RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2026, **16**, 7221

Structure prediction of porous organic crystals

Musiha Mahfuza Mukta,*[a] Romain Perriot,[b] Shinnosuke Hattori,*[c] Wei Zhou*[d] and Qiang Zhu[ae]

Received 3rd December 2025
Accepted 28th January 2026

DOI: 10.1039/d5ra09332e

rsc.li/rsc-advances

In this work, we explore the possibility of applying automated crystal structure prediction to reproduce the experimentally identified metastable porous polymorphs. Using our recently developed High-Throughput Organic Crystal Structure Prediction (HTOCSP) framework, we conducted a systematic study on five representative organic crystalline systems including hydrogen-bonded frameworks (HOFs), featured by the presence of significant porosity, in conjunction with different choices of energy models from classical, machine learning force fields, tight binding to density functional theory. Our results suggest that the current structure generation framework, with careful selection of symmetry conditions, is likely to generate rather complex and abundant metastable crystal candidates for porous crystals. In conjunction with the recent advance in universal machine learning force fields, it becomes possible to identify experimental structures as the energetically favorable candidates from a simple energy *versus* density analysis, thus paving the way for computational design of complex porous materials with the target systems prior to the experimental synthesis and characterization.

1 Introduction

In recent years, high-throughput computational screening of organic crystals has become a viable strategy for designing and searching new materials with improved physical properties.[1–6] However, most studies have focused on experimentally resolved structures drawn from existing databases, such as the Cambridge Structural Database (CSD)[7] and the Crystallography Open Database (COD).[8] In practice, the ability to screen likely crystal packings prospectively—prior to synthesis and characterization—would be highly valuable.[9,10]

Over the past two decades, crystal structure prediction (CSP) for small organic molecules has advanced rapidly.[11–17] In a typical CSP study, the objective is to generate a tractable set of low-energy (stable or metastable) crystal packings that are experimentally plausible, using efficient exploration algorithms.[10,18,19] Recently, we introduced the High-Throughput Organic Crystal Structure Prediction (HTOCSP) framework, an open-source platform that automates CSP workflows from minimal molecular input by integrating existing toolkits for force-field assignment, structure generation and energy ranking.[20] Accordingly, we demonstrated its performance across diverse molecular systems, with an emphasis on recovering dense packing motifs. As expected from earlier studies,[10] thermodynamic ground states overwhelmingly correspond to densely packed arrangements. For many application domains (*e.g.*, organic semiconductors and pharmaceuticals), focusing on dense forms is often sufficient.

In practice, however, there also exist many other applications that would benefit from metastable crystal forms, including low-density porous polymorphs that can exhibit desirable properties such as selective gas adsorption, molecular recognition, and catalytic activity. Porous organic crystals, constructed through weak intermolecular interactions (such as hydrogen bonds and van der Waals interactions), represent an emerging class of crystalline materials in which permanent porosity arises from the ordered packing of discrete organic molecules rather than from extended coordination or covalent networks. Within this family, hydrogen-bonded organic frameworks (HOFs) have attracted particular attention as a unique platform for designing lightweight, solution-processable porous solids with tunable functionality.[21] Early molecular-tectonics studies established that rationally designed hydrogen-bond donors and acceptors can assemble into robust three-dimensional porous networks with remarkable structural integrity, even through extensive guest–exchange cycles.[22] More recent HOFs, such as the flexible microporous framework HOF-5, exhibit substantial permanent porosity, guest-responsive lattice expansion/contraction, and selective gas sorption behavior, highlighting the rich structure–property landscape accessible through judicious hydrogen-bond design.[23] At the same time, the weak yet partly directional nature of hydrogen

[a]*Department of Mechanical Engineering and Engineering Science, Charlotte, NC, USA. E-mail: qzhu8@charlotte.edu*
[b]*Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA*
[c]*Advanced Research Laboratory, Research Platform, Sony Group Corporation, 4-14-1 Asahi-cho, Atsugi-shi 243-0014, Japan. E-mail: shinnosuke.hattori@sony.com*
[d]*NIST Center for Neutron Research, National Institute of Standards and Technology, Gaithersburg, MD, 20899-6102, USA. E-mail: wzhou@nist.gov*
[e]*North Carolina Battery Complexity, Autonomous Vehicle and Electrification (BATT CAVE) Research Center, Charlotte, NC 28223, USA*

bonding makes HOF packing highly sensitive to molecular geometry and conformation; consequently, even small structural variations can lead to distinct polymorphs with different pore architectures and properties.

From the CSP perspective, predicting porous organic crystals is inherently more challenging due to two main factors. First, the potential structural space for porous crystals is significantly larger than that of dense packings. The presence of large voids and channels leads to a rugged energy landscape with numerous local minima corresponding to different molecular orientations with similar energies. Consequently, the configurational space requiring exploration is substantially expanded. Second, practical applications often favor metastable polymorphs with low densities rather than the thermodynamically stable dense-packed ground state. This creates additional challenges in identifying which candidate structures among a large pool of generated metastable forms are most likely to be realized experimentally. While several CSP studies have addressed porous organic crystals,[24–26] this area remains largely underexplored.

In this work, we aim to explore the capability of the HTOCSP framework to predict porous organic crystals, building upon our earlier work on dense packings[20] and complementing previous CSP studies of porous molecular crystals.[24–26] We begin by presenting the selection criteria for our benchmark systems and providing an overview of the HTOCSP workflow. Subsequently, we evaluate these systems using complementary structure-sampling strategies combined with alternative energy models (including classical force fields, machine-learning potentials, and semi-empirical methods), and assess their effectiveness in recovering low-density polymorphs. We conclude with a discussion of new directions for improving accuracy in future studies.

2 Systems of choices

Fig. 1 displays the five porous organic crystalline systems selected for this study. These systems were chosen to represent a diverse range of structural complexity, pore architectures, and hydrogen-bonding motifs characteristic of HOFs:

• ZJU-HOF-60 and ZJU-HOF-62 are two recently reported HOFs, based on 5,5′-(1,2-ethynediyl)bis(1,3-benzene-dicarboxylic acid) and 5,5′-(1,3-butadiyne-1,4-diyl)bis(1,3-benzene-dicarboxylic acid) molecular units, respectively.[26] Due to their relatively simple planar molecular geometry, both HOFs have 2D layered structures, in which the molecules are connected through hydrogen bonds within individual layer and the interlayer stacking is enabled by π–π and van der Waals interactions. They exhibit great potential for certain hydrocarbon separation applications, thanks to their 1D channel-like pores along with the dialkynyl and carboxylic sites for selective gas adsorption.

• TTBI is a triptycene trisbenzimidazolone-based HOF that exhibits a three-dimensional porous framework stabilized by extensive hydrogen-bonding networks. The rigid triptycene scaffold enforces specific molecular orientations that facilitate permanent porosity.[24] Notably, TTBI has been reported to exhibit multiple polymorphic forms (α-, β-, and γ-phases) with varying pore sizes and shapes, making it an ideal candidate for studying the relationship between molecular design, packing motifs, and porosity in HOFs. Additionally, this system was explored in a previous CSP study,[27] providing a valuable benchmark for comparison with our CSP methodology.

• TCF-1 is a tetracarboxylic-based HOF, obtained by crystallization of methanetetrabenzoic acid.[25] Two structural forms have been reported. The first is called “porous-TCF-1”, which has moderate porosity and exhibits structural flexibility, characterized by reversible expansion and contraction of the pores upon guest adsorption and removal. The second form is a nonporous phase, called “dense-TCF-1”. The existence of both

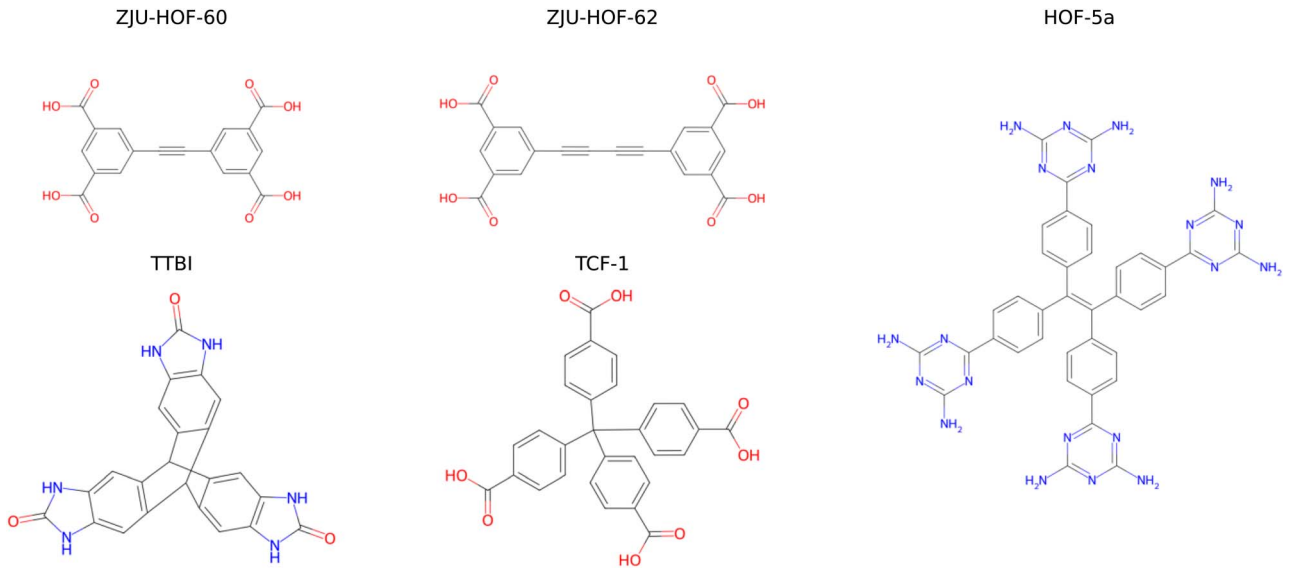

Fig. 1 The selected five molecular systems in the present work.

Table 1 Crystallographic data and porosity values of the eight porous organic crystals studied

System	Symmetry	Z'	Cell parameters (Å; °)	Density (g cm^{-3})	Porosity
ZJU-HOF-60 (ref. 26)	$I2/m$	1/4	$a = 3.65$, $b = 16.44$, $c = 20.95$, $\beta = 91.6$	0.937	46.5%
ZJU-HOF-62 (ref. 26)	$P3_1$	1	$a = 16.46$, $c = 10.10$	0.795	52.8%
α-TTBI[27]	$P4_2/m$	1/2	$a = 22.51$, $c = 7.34$	0.755	59.8%
β-TTBI[24]	$P\bar{1}$	1	$a = 7.25$, $b = 13.03$, $c = 20.66$, $\alpha = 72.5$, $\beta = 86.3$, $\gamma = 74.0$	0.782	55.4%
γ-TTBI[27]	$P6_3/mmc$	1/12	$a = 23.22$, $c = 7.29$	0.412	79.0%
Porous-TCF-1 (ref. 25)	$P4_2/n$	1/4	$a = 13.10$, $c = 8.08$	1.239	20.6%
Dense-TCF-1 (ref. 25)	$I\bar{4}$	1/4	$a = 12.53$, $c = 7.53$	1.394	1.2%
HOF-5a[23]	$C2/m$	1/4	$a = 14.35$, $b = 17.87$, $c = 12.26$, $\beta = 121.7$	0.954	41.1%

porous and dense polymorphs of this HOF presents a great opportunity for us to test our predictive capability as well.

• HOF-5a is a well-known prototypical flexible HOF, constructed from 4,4′,4″,4‴-*tetra*(2,4-diamino-1,3,5-triazin-6-yl)tetraphenylethene.[23] It exhibits moderately high porosity and interesting gas adsorption/separation properties. Upon guest inclusion, its channel-like pore can expand its cross-section, leading to a significant increase in pore volume, up to 70%. The bulk size and the intrinsic flexibility of the molecular building unit make this HOF a particularly challenging benchmark system.

In selecting these systems, we aimed to cover a spectrum of HOF architectures, from rigid to flexible frameworks, and from simple layered to complex hydrogen-bonding motifs, as well as varying degrees of porosity. This diversity allows us to systematically evaluate the performance of our CSP framework across different structural challenges inherent to porous organic crystals. As shown in Table 1, it is worth noting that most of the structures have been reported to adopt the structures with fractional Z' numbers, due to the presence of high molecular symmetry. In practical CSP applications, it is often necessary to convert these structures into equivalent subgroup representations with integer Z' numbers to facilitate structure generation and sampling.

3 Computational methodology

To predict porous organic crystals, we employed the `HTOCSP` framework,[20] which integrates various tools for structure generation, optimization, and analysis. Below, we outline the key components of our computational methodology.

3.1 The HTOCSP workflow

In the `HTOCSP` framework, it uses the list of chemical SMILES strings as the input to generate the 3D molecular structure using `RDKit`.[28] The generated molecule is then used to create trial crystal structures usin`PyXtal`,[29] which allows for flexible specification of space groups, Z', and other structural parameters, based on two common sampling strategies based on either the width-first search (WFS) or depth-first search (DFS).[20] Each of the generated structures are subsequently optimized using the classical force field powered by the CHARMM code,[30] and then more accurate machine learning force fields. Finally, the optimized structures are analyzed and ranked based on their relative stabilities and structural similarity to known polymorphs. For molecules with rotatable bonds, the initial structures are created with randomized torsional angles, so different conformations are sampled across the population. Each generated crystal is then fully relaxed under periodic boundary conditions, allowing both the lattice and all atomic coordinates, including intramolecular torsions to optimize. So, conformational changes are naturally driven by the crystal packing environment, without requiring a separate conformer library. However, PyXtal's default process is sampling-based, it samples conformers and uses tolerance rules (and compatibility with Wyckoff-site symmetry) to decide which orientations are acceptable. So, molecular flexibility is handled automatically during generation and relaxation.

```python
from pyxtal.optimize import DFS
from pyxtal import pyxtal

# Parameter Setup
# gen  : number of generations in the search.
# pop  : population size per generation.
# ncpu : number of CPU cores used in parallel for
    structure evaluation.

gen, pop, ncpu = 1500, 256, 48

# sg   : list of space groups sampled during generation.
sg =[1, 2, 4, 5, 7, 9, 12, 14, 15, 19, 29, 33, 60, 61,
    62, 75, 76, 77, 78, 143, 144, 145, 148]

# wdir : working directory for all intermediate and
    output files
wdir = "TTBI"
smiles = "O=c2[nH]c1cc5c(cc1[nH]2)C8c4cc3[nH]c(=O)[nH]
    c3cc4C5c7cc6[nH]c(=O)[nH]c6cc78"

# Sampling
go = DFS(smiles,
         wdir,
         sg,
         tag = "TTBI",
         N_gen = gen,
         N_pop = pop,
         N_cpu = ncpu,
         ff_style = "gaff")
go.run()
```

Listing 1 Python script to run HTOCSP for the TTBI system.

To illustrate the usage of HTOCSP, we provide a sample script in Listing 1 that demonstrates how to set up and run a CSP calculation for the TTBI using the DFS sampling strategy with GAFF force field. The script includes parameter setup, and the sampling process. After running the simulation, it generates random structures, and from those one can check for the experimental one.

3.2 Structural sampling strategies

In this study, we focus on two strategies for generating trial structures by setting different space groups:

3.2.1 Blind search on the known space group symmetries. This approach assumes prior knowledge of the space group. If the Z' is fractional, we convert it to an equivalent subgroup representation with $Z' = 1$ (corresponding to the case of molecules occupying the general Wyckoff position). The purpose of the test is to evaluate the efficiency of different sampling strategies and energy models in recovering the target structure within a limited number of sampled structures.

3.2.2 Blind search on a list common space group with $Z' = 1$. This approach mimics a more realistic CSP scenario where the space group is unknown. We select a list of 23 common space groups for organic crystals ($P1$, $P\bar{1}$, $P2_1$, $C2$, Pc, Cc, $P2/c$, $P2_1/c$, $C2/c$, $P2_12_12_1$, $Pna2_1$, $Pca2_1$, $Pbcn$, $Pbca$, $Pnma$, $P4$, $P4_1$, $P4_2$, $P4_3$, $P3$, $P3_1$, $P3_2$ and $R\bar{3}$) and set $Z' = 1$ for all trials. The goal is to assess the likelihood of identifying the target structure within a limited number of sampled structures across multiple space groups.

To check if the target structure is found in our search, the StructureMatcher module in Pymatgen (ref. 31) 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 0.25, the fractional site tolerance is less than 0.25, and the angle tolerance is less than 5°.

3.3 Energy model choices

After the structures are generated, they need to be ranked based on their relative stabilities. In the HTOCSP, we have implemented several energy models for geometry optimization and energy evaluation, including classical force fields (FFs), machine learning potentials (MLPs), semi-empirical tight binding (DFTB) methods, and density functional theory (DFT). In this work, we focus on the following energy models for CSP structure search.

3.3.1 Classical force fields for geometry relaxation. We employ GAFF[32] and OpenFF[33] for fast lattice and molecular relaxations. GAFF is a widely used general Amber force field that provides reasonable accuracy for a broad range of organic molecules. OpenFF uses direct chemical perception (SMIRNOFF) and often improves transferability over GAFF. In the HTOCSP, parameters are assigned *via* AmberTools (ref. 34) for GAFF and the OpenFF Toolkit for OpenFF, and structures are minimized with CHARMM.[30] Here, atomic partial charges are molecule-specific and consistent across the search.

3.3.2 Machine-learning force fields for refinement. MACE[35] is an equivariant message-passing neural network potential that achieves near-DFT accuracy at significantly reduced computational cost. We use MACE for post-optimization single-point energy evaluation and optional short local relaxations on GAFF/OpenFF-relaxed geometries to improve ranking accuracy at modest cost relative to DFT.

Additionally, we consider the following models for additional energy ranking of top candidates extracted from the CSP search.

3.2.2.1 Alternative MLPs for post-optimization ranking. We also consider MACE-OFF,[36] a MACE variant optimized for organic systems, and UMA,[37] a universal potential trained on a broad corpus of molecules and materials with small, fast variants suitable for CSP.

3.2.2.2 Semi-empirical tight binding. LATTE[38] is a self-consistent charge transfer density functional tight-binding (SCC-DFTB) code, which provides a semi-empirical tight-binding Hamiltonian. SCC-DFTB[39,40] can provide substantially better accuracy than classical FFs at a fraction of DFT cost. Here, we use the *lanl*31 parameterization developed for organic molecules,[41] with additional pairwise dispersion corrections.[42] Although the parameterization was fitted on small organic molecules in the gas phase, it has been demonstrated to provide an accurate description of organic crystals in the solid phase, reproducing lattice parameters and surface energies with good accuracy.[43]

3.2.2.3 Density functional theory. DFT serves as the highest-accuracy stage for validating and final re-ranking of top candidates. All computations were performed using VASP[44] 6.4.3, employing the GGA-PBE[45] functional with PAW-PBE pseudopotentials. Long-range dispersion interactions were included *via* Grimme's DFT-D3 (ref. 46) correction with zero damping function (IVDW = 11), ensuring an accurate description of van der Waals forces. Geometry optimizations were carried out using the conjugate-gradient algorithm (IBRION = 2) with full relaxation of both atomic positions and lattice parameters (ISIF = 3).

Using these models, we performed additional full structure relaxation (including both atomic coordinates and cell parameters) to obtain the final energy ranking.

4 Results and discussions

4.1 Blind search with the known space group

Fig. 2 summarizes the results for the case when the target space group symmetries are known. For ZJU-HOF-60 ($I2/m$) and ZJU-HOF-62 ($P3_1$), we attempted the structure search with $C2$ and $P3_1$, respectively. As shown in Fig. 2a and b, both simulations return high structure matches rate of 4–10 per 1000 sampled structures. We also emphasize that GAFF force field tends to generate the incorrect ground state geometry for both ZJU-HOF-60 and ZJU-HOF-62, while OpenFF provides better description of the geometry. This highlights the importance of using the correct force field even for the initial stage of structure screening. For the polymorphic cases of TTBI and TCF-1 in Fig. 2c and d, we conducted CSP simulations with ($P\bar{1}$, $P4_2$, $P2_12_12_1$) for TTBI and ($P2_1$, Pc) for TCF-1, the prediction of different

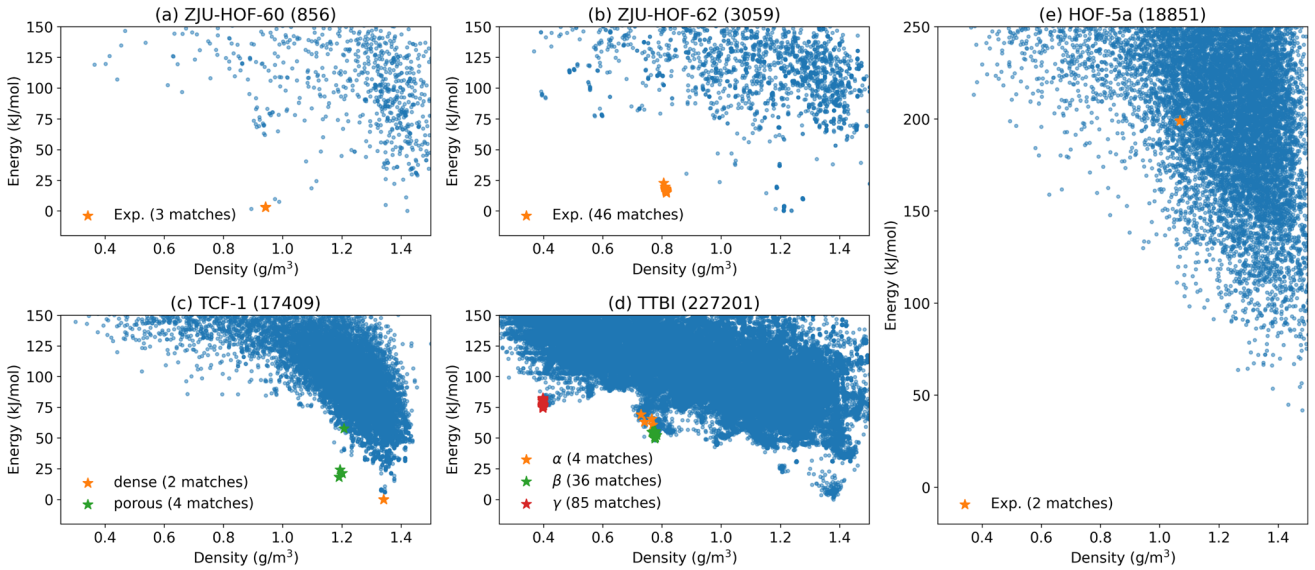

Fig. 2 Structure search based on the known space group symmetry for the five systems: (a) ZJH-HOF-60, (b) ZJU-HOF-62, (c) TCF-1, (d) TTBI and (e) HOF-5a. In the title of each plot the number of sampled structures are denoted within the parentheses. Here, 'matches' denote the generated structures that are identified as equivalent to the target experimental structure.

polymorphs requires more search efforts about 20–200k trial structures, but the cost are not significant since it is mostly about classical force field optimizations. For HOF-5a (see Fig. 2c), the simulation attempted at the space group *Cc*, returning a success rate of 1 per 10k structures that is similar to TCF-1.

In general, identifying the experimental structure is straightforward when the space group symmetry is known.

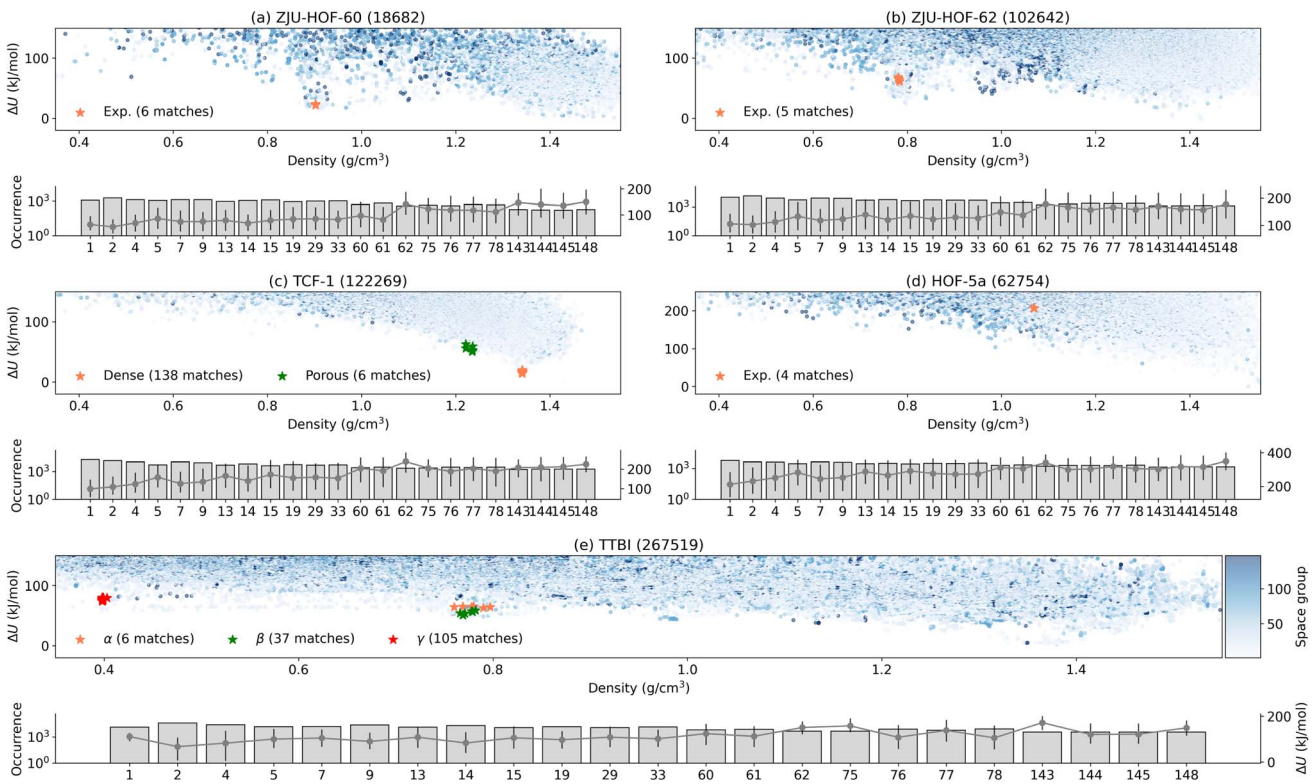

Fig. 3 Structure search performed within the common space group symmetries for the five systems: (a) ZJU-HOF-60, (b) ZJU-HOF-62, (c) TCF-1, (d) HOF-5a and (e) TTBI. For each system, the upper panel displays a scatter plot of energy *versus* density colored by the space group number, and the lower panel shows the occurrence of structures in each space group. In case of lower panel, bar height (plotted against the left *y*-axis) indicates the number of generated structures; and gray circles connected by lines (plotted against the right *y*-axis) denote the mean relative energy ΔU within each space group, while vertical whiskers represent the energy spread (10–90th percentile).

Using MACE for energy ranking, all experimental structures fall within 0–20 kJ mol^{-1} above the computed ground state. Even the higher-energy TTBI polymorphs (Fig. 2d) appear as distinct local minima within the nearby low-density basin. The sole exception is HOF-5a: the matched structure lies about 200 kJ mol^{-1} above the global minimum and about 100 kJ mol^{-1} above the lowest minimum at a similar density. This large discrepancy underscores a current limitation of the present energy models and this will be discussed later.

4.2 Blind search with 23 most common space groups

We next considered a realistic scenario when the space group is unknown, in which we performed blind searches with $Z' = 1$ over 23 common organic-crystal space groups. This setting is more challenging than the single known-space-group tests because (i) sampling capacity must be divided among symmetry classes and (ii) incorrect symmetry assumptions can delay convergence to the experimental packing. To improve efficiency, we employed the DFS strategy, in which each new trial selects a space group from the list and low-energy structures are preferentially retained and propagated.[20] Consequently, the effective space-group visitation frequencies become non-uniform and biased toward symmetries that support lower-energy packings for the given molecule.

Fig. 3 summarizes the aggregate outcomes. The overall search effort required to recover the experimental forms increased only modestly after introducing the multi–space-group pool. For example, the γ-TTBI polymorph (experimental $P6_3/mmc$, $Z' = \frac{1}{12}$) can be represented in several subgroup settings with $Z' = 1$ on a general Wyckoff position; it was matched 105 times in $\sim 2 \times 10^5$ trials—comparable to the rate observed in the focused search of Fig. 2. Similar behavior was found for α-TTBI and β-TTBI: match frequencies remained within the same order of magnitude as their known-symmetry counterparts. Furthermore, dense-TCF-1 exhibited an even more favorable outcome, with match frequencies improving notably under common-space-group sampling, indicating that the DFS energy bias quickly suppresses unproductive symmetry branches while amplifying access to the correct packing basin. However, the strategy may favor the formation of low energy structures. For ZJU-HOF-60 and ZJU-HOF-62, the layered low-density motifs emerged early and repeatedly, despite competition from low-symmetry triclinic and monoclinic settings. Hence, it is fair to conclude that blind search for porous structure with fractional Z' does not significantly increase the overhead as long as the true symmetry is included in the given list of space groups in the input.

A more significant challenge is the energy ranking. After search for more space groups, we have found that most of the target structures remain energetically favorable as either the global minimum or the local minimum in the surrounding density range. However, ZJU-HOF-62 in Fig. 3 is no longer favorable as compared to that Fig. 2. The case of HOF-5a is about the same. This imposes the additional challenge to identify the experimental structures if they do not appear as either the global or local energy minimum.

4.3 Energy ranking with different models

To address the energy-ranking challenges, we re-ranked representative structures (experimental form, matched instances, global minimum, and the nearest local minimum at comparable density) with MACE, MACE-OFF, UMA, LATTE, and DFT. The results are summarized in Fig. 4.

In general, MACE provides a reliable ordering for most systems at low cost as compared to the DFT results. MACE-OFF, while somewhat more expensive, did not consistently improve the rank correlation for the present porous sets—likely reflecting limited coverage of large low-density HOF topologies in its training data. LATTE (SCC-DFTB) is more costly than the ML single points yet still far cheaper than DFT; LATTE demonstrates an accuracy similar to MACE, and tends to overestimates the energy difference as compared do DFT. This comes as no surprise considering that HOF were not included in the development of the *lanl*31 parameterization used here.

Most importantly, UMA yields an ordering that is qualitatively consistent with the DFT re-ranking in all benchmark cases. Both UMA and DFT correct the MACE misordering for

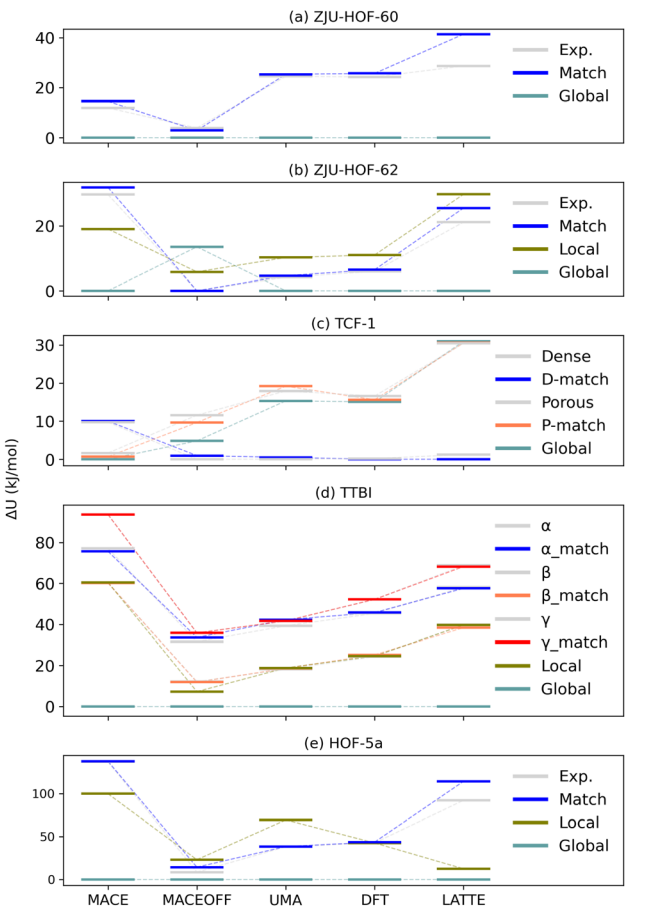

Fig. 4 Energy ranking across different methods for multiple representative structures found in each system: (a) ZJU-HOF-60, (b) ZJU-HOF-62, (c) TCF-1, (d) TTBI and (e) HOF-5a. Here, 'Global' refers to the structure with the lowest overall energy among all candidates found in the search, while 'Local' refers to the lowest-energy structure within the specific density region corresponding to the experimental structure.

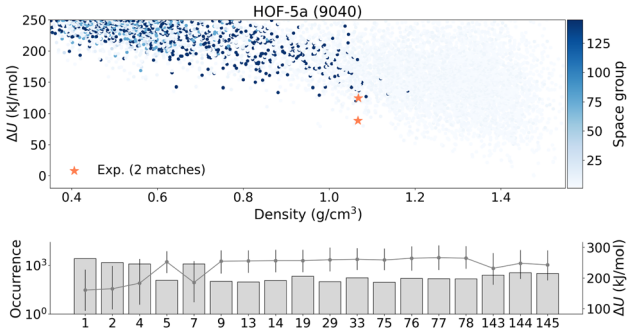

Fig. 5 Structure search performed within the common space group symmetries for HOF-5a. The upper panel displays a scatter plot of energy *versus* density colored by the space group number, and the lower panel shows the occurrence of structures in each space group; where bar height indicates the number of generated structures; gray circles connected by lines denote the mean relative energy ΔU within each space group, and vertical whiskers represent the energy spread.

HOF-5a, stabilizing the experimental porous form relative to the spurious nearby low-density minimum. This improvement directly increases the likelihood of retaining the true polymorph during down-selection. Given its accuracy – cost balance, UMA is a practical choice for intermediate (post-FF) energy refinement in future large CSP screenings, reserving DFT for only a small final candidate set.

4.4 Improved energy ranking of HOF-5a using UMA

Using UMA as the energy model substantially improved the energy ranking of the generated HOF-5a candidates, getting the matched structure within a local minimum region of the energy landscape. Moreover, we observed a clear improvement in efficiency: 2 matches were found within 9040 generated structures, compared with 4 matches within 62 754 structures in our previous setup. In Fig. 5, the two matched structures initially appear at different energies, but after proper relaxation they converge to the same final structure and overlap with each other. This indicates that they correspond to the same minimum and therefore have the same relaxed energy; the initial energy difference is due to imperfect starting geometries and model noise. Although UMA is more computationally expensive, we recommend using it for complex systems where MACE struggles to rank candidates correctly and fails to place the matched structures in the metastable region.

5 Conclusions

We have demonstrated that automated crystal structure prediction of porous organic crystals is achievable using the HTOCSP framework, despite the inherent challenges posed by their expanded configurational space and energetic preference for metastable low-density forms. Through systematic evaluation of five benchmark systems, we established that carefully designed sampling strategies combined with multi-stage energy refinement can reliably recover experimentally observed porous polymorphs.

When the experimental space group is known, the HTOCSP sampling generates the target structures efficiently, with match rates substantially improved by short classical FF relaxations. However, in realistic blind-search scenarios across common space groups with $Z' = 1$, the two-stage relaxation protocol of initial FF minimization followed by ML refinement—becomes essential for identifying true matches from the energy ranking perspective. In particular, UMA consistently rank experimental porous forms as energetically competitive within low-density basins, enabling effective post-screening to capture subtle dispersion and electrostatic balances. As such, CSP may be promising as a predictive tool for the rational design of functional porous organic materials prior to synthesis.

Author contributions

S. H., W. Z. and Q. Z. proposed this idea and supervised this research. M. M. M. performed the majority of materials simulations. R. P. participated on the tight binding simulation and structural analysis. All coauthors designed the research, analyzed the calculations and wrote this manuscript.

Conflicts of interest

There are no conflicts to declare.

Data availability

The data and scripts used in this study, are available in https://github.com/MaterSim/HTOCSP.

Acknowledgements

Q. Z. acknowledge the NSF (DMR-2410178) for the financial supports. R. P. acknowledges funding from the Laboratory Directed Research and Development program of Los Alamos National Laboratory under project no. 20260424ER. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (contract no. 89233218CNA000001). The computing resources are provided by ACCESS (TG-MAT230046).

References

1 S. Fratini, S. Ciuchi, D. Mayou, G. T. De Laissardière and A. Troisi, A map of high-mobility molecular semiconductors, *Nat. Mater.*, 2017, **16**(10), 998–1002.
2 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**(26), 1808256.
3 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**(SD), SD0801.
4 T. Nematiaram, D. Padula and A. Troisi, Bright frenkel excitons in molecular crystals: a survey, *Chem. Mater.*, 2021, **33**(9), 3368–3378.

5 A. Stuke, C. Kunkel, D. Golze, M. Todorović, J. T. Margraf, K. Reuter, *et al.*, Atomic structures and orbital energies of 61 489 crystal-forming organic molecules, *Sci. Data*, 2020, **7**(1), 1–11.
6 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**(3), 969–978.
7 R. Taylor and P. A. Wood, A Million Crystal Structures: The Whole Is Greater than the Sum of Its Parts, *Chem. Rev.*, 2019, **119**(16), 9427–9477.
8 A. Vaitkus, A. Merkys and S. Gražulis, Validation of the crystallography open database using the crystallographic information framework, *J. Appl. Crystallogr.*, 2021, **54**(2), 661–672.
9 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**(13), 4361–4371.
10 Q. Zhu and S. Hattori, Organic crystal structure prediction and its application to materials design, *J. Mater. Res.*, 2023, **38**(1), 19–36.
11 J. P. Lommerse, W. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Gavezzotti, D. W. Hofmann, *et al.*, A test of crystal structure prediction of small organic molecules, *Acta Crystallogr., Sect. B*, 2000, **56**(4), 697–714.
12 W. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Dzyabchenko, P. Erk, A. Gavezzotti, *et al.*, Crystal structure prediction of small organic molecules: a second blind test, *Acta Crystallogr., Sect. B*, 2002, **58**(4), 647–661.
13 G. M. Day, *et al.*, A third blind test of crystal structure prediction, *Acta Crystallogr., Sect. B*, 2005, **61**, 511–527.
14 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*, 2009, **65**, 107–125.
15 D. A. Bardwell, *et al.*, Towards crystal structure prediction of complex organic compounds—A report on the fifth blind test, *Acta Crystallogr., Sect. B*, 2011, **67**, 535–551.
16 A. M. Reilly, R. I. Cooper, C. S. Adjiman, S. Bhattacharya, A. D. Boese, J. G. Brandenburg, *et al.*, Report on the sixth blind test of organic crystal structure prediction methods, *Acta Crystallogr., Sect. B*, 2016, **72**(4), 439–459.
17 L. M. Hunnisett, J. Nyman, N. Francia, N. S. Abraham, C. S. Adjiman, S. Aitipamula, *et al.*, The seventh blind test of crystal structure prediction: structure generation methods, *Acta Crystallogr., Sect. B*, 2024, **80**(6), 517–547.
18 A. R. Oganov, C. J. Pickard, Q. Zhu and R. J. Needs, Structure prediction drives materials discovery, *Nat. Rev. Mater.*, 2019, **4**(5), 331–348.
19 S. L. Price, Predicting crystal structures of organic compounds, *Chem. Soc. Rev.*, 2014, **43**(7), 2098–2111.
20 Q. Zhu and S. Hattori, Automated high-throughput organic crystal structure prediction *via* population-based sampling, *Digital Discovery*, 2025, **4**(1), 120–134.
21 R. B. Lin, Y. He, P. Li, H. Wang, W. Zhou and B. Chen, Multifunctional porous hydrogen-bonded organic framework materials, *Chem. Soc. Rev.*, 2019, **48**(5), 1362–1389.
22 P. Brunet, M. Simard and J. D. Wuest, Molecular tectonics. Porous hydrogen-bonded networks with unprecedented structural integrity, *J. Am. Chem. Soc.*, 1997, **119**(11), 2737–2738.
23 H. Wang, B. Li, H. Wu, T. L. Hu, Z. Yao, W. Zhou, *et al.*, A flexible microporous hydrogen-bonded organic framework for gas sorption and separation, *J. Am. Chem. Soc.*, 2015, **137**(31), 9963–9970.
24 M. Mastalerz and I. M. Oppel, Rational construction of an extrinsic porous molecular crystal with an extraordinary high specific surface area, *Angew. Chem., Int. Ed.*, 2012, **51**(21), 5252–5255.
25 I. Bassanetti, S. Bracco, A. Comotti, M. Negroni, C. Bezuidenhout, S. Canossa, *et al.*, Flexible porous molecular materials responsive to CO_2, CH_4 and Xe stimuli, *J. Mater. Chem. A*, 2018, **6**(29), 14231–14239.
26 Y. B. Wang, Y. X. Lin, J. X. Wang, X. Zhang, H. Wu, W. Zhou, *et al.*, A Microporous Hydrogen-Bonded Organic Framework with Alkynyl Sites for Highly Efficient Propane/Propylene Separation, *J. Am. Chem. Soc.*, 2025, **147**, 24403–24412.
27 A. Pulido, L. Chen, T. Kaczorowski, D. Holden, M. A. Little, S. Y. Chong, *et al.*, Functional materials discovery using energy–structure–function maps, *Nature*, 2017, **543**(7647), 657–664.
28 *RDKit: Open-source cheminformatics*, nline; accessed 11-April-2013, https://www.rdkit.org.
29 S. Fredericks, K. Parrish, D. Sayre and Q. Zhu, Pyxtal: a python libraryfor crystal structure generation and symmetry analysis, *Comput. Phys. Commun.*, 2021, **261**, 107810.
30 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**(2), 187–217.
31 S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, *et al.*, Python Materials Genomics (pymatgen): a robust, open-source python library for materials analysis, *Comput. Mater. Sci.*, 2013, **68**, 314–319.
32 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**(9), 1157–1174.
33 S. Boothroyd, P. K. Behara, O. Madin, D. Hahn, H. Jang and V. Gapsys, *et al.*, *Development and Benchmarking of Open Force Field 2.0.0—the Sage Small Molecule Force Field*, 2023.
34 D. A. Case, H. M. Aktulga, K. Belfon, I. Ben-Shalom, S. R. Brozell and D. S. Cerutti, *et al.*, *Amber 2021*, University of California, San Francisco, 2021.
35 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. Oh A. H., Agarwal A., Belgrave D., Cho K., 2022.
36 D. P. Kovács, J. H. Moore, N. J. Browning, I. Batatia, J. T. Horton, Y. Pu, *et al.*, Mace-off: Short-range

transferable machine learning force fields for organic molecules, *J. Am. Chem. Soc.*, 2025, **147**(21), 17598–17611.

37 B. M. Wood, M. Dzamba, X. Fu, M. Gao, M. Shuaibi and L. Barroso-Luque, *et al.*, UMA: A Family of Universal Models for Atoms, *arXiv*, 2025, preprint, arXiv:250623971, DOI: 10.48550/arXiv.250623971.

38 N. Bock, M. J. Cawkwell, J. D. Coe, A. Krishnapriyan, M. P. Kroonblawd and A. Lang, *et al.*, LATTE, 2008, available from: https://github.com/lanl/LATTE.

39 M. Elstner and G. Seifert, Density functional tight binding, *Philos. Trans. R. Soc., A*, 2014, **372**, 20120483.

40 M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, *et al.*, Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1998, **58**(11), 7260–7268.

41 M. J. Cawkwell and R. Perriot, Transferable density functional tight binding for carbon, hydrogen, nitrogen, and oxygen: application to shock compression, *J. Chem. Phys.*, 2019, **150**(2), 024107.

42 N. Lease, L. M. Klamborowski, R. Perriot, M. J. Cawkwell and V. W. Manner, Identifying the Molecular Properties that Drive Explosive Sensitivity in a Series of Nitrate Esters, *J. Phys. Chem. Lett.*, 2022, **13**(40), 9422–9428.

43 H. Singh, C. F. A. Negre, A. Redondo and R. Perriot, Surface Studies of β-1,3,5,7-Tetranitro-1,3,5,7-Tetrazoctane and Pentaerythritol Tetranitrate from Density Functional Tight-Binding Calculations and Implications on Crystal Shape, *Cryst. Growth Des.*, 2024, **24**(9), 3681–3690.

44 G. Kresse and J. Furthmüller, Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1996, **54**, 11169–11186.

45 J. P. Perdew, K. Burke and M. Ernzerhof, Generalized gradient approximation made simple, *Phys. Rev. Lett.*, 1996, **77**(18), 3865.

46 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu, *J. Chem. Phys.*, 2010, **132**(15), 154104.