ROYAL SOCIETY
OF CHEMISTRY

PAPER

Check for updates

Automated high-throughput organic crystal structure prediction *via* population-based sampling

Qiang Zhu [*a] and Shinnosuke Hattori [*b]

With advancements in computational molecular modeling and powerful structure search methods, it is now possible to systematically screen crystal structures for small organic molecules. In this context, we introduce the Python package High-Throughput Organic Crystal Structure Prediction (HTOCSP), which enables the prediction and screening of crystal packing for small organic molecules in an automated, high-throughput manner. Specifically, we describe the workflow, which encompasses molecular analysis, force field generation, and crystal generation and sampling, all within customized constraints based on user input. We demonstrate the application of HTOCSP by systematically screening organic crystals for 100 molecules using different sampling strategies and force field options. Furthermore, we analyze the benchmark results to understand the underlying factors that may influence the complexity of the crystal energy landscape. Finally, we discuss the current limitations of the package and potential future extensions.

I. Introduction

Molecular solids, referring to substances consisting of discrete molecules that are held together by relatively weak intermolecular forces, have been extensively used in chemical,[1–4] pharmaceutical[5] and semiconductor industries.[6–8] In these fields, the development of new organic materials with targeted properties relies heavily on understanding and controlling intermolecular interactions within the crystal structure.

Nowadays, data-driven computer simulation has been playing an increasingly important role in materials development.[9] Specifically, computational high-throughput screening of the existing organic crystals has become popular for the design of new materials with improved physical properties.[10–15] However, most of the screening work has been limited by the availability of experimentally resolved structures. It has been found that crystal polymorphism can effectively modify the physical properties.[16,17] Nevertheless, to experimentally synthesize new polymorphs and characterize their structural properties is rather time-demanding and expensive. From a practical standpoint, having the capability to screen likely organic crystal formations before laboratory synthesis and characterization would be highly valuable.[18]

In the past decade, there has been rapid development in the community of crystal structure prediction (CSP) for small organic molecules.[19–24] The idea of CSP is to predict a short list of stable or metastable crystal packings that are likely to be observed in the experiment through powerful structure exploratory algorithms.[25–28] Practically speaking, conducting a successful organic crystal prediction requires the integration of several different computational pipelines including force field generation, structure sampling and post-process energy ranking. Most practitioners rely on subscriptions to commercial licenses to run simulations and analyze the results. Due to the license restriction, it is often hard to reproduce previously published results, even for an experienced researcher.

In the field of CSP for inorganic materials, there have been plenty of code choices (*e.g.*, USPEX,[29] AIRSS,[30] CALYPSO,[31] XtalOpt,[32] and GASP[33]) that are either completely open source or free for academic researchers. The availability of these tools has enabled significant progress in the field, allowing researchers to explore complex systems and improve their understanding of inorganic material behavior at a fundamental level.[25] While the field of CSP has been progressing rapidly, the code choices[29,34,35] for the prediction of organic materials are still limited. More importantly, to the best of our knowledge, there is currently no open-source CSP code specifically designed for the high-throughput prediction of organic crystals. As organic crystals are increasingly relevant in fields like pharmaceuticals, organic electronics, and molecular materials, the development of such tools is becoming ever more critical to support the growing demand for rapid and accurate crystal structure prediction.

To promote the open-source activity in organic CSP, we have developed an open source code High-Throughput Organic Crystal Structure Prediction (HTOCSP) that allows the automated prediction of organic crystal structures with a minimal input, by leveraging several existing open-source infrastructures[36–40] in molecular modeling. In the following section, we will begin by detailing the workflow, covering molecular analysis, force field generation, and crystal sampling. This will be followed by

aDepartment of Mechanical Engineering and Engineering Science, University of North Carolina at Charlotte, Charlotte, NC 28223, USA. E-mail: qzhu8@uncc.edu

bAdvanced Research Laboratory, Research Platform, Sony Group Corporation, 4-14-1 Asahi-cho, Atsugi-shi 243-0014, Japan. E-mail: shinnosuke.hattori@sony.com

a benchmark study on 100 molecules, utilizing different sampling strategies and force field options. Finally, we will analyze the test results and conclude with a discussion of the current limitations and potential future extensions.

II. The HTOCSP workflow

In practical crystal structure prediction, the most common objective is to generate a shortlist of plausible crystal packings for a given molecular system. To address this challenge, multiple computational pipelines are often required to manage the system at both the molecular and crystal levels. In HTOCSP, we divide the entire crystal structure prediction process into six sequential tasks, as illustrated in Fig. 1.

A Molecular analyzer

In organic chemistry, the most common way to represent a molecule is by using the SMILES (Simplified Molecular Input Line Entry System) string. SMILES strings use standard ASCII characters to denote atoms and bonds in a molecule. In HTOCSP, we utilize the RDKit library[38] to convert a SMILES string into 3D coordinates and analyze the flexible dihedral angles in the input molecule. For multicomponent systems, such as cocrystals, salts, and hydrates, each molecule must be handled separately.

In many cases, molecular SMILES strings and crystallographic information can be found through public online repositories such as PubChem[41] and the Cambridge Crystallographic Data Center (CCDC). However, it is important to note that some erroneous entries may exist. Therefore, the user is strongly recommended to verify the correctness of the retrived data to ensure it accuracy and reliability.

B Force field maker

Once the molecular information is known, we can proceed to extract the force field (FF) parameters. Currently, HTOCSP supports two main types of FF models based on the AMBER-TOOLS[39] utility. The first type is the widely used General Amber Force Field (GAFF),[42] which covers C–H–O–N–S–P–F–Cl–Br–I elements to model small organic molecules. The second type is the SMIRNOFF (SMIRKS Native Open Force Field), derived from the OpenFF initiative,[40] which offers a more flexible and extensible way to describe force fields compared to traditional atom type-based methods. The OpenFF types support elements including C–H–O–N–S–P–F–Cl–Br–Li–Na–K–Rb–Cs. Due to its flexibility, it is also more convenient to extend support to other elements if needed. In addition, AMBERTOOLS is used to compute atomic partial charges using various schemes, such as Gasteiger, MMFF94, and AM1-BCC.[43]

The extracted force field parameters, including descriptions of charges, bonds, angles, torsions, and nonbonded van der Waals interactions, can be saved as an Extensible Markup Language (XML) file according to the OpenFF standard. Optionally, companion topology files can also be generated in different formats for later use in structural relaxation codes. While we recommend the use of fixed GAFF/OpenFF parameters for most
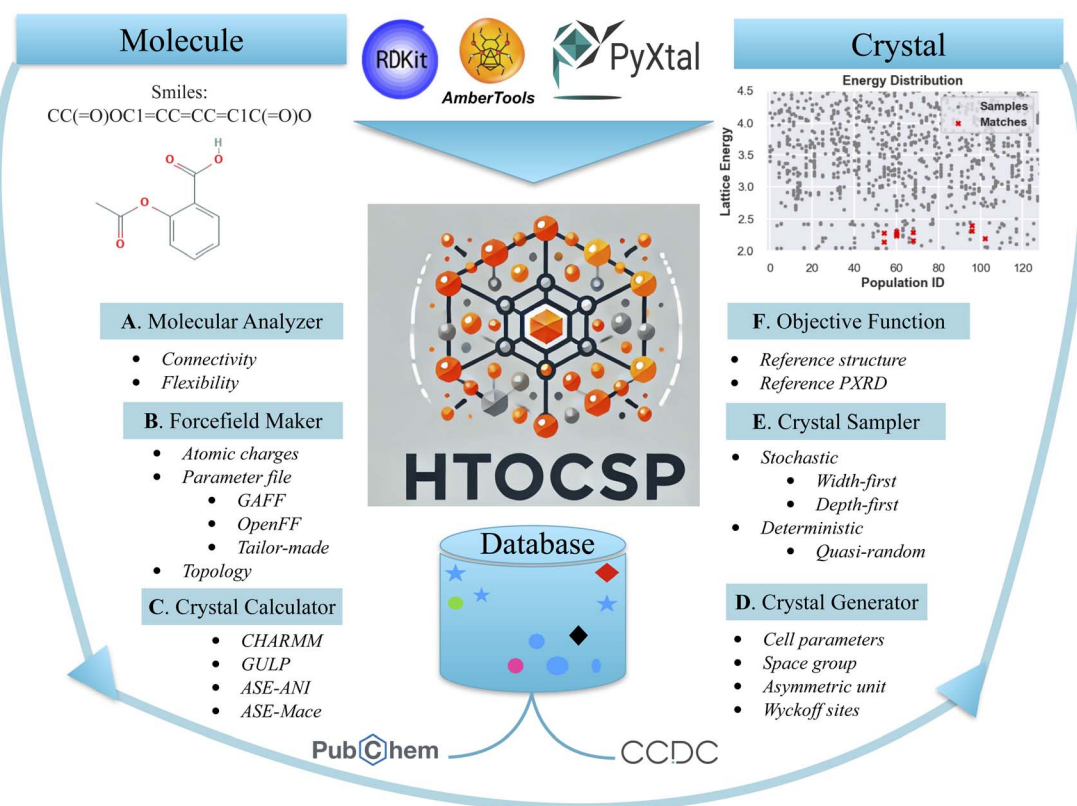

Fig. 1   The overview of the HTOCSP workflow.

CSP applications, it is possible that the default GAFF or OpenFF parameters may poorly describe the system. In such cases, we allow users to retrain the parameters for better accuracy, similar to the tailor-made force field concept proposed by Neumann and coworkers[44] (see an extended discussion in Section V).

Finally, all the aforementioned functions have been implemented to the pyocse framework[37] based on our earlier work.[45] It is important to note that the current force field parameters were mainly fitted by using data under standard temperature/pressure conditions. Therefore, it is not recommended to use it for the search for polymorphs under extreme conditions.

C Symmetry-constrained structure calculators

With the classical force field in place, we are ready to build the structure calculator to perform geometry optimization and evaluate the energy of candidate crystals. Unlike ordinary geometry optimization tasks in molecular modeling, our goal is to optimize the crystal structure without breaking its symmetry or disrupting molecular connectivity.[46] Currently, HTOCSP supports two symmetry-adapted molecular simulation codes suitable for CSP: GULP[47] and CHARMM.[48] In both codes, it is necessary to specify molecular topology and symmetry operations, and then instruct the calculator to optimize the cell parameters and molecular coordinates within the asymmetric unit. Optionally, the cell parameters can be fixed during optimization if they are already known.

In our earlier studies, GULP has been successfully used to address a range of practical CSP challenges.[46,49–52] However, CHARMM generally offers faster performance than GULP, as it provides better support for the symmetry-adapted implementation of the Particle Mesh Ewald (PME) method.[53] Consequently, CHARMM is the default FF calculator choice in HTOCSP due to its faster computational speed.

Beyond classical force field models, there are ongoing efforts to develop generic or universal machine learning force fields for material and molecular screenings. Notable examples include ANI[54] and MACE.[55,56] These models have been trained comprehensively on extensive datasets, and the resulting parameters are available as an open source for public access. Recent surveys[28,57] indicate that machine learning force fields (MLFFs) generally reproduce experimental geometries more accurately and provide more reasonable rankings of crystal structures.

Despite these promising results, it's important to note that most generic or universal MLFFs are trained primarily on known experimental structures and their nearby basins of attraction. Consequently, they may struggle with geometry relaxation if the initial structure is far from equilibrium or has a high energy. Due to these limitations, HTOCSP currently supports the use of ANI and MACE only for post-energy re-ranking on pre-optimized crystals generated by GAFF or OpenFF. These structure optimizations can be performed using a customized ASE calculator,[58] with the given space group symmetry constraints.

D Crystal generator

In a standard CSP task, the primary objective is often to search for plausible crystal packings within a shortlist of common space groups.[26,28,46] Additional information, such as cell parameters pre-determined from experimental powder X-ray diffraction, can also be incorporated as input. This reduces the goal of CSP to generating multiple trial crystal packings within these constraints. To address this challenge, a structure engine is needed to generate random symmetric crystals.

Recently, we have developed a Python program called PyXtal that handles various structure generation and symmetry analysis tasks for 0D/1D/2D/3D atomic and molecular crystals.[59] For the module dedicated to 3D molecular crystals, PyXtal enables the generation of trial structures with a customized asymmetric unit within a specified space group. In the asymmetric unit, one can place one or multiple molecules at the general Wyckoff positions. If the symmetry of the input molecule is compatible with the site symmetry of a given Wyckoff position, PyXtal also supports the assignment of the molecule to a special Wyckoff position with higher symmetry.[28] This feature is particularly useful for generating crystal structures of high-symmetry molecules with a fractional Z' number (i.e., less than one molecule per asymmetric unit).

Additionally, it is quite common to encounter molecular crystal data labeled with a space group in non-standard settings. For instance, the standard setting for the most common molecular space group 14 is labeled as $P12_1/c1$ (shortened as $P2_1/c$). However, this space group can also be represented as $P12_1/n1$ or $P12_1/a1$, where the glide plane translates molecules in the diagonal direction or along the a-axis. Alternatively, one can choose the unique axis to be along a ($P2_1/c11$) or c ($P112_1/c$). Taking into account variations in space group notations due to the choice of origin, there are 530 concise space group settings according to Hall.[60] In PyXtal, we allow the use of Hall notations to generate crystals in non-standard space group settings.

To illustrate the use of PyXtal, we provide a few examples in the following listing.

```python
from pyxtal import pyxtal
xtal = pyxtal(molecular=True)

# A random crystal in the 4e site of spg P21/c
xtal.from_random(3, 14, ['benzene'], [4],
                sites=[["4e"]])

# A random crystal in the 2a site of spg P21/c
xtal.from_random(3, 14, ["benzene"], [2],
                sites=[["2a"]])

# A random crystal in the '4e' site of spg P21/n
# P21/n correspond to Hall number 82
xtal.from_random(3, 82, ["benzene"], [4],
                sites=[["4e"]],
                use_hall=True)

# Random crystal from smiles
# Make sure rdkit is installed
mol = ["CC(=O)OC1=CC=CC=C1C(=O)O.smi"] # aspirin
xtal.from_random(3, 14, mol, [4],
                sites=[["4e"]])
```

Listing 1 PyXtal script to organic crystals.

E Population-based sampling methods

After a random trial crystal is built, it needs to be relaxed to a configuration corresponding to a local energy minimum on

the potential energy surface. By repeating this procedure many times, one can identify a complete set of plausible crystal packings that may exist in reality. Following this, their relative thermodynamic (and kinetic) stabilities must be evaluated using an accurate energy model.

With the aforementioned modules, one is ready to run a CSP calculation using a brute-force random search approach. This involves repeatedly generating random trial crystals and then using a structure calculator to perform geometry minimization, thereby obtaining low-energy structures. However, the success rate of finding the desired target may decrease exponentially as the search space becomes larger.

To address this challenge, HTOCSP implements several population-based methods to achieve more efficient structure sampling. In global optimization, population-based methods refer to a class of optimization techniques that maintain and evolve a population of candidate solutions throughout the optimization process. Unlike traditional optimization methods that work with a single solution (*e.g.*, gradient descent), population-based methods explore the search space using multiple solutions simultaneously. This approach helps to avoid local minima and increases the likelihood of finding a global optimum.

In the past, several population-based methods have been successfully applied to address challenges in both inorganic and organic crystal structure prediction, utilizing genetic/evolutionary algorithms[32,33,35,61] and particle swarm optimization (PSO).[31] These approaches often require mechanisms to generate new solutions by sharing information from multiple sources (*e.g.*, the crossover operation in genetic algorithms or velocity updates *via* the Cognitive Social Component in PSO). However, handling such processes without breaking crystal symmetry is often challenging. Therefore, we decided to adopt a simplified population-based approach that avoids information sharing between individuals within the same generation during the evolution process.

Specifically, we begin with a population of randomly generated trial crystal structures and update each individual based on its own evolution history. For a given chemical, a large number of hypothetical structures can exist. After relaxation, their trajectories form basins of attraction on the potential energy surface. The probability of successfully finding a low-energy structure depends on the shapes and sizes of the hyper-volumes of these basins of attraction and the details of the search.[30,62,63] And we hypothesize that the target crystals (*i.e.*, the likely observed crystal forms in experiments) form stable basins on the energy surface. However, due to artifacts in force fields and the constraints of symmetry and zero temperature during geometry optimization, these basins are often surrounded by many shallow energy minima. Consequently, a true basin and its surrounding shallow minima collectively form a meta-basin (as illustrated in Fig. 2).

In a purely random search, there is a good chance of sampling within the meta-basin and landing in one of the surrounding shallow minima, but this may miss the true target structure (*i.e.*, the center of the meta-basin) within a limited number of sampling attempts. Therefore, a wise approach is to
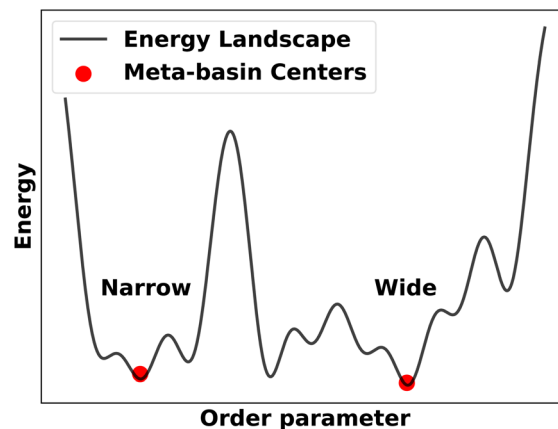

Fig. 2 The schematic representation of two meta-basins.

perform a further local search to locate the centers of these meta-basins based on historically explored energy minima. Depending on the extent of effort in utilizing historical information, we devised two stochastic sampling strategies and one deterministic strategy as follows.

1 Stochastic width-first sampling. This approach resembles random sampling but with a structured twist.[30] Initially, all structures are randomly generated and then relaxed using the structure calculator. These structures are ranked by their final energies. In subsequent generations, a portion of the high-ranked (low-energy) structures undergo mutation by perturbing the cell parameters and molecular Wyckoff sites while preserving crystal symmetry. Conversely, the high-energy structures are discarded, and new random structures are generated in their place. The fraction of structures subject to mutation is a hyperparameter ranging between 0 and 1. When this parameter is set to 0, the method reverts to a pure random sampling algorithm.

Compared to pure random sampling, this method dedicates additional effort to exploring the surroundings of previously visited low-energy regions, thereby increasing the success rate of identifying more promising low-energy basins. Despite these extra efforts, the majority of the sampling still focuses on generating new random solutions, which is why we refer to this method as *Width-First Sampling* (WFS). This type of sampling is particularly useful for detecting the target structure within a *narrow meta-basin* with limited spread width, as illustrated in Fig. 2.

2 Stochastic depth-first sampling. In a real system, the target structure may be located within a *wide meta-basin* surrounded by numerous shallow minima (see Fig. 2). In such cases, WFS may repeatedly visit some of these surrounding shallow minima without successfully identifying the true target state. To address this issue, we propose an alternative sampling strategy that prioritizes exploring the surroundings of already identified low-energy regions.

In this strategy, for each individual, we select either the most recently optimized structure (with a probability of 0.7) or the previously identified lowest-energy structure (with a probability

of 0.3) to perform mutations and generate new samples. This process continues for 20 generations before restarting with new random samples. This mechanism is analogous to performing parallel basin hopping[64] on multiple random samples. As a result, it achieves better sampling within each individual meta-basin, increasing the likelihood of identifying a true target surrounded by many shallow energy minima. Unlike WFS, which focuses on breadth, this approach is more focused on learning the shape of each meta-basin, which is why we refer to it as *Depth-First Sampling* (DFS).

3 Deterministic quasi-random sampling. In the CSP community, the deterministic quasi-random sampling (QRS) method is also popular.[34] Quasi-random sequences are deterministic sequences of numbers designed to cover a multidimensional space more uniformly than uncorrelated random sequences. Unlike purely random sequences, which can result in clustering and gaps, quasi-random sequences aim to fill the space evenly, ensuring better coverage of the domain. Due to their deterministic nature, quasi-random sequences can be used to systematically cover the space and test the convergence of results without introducing randomness.

To perform QRS on molecular crystals, we consider cell parameters and molecular positions separately. For example, the aspirin crystal can be represented by the following variables (Table 1).

In the context of population-based methods, we generate a valid QRS sample for the cell parameters in each generation, followed by sampling the molecular positions, orientations, and dihedral angles with a sufficiently large population size. The user must define the lower and upper bounds for cell lengths to constrain the range of QRS samples, as well as specify a range of acceptable volume values to filter out unit cells that are either too large or too small.

4 Termination mechanisms. Finally, all three sampling algorithms have been implemented in the *pyxtal.optimize* module, and they can be conveniently accessed through HTOCSP. In practical CSP, we recommend a minimum population size of 256 in each generation. For the initial guess, the Bravis lattices are generated randomly within a range of allowed volumes. The crystal volumes (V) are estimated based on the molecular volumes and multiplicities. Then, two numerical factors (*e.g.*, 0.7 and 1.3) are applied to define the lower and upper bound. The minimum and maximum vector lengths are applied based on its molecular shape and crystal symmetry, and the minimum and maximum angles are 30° and 150°. All newly generated structures undergo multiple steps of geometry optimization using either GAFF or OpenFF. At the end of each optimization, it is possible that the unit cell may have very acute or obtuse angles, particularly in triclinic or monoclinic systems. In such cases, the angles are transformed to approach 90° whenever possible.

In principle, all sampling algorithms should run for a sufficiently large number of generations to ensure convergence. There are two mechanisms to terminate the sampling. If the sampling is run in *production* mode (*i.e.*, without knowledge of the target structure), it will stop once the maximum number of generations is reached. If the sampling is run in *validation* mode with the target structure provided, the sampling may stop earlier as long as a good match with the target structure is detected 10 times.

To determine if a good matched structure is found, the StructureMatcher module in Pymatgen[65] is used. In this module, we ignore all H atoms and build a one-to-one map between each molecule in the unit cell, then check the largest root mean squared error (RMSE) between each atomic pair. By default, two structures are considered identical if the fractional length tolerance is 0.3, the site tolerance is less than 0.3 Å, and the angle tolerance is less than 5°. It's worth noting that Mercury[66] also offers a mechanism to compute the RMSE for 15–20 packed molecules *via* the COMPACK algorithm.[67] After extensive comparison, we found that these two functions produce nearly identical results. For ease of code implementation and accessibility, we have chosen to use Pymatgen's StructureMatcher module in HTOCSP.

F Objective functions

In both WFS and DFS runs, the structures need to be ranked according to an objective function. By default, lattice energy is used as the optimization objective. However, it is also possible to choose an alternative objective function based on a similarity measure with respect to a reference powder X-ray diffraction (PXRD) pattern. To use this feature, users simply need to provide the experimentally measured PXRD raw data. During the search process, each structure will be used to compute a normalized similarity score (S) with the given reference PXRD using PyXtal, where a score of 1 indicates a perfect match and a score of 0 indicates no correlation. These S values will be used for ranking the structures and will be recorded in the exported CIF files. This feature can guide the search toward the desired structure if the PXRD data are known. Currently, HTOCSP only supports the single objective optimization based on the PXRD match. Enabling the dual objective Pareto optimization[25] is planned in the next release.

III. Software installation and example usage

As described in the previous section, most of the structure generation and sampling algorithms have been implemented in PyXtal. However, a complete CSP calculation requires additional, more complex software dependencies, such as RDKit, AmberTools, and CHARMM. Therefore, we have integrated all these packages into the HTOCSP platform for the convenience of CSP practitioners.

Table 1 The representation of $P2_1/c$ aspirin form I in QRS

Hall space group number	81
Cell parameters	11.43, 6.49, 11.19, 83.31
Wyckoff index	4e
Molecular XYZ	0.77, 0.57, 0.53
Molecular orientation	48.55, 24.31, 145.9
Molecular dihedral angles	−77.85, −4.40, 170.9

For practical usage, one should first install the required Python packages *via conda install*. Then, download and compile the source codes of CHARMM or GULP. The executable paths and necessary libraries must be added to the environmental variables.

To run the simulation, one typically needs to setup the following Python script.

```
wdir = 'aspirin-simple'
smiles = "CC(=O)OC1=CC=CC=C1C(=O)O"

# Obtain the reference crystal from 1d string
x = "81 11.38  6.48 11.24  96.9 1 0 0.23 0.43 0.03
    -44.6   25.0   34.4  -76.6  -5.2 171.5 0"
rep = representation.from_string(x, [smiles])
xtal = rep.to_pyxtal()

# Sampling
go = WFS(smiles,
        wdir,
        sg,  # list of space groups
        tag = 'aspirin',
        N_gen = gen,
        N_pop = pop,
        N_cpu = ncpu,
        ff_style = 'openff')
go.run(ref_pmg=xtal.to_pymatgen())
go.print_matches(header='Ref-DFS-gaff')
go.plot_results()
```

Listing 2 PyXtal script to run HTOCSP for the aspirin system.

This script will sample the aspirin crystals in *WFS* mode using the OpenFF force field, with a population size of 256 over 20 generations. With 128 parallel cores, the total sampling time is approximately 2.5 generations per minute. Therefore, the entire run of 20 generations can be completed within 10 minutes. When compared to the experimental aspirin form I in space group 14 ($P2_1/c$) with $Z' = 1$, we found a total of 7 matches out of 5120 sampled structures, as shown in Fig. 3. We also found another structure with lower energy corresponding to the form II of aspirin. These two forms have been well known due to very close lattice energy ranking and their relative stability can only be distinguished with the advanced free energy calculation method.[68] Last, it is important to note that the number of matched structures may vary due to the stochastic nature of the sampling algorithm. The confidence in these sampling results will be further discussed in the following section.

In a standard CSP run *via* HTOCSP, one only needs to provide a SMILES string to define the target molecule (either real or hypothetical), a list of space group choices, and a few sampling parameters related to population and generation sizes. If the target system is a cocrystal or salt, a list of SMILES strings and the compositional ratio must be provided. For further instructions, users can refer to the online documentation of HTOCSP[69] to learn how to set up other types of calculations, such as structure searches with respect to the reference PXRD data.

IV. Benchmark results and discussion

In this section, we expand upon the example discussed in Fig. 3. In a typical CSP calculation, the search space grows exponentially with the number of variables, including unit cell vectors, symmetry operations, molecular positions, conformations, orientations, and the number of molecules per asymmetric unit. Under blind test conditions, it is often difficult to be confident in the results obtained from limited sampling efforts. Clearly, some molecules may have more complex and fuzzy energy landscapes compared to others.[26]

To validate the concept of energy landscape complexity, we performed a systematic benchmark test on a set of 100 experimentally reported crystals. We selected these systems to encompass a range of challenges encountered in realistic CSP simulations. The set includes 30 systems from past blind tests,[19–24] 38 representative organic semiconducting (OSC) molecules, primarily consisting of thiophene rings, 14 polynuclear aromatic hydrocarbons (PAHs),[70] and 18 other systems based on our previous research for various purposes (Misc). These systems vary widely, including single-component systems, multi-component cocrystals/salts, fractional to multiple Z', and from rigid molecules to complex molecules with multiple flexible dihedral angles. In addition, a large fraction of the structures are not the most stable polymorph but are metastable. A complete list of their associated Cambridge Structural Database (CSD) entries and basic descriptions is provided in Appendix Table 2.

For each system, we performed four types of simulations: (1) WFS with GAFF (WFS-GAFF), (2) DFS with GAFF (DFS-GAFF), (3) WFS with GAFF relaxation and ANI energy evaluation (WFS-GAFF-ANI), and (4) DFS with GAFF relaxation and ANI energy evaluation (DFS-GAFF-ANI). In each run, we limited the search space based on the experimentally determined space group and Z'. If the target crystal had a fractional Z' due to occupation of a special Wyckoff site, we followed the symmetry relation to transform it into a subgroup representation where the molecules occupy the general Wyckoff site.[59] In all calculations, we set a population size of 256 with a maximum of 500 generations. However, the calculation could be terminated earlier if a total of at least 10 matched structures were found. In each simulation, we defined the success rate (SR) as the ratio of the number of
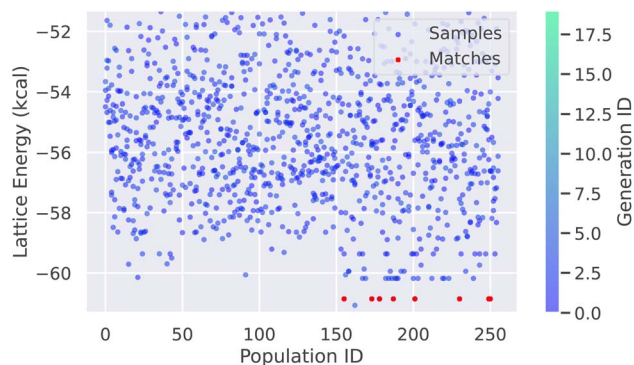
Fig. 3 The CSP results based on WFS and OpenFF for the aspirin crystal in space group 14 with $Z' = 1$ are presented in this plot. Each dot represents a sampled individual structure, denoted by a tuple of (generation ID, population ID, lattice energy). The matched structures are highlighted as red crosses for clarity.

matching structures to the total number of sampled structures. In each WFS run, the fraction of mutation was set to be 0.4.

A Categorization by using the sampling success rate

Fig. 6 summarizes the whole results for all 100 systems. According to the averaged SR value of each system from four different sampling strategies, we divide them into four tiers as follows,

(1) Tier I (SR > 0.5%). This tier consists of 56 systems that can be considered relatively easy CSP challenges. Statistically, one only needs to sample 2–3k structures to hit the target structure at least 10 times. At the molecular level, Tier I systems are mostly composed of symmetric molecules with low mass (<300) and only a few flexible dihedral angles (<3). Representative molecules in this tier include many PAH/OSC polymorphs (e.g., BENZEN: 27.5390%; ADAMAN01: 24.3164%; NAPHTA: 19.1406%; UVAHEF: 3.0835%; XELYUJ: 1.9458%) and simple targets (e.g., AXOSOW01: 10.3516%; BOQQUT: 3.2715%; GUF-JOG: 0.9729%; NACJAF: 0.6925%). At the crystal level, the Z' number is no more than 1. AFIGIH (2.2326%) is an outlier in this group, as it has only C1 symmetry, a relatively high mass (303.4), and 6 flexible dihedral angles. However, as observed in other systems, crystal structures with $P\bar{1}$ symmetry are generally easier to predict than those with monoclinic or orthorhombic symmetries with more symmetry operation constraints.

(2) Tier II (0.05% < SR < 0.5%). This tier includes 33 systems that present relatively moderate CSP challenges. Typical molecules in this tier include 14 blind test targets (e.g., WIDBAO: 0.2650%; OBEQOD: 0.1755%), two famous aspirin polymorphs (ACSALA13: 0.2777%; ACSALA: 0.1617%), and many OSC polymorphs. To ensure convergence of 10 hits for the target structure, up to 20–30k structural samples are required. Most systems in this group have either fairly flexible molecules (with 2–6 rotors) or more than one molecule in the asymmetric unit ($Z' \geq 1$). Notably, the blind test target XXVI XAFQIH, previously considered a challenging example, also falls into this group with a modest SR value of 0.1486%.

(3) Tier III (0.001% < SR < 0.05%). This tier includes 6 systems (JOHSOP, WICZUF, KONTIQ, LEVJON, XAFPAY, and XAFPAY03). These systems typically require over 100k structural samples (approximately 4–6 hours with 128 parallel CPU processes). Systems in this tier are often characterized by either $Z' = 2$ or a high number of rotors (4–6). As we will discuss in the following section, the sampling efficiency strongly depends on the choice of algorithms. An efficient algorithm can significantly speed up the finding of target structures.

(4) Tier IV (SR < 0.001%). This tier consists of 4 examples that represent the most challenging CSP cases. Despite running 128k samples with four different strategies, we observed only a few hits for OBEQIX (7), XAFQON (4), and YOKBIK (2), and we were unable to find the target structures of XAFPAY02 and UJIRIO02 within several runs of 500 generations. UJIRIO02 was found a couple of times when the maximum number of generations is 3000. To reliably predict these systems, one either needs to spend significantly more CPU hours or more efficient sampling algorithms are needed.

In practical CSP applications, it is popular to search for plausible crystal packing with 10 to 15 common space groups.[28] As a result, the calculated success rate may decrease by 10–20 times in read-world scenarios. For clarity, we also provide the molecular diagrams and the corresponding space group/Z' numbers for several representative systems from each tier in Fig. 4. One may use the complete data in Table 2 as a test bed to develop a predictive model to infer the CSP landscape complexity purely from the molecular and crystal information for any arbitrary system.

B Impacts of sampling strategies

We now proceed to analyze the impact of the sampling strategy on the success rate. In Tier I, around 25 experimental targets, such as benzene (BENZEN), naphthalene (NAPHTA15), biphenyl (BIPHEN), and urea (UREAXX02), can be trivially found multiple times within just two generations of 512 structural samplings. For these systems, all four sampling strategies are expected to yield very similar statistical results. For clarity, we have excluded these systems from our main text and instead focus on analyzing the remaining 75 systems that require sampling over multiple generations, as shown in Fig. 5.

Comparing Fig. 5a and b, we first observe that adding ANI energy evaluation does not improve the efficiency of finding the target structures for the majority of the 100 systems. A few exceptional cases include KONTIQ, WICZUF, PYRZIN01, OBE-QUJ, and XAFQON. However, the inclusion of ANI adds about 50–100% overhead to the calculation time. Therefore, it is recommended to run these samplings with pure GAFF or OpenFF to reduce computational costs.

Second, the comparison between WFS and DFS in Fig. 5a shows that the success rates do not vary significantly for Tier I and most Tier II structures. For these structures, sampling at most 10 000 structures should be sufficient to obtain at least 10 hits of the target structure, ensuring successful sampling. However, there are a few exceptions in Tier II where WFS tends to outperform DFS. For example, in FLUANT (a rigid PAH molecule with $Z' = 2$ in $P2_1/n$) and QQQCIG04 (the semi-conducting rubrene with $Z' = 0.25$ in $Cmca$), WFS can identify the targets about 5–10 times faster than DFS. This suggests that the experimental structures in these systems are characterized by relatively isolated, narrow meta-basins.

On the other hand, DFS tends to be more efficient for Tier III/IV systems. For instance, the success rate of DFS-GAFF for JOHSOP is 0.0551%, which is significantly higher than the 0.0095% achieved by WFS-GAFF. Additionally, DFS produces relatively decent SR values for LEVJON (0.0055%), OBEQIX (0.0016%), XAFQON (0.0008%), and YOKBIK (0.0008%), whereas WFS fails completely. These cases suggest that the target structures are likely characterized by wide meta-basins, making an in-depth exploration strategy more efficient.

Clearly, there is a trade-off between using WFS and DFS. In the future, more in-depth studies will be conducted to develop a predictive model that infers CSP complexity and further improves the success rate by enhancing sampling methods.
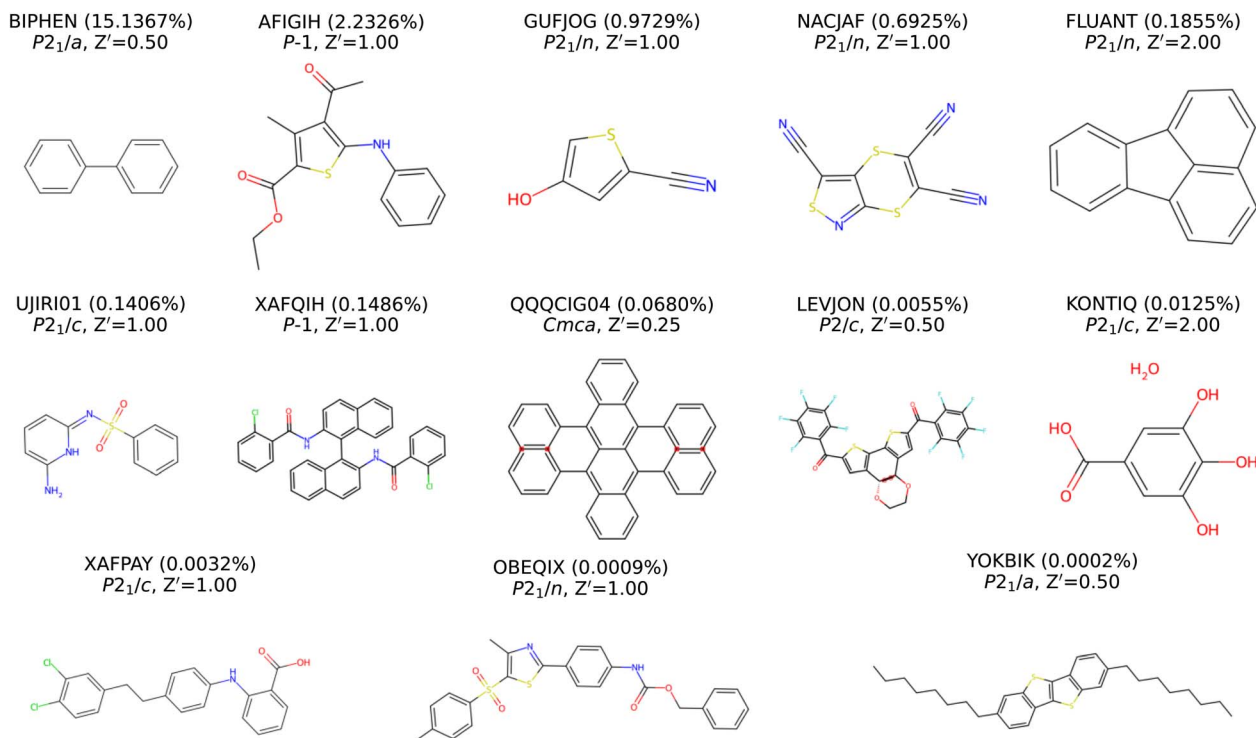
Fig. 4 The list of 13 representative molecules sorted by using the sampling success rate values. The molecules are listed based on their entry names as deposited in the Cambridge Structural Database as shown in https://www.ccdc.cam.ac.uk.
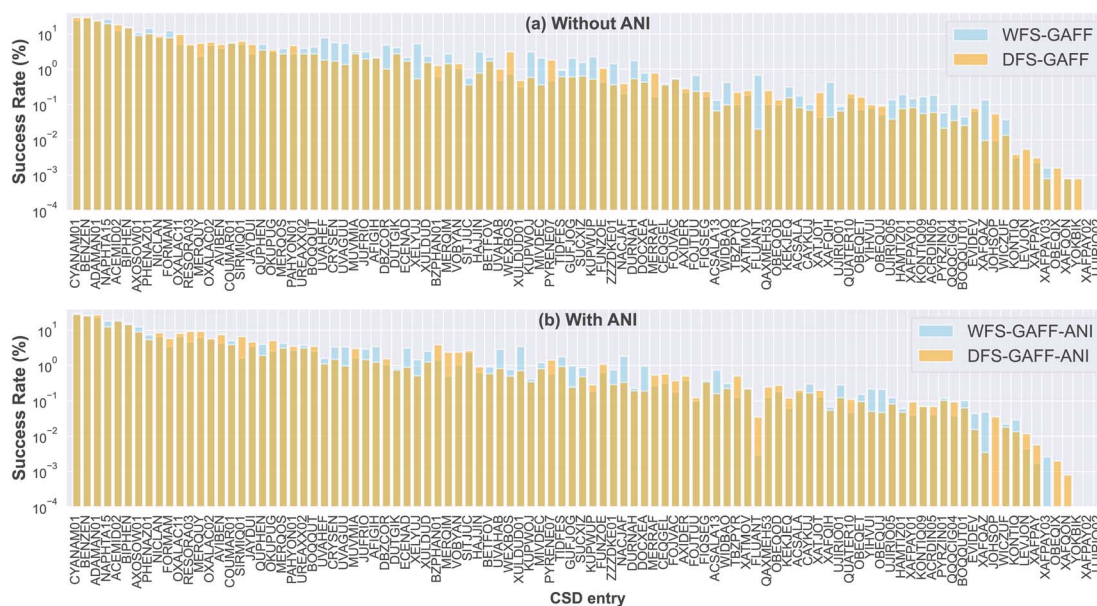
Fig. 5 The success rate distribution of 75 systems from four different sampling strategies, including (a) compares the results between WFS-GAFF and DFS-GAFF without applying the ANI model correction, and (b) compares WFS-GAFF and DFS-GAFF results after incorporating the ANI model correction in the final step.

V. Remaining challenges

While the current HTOCSP provides a general framework for automated crystal sampling in a high-throughput manner, there is certainly room for further improvements. Below, we briefly discuss several immediate challenges.

A More efficient sampling strategy

Based on our past experiences, it appears advantageous to separate cell parameters from the rest of the crystal variables during sampling. If the cell parameters are known, the CSP task becomes significantly easier, often by several orders of magnitude. In early CSP studies by Desiraju and Gavezzotti,[70] it was found that the

crystal cell shapes of similar molecules could often be fitted using empirical relations. With the rapid progress in deep learning, it is reasonable to anticipate that predicting cell parameters from molecular and space group information could become achievable in the near future. Therefore, it may be worthwhile to bias structure sampling toward a reduced set of more promising cell parameters to further improve sampling efficiency.

Additionally, data-mining-based approaches could supplement the current *ab initio* methods. In inorganic crystal prediction, chemical substitution on well-known prototypes is widely used to generate new candidate crystals. If the packing motifs of organic crystals can be effectively grouped,[71] one could analyze the occurrence of each packing prototype and use them as templates to generate new structures more efficiently.

B Uncertainty quantification

Our benchmark results on 100 systems suggest a correlation between crystal landscape complexity and the given input molecular/crystal variables. In a standard CSP run, the goal is typically to generate as many trial crystal structures as possible. However, the development of organic materials may require consideration of many molecules, necessitating high-throughput polymorph screening within a reasonable time frame. Ideally, we want set a smaller number of generations for simple systems and more generations for molecules with complex landscapes. To achieve this, a more rigorous uncertainty model should be developed to automate the setup of the sampling efforts based on the input molecular and crystallographic information.

C CSP post-analysis and iterative FF optimization

So far, the primary focus of the current HTOCSP development has been on identifying the most promising candidate crystal forms through extensive structure sampling. However, a significant challenge remains in extracting the target structure from the sampled structure pool. As observed in Table 2, it often requires sampling 1–100k structures to ensure success. Due to the limitations of force fields, it is quite possible that the true target structures have an unfavorable energy ranking at the force field level. This trend of energy misrank is likely to be more pronounced when predicting the polymorphs under non-standard pressure and temperature conditions. To address this challenge, one possible approach is to implement a multi-stage ranking process that refines the selection of candidate structures. This process could involve additional criteria, such as energy calculations using more accurate methods like DFT, to filter out less likely candidates. The development of a comprehensive post-analysis module will be the subject of our future work.

Additionally, one can start a massive CSP calculation with the generic FF, label the energy and forces for already sampled configurations with more accurate DFT or machine learning models, and then reparameterize the original force field parameters. Such an iterative strategy has been proved successful in past blind tests and many pharmaceutical applications.[5,44,72] The current HTOCSP has begun to implement this function and it should be available in the near future.

VI. Conclusions

In summary, we have presented HTOCSP, a comprehensive platform designed to facilitate high-throughput crystal structure prediction for organic molecules. By integrating various tools such as PyXtal, Pyocse, RDKit, AmberTools, and CHARMM, HTOCSP enables automated sampling and structure optimization, making it accessible for CSP practitioners to predict and analyze potential crystal forms efficiently.

Our systematic benchmark on 100 experimentally reported crystals demonstrates the effectiveness of different sampling strategies. We have shown that the choice of sampling strategy can significantly impact the success rate, especially for systems with varying degrees of landscape complexity. While width-first sampling proves to be more effective for systems characterized by narrow meta-basins, the depth-first sampling shows its strength in exploring wide meta-basins, particularly in more challenging CSP cases. In addition, these statistical results may serve as a ground for more efficient sampling method development in the future.

Despite these advancements, challenges remain, particularly in improving the efficiency of sampling strategies and developing robust post-analysis tools to accurately identify target structures from large datasets. The current work lays the foundation for future enhancements, including the integration of machine learning techniques to predict cell parameters, the implementation of more rigorous uncertainty quantification methods, and the development of advanced post-analysis modules to refine structure ranking. And the continuous development of HTOCSP will focus on addressing these challenges, with the aim of further improving the reliability and efficiency of CSP simulations.

Data availability

The HTOCSP source code, instructions, as well as scripts used in this study, are available at **https://github.com/MaterSim/HTOCSP**. The results from this work were obtained using the HTOCSPv-1.0 as released in 2024/10. The 100 crystal structures as discussed in this study can be directly acquired from **https://www.ccdc.cam.ac.uk** by searching for their corresponding CSD entries (*e.g.*, ACSALA).

Author contributions

QZ proposed this idea. Both QZ and SH designed the research, analyzed the calculations and wrote this manuscript.

Conflicts of interest

There are no conflicts to declare.

Appendix

Table 2 The summary of 100 molecular systems used in the benchmark test. The success rates are averaged from all four different sampling strategies, while the energy rank is obtained from the DFS-GAFF calculation as mentioned in Sec. IV. The T_{WFS} and T_{DFS} denote the time costs for each WFS/DFS calculation with 128 parallel CPU processes based on the model 3rd Gen AMD EPYC™ CPUs (AMD EPYC 7763) in the Anvil Cluster at Purdue University. In the remarks, BT stands for the blind test, OSC standards for an organic semiconductor, PAH stands for polynuclear aromatic hydrocarbons, and Misc stands for a miscellaneous group

CSD_entry	Space group symbol	Z'	Rotors/unit	Mass	Success rate (%)	Energy rank	T_{WFS} (min)	T_{DFS} (min)	Tier	Remarks
CYANAM01	$Pbca$	1.00	0	42.0	27.9889	1/256	1	1	1	Misc-cyanamide
BENZEN	$Pbca$	0.50	0	78.1	27.5391	1/256	1	1	1	PAH-benzene
ADAMAN01	$P2_1$	1.00	0	136.2	24.3164	1/256	1	1	1	Misc-adamantane
NAPHTA15	$P2_1/c$	0.50	0	128.2	19.1406	1/256	1	1	1	PAH-naphthalene
ACEMID02	$R3c$	1.00	0	59.1	17.2191	1/256	1	1	1	Misc-acetamide
BIPHEN	$P2_1/a$	0.50	1	154.2	15.1367	1/256	1	1	1	PAH-biphenyl
AXOSOW01	$Pbca$	1.00	1	56.1	10.3516	57/256	1	1	1	BT-XI
PHENAZ01	$P2_1/a$	0.50	0	180.2	9.3750	1/256	1	1	1	Misc-phenazine
NICLAN	$P2_1$	1.00	0	340.5	7.9102	39/256	1	1	1	OSC-DNTT
FORMAM	$P2_1/c$	1.00	0	45.0	7.4219	54/256	1	1	1	Misc-formamide
OXALAC11	$Pcab$	0.25	1	90.0	7.4219	22/256	1	1	1	Misc-oxalic acid
RESORA03	$Pna2_1$	1.00	0	110.1	5.9570	9/256	1	1	1	Misc-resorcinol
MERQUY	$P2_1/n$	0.50	0	334.5	5.8594	17/256	1	1	1	OSC-BDT
OXALAC02	$Pcab$	0.50	1	90.0	5.4688	19/256	1	1	1	Misc-oxalic acid
AVIBEN	$P2_1$	1.00	0	440.6	5.2094	21/256	1	1	1	OSC-DPhNDT
COUMAR01	$Pca2_1$	1.00	0	146.1	5.0781	1/256	1	1	1	Misc-coumarin
SIRMIQ01	$P2_1/n$	0.50	0	290.4	4.9020	1/256	1	1	1	OSC-BBBT
JAYDUI	$P2_1/n$	1.00	0	44.1	4.0039	1/256	1	1	1	BT-VII
QUPHEN	$P2_1/a$	0.50	3	306.4	3.6303	1/256	1	1	1	PAH-p-quaterphenyl
OKUPUG	$P2_1/a$	0.50	2	392.5	3.5909	4/256	1	1	1	OSC-DPh-NDT
MERQOS	$Pbca$	0.50	0	370.6	3.5156	15/256	1	1	1	OSC-BEDT-BDT
PAHYON01	$C2/c$	1.00	0	100.1	3.4180	1/256	1	1	1	BT-VIII
UREAXX02	$P\bar{4}2_1m$	0.25	0	60.1	3.3933	30/256	1	1	1	Misc-urea
BOQQUT	$P2_1/a$	1.00	0	153.2	3.2715	12/256	1	1	1	BT-IV
UVAHEF	$P2_1/a$	0.50	2	392.5	3.0835	6/768	1	1	1	OSC-DPh-BTBT
CRYSEN	$I2/c$	0.50	0	228.3	3.0570	1/768	1	1	1	PAH-chrysene
UVAGUU	$P2_1/c$	1.00	2	392.5	2.7986	1/1024	1	2	1	OSC-Ph-NDT
MUVMIA	$P2_1/n$	0.50	0	368.5	2.6348	1/256	1	1	1	OSC-3,10-DMeDNTT
JUFRIO	$P2_1/c$	0.50	2	382.6	2.3788	1/768	1	1	1	OSC-MT-ADT
AFIGIH	$P\bar{1}$	1.00	6	303.4	2.2327	1/768	1	1	1	Misc-thiophene deriv.
DBZCOR	$C2/c$	0.50	0	400.5	2.1304	1/1280	1	2	1	PAH-dibenzocoronene
DUTGIK	$P2_1/a$	0.50	2	604.8	2.1016	1/256	1	1	1	OSC-DPh-BBTNDT
ECENAD	$P2_1/c$	0.50	0	402.6	1.9637	82/768	1	1	1	OSC-BNTBDT
XELYUJ	$P2_1$	1.00	5	482.7	1.9458	1/2048	1	4	1	OSC-dithiophenyl deriv.
XULDUD	$Pbca$	1.00	0	94.1	1.9317	422/768	1	1	1	BT-I
BZPHAN01	$P2_12_12_1$	1.00	0	228.3	1.8069	48/1024	5	2	1	PAH-benzo(c)phenanthrene
MERQIM	$P2_1/a$	0.50	0	334.5	1.7658	37/768	1	1	1	OSC-HTPBDT
VOBYAN	$P2_1/n$	0.50	0	402.6	1.4879	53/768	1	1	1	OSC-anthracene-deriv.
SITJUC	$P2_1/c$	0.50	2	600.8	1.4861	54/3054	6	6	1	OSC-IDBT-7d

Table 2 (*Contd.*)

CSD_entry	Space group symbol	Z′	Rotors/unit	Mass	Success rate (%)	Energy rank	T_{WFS} (min)	T_{DFS} (min)	Tier	Remarks
HAJJIN	$P2_1$	1.00	6	400.6	1.3669	1/1536	1	2	1	OSC-DEP-DTT
BETFOV	$P\bar{1}$	1.00	0	274.4	1.3546	31/766	1	3	1	Misc
UVAHAB	$P2_1/c$	0.50	2	392.5	1.3043	94/1280	4	2	1	OSC-DPh-NDT4
WEXBOS	$P2_1/c$	0.50	4	444.6	1.2810	3/256	3	1	1	OSC-thiophene oligomer
XULDUD01	$P2_1/c$	1.00	0	94.1	1.2503	908/2304	6	3	1	BT-I
KUPWOJ	$P2_1/c$	0.50	2	492.7	1.1189	2/2048	1	4	1	OSC-2,9DPhDNTT
MIVDEC	$Pna2_1$	1.00	4	560.7	1.1186	4/3072	8	15	1	OSC-rubrene-deriv.
PYRENE07	$P2_1/a$	1.00	0	202.3	1.0785	162/768	4	1	1	PAH-pyrene
TIDFES	$P\bar{1}$	1.00	3	428.5	1.0664	9/1790	6	7	1	OSC-tetracene-deriv.
GUFJOG	$P2_1/n$	1.00	0	125.1	0.9729	7/2304	1	3	1	BT-II
SUCXIZ	$P22_12_1$	1.00	2	286.4	0.8888	1/2048	1	3	1	OSC-benzothiophene
KUPWUP	$Pbca$	0.50	2	492.7	0.8155	1/2304	3	5	1	OSC-3,10-DPhDNTT
FUNZOE	$P2_1$	1.00	4	268.3	0.8084	29/1024	4	1	1	Misc
ZZZDKE01	$P\bar{1}$	1.00	0	328.4	0.7102	22/3324	2	9	1	PAH-hexacene
NACJAF	$P2_1/n$	1.00	0	248.3	0.6925	415/3072	7	4	1	BT-XXII
DURNAH	$P2_1$	1.00	4	372.5	0.6810	6/2048	1	4	1	OSC-Ph-BTBT-C4
DOCXEA	$P2_1/n$	1.00	0	452.6	0.5930	152/2560	3	6	1	OSC-BBTNDT
MERRAF	$P2_1/c$	0.50	0	342.6	0.4413	213/1536	9	2	2	OSC-BMBT-BDT
CEQGEL	$P\bar{1}$	2.00	0	252.3	0.4096	1/3072	7	8	2	PAH-benzo(e)pyrene
FOJVAC	$P2_1$	1.00	4	518.7	0.4053	9/2304	5	5	2	OSC-BBPh-NDT
AXIDER	$P2_1/c$	0.50	4	604.6	0.3495	49/3840	10	8	2	OSC-rubrene-deriv.
FOJTUU	$Pbca$	0.50	2	466.6	0.2839	1/4608	4	10	2	OSC-DNap-NDT
FIQSEG	$P2_1/c$	0.50	2	282.5	0.2827	1963/4608	10	7	2	OSC-MT-BDT
ACSALA13	$P2_1/c$	1.00	3	180.2	0.2777	261/16 384	14	26	2	Misc-aspirin
WIDBAO	$P2_1/c$	1.00	2	254.4	0.2651	1775/11 008	5	17	2	BT-XIV
TBZPYR	$Pn2_1m$	0.50	0	352.4	0.2413	2107/4864	25	10	2	PAH-b-tribenzopyrene
XATMOV	$P2_1/c$	2.00	0	57.1	0.2224	86/4352	9	7	2	BT-XI
FLUANT	$P2_1/n$	2.00	0	202.3	0.1855	1/54 016	6	165	2	PAH-fluoranthene
QAXMEH53	$P\bar{1}$	1.00	3	259.3	0.1841	1828/4352	28	7	2	Misc-ROY
OBEQOD	$P2_1/c$	1.00	2	237.0	0.1755	4245/7936	16	12	2	BT-XVII
KEKJEQ	$P2_1/n$	0.50	2	418.4	0.1642	230/6912	7	13	2	OSC-DP-NTCDI
ACSALA	$P2_1/c$	1.00	3	180.2	0.1617	753/13 312	9	21	2	Misc-aspirin
CAYKUJ	$P2_1$	1.00	5	472.7	0.1601	23/15 616	26	34	2	Misc
XATJOT	$Pca2_1$	2.00	3	246.2	0.1504	71/5376	54	8	2	BT-XIX-cocrysta
XAFQIH	$P\bar{1}$	1.00	7	561.5	0.1486	9/24 576	8	64	2	BT-XXVI
UJRIOO1	$P2_1/c$	1.00	3	249.3	0.1406	560/16 640	19	27	2	BT-VI
QUATER10	$P2_1/a$	1.00	0	500.6	0.1311	1128/5374	36	30	2	PAH-quaterrylene
OBEQET	$Pbca$	1.00	3	258.7	0.1227	166/6656	24	10	2	BT-XVIII
YIHVUI	$P2_1/c$	0.50	1	490.7	0.1131	2729/11 008	26	21	2	OSC-BTDT
OBEQUJ	$Pbca$	1.00	0	120.1	0.1017	11 461/12 032	31	17	2	BT-XVI

Table 2 *(Contd.)*

| CSD_entry	Space group symbol	Z'	Rotors/unit	Mass	Success rate (%)	Energy rank	T_{WFS} (min)	T_{DFS} (min)	Tier	Remarks
UJIRIO05	$Pbca$	1.00	3	249.3	0.0962	637/28 160	14	48	2	BT-VI
HAMTIZ01	$P2_1/n$	1.00	4	239.2	0.0957	478/14 080	9	24	2	BT-X
XAFPAY01	$P\bar{1}$	1.00	6	386.3	0.0917	43/13 310	21	32	2	BT-XXIII
KONTIQ09	$P2_1/c$	2.00	1	188.1	0.0910	1895/19 456	19	32	2	BT-XXI-hydrate
ACRDIN05	Cc	2.00	0	179.2	0.0897	1/18 176	11	36	2	Misc-acridine
PYRZIN01	$P2_1/c$	1.00	1	123.1	0.0770	50 565/51 200	28	79	2	Misc-pyrazinamide
QQQCIG04	$Cmca$	0.25	4	532.7	0.0680	256/30 976	44	113	2	OSC-ruberene
BOQQUT01	$P2_1/c$	2.00	0	153.2	0.0597	602/42 752	53	82	2	BT-IV
EVIDEV	$P2_1/c$	0.50	2	344.6	0.0512	4788/13 824	27	23	2	OSC-4TA-2
XAFQAZ	$P2_1/c$	2.00	3	462.5	0.0501	13/113 914	18	284	2	BT-XXV-cocrystal
JOHSOP	$P2_1$	2.00	0	284.4	0.0251	914/19 968	265	55	3	OSC-ABT
WICZUF	$P2_1/n$	2.00	1	245.3	0.0230	77 460/80 384	52	155	3	BT-XV
KONTIQ	$P2_1/c$	2.00	1	188.1	0.0125	18 876/128 000	230	231	3	BT-XXI-hydrate
LEVJON	$P2/c$	0.50	4	696.5	0.0055	143/127 992	348	383	3	OSC
XAFPAY	$P2_1/c$	1.00	6	386.3	0.0032	65/127 958	312	286	3	BT-XXIII
XAFPAY03	$P2_1/n$	1.00	6	386.3	0.0013	53/127 938	312	284	3	BT-XXIII
OBEQIX	$P2_1/n$	1.00	8	478.6	0.0009	5/127 998	362	349	4	BT-XX
XAFQON	$P2_1/c$	3.00	4	201.7	0.0006	28 062/127 994	219	209	4	BT-XXIV
YOKBIK	$P2_1/a$	0.50	14	464.8	0.0002	631/127 992	481	384	4	OSC-C8-BTBT
XAFPAY02	$P\bar{1}$	2.00	12	386.3	0.0000	0/127 998	367	403	4	BT-XXIII
UJIRIO02	$P2_1/c$	2.00	6	249.3	0.0000	0/127 998	273	297	4	BT-VI
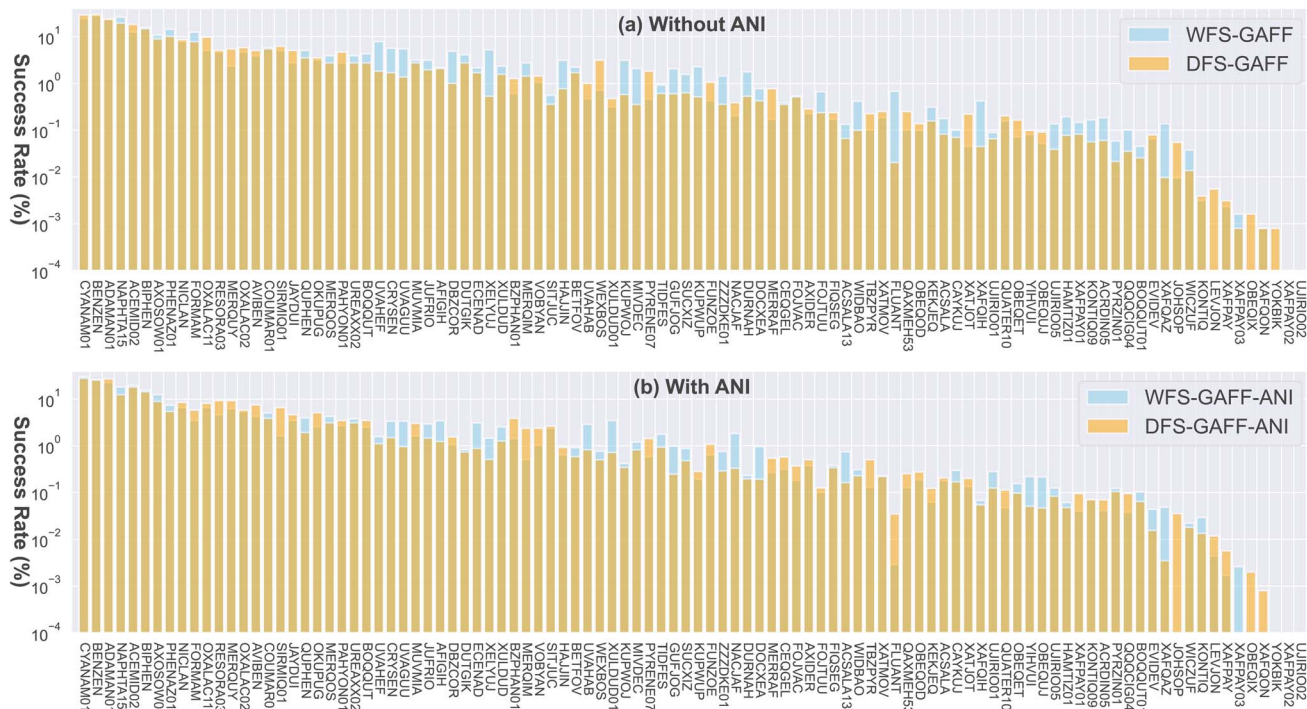
Fig. 6 The success rate distribution of 100 systems from four different sampling strategies, including (a) compares the results between WFS-GAFF and DFS-GAFF without applying the ANI model correction, and (b) compares WFS-GAFF and DFS-GAFF results after incorporating the ANI model correction in the final step.

Acknowledgements

References

1 A. Y. Lee, D. Erdemir and A. S. Myerson, Crystal polymorphism in chemical process development, *Annu. Rev. Chem. Biomol. Eng.*, 2011, **2**, 259–280.

2 Z. Zhuo, C. Wei, M. Ni, J. Cai, L. Bai, H. Zhang, Q. Zhao, L. Sun, J. Lin, W. Liu, *et al.*, Organic molecular crystal with a high ultra-deep-blue emission efficiency of 85% for low-threshold laser, *Dyes Pigm.*, 2022, 110425.

3 J. Yang, C. Hu, X. Zhu, Q. Zhu, M. D. Ward and B. Kahr, Ddt polymorphism and the lethality of crystal forms, *Angew. Chem.*, 2017, **129**, 10299–10303.

4 G. Liu, R. Gou, H. Li and C. Zhang, Polymorphism of energetic materials: a comprehensive study of molecular conformers, crystal packing, and the dominance of their energetics in governing the most stable polymorph, *Cryst. Growth Des.*, 2018, **18**, 4174–4186.

5 M. Neumann, J. Van De Streek, F. Fabbiani, P. Hidber and O. Grassmann, Combined crystal structure prediction and high-pressure crystallization in rational pharmaceutical polymorph screening, *Nat. Commun.*, 2015, **6**, 7793.

6 H. Kallmann and M. Pope, Bulk conductivity in organic crystals, *Nature*, 1960, **186**, 31–33.

7 Q. Li and Z. Li, Molecular packing: another key point for the performance of organic and polymeric optoelectronic materials, *Acc. Chem. Res.*, 2020, **53**, 962–973.

8 P. Yu, Y. Zhen, H. Dong and W. Hu, Crystal engineering of organic optoelectronic materials, *Chem*, 2019, **5**, 2814–2853.

9 A. Jain, Y. Shin and K. A. Persson, Computational predictions of energy materials using density functional theory, *Nat. Rev. Mater.*, 2016, **1**, 15004.

10 S. Fratini, S. Ciuchi, D. Mayou, G. T. De Laissardière and A. Troisi, A map of high-mobility molecular semiconductors, *Nat. Mater.*, 2017, **16**, 998–1002.

11 P. Friederich, A. Fediai, S. Kaiser, M. Konrad, N. Jung and W. Wenzel, Toward design of novel materials for organic electronics, *Adv. Mater.*, 2019, **31**, 1808256.

12 A. Saeki and K. Kranthiraja, A high throughput molecular screening for organic electronics via machine learning: present status and perspective, *Jpn. J. Appl. Phys*, 2019, **59**, SD0801.

13 T. Nematiaram, D. Padula and A. Troisi, Bright frenkel excitons in molecular crystals: A survey, *Chem. Mater.*, 2021, **33**, 3368–3378.

14 A. Stuke, C. Kunkel, D. Golze, M. Todorović, J. T. Margraf, K. Reuter, P. Rinke and H. Oberhofer, Atomic structures and orbital energies of 61,489 crystal-forming organic molecules, *Sci. Data*, 2020, **7**, 1–11.

15 C. Kunkel, C. Schober, J. T. Margraf, K. Reuter and H. Oberhofer, Finding the right bricks for molecular legos: A data mining approach to organic semiconductor design, *Chem. Mater.*, 2019, **31**, 969–978.

16 Y. Diao, K. M. Lenn, W.-Y. Lee, M. A. Blood-Forsythe, J. Xu, Y. Mao, Y. Kim, J. A. Reinspach, S. Park, A. Aspuru-Guzik, *et al.*, Understanding polymorphism in organic semiconductor thin films through nanoconfinement, *J. Am. Chem. Soc.*, 2014, **136**, 17046–17057.

17 M. K. Corpinot and D.-K. Bučar, A practical guide to the design of molecular crystals, *Cryst. Growth Des.*, 2018, **19**, 1426–1453.

18 J. Yang, S. De, J. E. Campbell, S. Li, M. Ceriotti and G. M. Day, Large-scale computational screening of molecular organic semiconductors using crystal structure prediction, *Chem. Mater.*, 2018, **30**, 4361–4371.

19 J. P. Lommerse, W. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Gavezzotti, D. W. Hofmann, F. J. Leusen, W. T. Mooij, S. L. Price, B. Schweizer, *et al.*, A test of crystal structure prediction of small organic molecules, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2000, **56**, 697–714.

20 W. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Dzyabchenko, P. Erk, A. Gavezzotti, D. W. Hofmann, F. J. Leusen, J. P. Lommerse, W. T. Mooij, *et al.*, Crystal structure prediction of small organic molecules: a second blind test, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2002, **58**, 647–661.

21 G. M. Day, *et al.*, A third blind test of crystal structure prediction, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2005, **61**, 511–527.

22 G. M. Day, *et al.*, Significant progress in predicting the crystal structures of small organic molecules - a report on the fourth blind test, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2009, **65**, 107–125.

23 D. A. Bardwell, *et al*, Towards crystal structure prediction of complex organic compounds—A report on the fifth blind test, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2011, **67**, 535–551.

24 A. M. Reilly, R. I. Cooper, C. S. Adjiman, S. Bhattacharya, A. D. Boese, J. G. Brandenburg, P. J. Bygrave, R. Bylsma, J. E. Campbell, R. Car, *et al.*, Report on the sixth blind test of organic crystal structure prediction methods, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**, 439–459.

25 A. R. Oganov, C. J. Pickard, Q. Zhu and R. J. Needs, Structure prediction drives materials discovery, *Nat. Rev. Mater.*, 2019, **4**, 331–348.

26 S. L. Price, Predicting crystal structures of organic compounds, *Chem. Soc. Rev.*, 2014, **43**, 2098–2111.

27 A. Gavezzotti, *The Crystalline States of Organic Compounds*, Elsevier, 2021, vol. 20.

28 Q. Zhu and S. Hattori, Organic crystal structure prediction and its application to materials design, *J. Mater. Res.*, 2023, **38**, 19–36.

29 A. R. Oganov and C. W. Glass, Crystal structure prediction using ab initio evolutionary techniques: Principles and applications, *J. Chem. Phys.*, 2006, **124**, 244704.

30 C. J. Pickard and R. Needs, Ab initio random structure searching, *J. Phys.: Condens. Matter*, 2011, **23**, 053201.

31 Y. Wang, J. Lv, L. Zhu and Y. Ma, Crystal structure prediction via particle-swarm optimization, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2010, **82**, 094116.

32 D. C. Lonie and E. Zurek, Xtalopt: An open-source evolutionary algorithm for crystal structure prediction, *Comput. Phys. Commun.*, 2011, **182**, 372–387.

33 B. C. Revard, W. W. Tipton, and R. G. Hennig, *Genetic algorithm for structure and phase prediction*, 2018.

34 D. H. Case, J. E. Campbell, P. J. Bygrave and G. M. Day, Convergence properties of crystal structure prediction by quasi-random sampling, *J. Chem. Theory Comput.*, 2016, **12**, 910–924.

35 F. Curtis, X. Li, T. Rose, Á. Vázquez-Mayagoitia, S. Bhattacharya, L. M. Ghiringhelli and N. Marom, Gator: A first-principles genetic algorithm for molecular crystal structure prediction, *J. Chem. Theory Comput.*, 2018, **14**, 2246–2264.

36 Q. Zhu, S. Fredericks, and K. Parrish, *Pyxtal: Python xtallography toolkit*, 2024.

37 Q. Zhu and S. Hattori, *Pyocse: Python organic crystal simulation environment*, 2024.

38 RDKit, online, *RDKit: Open-source cheminformatics*, **https://www.rdkit.org**, accessed 11-April-2013.

39 D. A. Case, H. M. Aktulga, K. Belfon, I. Ben-Shalom, S. R. Brozell, D. S. Cerutti, T. E. Cheatham III, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, *et al.*, *Amber 2021*, University of California, San Francisco, 2021.

40 S. Boothroyd, P. K. Behara, O. Madin, D. Hahn, H. Jang, V. Gapsys, J. Wagner, J. Horton, D. Dotson, M. Thompson, J. Maat, T. Gokey, L.-P. Wang, D. Cole, M. Gilson, J. Chodera, C. Bayly, M. Shirts, and D. Mobley, *Development and benchmarking of open force field 2.0.0—the sage small molecule force field*, 2023.

41 S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, *et al.*, Pubchem in 2021: new data content and improved web interfaces, *Nucleic Acids Res.*, 2021, **49**, D1388–D1395.

42 J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, Development and testing of a general amber force field, *J. Comput. Chem.*, 2004, **25**, 1157–1174.

43 A. Jakalian, B. L. Bush, D. B. Jack and C. I. Bayly, Fast, efficient generation of high-quality atomic charges. am1-bcc model: I. method, *J. Comput. Chem.*, 2000, **21**, 132–146.

44 M. A. Neumann, F. J. Leusen and J. Kendrick, A major advance in crystal structure prediction, *Angew. Chem.*, 2008, **120**, 2461–2464.

45 P. A. Santos-Florez, S. Hattori and Q. Zhu, Bending deformation driven by molecular rotation, *Phys. Rev. Res.*, 2023, **5**, 033185.

46 Q. Zhu, A. R. Oganov, C. W. Glass and H. T. Stokes, Constrained evolutionary algorithm for structure prediction of molecular crystals: methodology and applications, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2012, **68**, 215–226.

47 J. D. Gale, Gulp: A computer program for the symmetry-adapted simulation of solids, *J. Chem. Soc., Faraday Trans.*, 1997, **93**, 629–637.

48 B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, Charmm: a program for macromolecular energy, minimization, and dynamics calculations, *J. Comput. Chem.*, 1983, **4**, 187–217.

49 A. G. Shtukenberg, Q. Zhu, D. J. Carter, L. Vogt, J. Hoja, E. Schneider, H. Song, B. Pokroy, I. Polishchuk, A. Tkatchenko, *et al.*, Powder diffraction and crystal structure prediction identify four new coumarin polymorphs, *Chem. Sci.*, 2017, **8**, 4926–4940.

50 A. G. Shtukenberg, C. T. Hu, Q. Zhu, M. U. Schmidt, W. Xu, M. Tan and B. Kahr, The third ambient aspirin polymorph, *Cryst. Growth Des.*, 2017, **17**, 3562–3566.

51 M. Tan, A. Shtukenberg, S. Zhu, W. Xu, E. Dooryhee, S. M. Nichols, M. D. Ward, B. Kahr and Q. Zhu, Roy revisited, again: The eighth solved structure, *Faraday Discuss.*, 2018, **211**, 477–491.

52 J. Yang, X. Zhu, C. T. Hu, M. Qiu, Q. Zhu, M. D. Ward and B. Kahr, Inverse correlation between lethality and thermodynamic stability of contact insecticide polymorphs, *Cryst. Growth Des.*, 2019, **19**, 1839–1844.

53 T. Darden, D. York and L. Pedersen, Particle mesh ewald: An nlog (n) method for ewald sums in large systems, *J. Chem. Phys.*, 1993, **98**, 10089–10092.

54 X. Gao, F. Ramezanghorbani, O. Isayev, J. S. Smith and A. E. Roitberg, Torchani: A free and open source pytorch-based deep learning implementation of the ani neural network potentials, *J. Chem. Inf. Model.*, 2020, **60**, 3408–3415.

55 I. Batatia, D. P. Kovacs, G. N. C. Simm, C. Ortner, and G. Csanyi, MACE: Higher order equivariant message passing neural networks for fast and accurate force fields, in *Advances in Neural Information Processing Systems*, ed. A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, 2022.

56 I. Batatia, S. Batzner, D. P. Kovács, A. Musaelian, G. N. C. Simm, R. Drautz, C. Ortner, B. Kozinsky, and G. Csányi, The design space of e(3)-equivariant atom-centered interatomic potentials, *arXiv*, 2022, preprint, arXiv:2205.06643, DOI: **10.48550/arXiv.2205.06643**.

57 A. Kadan, K. Ryczko, A. Wildman, R. Wang, A. Roitberg and T. Yamazaki, Accelerated organic crystal structure prediction with genetic algorithms and machine learning, *J. Chem. Theory Comput.*, 2023, **19**, 9388–9402.

58 A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, *et al.*, The atomic simulation environment—a python library for working with atoms, *J. Phys.: Condens. Matter*, 2017, **29**, 273002.

59 S. Fredericks, K. Parrish, D. Sayre and Q. Zhu, Pyxtal: A python library for crystal structure generation and symmetry analysis, *Comput. Phys. Commun.*, 2021, **261**, 107810.

60 S. Hall, Space-group notation with an explicit origin, *Acta Crystallogr., Sect. A*, 1981, **37**, 517–525.

61 A. O. Lyakhov, A. R. Oganov, H. T. Stokes and Q. Zhu, New developments in evolutionary structure prediction algorithm uspex, *Comput. Phys. Commun.*, 2013, **184**, 1172–1182.

62 S. Martiniani, K. J. Schrenk, J. D. Stevenson, D. J. Wales and D. Frenkel, Structural analysis of high-dimensional basins of attraction, *Phys. Rev. E*, 2016, **94**, 031301.

63 V. Stevanović, Sampling polymorphs of ionic solids using random superlattices, *Phys. Rev. Lett.*, 2016, **116**, 075503.

64 D. J. Wales and J. P. Doye, Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms, *J. Phys. Chem. A*, 1997, **101**, 5111–5116.

65 S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, Python materials genomics (pymatgen): A robust, open-source python library for materials analysis, *Comput. Mater. Sci.*, 2013, **68**, 314–319.

66 C. F. Macrae, I. Sovago, S. J. Cottrell, P. T. Galek, P. McCabe, E. Pidcock, M. Platings, G. P. Shields, J. S. Stevens, M. Towler, *et al.*, Mercury 4.0: From visualization to analysis, design and prediction, *J. Appl. Crystallogr.*, 2020, **53**, 226–235.

67 J. A. Chisholm and S. Motherwell, Compack: a program for identifying crystal structure similarity using distances, *J. Appl. Crystallogr.*, 2005, **38**, 228–231.

68 S. Hattori and Q. Zhu, Revisiting aspirin polymorphic stability using a machine learning potential, *ACS Omega*, 2024, **9**, 36589–36599.

69 Q. Zhu and S. Hattori, *Htocsp: Automated high-throughput organic crystal structure prediction*, 2024.

70 G. R. Desiraju and A. Gavezzotti, Crystal structures of polynuclear aromatic hydrocarbons. classification, rationalization and prediction from molecular structure, *Acta Crystallogr., Sect. B: Struct. Sci.*, 1989, **45**, 473–482.

71 Q. Zhu, W. Tang and S. Hattori, Quantification of crystal packing similarity from spherical harmonic transform, *Cryst. Growth Des.*, 2022, **22**, 7308–7316.

72 M. A. Neumann, Tailor-made force fields for crystal-structure prediction, *J. Phys. Chem. B*, 2008, **112**, 9810–9829.