production conditions. Such forms may lead to the inability to obtain a crystal form that was previously prepared for pharmaceutical or other applications, thus necessitate the redesign of the production process[2]. They can alter the solubility, bioavailability, stability, and dissolution rate of the active pharmaceutical ingredient (API), significantly impacting the quality, efficacy, and safety of pharmaceutical products. The pharmaceutical industry has suffered a few disastrous issues due to late-appearing polymorphs, leading to patent disputes, regulatory issues, and even market recalls, including the famous cases of ritonavir[3], rotigotine[4], and many others[2]. Therefore, it is crucial to identify and characterize all possible polymorphs of a given API and understand the factors that influence their formation and transformation.

The conventional process for designing a clinical formulation of a small molecule drug typically begins with experimental polymorph screening and scale-up studies. This process aims to identify and characterize the different polymorphs of the API and to select the most suitable one for development. However, this process can be time consuming and may miss some important low energy polymorphs due to the inability to exhaust all crystallization conditions. As such, inexhaustive polymorph screening poses serious challenges for drug development and manufacturing.

Computational polymorph prediction can complement experiments to de-risk unexpected polymorphic changes during drug development[5–9]. Unlike experiments, computational methods, in principle, can enable identification of all low-energy polymorphs of the API, including those that may not be easily accessible by conventional experimental methods, or that may only appear under specific isolation conditions. This can help avert discovery of new polymorphs in late stage development that could potentially affect the quality, efficacy, and safety of the drug product.

Motivated by the crystal structure prediction (CSP) blind test challenge organized by CCDC[10–13], the field of computational polymorph prediction has made large leaps in the past two decades[11–13]. Several studies have demonstrated the ability of computational methods to accurately predict the crystal structures of small molecules, including flexible molecules of comparable complexity to typical modern small molecule drugs[5,14–18]. These studies have provided valuable insights and examples for the field of computational polymorph screening. However, most of these previous CSP studies have only investigated a small number of molecules to demonstrate the potential of such calculations. A large-scale validation of the proposed methods for general small molecule crystal structure prediction has yet to be reported.

In this paper, we report a novel crystal structure prediction method and demonstrate its accuracy on a large set of diverse molecules. The method integrates a novel systematic approach to search the crystal packing parameters and a hierarchical energy ranking method that balances accuracy and cost (Fig. 1). The new packing search method uses a divide-and-conquer strategy to break down the parameter space into subspaces based on space group symmetries. Each subspace is then searched consecutively. The energy ranking method combines molecular dynamics (MD) simulations using a classical force field (FF), structure optimization and reranking using a machine learning force field (MLFF) with long range electrostatic and dispersion interactions[19], and periodic density functional theory (DFT) calculations for ranking the final shortlist. The temperature-dependent stability of different polymorphs is evaluated with free energy calculations using previously established methods[20,21].

Currently focused on searching crystal structures with one molecule in the asymmetric unit (ASU), corresponding to the $Z' = 1$ search space, the method was validated on a large set of 66 molecules with 137 unique crystal structures, including all relevant molecules from the first six CCDC CSP blind tests, Target XXXI from the seventh

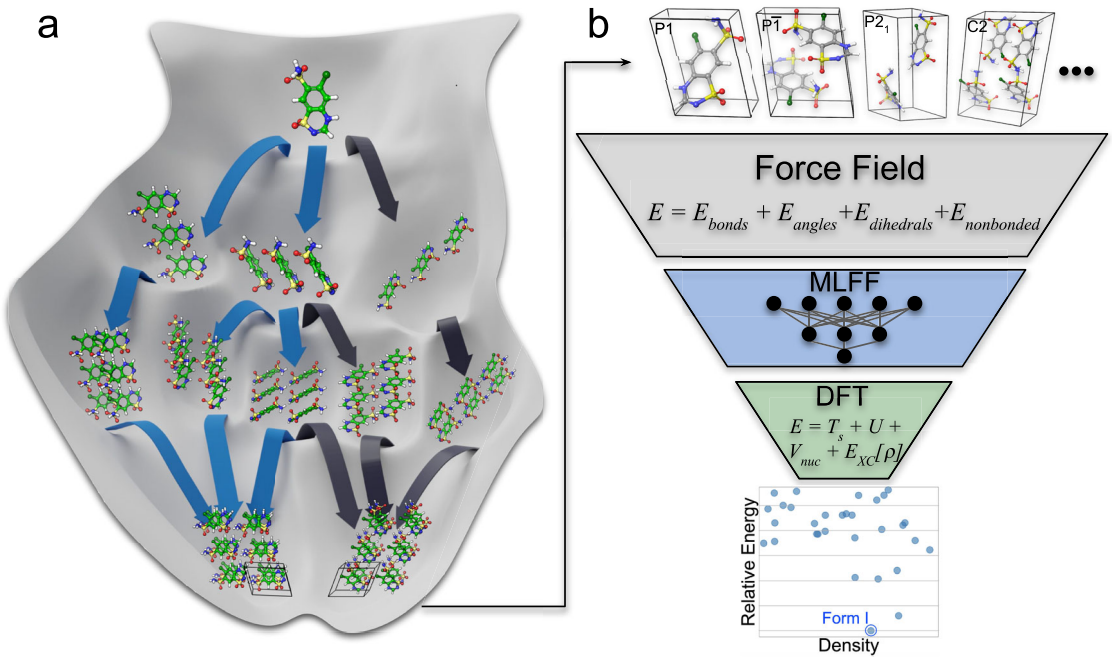

Fig. 1 | Overview of the computational polymorph prediction method. a The novel packing search method uses a divide-and-conquer strategy to break down the parameter space into subspaces based on space group symmetries. 3D candidate crystal structures are built step by step from low energy conformers to different sized clusters with favorable interactions following the funnel shaped energy landscape mimicking the crystal nucleation process. Each conformer can generate multiple candidate structures via different pathways (the entire sampling tree), and multiple pathways can lead to the same candidate structure (the blue colored paths generate the same candidate structure). **b** The energy ranking method integrates a multi-stage energy relaxation and filtering process with increasing accuracy and cost, including molecular dynamics simulations using a classical force field, structure optimization and reranking using a machine learning force field (MLFF) with long range electrostatic and dispersion interactions, and periodic density functional theory (DFT) calculations for ranking the final shortlist.

https://doi.org/10.1038/s41467-025-57479-1

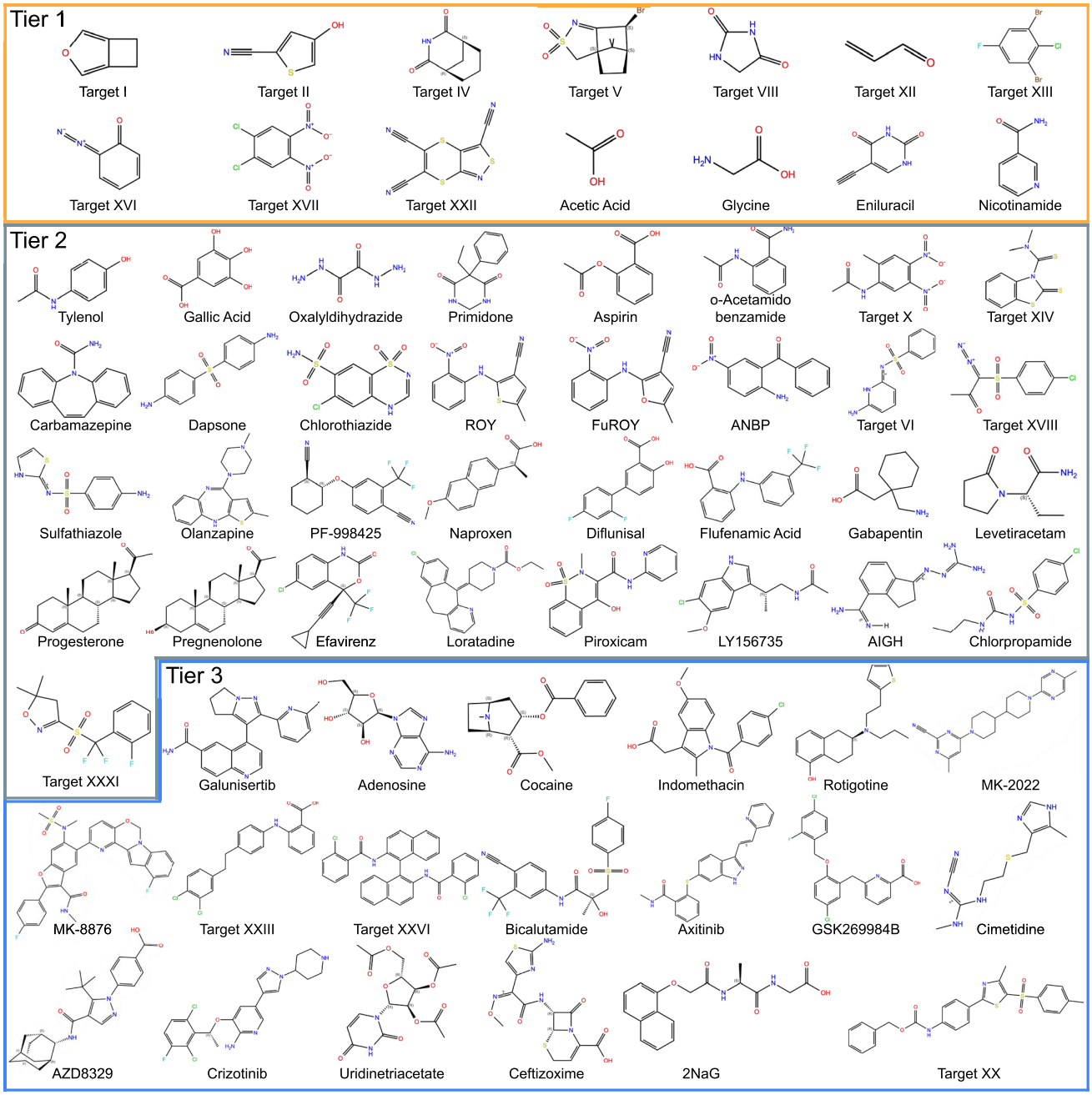

Fig. 2 | 2D diagram and common names of 66 molecules for retrospective validation of crystal structure prediction (CSP). A dataset of 66 molecules split into three tiers, orange for relatively rigid molecules, gray for small drug-like molecules with two to four rotatable bonds, and blue for larger drug-like molecules with five to ten rotatable bonds.

blind test, other molecules studied by previous CSP methods, and several molecules from modern drug discovery programs. For all the molecules in this large test set, the experimentally known polymorphs are correctly predicted by our method and are ranked among the top candidate structures. For several molecules, our prediction suggests new low energy polymorphs yet to be discovered by experiment, implying potential risks that could jeopardize the development of the currently known forms of these compounds. Comparisons to the results of other computational approaches are made when feasible.

We also report the accurate crystal structure prediction on an agrochemical molecule in a blinded study and discuss the limitations and future directions of our method for molecular crystal polymorph prediction.

Results

Comprehensive set of molecules for method validation

A comprehensive set of molecules for CSP method validation was compiled. The dataset is divided into three tiers following definitions established by previous CCDC CSP blind tests[13]. The first tier consists of mostly rigid molecules up to 30 atoms. The second tier consists of small drug-like molecules with around two to four rotatable bonds, and up to approximately 40 atoms. The third tier is composed of large drug-like molecules with five to ten rotatable bonds, usually containing 50 to 60 atoms. Additional tiers will be defined in the future as the complexity of CSP targets increase. The entire collection of test molecules is shown in Fig. 2.

The dataset construction began by including all of the $Z' = 1$ cases from the first six CCDC CSP blind tests with a few exceptions[13]. The

https://doi.org/10.1038/s41467-025-57479-1

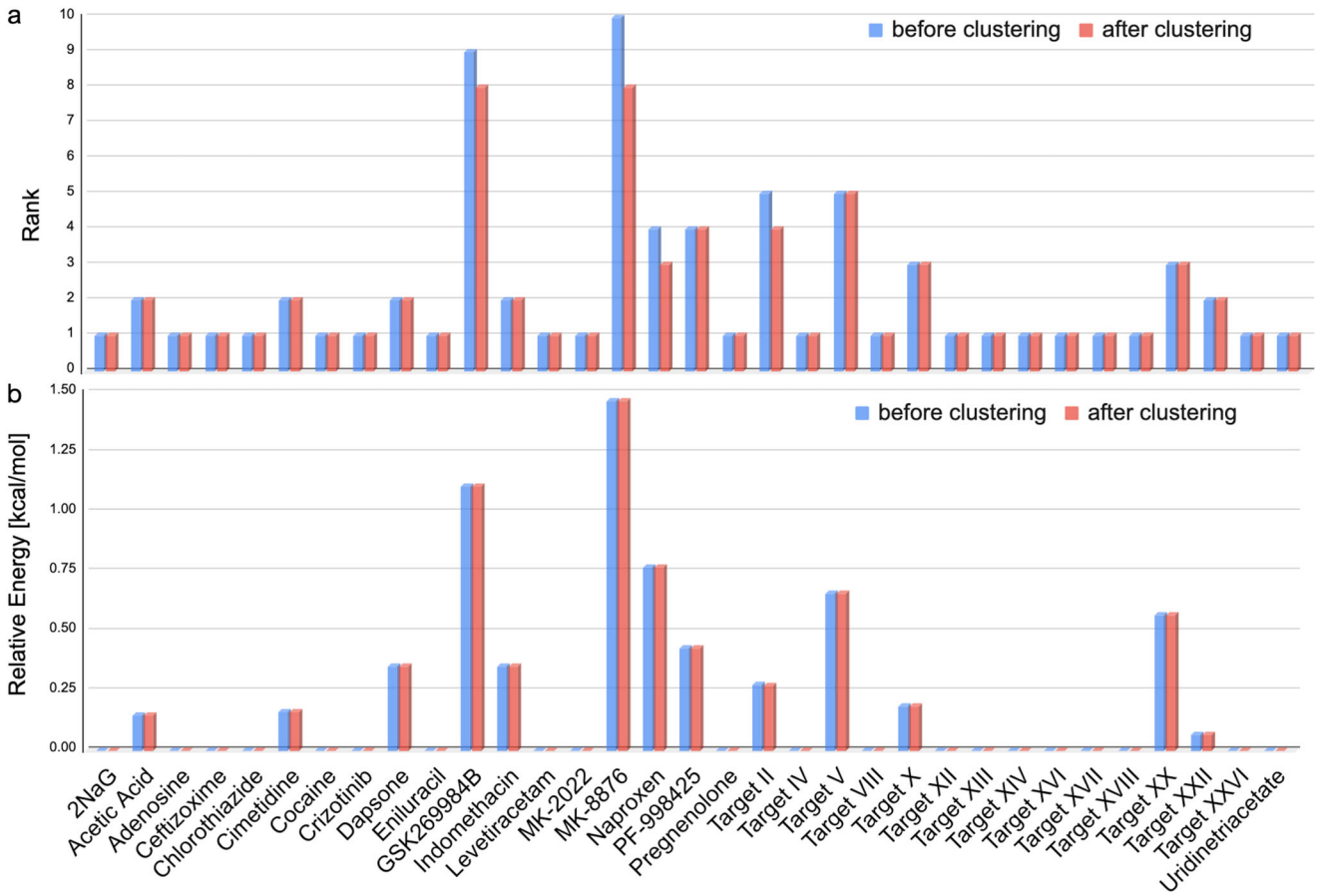

Fig. 3 | Summary of the polymorph energy landscape for a subset of 33 molecules each with one known Z′ = 1 experimental structure. All known Z′ = 1 crystal structures are sampled and ranked well from the crystal structure prediction (CSP) calculations. The rank of the predicted crystal structure that best matches each experimental crystal structure is shown in (**a**) and the relative energy is shown in (**b**). The blue and red bars in (**a**) and (**b**) represent the results before and after the clustering analysis to remove non-trivial duplicates. The results are obtained from r^{2}SCAN-D3 energy landscape computed on PBE-D3 optimized crystal structures. Source data are provided as a Source Data file.

excluded molecules were either not drug-like or contained elements not supported by the pretrained MLFF, the charge recursive neural network (QRNN)[19]. These were Target IX, which contains iodine and resembles an organic semiconductor, and Target III, which contains boron. Targets V and VIII were able to be included by replacing bromine atoms with chlorine, for the purposes of QRNN, which were then replaced back with bromine for DFT ranking. A total of 17 molecules from the first six CCDC CSP blind tests were included.

Added to the dataset were well studied molecules with many known Z′ = 1 polymorphs. Examples include ROY, Olanzapine, Galunisertib, Axitinib, Chlorpropamide, Flufenamic acid, and Piroxicam. Accurate prediction for these molecules requires CSP methods to produce a diversity of molecular packing solutions for each molecule and achieve accurate relative energy evaluations.

The dataset was further enriched by molecules with previously published CSP and experimental results with relevance to small-molecule pharmaceuticals. The final dataset, comprising 66 molecules, covers a diversity of functional groups, including polar groups such as amide, urea, pyridine, sulfonamide, hydroxyl, nitro, carboxylate, cyano, nonpolar groups such as phenyl and alkane chains, and various substituted aromatic and nonaromatic rings. The functional group diversity present in the dataset requires high accuracy of the energy models across chemical space, particularly as it applies to intra- and inter-molecular interactions relevant for molecular crystal packing.

The experimental crystal structures for the molecular crystal polymorphs were obtained from the CSD[10] with a few exceptions when the crystal structure data only existed in literature, such as Form D of Cimetidine. When multiple data entries exist for a polymorph in the CSD, the most reliable one was selected. The preference was as follows: neutron diffraction studies, followed by low temperature single-crystal X-ray diffraction (XRD), and room temperature powder X-ray diffraction (PXRD) studies were considered the least reliable. When all other experimental conditions were equal across multiple entries, the crystal structure with the smallest R-factor was used.

Retrospective validation on a comprehensive set of 66 molecules

Out of the 66 molecules in the test set, 33 molecules have only one experimentally known crystalline form with Z′ = 1. The remaining 33 molecules have multiple experimentally known polymorphs with Z′ = 1, including molecules with very complex polymorphic landscapes such as ROY and Galunisertib. In the following, we will first present the summary of the calculated polymorphic landscapes on these two subsets of molecules and then discuss the details on a few representative examples.

For the 33 molecules with one target crystalline form, Fig. 3a shows the final rankings of the predicted crystal structures that best match the known experimental structures using the r^{2}SCAN-D3 functional. In all cases, a predicted structure with RMSD$_N$ (RMSD of a spherical cluster of N molecules following the CSD standard) better than 0.50 Å for a cluster of at least 25 molecules was sampled and ranked among the top 10 of the predicted structures. For 26 out of the 33 molecules, the best match candidate structures was ranked among the top 2.

Upon careful inspection of the top ranked predicted candidate structures, we found that some of them adopt very similar conformers

https://doi.org/10.1038/s41467-025-57479-1

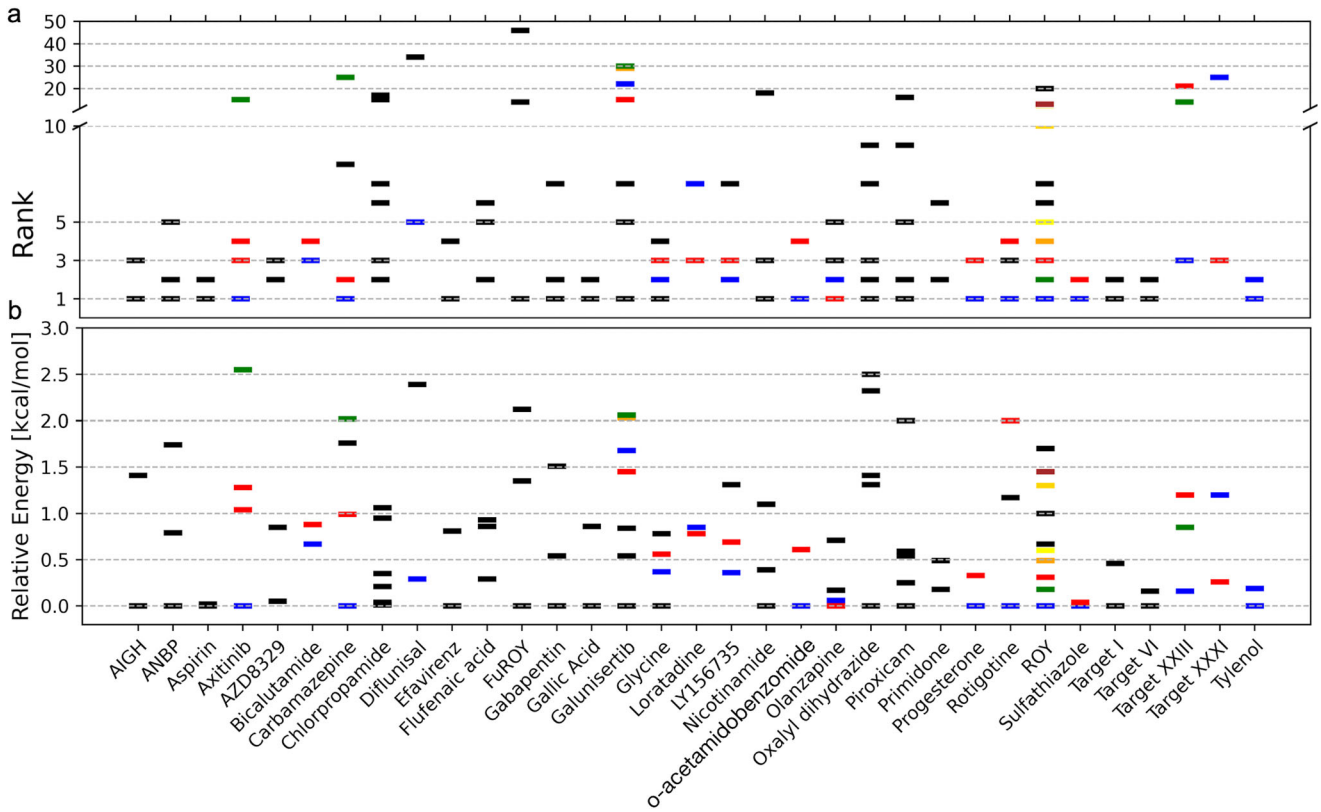

Fig. 4 | Summary of the polymorph energy landscape of the subset of 33 molecules with multiple Z′ = 1 experimental structures. The energy landscapes are obtained with r^{2}SCAN-D3, except for Axitinib, Galunisertib, o-acetamidobenzomide, and ROY, in which additional conformational energy corrections are incorporated using ωB97X-D3 (i.e. r^{2}SCAN-D3 + Δ ωB97X-D3, see SI). All known Z′ = 1 crystal structures are sampled and ranked well from the crystal structure prediction (CSP) calculations. The rank of the predicted crystal structure that best matches each experimental crystal structure is shown in (**a**) and the relative energy is shown in (**b**). The color codes are used to indicate experimental relative stability ordering if available. The experimental relative stability ordering is denoted as: blue > red > green > orange > brown > yellow > gold. The black lines indicate experimental relative stability is unknown. Source data are provided as a Source Data file.

and packing patterns. These structures correspond to different local minima of the quantum chemical potential energy surface at 0 K, but may interconvert at room temperature since the corresponding energy barriers might be comparable to thermal fluctuation[22,23]. This could be related to disordered structures[24] and has been proposed to be one of the common reasons for the well known over-prediction problem in CSP calculations[25]. To remove this type of non-trivial duplicate from the static landscapes, we clustered similar structures (with $RMSD_{15}$ better than 1.2 Å) into a single representative structure with the lowest energy among the cluster, similar to what has been done in earlier studies[23,25]. The rankings of the best matched structures after clustering are shown in Fig. 3a (red bars). This improved the rankings, for example, for MK-8876, Target V, and naproxen.

Figure 3b shows the relative energy between the candidate structure that best matches the experimental structure and the lowest energy predicted structure for each molecule in the subset of molecules with only a single experimentally known crystal structure. The blue and red bars show the results before and after clustering analysis. For most molecules, the known experimental structure corresponds to the lowest energy predicted structure using the r^{2}SCAN-D3 functional, thus the relative energy is zero. For a few molecules, the relative energy is less than 0.5 kcal/mol, which is likely within the error bars of the DFT calculations[7]. However, in two systems we observe candidate structures with significantly lower energies (1 kcal/mol) than any experimentally observed polymorph. These two molecules are GSK268499B and MK-8876. Such low energy structures indicate the possibility for stable polymorphs that are yet to be discovered by experiment. A detailed analysis for these molecules will be presented in the following section.

For molecules with multiple known experimental structures, Fig. 4a, b shows the rankings and relative energies of the predicted candidate structures best matching the corresponding known experimental structures after clustering. In all cases, the known experimental structures were sampled and ranked well. 80% of the candidate structures matching known experimental structures were ranked among the top 10 of the predictions with relative energies less than 1.0 kcal/mol as compared to the lowest energy predicted structures. Among this subset of molecules with multiple known experimental structures, the energy gap between the most stable known polymorph and the lowest energy predicted structure is only about 0.5 kcal/mol, which indicates that extensive experimental screening might have already identified the most stable polymorphs for these molecules.

In the following sections, we will present a few representative examples of polymorphic energy landscapes from our calculations and compare them with prior experimental and CSP studies.

Cocaine

The CSP generated energy landscape of cocaine is shown in Fig. 5a. The predicted structure with the lowest energy matches the experimental structure (CSD refcode COCAIN10) with a $RMSD_{25}$ of 0.07 Å. Its energy is around 0.8 kcal/mol lower than the second lowest energy predicted structure using PBE-D3, which is in agreement with previously reported energy landscapes using the PBE-D functional[26]. However, the energy gap between the known experimental structure and second lowest energy candidate structures reduces to 0.3 kcal/mol using r^{2}SCAN-D3, indicating possible limitations in PBE-D3 for the purpose of high accuracy ranking. A few benchmark studies have shown that

https://doi.org/10.1038/s41467-025-57479-1

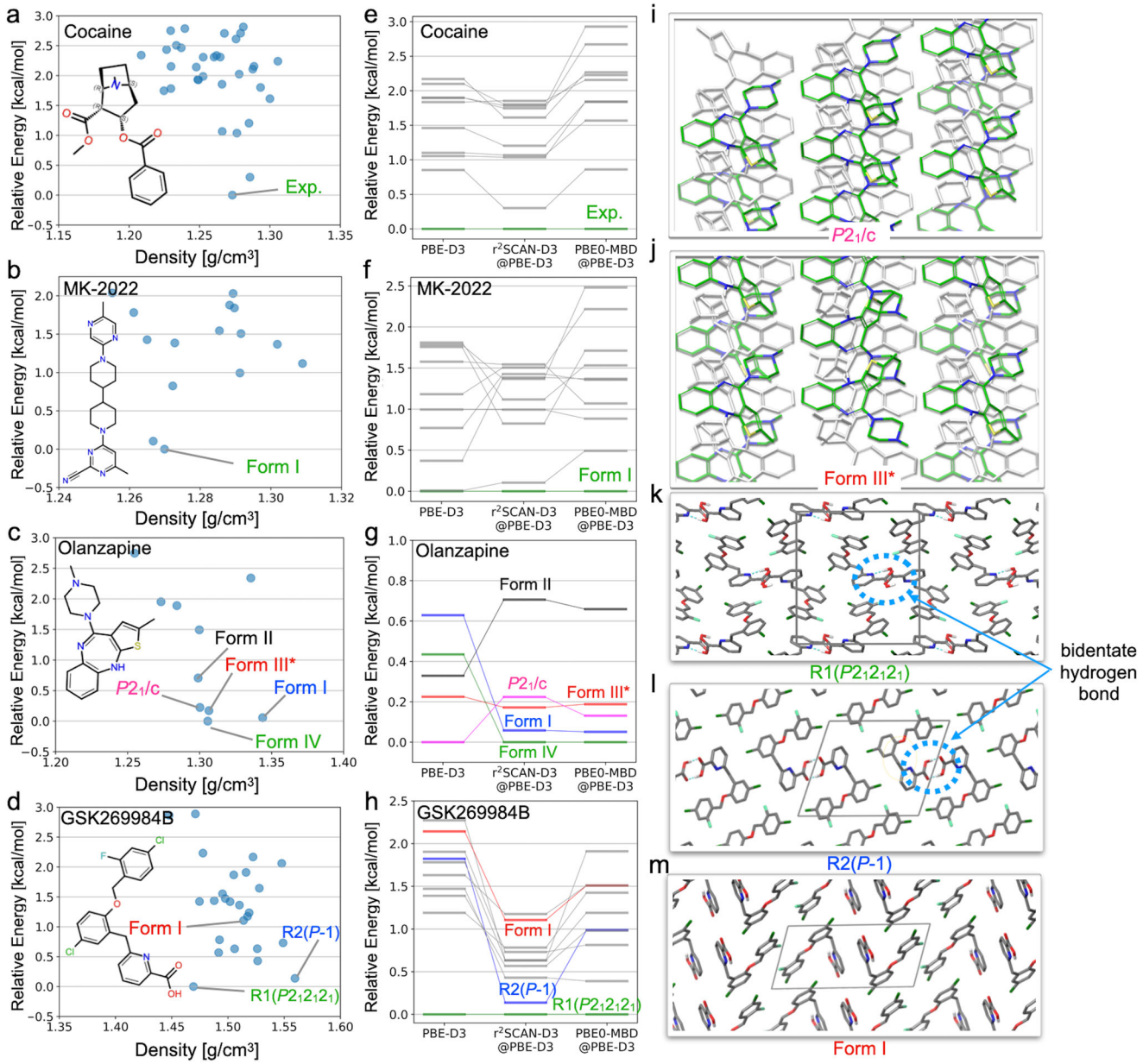

Fig. 5 | Energy landscape and structure of four representative drug molecules. Predicted polymorphic energy landscapes for (**a**) cocaine, (**b**) MK-2022 (**c**) olanzapine and (**d**) GSK269984B using r^{2}SCAN-D3 energy computed on PBE-D3 optimized crystal structures. The corresponding energy landscape dependence on density functional theory (DFT) functionals are shown in (**e**–**h**). For cocaine (**a**), the only known experimental structure is predicted to have much lower energy than all other predicted candidate structures with an energy gap of about 1.0 kcal/mol by PBE-D3 and PBE0-MBD functionals (**e**). For MK-2022 all functionals predict experimentally known form I to be most stable (**f**). For olanzapine, the three resolved experimental structures are predicted among the top five candidate structures with three different DFT functionals (**g**). For the unresolved form III, two predicted candidate structures have the spectra characteristics of form III from experiment, one corresponding to a previously identified structure (III*) in the Pbca space group (**j**), and one novel structure in the P2$_1$/c space group with an identical 2D layer (**i**). For GSK269984B, candidate structures with favorable intermolecular hydrogen bonds (**k**) and (**l**), including the typical bidentate hydrogen bonds for carboxylic acids, as highlighted in (**k**) and (**l**), are predicted to have lower energy than the experimental structure with only intramolecular hydrogen bonds (**m**), indicating the possibility of other polymorphs with competitive stability yet to be discovered by experiment. The grey frames in (**k**), (**l**), and (**m**) represent the unit cell of the crystals. Source data are provided as a Source Data file.

r^{2}SCAN-D3 describes the energy and structure in a variety of chemical environments more accurately than PBE-D3, in particular, for hydrogen bonding and dispersion interactions that are typically present in organic molecular crystals[27–29]. The top 7 predicted structures, which are within the lowest 1.8 kcal/mol energy window from our calculations, match those from a previous study within a RMSD$_{25}$ of 0.2 Å[26].

The energy gap between the known experimental structure and other top candidate structures is further increased when using the PBE0-MBD functional (Fig. 5e), often considered as a more reliable method for calculating crystal lattice energies according to prior studies[15,30]. The relative stabilities of the top predicted structures vary slightly as a function of temperature, but the experimental form remains the most stable at all temperatures (Fig. S18a in SI). This further indicates the low likelihood of observing any additional more stable Z′ = 1 polymorph of cocaine. Such a CSP landscape, with the global minimum separated by about 1.0 kcal/mol from all other candidate structures, is uncommon in our experience. Most drug-like

molecules we have studied have many predicted structures within 1.0 kcal/mol of the global minimum using accurate periodic DFT.

MK-2022

MK-2022, a GPR119 agonist[31], was developed by Merck & Co., Inc. for potential applications in metabolic diseases, including type-2 diabetes. A recent study used the GRACE software to predict the crystal polymorphs of this molecule and compared the prediction results with experiments[31]. The GRACE software correctly identified a candidate structure (ranked 6th with around 1.0 kcal/mol higher energy compared to the global minimum) matching the experimentally known polymorph for MK-2022. Our calculations also correctly predicted this experimental structure, with the lowest energy predicted structure matching experiment with an $RMSD_{25}$ of 0.23 Å (Fig. 5b). The relative stabilities of the top 5 predicted structures vary slightly as a function of temperature (Fig. S19d in SI), though the rank II predicted structure has slightly lower free energy than the experimental structure at room temperature.

Olanzapine

Olanzapine is used as an atypical antipsychotic agent for the treatment of bipolar disorder and schizophrenia[32]. Despite having only one rotatable bond, it exhibits a high degree of polymorphism in its solid state, giving rise to about 60 known distinct solid forms including four anhydrous polymorphs, 56 crystalline solvates, and an amorphous phase[33,34]. Here, we focus on the four $Z' = 1$ anhydrous forms. Among them, only three forms I, II, and IV have been characterized as single crystals, whereas form III was concomitantly found with form II. Form III was identified by comparing the PXRD patterns of form II and a mixed phase crystal of forms II and III. The presence of form III was also detected in a combined solid state NMR and CSP study[34], where a CSP generated structure in the Pbca space group was identified to possess the spectra characteristics of form III. We will refer to this structure as form III* in the following discussion.

From our CSP calculations, candidate structures matching well with published structures for forms I, II and IV ($RMSD_{26}$ of 0.04, 0.13 and 0.20 Å in reference to CSD structures UNOGIN03, UNOGIN04 and UNOGIN05) are sampled and ranked among the top 5 of the candidate list (Fig. 5c). In agreement with the previous study[35], we find that forms I and IV are highly competitive in energy with an energy difference of only around 0.1 kcal/mol as obtained from using both the r^{2}SCAN-D3 and the PBE0-MBD methods. In addition, the lattice parameters of one of the low-lying predicted structures in the Pbca space group (Rank 3) has very similar lattice parameters and structure ($RMSD_{28}$ of 0.32 Å) as compared to form III*[34]. The simulated PXRD data of our best match Pbca structure show many of the characteristic peaks which were assigned to form III (see Fig. S5 in SI). These comparisons indicate that the rank 3 predicted structure in the Pbca space group might be a putative structure of form III (Following the nomenclature in ref. 34., we refer our best match Pbca structure as form III* in Fig. 5c).

In addition to these four structures identified in earlier experimental and computational studies, our calculations also predicted a low-energy candidate structure in the $P2_1/c$ space group (Rank 4 prediction). Both the putative form III* in the Pbca space group and the newly identified $P2_1/c$ structure have identical 2D layers as the experimental form II, but the packing along the third dimension differs, similar to Aspirin's crystalline forms I and II[36]. The simulated PXRD data of this $P2_1/c$ structure matched better with the characteristic peaks assigned to experimental form III (Fig. S5 in SI). These results challenge the conclusion from ref. 34. that the experimentally observed form III corresponds to the Pbca structure, as it appears that our predicted $P2_1/c$ candidate structure can better explain the experimental data. Furthermore, since the lattice parameters of the $P2_1/c$ structure are very close to those of form II, it is more likely than form III* to grow together with form II. It is also possible that the experimentally observed form III might be a mixture of the predicted $P2_1/c$ and Pbca candidate structures.

While the paper was under review, an experimental group resolved the structure of olanzapine form III via microED[37] which confirmed the prospectively predicted structure reported in our manuscript. In our initially submitted manuscript in Jan, 2024, we reported the $P2_1/c$ structure from our predictions as a novel structural model for the then unresolved experimental form III. The newly resolved experimental structure of olanzapine form III confirmed our prospective predictions with a $RMSD_{33}$ of 0.2 Å between predicted and experimental structures.

GSK269984B

GSK268499B, a drug candidate developed by GlaxoSmithKline, has been reported as an EP(1) receptor antagonist for the treatment of inflammatory pain[38]. Previous studies combining experimental screening and CSP have concluded that form I is the most stable anhydrous crystal structure[39]. Interestingly, form I does not possess strong intermolecular hydrogen bonding apart from intermolecular halogen bonds and pi-pi interactions, as shown in Fig. 5m. Instead, it contains intramolecular hydrogen bonds between the carboxylic acid proton and pyridine nitrogen.

A prior CSP study of GSK268499B predicted form I as the global minimum with other low-lying crystals possessing intermolecular hydrogen bonding[39]. In that study, a mixed DFT intramolecular conformational energy and empirical intermolecular electrostatic and dispersion model was used to evaluate relative stability of candidate crystal structures. From our CSP calculations using accurate periodic DFT calculations, at 0 K and without zero-point energies, the candidate structure matching the experimental form I ($RMSD_{32}$ of 0.09 Å in reference to CSD structure BIFHOP) is ranked 9th (Fig. 5d) using the r^{2}SCAN-D3 method, more than 1.0 kcal/mol less stable compared to the lowest energy candidate structure.

The two lowest energy candidate structures, referred to as R1($P2_12_12_1$) and R2(P-1) (Fig. 5k, l), both have the carboxylic acid in the *cis*-isomer conformation forming strong intermolecular hydrogen bonds with nearby molecules in the crystals. The intermolecular hydrogen bonding in R1($P2_12_12_1$) is between the carboxylic acid proton and the pyridine nitrogen from a nearby molecule, similar to the hydrogen bond networks in the rank 2 predicted structure from the prior study[39], whereas the R2(P-1) structure exhibits the typical bidentate hydrogen bonds between *cis* carboxylic acids existing in many other crystal structures with the same chemical group. Eleven molecules in our validation set contain a carboxylic acid group; among those, nine form a double hydrogen bond motif with the carboxylic acid of another molecule in at least one experimentally observed form. The remaining two cases are either zwitterionic (Gabapentin) or form another favorable double hydrogen bond involving the carboxylic acid and another polar chemical group (Ceftizoxime). All candidate structures within the 3.0 kcal/mol energy window have similar hydrogen bond networks in R1($P2_12_12_1$) and/or R2(P-1).

Although the experimental structure was predicted to have higher energy than a few other predicted structures at 0 K, its relative stability improved substantially with increased temperature (Fig. S18c in SI). At room temperature, the rank I predicted structure and the experimental structure have a free energy difference of about 0.3 kcal/mol while all other predicted structures have higher free energies. These results highlight the important role of entropy in accurately evaluating the relative stabilities of crystal structures at room temperature.

Other interesting molecules: ROY, Galunisertib, Rotigotine, MK-8876,LY-156735, Bicalutamide, XXIII, Axitinib and AZD8329, Target XXXI

Among all the small organic molecules exhibiting multiple crystal polymorphs, ROY and Galunisertib have been shown to exhibit the

https://doi.org/10.1038/s41467-025-57479-1

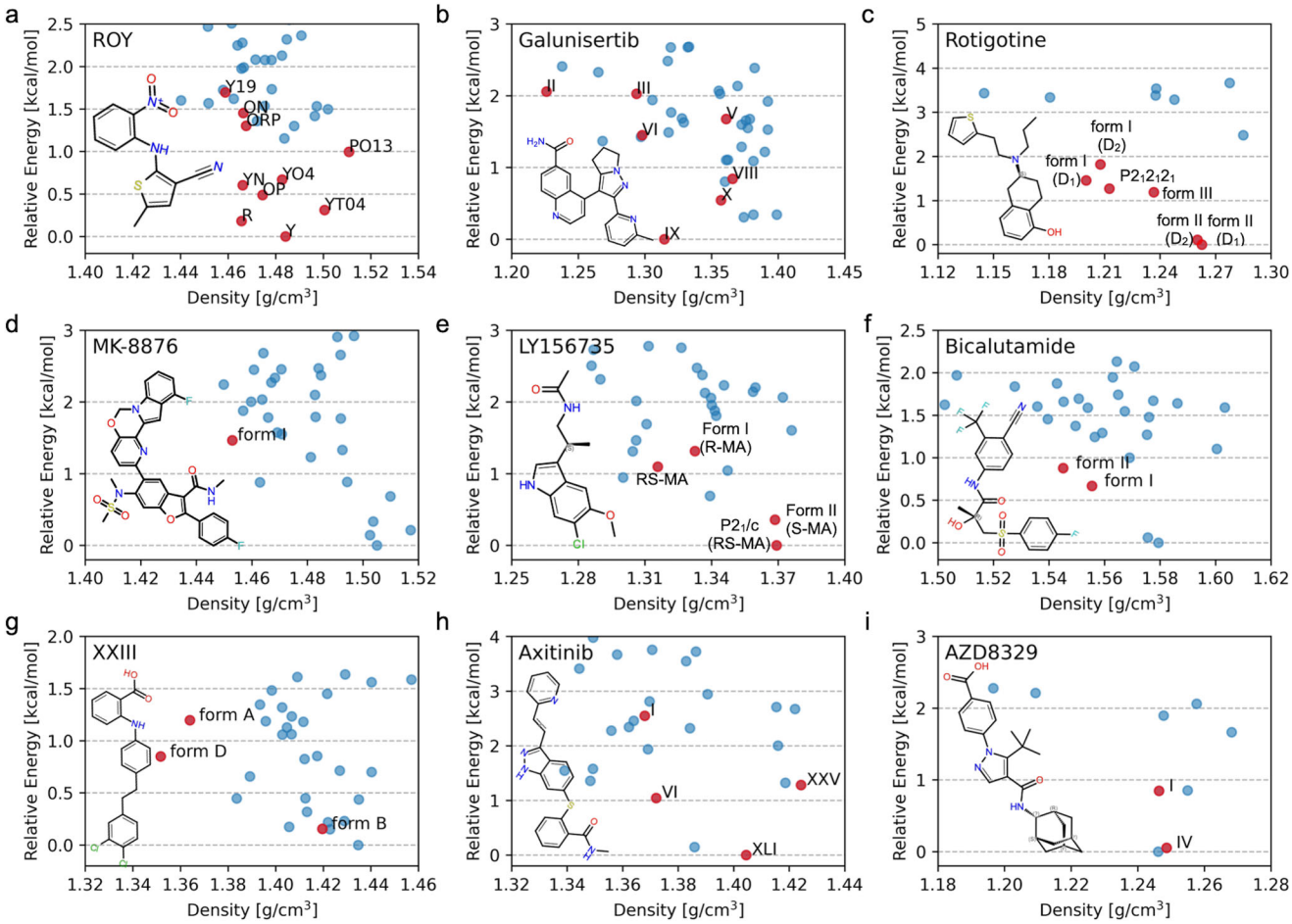

Fig. 6 | Energy landscapes for six representative drug molecules. Predicted polymorphic energy landscapes for six representative drug molecules, (**a**) ROY, (**b**) Galunisertib, (**c**) Rotigotine, (**d**) MK-8876, (**e**) LY156735, (**f**) Bicalutamide, (**g**) XXIII, (**h**) Axitinib, and (**i**) AZD8329. The predicted crystal structures that best match the known experimentally crystal structures are highlighted in red circles and others are shown in blue circles. The energy landscapes are obtained with r²SCAN-D3, except in (**a**), (**b**), and (**h**), in which additional conformational energy corrections are incorporated using ωB97X-D3 (i.e. r²SCAN-D3 + Δ ωB97X-D3). For (**a**) ROY and (**b**) Galunisertib, molecules holding the current record for the largest number of fully characterized polymorphs, our crystal structure prediction (CSP) approach correctly predicted all the known Z′ = 1 polymorphs. For rotigotine (**c**), where the late appearance of the more stable form II disrupted the clinical usage of the original form I formulation, our CSP calculations correctly predicted all the experimentally known forms I and II, the candidate form III from a prior CSP study, and another novel candidate crystal in the $P2_12_12_1$ space group with greater stability than form I. For MK-8876 (**d**) and MK-2202 (**e**), a previous CSP study using the GRACE software correctly predicted a candidate structure matching the experimentally known polymorph for MK-2022, but failed for MK-8876. Our CSP calculations correctly predicted both experimental structures with the experimental structure ranking 1st for MK-2022 and 2nd for MK-8876. For LY156735 (**f**), a potent and selective melatonin agonist with some unusual crystallization behavior among the R and S enantiomers, our calculations correctly predicted all three known experimental forms, and suggested a few other competitive structures, indicating the complex polymorphic landscape of this molecule may remain to be extensively explored experimentally. For XXIII (**g**), a predicted P-1 crystal structure not discovered by experiment has lowest energy at 0K. For axitinib (**h**), form XLI was correctly predicted to be the most stable form, matching experiment. For AZD8329 (**i**), form IV is predicted to be more stable than form I at 0 K, in agreement with the experimental stability ordering at ambient conditions. Source data are provided as a Source Data file.

largest number of fully characterized polymorphs[40,41]. For ROY, twelve fully characterized crystal structures, Y, ON, R, OP, YN, ORP, YT04, Y04, R05, PO13, R18, and Y19, have been discovered[40]. These polymorphs stem from a broad set of conformers responsible for their red, orange, and yellow colors. For Galunisertib, solid form screening experiments produced many solvates and ten anhydrous forms (form I – form X), with forms IV-VII found to be the most stable, having experimentally measured melting enthalpies within the margin of error of one another[40–42]. Our approach correctly predicted all the known Z′ = 1 polymorphs for ROY and Galunisertib (Fig. 6a, b). For ROY, the intramolecular torsion energies are known to be challenging for typical functionals used in periodic DFT and corrections to periodic DFT energies are needed to accurately rank order the relative stabilities of these polymorphs[40]. The raw rankings from periodic DFT are shown in Fig. 6 with intramolecular torsion energy corrections discussed in Fig. S6 of SI. By incorporating conformer energy corrections using the ωB97X-D3/def2-TZVP functional, Y, R, YT04 and OP were correctly ranked to be the four most stable polymorphs at 0 K matching experiments, although the relative stability between R and YT04 is the opposite compared to experiment by a very small energy difference. The relative stabilities changed slightly as a function of temperature but the energy gap remains small (Fig. S18d in SI).

For Galunisertib, forms VIII, IX, X, were predicted to be more stable than forms V and VI according to the r²SCAN-D3 functional. Our second most stable candidate structure in the $P2_12_12_1$ space group, is yet to be observed experimentally, and matches the structure reported as the global minimum in prior CSP landscapes[41]. Using the conformer energy correction, the order of stabilities of the Z′ = 1 experimentally known forms remains largely unchanged (Fig. S7), in agreement with prior computed results[43]. Adding room temperature free energy

calculations (Fig. S19a in SI) indicates that form VIII becomes less stable with increasing temperature and forms IX and X remain more stable than forms V and VI across the computed temperature range.

Rotigotine is a dopamine agonist for the treatment of Parkinson's disease and restless legs syndrome. The late appearing, more stable form II disrupted the clinical usage of its initial formulation of form I[4,44]. A previous CSP study predicted another candidate structure in the $P3_2$ space group, referred to as form III in our discussion and in Fig. 6c, with stability between the experimentally known forms I and II[30]. From our CSP calculations, all 3 forms are sampled and ranked well. Form II is correctly predicted to be about 2.0 kcal/mol (Fig. 6c) lower in energy than form I according to the r^{2}SCAN-D3 and the PBE0-MBD functionals, in agreement with the energy gap previously reported with PBE0-MBD (1.7 kcal/mol) and differential scanning calorimetry (DSC) measurements (1.8 kcal/mol)[30]. In addition to form III, we find another candidate crystal structure in the $P2_12_12_1$ space group with stability comparable to form I. The additional $P2_12_12_1$ crystal structure was also predicted in the prior study, although the relative energy by their computed method was slightly higher. Detailed structural and energetic comparison with the earlier CSP study is given in Figs. S8 and S9 of the SI.

The recent study by Merck & Co, Inc. also investigated the application of GRACE to predict the crystal structures of MK-8876, an HCV site NS5B site D inhibitor[31,45]. The GRACE software failed in predicting the crystal structure for MK-8876 since the candidate structure matching the experimental form was ranked 2290th by the tailor-made force field and was dropped from the workflow. Our calculations correctly predicted the experimental structure, corresponding to rank 10 of the predicted structure (Fig. 6d). The relative energy of the experimental structure of MK-8876 compared to the predicted global minimum varies between 0.1 to 1.5 kcal/mol with different DFT functionals (Fig. S13), and the gap increased slightly at room temperature (Fig. S19c in SI), indicating the possibility for the existence of other polymorphs with competitive stability. This is in line with experimental evidence suggesting other metastable forms.

LY156735, a potent and selective melatonin agonist (MA) investigated for the treatment of insomnia and circadian rhythm disorders, showed some unusual crystallization behavior among the R and S enantiomers[17,46]. While the inactive enantiomer, S-MA, was found to crystallize in at least two anhydrous polymorphic forms (form 1 and 2), with form 2 about 0.7 kcal/mol more stable than form 1, the active enantiomer, R-MA, was found to crystallize only in form 1 despite extensive efforts. In addition, another racemate structure, form RS-MA, was also characterized in the prior work. A prior CSP study predicted form 2 to be more stable than form 1, with the racemate form in between[17]. Our CSP calculations predicted all three forms and ranked them among the top ten lowest energy structures (Fig. 6e). Form 2 is correctly predicted to be about 1.0 kcal/mol more stable than form 1 according to r^{2}SCAN-D3 and PBE0-MBD functionals, with the energy of the racemate structure in between (Fig. S13). Interestingly, our calculations also suggested a few other competitive structures, some of which were also predicted in the prior study. A novel racemate candidate structure in the $P2_1/c$ space group with different hydrogen bond networks was predicted to be the most stable form at 0 K but became less stable at room temperature (Fig. S20a in SI). These results indicate that the polymorphic landscape of this molecule is perhaps more complex than what is currently known experimentally.

Bicalutamide is an anti-androgen medication that is primarily used to treat prostate cancer[47]. Extensive experimental studies have identified two Z′ = 1 forms, I and II, with form I found to be more stable[48]. From our CSP calculations, candidate structures matching well with experimental structures for forms I and II ($RMSD_{30}$ of 0.09 and 0.19 Å in reference to CSD structures JAYCES and JAYCES02) are sampled and ranked among the top 5 (Fig. 6f). Form I is correctly predicted to be slightly more stable than form II according to r^{2}SCAN-D3 although the relative ranking is sensitive to different DFT functionals due to the relatively small energy difference (Fig. S14). The relative stability between these two forms does not change significantly as a function of temperature (Fig. S20d in SI).

The sixth CCDC blind test featured molecule XXIII submitted by Pfizer Inc. which was a former research compound aimed at treating Alzheimer's disease[49]. Three Z′ = 1 forms (A, B, D) and two Z′ = 2 forms (C, E) have been experimentally characterized and their relative stabilities are strongly dependent on temperature. The thermodynamically most stable form changes over a small range of temperatures in the following way: form C below 253 K, form B in 273 – 288 K, form A in 290 – 294 K, and form D at 295 K and higher[50]. Many CSP approaches in the sixth blind test predicted form B to be the most stable among the three Z′ = 1 polymorphs. From our CSP calculations, candidate structures that match well with published structures for the three Z′ = 1 forms A, B, D ($RMSD_{30}$ of 0.39, 0.23, and 0.18 Å in reference to CSD structures XAFPAY, XAFPAY01, XAFPAY03, respectively) are sampled and ranked among the top 30. We find that PBE-D3 favors form B as the most stable structure, similar to previous findings, and that r^{2}SCAN-D3 and PBE0-MBD favor a predicted P-1 crystal structure not discovered by experiment (Fig. S15). Our computed temperature dependent stabilities correctly predicted form D to be most stable above room temperature, with form A very similar in energy as compared to form D across the temperature range (Fig. S20b in SI).

Pfizer's anticancer drug axitinib has 5 known neat polymorphs and 66 solvates to date. Form IV was initially targeted for development but more stable forms XXV and XLI were fortuitously discovered later during the manufacturing campaign[51,52]. DSC and solubility experiments indicate the stability in this order: XLI > XXV, VI > IV > I, where the energy ordering of XXV compared to VI is uncertain[51,52]. A previous CSP study ranked the relative stabilities among these forms incorrectly, with XLI predicted to have a 2.4 kcal/mol higher energy than form VI[53]. From our CSP calculations, we correctly predicted all four Z′ = 1 experimentally known structures ($RMSD_{29}$ within 0.13–0.29 Å in reference to CSD structures VUSDIX, VUSDIX03, VUSDIX04, VUSDIX06) among the top 20 candidate structures. Contrary to experiment, form VI was predicted to be more stable than form XLI according to the r2SCAN-D3 functional, but the energy difference decreased to within 0.2 kcal/mol using the PBE0-MBD functional. By incorporating conformer energy corrections using the ωB97X-D3/def2-TZVP functional in conjunction with r^{2}SCAN-D3 crystal energies (referred to as r^{2}SCAN-D3 + Δ ωB97X-D3 in Figure S16 in SI), form XLI was correctly predicted to be the most stable form, matching experiment and a previous computational study[43]. The temperature dependent free energy calculation indicates the stability of forms XLI and VI gets closer as temperature increases, with form XLI remaining more stable at temperatures below 200 K (Fig. S21a in SI).

AZD8329 is a pharmaceutical compound under development for the potential treatment of metabolic disorders including type 2 diabetes[54,55]. Among the seven known crystal forms, the Z′ = 1 forms I and IV have been chosen for development due to their suitable material properties. The two forms are enantiotopically related, with form IV more stable at ambient conditions and form I more stable at high-temperature[55]. From our CSP calculations, candidate structures that match well with published structures for forms I and IV ($RMSD_{27}$ of 0.23 Å and 0.64 Å, in reference to structures published in ref. 55) are sampled and ranked among the top 5 of the candidate list. Form IV is found to be more stable than form I at 0 K by the r^{2}SCAN-D3 functional, which is in agreement with the reported stability ordering at ambient conditions. In addition, the temperature dependent stability calculations indicate form I becomes more stable than form IV with increasing temperature, in qualitative agreement with experiments.

While this paper was under review, reports of the seventh blind test results of crystal structure prediction organized by the CCDC were published[11,12]. Compound XXXI is the only molecule in this blind test

https://doi.org/10.1038/s41467-025-57479-1

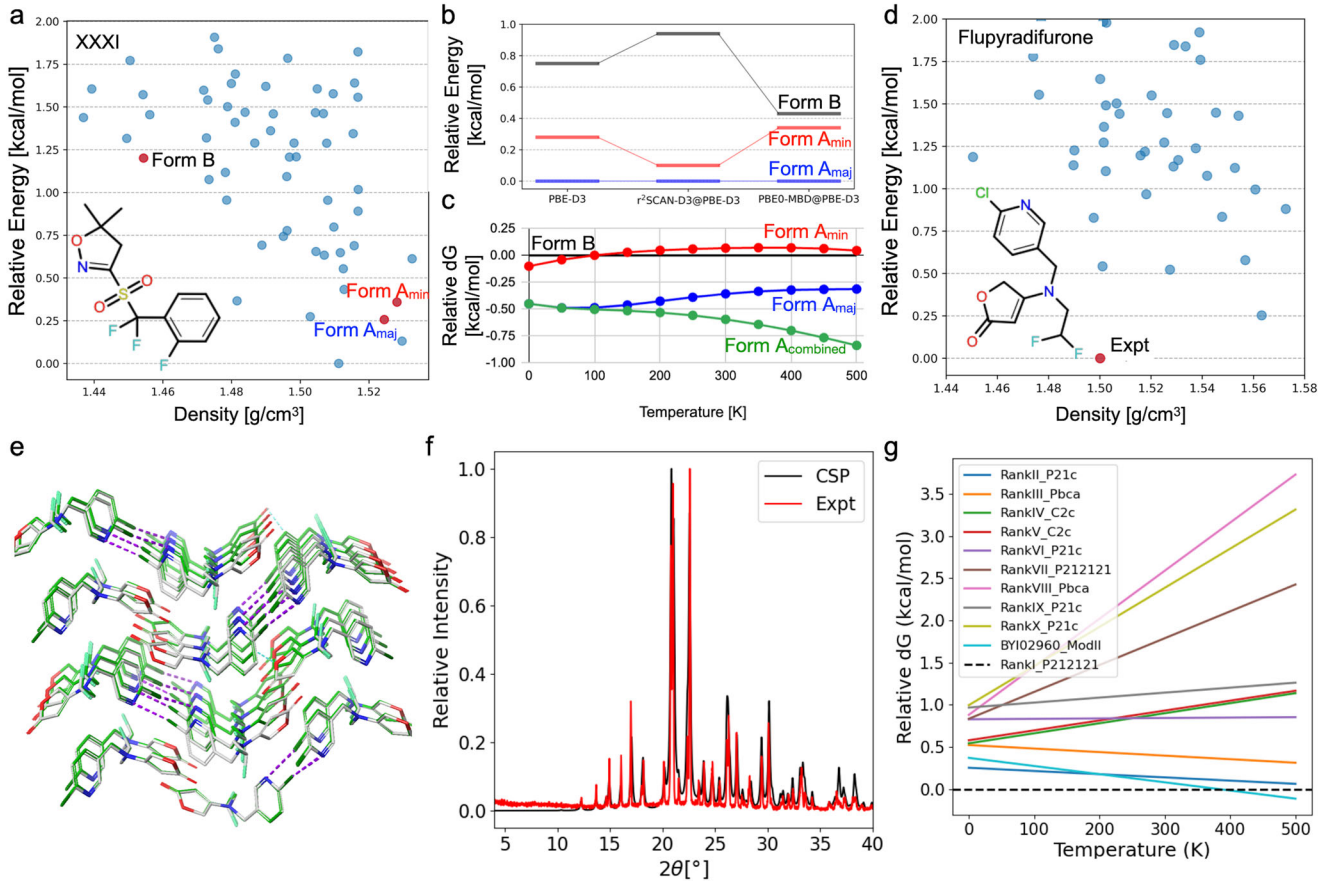

Fig. 7 | Energy landscape of target XXXI and Flupyradifurone. a Energy landscape of target XXXI, used in the 7th blind test, using the r^{2}SCAN-D3 functional. The known experimental forms Form A_{maj} Form A_{min}, and Form B are highlighted. **b** The relative energy among these three crystal structures at 0 K with different density functional theory (DFT) functionals. **c** Temperature dependent free energy (dG) of Form A relative to Form B. The 0 K relative energies are from PBE0-MBD. **d** Energy landscape of Flupyradifurone molecule provided by Bayer using the r^{2}SCAN-D3 functional. The lowest energy structure marked in red in the P_{212121} space group agrees well with the only known experimental (Expt) structure in the Z′ = 1 space. **e** 3D overlap between the predicted lowest energy structure in gray and the experimental structure in green ($RMSD_{32}$ of 0.12Å). **f** Comparison of experimental PXRD and crystal structure prediction (CSP) generated PXRD of the predicted lowest energy structure. **g** Temperature dependence of the relative free energy of the lowest 10 predicted structures along with the Z′ = 2 experimental structure denoted as BYI02960_ModII. The predicted lowest energy structure matching the experimental Z′ = 1 structure of Mod I is most stable at all temperatures below 400 K according to the calculations, matching experimental results that Mod I is the most stable form at all temperatures below its melting point of 361 K. Source data are provided as a Source Data file.

that is relevant for agrochemical and pharmaceutical development and has experimental forms with only Z′ = 1 structures. Three experimental structures of XXXI were determined. However, Form C is a channel-type solvate containing unresolved solvent and therefore falls outside the scope of prediction. Form A contains two disorder components resulting from the flipping of the fluorinated ring, and form B was determined by competitive slurry experiments as the most stable form below 55 °C with form A becoming more stable for temperatures above 55 °C.

The predicted energy landscape of XXXI at 0 K from our calculations and the temperature dependent relative stability between forms A and B are reported in Fig. 7a–c. Our calculations correctly predicted form B and the two disordered forms of form A among the top 25 predicted structures. The temperature dependent relative stability calculations indicate that with increased temperature form A, with its two disordered structures, become more stable as compared to form B, consistent with the experimental findings that form A is more stable at temperatures higher than 55 °C. However, the relative free energy difference between the two forms at room temperature is small (about 0.4 kcal/mol), within the errors of the calculations, and therefore it can not be predicted with confidence which form is more stable at room temperature. It is also interesting to note that the relative lattice energy difference (stability at 0 K reported in Fig. 7a) among these three structures from our calculations are consistent with the results from all groups participating in seventh blind test that used similar DFT calculations to rank order the structures (including groups 3, 20), except for group 10 who reported a different order than others[12].

Blinded study of an agrochemical molecule

To further substantiate the reliability and accuracy of our polymorph prediction method, we recently applied it blindly on an insecticide developed by Bayer Crop Science and present the results in the following section.

Flupyradifurone is a systemic insecticide developed by Bayer CropScience that protects crops from sap-feeding pests like aphids and whiteflies. Flupyradifurone was approved for use in plant protection products by the EU in 2015, and extensive experimental polymorph screening was performed by Bayer during the product development. The Schrödinger team were provided with only the 2D structure of flupyradifurone and performed polymorph predictions without any experimental data. Upon completion of the presented CSP workflow, prediction results were sent to Bayer to compare with the experimental polymorph screening results. The predictions were limited to Z′ = 1 space due to the current limitation of the method. The

Table 1 | Compute time (in CPU hour units for all except MD relaxation in GPU hour units) for all steps of the crystal structure prediction workflow on a set of representative molecules with varying complexity

	# conf	Packing search (CPU hour)	MD (GPU hour)	# Crys (FF)	QRNN (CPU hour)	# Crys (QRNN)	QRNN worst rank of exp.	DFT (CPU hour)
glycine	43	283	87	23.1 K	1.0 K	1662	12	2.3 K
PF-998425	30	609	21	21.1 K	2.8 K	2809	24	19.0 K
cocaine	35	995	21	9.7 K	2.0 K	1827	1	23.7 K
MK-8876	2632	24.6 K	192	7.3 K	10.5 K	237	39	58.7 K
rotigotine	1027	12.0 K	118	5.9 K	1.5 K	185	24	11.8 K
GSK269984B	1680	58.6 K	317	16.7 K	10.5 K	4821	27	28.7 K
Target XXXI	123	36.1 K	200	75.8 K	31.7 K	7658	83	25.5 K

The number of candidate crystal structures within 10.0 kcal/mol by the FF and 6.0 kcal/mol by QRNN are also reported. DFT calculations are run for up to 300 candidate structures from QRNN in the 6.0 kcal/mol energy window. Energy ranking via periodic DFT calculations are the most time-consuming step for small rigid molecules while packing search and DFT calculations consume roughly equal amounts of time for larger flexible molecules. Force field parametrization and conformer generation costs are omitted due to minimal computational expense (tens of CPU hours at most).

predicted polymorph landscape in Z′ = 1 space is shown in Fig. 7d. Using r^{2}SCAN-D3, flupyradifurone has a very dense energy landscape with 42 predicted structures within the lowest 2.0 kcal/mol energy window. The lowest energy predicted structure matches very well with the only known Z′ = 1 experimental structure Mod I ($RMSD_{32}$ of 0.12Å between predicted and experimental structure as shown in Fig. 7e), and simulated PXRD also agrees well with the experimental spectra (Fig. 7f). In addition to Mod I, flupyradifurone also has a Z′ = 2 experimental structure (Mod II) that was not attempted in the current round of calculations. We also performed free energy calculations of the top predicted Z′ = 1 structures along with the Z′ = 2 experimental Mod II to obtain their relative stabilities as a function of temperature. As shown in Fig. 7g, according to the calculations, the lowest energy predicted structure at 0 K (corresponding to experimental Mod I) is most stable at all temperatures below 400 K. These results are consistent with Bayer's experimental findings that Mod II and Mod I are monotropic polymorphs with Mod I more stable at all temperatures below its melting point of 361 K.

Discussion

Like many other CSP approaches, we separate the sampling of molecular conformational degrees of freedom from that of the lattice degrees of freedom[11,18,56]. To validate the conformer generation method, the CSP test set was enriched with examples from the CSD drug subset with 9 or fewer rotatable bonds from Z′ = 1 crystal structures. In total, the dataset consists of 430 experimental crystal conformers. The RMSD of the results from the conformer generation protocol are presented in Table S1 in SI. The conformer generation protocol reliably samples conformations within a RMSD of 0.40 Å compared to the experimental ASU for all molecules in the Z′ = 1 CSP test set. Due to the exhaustive nature of our method, high quality conformers are always generated across the chemical space tested, even for large flexible molecules.

To benchmark the packing search protocol, we curated a dataset of 426 neat Z′ = 1 crystal polymorphs from the CSD drug subset (Table S2 in SI). These crystals are relaxed with symmetry constrained MD using the OPLS4 force field. Good crystal similarity is achieved for the full dataset, demonstrating the accuracy of the OPLS4 in describing molecular crystal packing interactions. The crystal-relaxed ASU is then used to search for the relaxed crystal structure. For the 426 crystals used in this benchmark, a candidate structure matching experiment was generated with close to a 100% success rate (Table S2 in SI).

With the comprehensive data set for 66 molecules, we demonstrated accurate and reliable crystal structure prediction for small molecule drugs with Z′ = 1 and Z′ = 0.5. Representative timing information and throughput metrics are listed in Table 1 for systems of varying complexity.

In Table 1, we list the compute cost for each step of the polymorph prediction workflow for a representative set of molecules with increasing complexity. Our CSP workflow is significantly more computationally efficient than CPU costs reported previously by other approaches. To facilitate comparisons, we convert our GPU costs to CPU costs using a 1:10 conversion. For PF-998245, our CPU cost is 22.6 K hours, whereas approximately 200 K hours was reported in ref. 14. For rotigotine, our CPU cost is 26.5 K hours whereas 125 K CPU hours was reported in ref. 30. For XXXI, our computational cost is 93 K CPU hours and 200 GPU hours, whereas most participants in the seventh blind test who successfully predicted the experimental structures used significantly more CPU hours. It should be noted that our calculations were limited to Z′ = 1 structure predictions, whereas these participants might have spent a lot of effort for higher Z′ structure predictions, so these CPU hours might not be directly comparable.

For small or rigid molecules, the conformer pool is small and DFT relaxation takes most of the computational time. When the molecule contains more conformational degrees of freedom (ring states, nitrogen inversion centers, flexible torsions), the number of conformers increases and so does the packing search time. For very large and flexible molecules like ritonavir, the conformer and crystal packing sampling would take much more computation resources than DFT relaxation. This also poses a challenge for composite systems such as Z′ > 1, hydrates, salts, and cocrystals. An effort is underway to further improve the performance of the crystal packing search. These improvements include workflow optimization, pre-computing and storing the geometry dependent molecular interactions in the memory to reuse them during geometry optimizations. This type of optimizations led to orders of magnitude performance improvements in molecular docking[57] and preliminary results suggest such optimizations could boost the performance of crystal packing search by a factor of 3-5.

The systematic and hierarchical crystal packing approach (see Method Summary section) provides a significant efficiency improvement as compared to the stochastic approach employed by other CSP methods. Most groups participating in the 7th CSP Blind Test used methods that parameterize the search space via the six unit cell lattice parameters, the six parameters for ASU location and orientation, and the conformational degrees of freedom, and search these degrees of freedom simultaneously[11,12]. These methods require substantial computational effort to adequately cover the search space due to the size and complexity of the parameter space. Alternative methods included making perturbations to existing crystal structures and choosing to accept or reject the new candidate using a Monte Carlo scheme as well as the generative adversarial network of Group 10. To manage computational cost, many participants chose to terminate their program's execution after a certain number of structures or generation attempts were made. Some participants included convergence criteria by monitoring repeated generation of the lowest energy structures or by monitoring the number of generation attempts without finding new low-energy candidates. In contrast, the method presented here

efficiently covers the search space for molecular crystals using a divide-and-conquer strategy that partitions the parameter space into subspaces based on space group symmetries, eliminating the need for the convergence checks required by stochastic approaches.

We have presented the use of QRNN, a machine learned force field (MLFF), to perform structural relaxation and energy ranking of molecular crystals in our CSP workflow prior to DFT. Innovations in the successful use of machine learning methods were also demonstrated by participants in the 7th CSP Blind Test[11,12]. MLFFs were developed and utilized in the CSP protocols of groups 12, 15, 16, and 23. In the second part of the seventh blind test, Group 16 demonstrated energy ranking results with system specific MLFFs comparable to accurate DFT methods. Additionally, groups 10 and 20 made use of machine learning to reduce the number of structures required to be evaluated by costly methods, such as DFT. Future developments of QRNN in our CSP workflow will include improving molecule specific fine-tuning protocols to further reduce the cost of DFT evaluations.

A disadvantage of the packing search method described here is the significant effort required to implement it. This challenge arises because the symmetry operations differ for each space group, resulting in unique intermediate cluster constructions for each space group. Our current sampling covers the top 22 most common space groups, but extending the approach to all 230 space groups would require significant effort. Although, the choice to focus efforts on only a subset of space groups was universal among participants in 7th CSP Blind Test[11].

Although the current packing search aims for crystals of $Z' = 1$, coverage of $Z' < 1$ crystals is inherent to the method. This is because the point group symmetry of the ASU together with a lower symmetry space group can form a higher symmetry space group. If the corresponding lower symmetry space groups are already searched, no additional development is needed for $Z' < 1$ systems[58]. Extension to other complex systems such as hydrates, salts, and cocrystals, is straightforward. However, a naive implementation may present a challenge for workflow throughput.

Through careful comparison of relative stabilities across different DFT functionals for the set of molecules studied, we found the relative energies between different DFT functionals can vary by around 0.5 kcal/mol or even up to 1.0 kcal/mol in some cases, in agreement with the error estimates from previous studies[7,59]. Future work on more accurate DFT methods, particularly DFT methods that more accurately describe the conformational energies could further improve the reliability of the ranking among crystals with competitive stabilities[40]. The temperature dependent free energy calculations also improved the relative stabilities as compared to experiment in many cases. We hope the large number of predicted solid form landscapes we have provided, including experimental relative energy rankings where available, can aid in the investigation and further development of more accurate energy ranking methods, including DFT functional selection, conformer energy correction protocols, and temperature dependent free energy evaluations.

We have presented a reliable computational method with state of the art accuracy for predicting molecular crystal polymorphs validated on a large and diverse set of molecules including a blinded study. Our validations include all of the relevant molecules from the first six CCDC blind test, other molecules studied by previous computational polymorph prediction methods, and several molecules from modern drug discovery programs. Our method not only reproduces the experimentally known polymorphs, but also suggests new low energy polymorphs that have not been observed experimentally, posing potential risks to developing the currently known forms of these compounds without carefully considering these alternative low energy structures from our calculations.

Our method has several novel features and clear advantages over existing CSP methods and protocols for predicting molecular crystal polymorphs. First, our method uses a novel algorithm that systematically and efficiently explores the vast space of possible crystal packing parameters, overcoming the exponential scaling challenge of multi-parameter sampling for Monte-Carlo or other stochastic methods. Second, our method effectively balances the accuracy and throughput for candidate structures ranking by using a hierarchical approach that incorporates different levels of accuracy and computational cost, from empirical force fields to DFT calculations. Third, our method leverages the power of a pretrained, transferable MLFF which predicts the relative stability of different polymorphs with sufficient accuracy to require only a practical number of DFT calculations. The pretrained MLFF can be further customized on a per molecule basis with molecule specific DFT calculations to further improve the reliability and further decrease the number of expensive periodic DFT calculations. Considering the well documented high accuracy requirements of successful crystal structure ranking, neither of the current MLFF models used in this work are of sufficient accuracy to completely remove the need for periodic DFT refinement.

The high accuracy and reliability of our method, as demonstrated here, position it for routine crystal structure prediction in drug formulation. In the future, we plan to extend our method to support more complex and/or multi-component systems, such as co-crystals, solvates, hydrates and salts. We also plan to integrate our method with other computational tools, such as solubility, permeability, mechanical properties and crystal morphology predictions, to enable a comprehensive analysis of polymorphs in terms of their structure, stability, function, and performance.

Methods

Our CSP workflow begins by converting a molecule's bonding information to a single, canonicalized three dimensional geometry. This geometry is input to the Schrödinger Force Field Builder (FFBuilder), which examines the coverage of the OPLS4 force-field for the molecule and refits torsion parameters which are not well represented in the default training set with additional QM calculations when needed. This approach differs from other CSP approaches where customized parameters are trained specifically on one molecule or a set of molecules[26,60]. The resulting parameters from our method are transferable to different chemistries and for different applications, whereas the customized parameters used in other CSP methods only have limited scope of application. More detailed discussion on the construction of OPLS4 and automated torsion parameter fitting is reported elsewhere[61,62].

To generate a pool of conformers for molecular crystal packing, the molecule is fragmented, low energy ring states are determined, pyramidal nitrogens with unique inversion states are identified, and energy profiles of each flexible torsion are obtained. Sample angles for each flexible torsion are determined using a combination of simple chemical rules and automated detection of low energy regions. All combinations of the chosen torsion angles, ring states, and pyramidal nitrogen states are placed onto the molecule fragments. Fragments with low energy according to the OPLS4 force field are combined to produce a diverse pool of unrelaxed conformers. The conformers within 8 kcal/mol of the global minimum are then relaxed, combined with low-energy unrelaxed conformers, and deduplicated to produce the conformer pool[62,63].

Our crystal packing search treats the sampled conformers as rigid building blocks and only targets crystals with $Z' = 1$ currently. Molecular clusters constructed from initial guesses of unevenly distributed parameters corresponding to basic symmetry elements, such as translation, inversion, rotation, screw, and glide are minimized against these parameters. Additional symmetry elements are searched sequentially for each cluster until unit cells can be constructed for the space group under consideration. The cluster size ranges from 2 for an inversion dimer to 36 for the final representation of the 3D crystal in

some space groups. For example, the P1 space group contains three translation symmetries, and we effectively search low-energy periodic structures in one, two, and three dimensions consecutively corresponding to clusters with 3, 6, and 12 molecules (Fig. 1a). When multiple different symmetries exist, the search order matters. Therefore we search with different symmetry orderings for these space groups. The search scope is determined by the energy tolerance of the symmetrized clusters at each search step, and intermediate clusters with energies lower than 6 kcal/mol per molecule from the global minimum of that step are retained for the following steps. As a result, this systematic and hierarchical approach does not need a feedback loop to determine a termination point, unlike other CSP methods based on random sampling[64–66].

After the packing search, the predicted crystals pass through a multi-stage energy relaxation and filtering of increasing accuracy and cost, including MD, QRNN, and DFT stages where all structural degrees of freedom are fully flexible to relax. At each stage, high energy configurations are discarded. The crystal candidates first go through structural relaxation and filtering using symmetry constrained molecular dynamics with the OPLS4 force field. Then, the machine learning force field (MLFF) is used for relaxation and reranking of the remaining crystals. Following this, a coarse DFT relaxation using PBE-D3 is performed followed by more DFT evaluations with increasingly accurate settings and functionals to obtain the final energy ranking. On average, it takes about 20 GPU seconds to relax a crystal with MD, 20 CPU minutes with QRNN, and 60 CPU hours with DFT.

The temperature dependent free energy calculation uses the previously established pseudo super critical path (PSCP) and temperature replica exchanges (REMD) method[20,21]. The reference state for the PSCP free energy calculation is performed at 300 K, and the temperature REMD is run in the range of 100-500 K under the NPT ensemble. The free energy changes along the temperatures are extrapolated to 0 K, and the 0 K enthalpies are corrected to be the same as the periodic DFT relative energy results. All the free energy calculations are done with Desmond on GPU with the OPLS4 force field.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The top predicted candidate structures for all 66 molecules in the validation set within 3 kcal/mol energy window ordered by the r^{2}SCAN-D3 functional generated in this study have been deposited in the github repository [https://github.com/bsantra/csp-validation-data]. The .cif files are indexed with the ranking predicted by r^{2}SCAN-D3. Detailed information about the MLFF and DFT calculations, and the comparison between predicted versus experimentally available crystals structures are available in the supplementary information. Source data are provided with this paper.

Code availability

The periodic DFT calculations were performed using Quantum Espresso version 7.2 available via Schrodinger Materials Sciences Suite version 2025-1. The candidate crystal structures were generated using the systematic packing search method described in the methods section of the paper, with the pseudo code in the supplementary information, and available via the crystal structure prediction software package in Schrodinger's software suite version 2025-1. The OPLS4 force field, Desmond, and QRNN packages are available in Schrodinger's software suite version 2024-2 and forward. An evaluation license to reproduce the work reported here is available upon request.

References

1. Bernstein, J. *Polymorphism in Molecular Crystals*; Oxford University Press, https://doi.org/10.1093/oso/9780199655441.001.0001 (2020).
2. Bučar, D.-K., Lancaster, R. W. & Bernstein, J. Disappearing polymorphs revisited. *Angew. Chem. Int. Ed.* **54**, 6972–6993 (2015).
3. Chemburkar, S. R. et al. Dealing with the impact of ritonavir polymorphs on the late stages of bulk drug process development. *Org. Process Res. Dev.* **4**, 413–417 (2000).
4. Wolff, H.-M., Quere, L. & Riedner, J. Polymorphic form of rotigotine. European patent EP2215072B1 (2010).
5. Price, S. L., Braun, D. E. & Reutzel-Edens, S. M. Can computed crystal energy landscapes help understand pharmaceutical solids? *Chem Commun* **52**, 7065–7077 (2016).
6. Abramov, Y. A., Sun, G. & Zeng, Q. Emerging landscape of computational modeling in pharmaceutical development. *J. Chem. Inf. Model.* **62**, 1160–1171 (2022).
7. Firaha, D. et al. Predicting crystal form stability under real-world conditions. *Nature* **623**, 324–328 (2023).
8. Sun, G. et al. Current state-of-the-art in-house and cloud-based applications of virtual polymorph screening of pharmaceutical compounds: a challenging case of AZD1305. *Cryst. Growth Des.* **21**, 1972–1983 (2021).
9. Broo, A. & Nilsson Lill, S. O. Transferable force field for crystal structure predictions, investigation of performance and exploration of different rescoring strategies using DFT-D methods. *Acta Crystallogr. Sect. B* **72**, 460–476 (2016).
10. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The cambridge structural database. *Acta Crystallogr. Sect. B* **72**, 171–179 (2016).
11. Hunnisett, L. M. et al. The seventh blind test of crystal structure prediction: structure generation methods. *Acta Crystallogr. Sect. B* **80**, 517–547 (2024).
12. Hunnisett, L. M. et al. The seventh blind test of crystal structure prediction: structure ranking methods. *Acta Crystallogr. Sect. B* **80**, 548–574 (2024).
13. Reilly, A. M. et al. Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallogr. Sect. B* **72**, 439–459 (2016).
14. Zhang, P. et al. Harnessing cloud architecture for crystal structure prediction calculations. *Cryst. Growth Des.* **18**, 6891–6900 (2018).
15. Hoja, J. et al. Reliable and practical computational description of molecular crystal polymorphs. *Sci. Adv.* **5**, eaau3338. https://doi.org/10.1126/sciadv.aau3338.
16. Chan, E. J., Shtukenberg, A. G., Tuckerman, M. E. & Kahr, B. Crystal structure prediction as a tool for identifying components of disordered structures from powder diffraction: a case study of benzamide II. *Cryst. Growth Des.* **21**, 5544–5557 (2021).
17. Kendrick, J., Stephenson, G. A., Neumann, M. A. & Leusen, F. J. J. Crystal structure prediction of a flexible molecule of pharmaceutical interest with unusual polymorphic behavior. *Cryst. Growth Des.* **13**, 581–589 (2013).
18. Nyman, J. & Reutzel-Edens, S. M. Crystal structure prediction is changing from basic science to applied technology. *Faraday Discuss* **211**, 459–476 (2018).
19. Jacobson, L. D. et al. Transferable neural network potential energy surfaces for closed-shell organic molecules: extension to ions. *J. Chem. Theory Comput.* **18**, 2354–2366 (2022).
20. Abraham, N. S. & Shirts, M. R. Statistical mechanical approximations to more efficiently determine polymorph free energy differences for small organic molecules. *J. Chem. Theory Comput.* **16**, 6503–6512 (2020).
21. Yang, M. et al. Prediction of the relative free energies of drug polymorphs above zero kelvin. *Cryst. Growth Des.* **20**, 5211–5224 (2020).

22. Francia, N. F., Price, L. S. & Salvalaglio, M. Reducing crystal structure overprediction of ibuprofen with large scale molecular dynamics simulations. *CrystEngComm*. **23**, 5575–5584 (2021).
23. Francia, N. F., Price, L. S., Nyman, J., Price, S. L. & Salvalaglio, M. Systematic finite-temperature reduction of crystal energy landscapes. *Cryst. Growth Des.* **20**, 6847–6862 (2020).
24. Putra, O. D., Ottosson, J., Nilsson Lill, S. O. & Pettersen, A. Understanding crystal structures to guide form selection of active pharmaceutical ingredients: a case study of AZD9567. *Cryst. Growth Des.* **22**, 535–546 (2022).
25. Price, S. L. Why Don't We Find More Polymorphs? *Acta Crystallogr. Sect. B* **69**, 313–328 (2013).
26. Li, X., Neumann, M. A. & van de Streek, J. The application of tailor-made force fields and molecular dynamics for NMR crystallography: a case study of free base cocaine. *IUCrJ* **4**, 175–184 (2017).
27. Ehlert, S. et al. r2SCAN-D4: dispersion corrected meta-generalized gradient approximation for general chemical applications. *J. Chem. Phys.* **154**, 061101 (2021).
28. Furness, J. W., Kaplan, A. D., Ning, J., Perdew, J. P. & Sun, J. Accurate and numerically efficient r2SCAN meta-generalized gradient approximation. *J. Phys. Chem. Lett.* **11**, 8208–8215 (2020).
29. Grimme, S., Hansen, A., Ehlert, S. & Mewes, J.-M. r2SCAN-3c: A "Swiss Army Knife" Composite Electronic-Structure Method. *J. Chem. Phys.* **154**, 064103 (2021).
30. Mortazavi, M. et al. Computational polymorph screening reveals late-appearing and poorly-soluble form of rotigotine. *Commun. Chem.* **2**, 70 (2019).
31. Newman, J. A. et al. From powders to single crystals: a crystallographer's toolbox for small-molecule structure determination. *Mol. Pharm.* **19**, 2133–2141 (2022).
32. Fulton, B. & Goa, K. L. Olanzapine. *Drugs* **53**, 281–298 (1997).
33. Reutzel-Edens, S. M. & Bhardwaj, R. M. Crystal forms in pharmaceutical applications: olanzapine, a gift to crystal chemistry that keeps on giving. *IUCrJ* **7**, 955–964 (2020).
34. Bhardwaj, R. M. et al. Exploring the experimental and computed crystal energy landscape of olanzapine. *Cryst. Growth Des.* **13**, 1602–1617 (2013).
35. LeBlanc, L. M. & Johnson, E. R. Crystal-energy landscapes of active pharmaceutical ingredients using composite approaches. *CrystEngComm* **21**, 5995–6009 (2019).
36. Bond, A. D., Boese, R. & Desiraju, G. R. On the polymorphism of aspirin: crystalline aspirin as intergrowths of two "Polymorphic" domains. *Angew. Chem. Int. Ed.* **46**, 618–622 (2007).
37. Anyfanti, G., Husanu, E., Andrusenko, I., Marchetti, D. & Gemmi, M. The crystal structure of olanzapine form III. *IUCrJ* **11**, 843–848 (2024).
38. Hall, A. et al. Discovery of Sodium 6-[(5-Chloro-2-{[(4-Chloro-2-Fluorophenyl)Methyl]Oxy}phenyl)Methyl]-2-Pyridinecarboxylate (GSK269984A) an EP1 Receptor Antagonist for the Treatment of Inflammatory. *Pain. Bioorg. Med. Chem. Lett.* 19, 2599–2603 (2009).
39. Ismail, S. Z., Anderton, C. L., Copley, R. C. B., Price, L. S. & Price, S. L. Evaluating a crystal energy landscape in the context of industrial polymorph screening. *Cryst. Growth Des.* **13**, 2396–2406 (2013).
40. Beran, G. J. O. et al. How many more polymorphs of ROY remain undiscovered. *Chem Sci* **13**, 1288–1297 (2022).
41. Bhardwaj, R. M. et al. A prolific solvate former, galunisertib, under the pressure of crystal structure prediction, produces ten diverse polymorphs. *J. Am. Chem. Soc.* **141**, 13887–13897 (2019).
42. Nyman, J., Yu, L. & Reutzel-Edens, S. M. Accuracy and reproducibility in crystal structure prediction: the curious case of ROY. *CrystEngComm* **21**, 2080–2088 (2019).
43. Greenwell, C. & Beran, G. J. O. Inaccurate conformational energies still hinder crystal structure prediction in flexible organic molecules. *Cryst. Growth Des.* **20**, 4875–4881 (2020).
44. Rietveld, I. B. & Céolin, R. Rotigotine: unexpected polymorphism with predictable overall monotropic behavior. *J. Pharm. Sci.* **104**, 4117–4122 (2015).
45. Williams, M. J. et al. Process development of the HCV NS5B site D inhibitor MK-8876. *Org. Process Res. Dev.* **20**, 1227–1238 (2016).
46. Stephenson, G. A., Kendrick, J., Wolfangel, C. & Leusen, F. J. J. Symmetry breaking: polymorphic form selection by enantiomers of the melatonin agonist and its missing polymorph. *Cryst. Growth Des.* **12**, 3964–3976 (2012).
47. Goa, K. L. & Spencer, C. M. Bicalutamide in advanced prostate cancer. *Drugs Aging* **12**, 401–422 (1998).
48. Vega, D. R., Polla, G., Martinez, A., Mendioroz, E. & Reinoso, M. Conformational polymorphism in bicalutamide. *Int. J. Pharm.* **328**, 112–118 (2007).
49. Simons, L. J. et al. The synthesis and structure–activity relationship of substituted N-phenyl anthranilic acid analogs as amyloid aggregation inhibitors. *Bioorg. Med. Chem. Lett.* **19**, 654–657 (2009).
50. Samas, B., Clark, W. D., Li, A.-F., Pickard, F. C. I. & Wood, G. P. F. Five degrees of separation: characterization and temperature stability profiles for the polymorphs of PD-0118057 (molecule XXIII). *Cryst. Growth Des.* **21**, 4435–4444 (2021).
51. Campeta, A. M. et al. Development of a targeted polymorph screening approach for a complex polymorphic and highly solvating API. *J. Pharm. Sci.* **99**, 3874–3886 (2010).
52. Chekal, B. P. et al. The challenges of developing an API crystallization process for a complex polymorphic and highly solvating system. Part I. *Org. Process Res. Dev.* **13**, 1327–1337 (2009).
53. Vasileiadis, M., Pantelides, C. C. & Adjiman, C. S. Prediction of the crystal structures of axitinib, a polymorphic pharmaceutical molecule. *Chem. Eng. Sci.* **121**, 60–76 (2015).
54. Blade, H. et al. Conformations in solution and in solid-state polymorphs: correlating experimental and calculated nuclear magnetic resonance chemical shifts for tolfenamic acid. *J. Phys. Chem. A* **124**, 8959–8977 (2020).
55. Baias, M. et al. 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**, 17501–17507 (2013).
56. Case, D. H., Campbell, J. E., Bygrave, P. J. & Day, G. M. Convergence properties of crystal structure prediction by quasi-random sampling. *J. Chem. Theory Comput.* **12**, 910–924 (2016).
57. Friesner, R. A. et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **47**, 1739–1749 (2004).
58. Nespolo, M. Teaching Edition of It International Tables for Crystallography: Crystallographic Symmetry. Edited by Mois I. Aroyo. IUCr/Wiley, 2021. Softcover, Pp. Xii + 236. ISBN 978-0-470-97422-3. Price GBP 29.99. *Acta Crystallogr. Sect. A* **77**, 506–508 (2021).
59. O'Connor, D., Bier, I., Hsieh, Y.-T. & Marom, N. Performance of dispersion-inclusive density functional theory methods for energetic materials. *J. Chem. Theory Comput.* **18**, 4456–4471 (2022).
60. Neumann, M. A. Tailor-made force fields for crystal-structure prediction. *J. Phys. Chem. B* **112**, 9810–9829 (2008).
61. Harder, E. et al. OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. *J. Chem. Theory Comput.* **12**, 281–296 (2016).
62. Lu, C. et al. OPLS4: improving force field accuracy on challenging regimes of chemical space. *J. Chem. Theory Comput.* **17**, 4291–4300 (2021).
63. Watts, K. S. et al. ConfGen: a conformational search method for efficient generation of bioactive conformers. *J. Chem. Inf. Model.* **50**, 534–546 (2010).
64. Kitaigorodskii, A. I. (Aleksandr I., 1914-. *Molecular Crystals and Molecules [by] A. I. Kitaigorodsky;* Physical chemistry; v. 29.; Academic Press: New York, 1973.

65. Perlstein, J. Molecular Self-Assemblies. 2. A Computational Method for the Prediction of the Structure of One-Dimensional Screw, Glide, and Inversion Molecular Aggregates and Implications for the Packing of Molecules in Monolayers and Crystals. *J. Am. Chem. Soc.* **116**, 455–470 (1994).
66. Gavezzotti, A. Generation of possible crystal structures from molecular structure for low-polarity organic compounds. *J. Am. Chem. Soc.* **113**, 4622–4629 (1991).

Acknowledgements

The authors thank Drs Casey Brock, Alexandr Fonari and Wei Chen for many helpful discussions, and thank Drs Shiva Sekharan, Peter Skrdla and Paul Devine for proofreading the manuscript.

Author contributions

D.Z., I.B., B.S., L.D.J., C.W., H.Y., R.A., R.A.F. and L.W. designed the research; D.Z., I.B., B.S., L.D.J., C.W., H.Y. and L.W. performed the research; D.Z., I.B., B.S., L.D.J., C.W., and L.W. wrote the initial version of the manuscript.; A. G.S. and B.R.A. contributed to the blind study on the flupyradifurone polymorph prediction; all coauthors revised the manuscript and approved the final version of the manuscript.

Competing interests

The authors declare the following competing interests: R.A.F. has a significant financial stake in, is a consultant for, and is on the Scientific Advisory Board of Schrodinger, Inc. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-57479-1.

Correspondence and requests for materials should be addressed to Lingle Wang.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.