PNAS RESEARCH ARTICLE | CHEMISTRY OPEN ACCESS

A data-driven and topological mapping approach for the a priori prediction of stable molecular crystalline hydrates

Richard S. Hong[a,b], Alessandra Mattei[a], Ahmad Y. Sheikh[a], and Mark E. Tuckerman[b,c,d,e,1]

Edited by Berend Smit, University of California, Berkeley, CA; received March 12, 2022; accepted September 5, 2022 by Editorial Board Member Daan Frenkel

Predictions of the structures of stoichiometric, fractional, or nonstoichiometric hydrates of organic molecular crystals are immensely challenging due to the extensive search space of different water contents, host molecular placements throughout the crystal, and internal molecular conformations. However, the dry frameworks of these hydrates, especially for nonstoichiometric or isostructural dehydrates, can often be predicted from a standard anhydrous crystal structure prediction (CSP) protocol. Inspired by developments in the field of drug binding, we introduce an efficient data-driven and topologically aware approach for predicting organic molecular crystal hydrate structures through a mapping of water positions within the crystal structure. The method does not require *a priori* specification of water content and can, therefore, predict stoichiometric, fractional, and nonstoichiometric hydrate structures. This approach, which we term a *mapping approach for crystal hydrates* (MACH), establishes a set of rules for systematic determination of favorable positions for water insertion within predicted or experimental crystal structures based on considerations of the chemical features of local environments and void regions. The proposed approach is tested on hydrates of three pharmaceutically relevant compounds that exhibit diverse crystal packing motifs and void environments characteristic of hydrate structures. Overall, we show that our mapping approach introduces an advance in the efficient performance of hydrate CSP through generation of stable hydrate stoichiometries at low cost and should be considered an integral component for CSP workflows.

Significance

The problem of predicting molecular crystalline hydrates through crystal structure prediction (CSP) is a significant challenge due to the range of possible stoichiometric and nonstoichiometric water contents and the large number of degrees of freedom that need to be considered during the crystal structure search. The algorithm introduced here represents an advance in this area that allows for quick and efficient mapping and prediction of hydrate crystal structures from an anhydrous framework without requiring a priori specification of the water content. Overall, this approach can serve as a computationally inexpensive tool to complement standard CSP for uncovering stable hydrate structures.

crystal structure prediction | hydrate polymorphs

The unique polar characteristics of water allow it to form a variety of strong, directional interactions in the solid state, including hydrogen-bonding and ion-coordinated interactions. As such, organic molecules often crystallize as or convert to hydrates under humid conditions during storage and/or downstream processing (1) through the incorporation of water into the crystal lattice. These organic crystalline hydrates are of particular interest to the pharmaceutical industry as they can significantly impact bioavailability (2), manufacturability (3), mechanical properties (4, 5), and/or chemical stability (6) of the active pharmaceutical ingredient (API). In fact, it is estimated that up to a third of commercially marketed pharmaceuticals contain APIs in hydrated forms. Hydrates also represent a very significant portion of patented solid forms of pharmaceuticals (7). In addition, in the fields of organic electronics, understanding hydrate formation can be essential in designing, tuning, and rationalizing the performance of luminescent materials (8). Given the prevalence and importance of these hydrate materials, many efforts have been made to obtain a thorough understanding of their structural patterns (9), to rationalize why and when they form, and to predict their structures (10–12) and physical and chemical properties (13).

While computational tools, such as crystal structure prediction (CSP), have occasionally been utilized to predict structures of crystal hydrates, these predictions are significantly more computationally intensive than anhydrous CSP searches due to the additional degrees of freedom that need to be considered when water molecules are present (10). Without *a priori* knowledge of the thermodynamically stable hydrate stoichiometry, hydrate CSP requires separate "brute force" searches for each plausible stoichiometry or fractional stoichiometry. This inadvertently leads to high computational overhead and still leaves the additional and nontrivial task of determining the most stable hydrate stoichiometry.

While CSP for hydrates is often expensive, the guest-free hydrate (or solvate) framework structures may often be captured via an anhydrous CSP search (14–17). However, an inability to predict and determine potential uptake and incorporation of water in these predicted framework structures can be a significant and costly gap in the overall understanding of the solid form landscape. Even when these anhydrous framework structures

Author contributions: A.M., A.Y.S., and M.E.T. designed research; R.S.H. performed research; R.S.H. analyzed data; and R.S.H., A.M., A.Y.S., and M.E.T. wrote the paper.

Competing interest statement: AbbVie sponsored and partially funded the study; contributed to the design; and participated in the collection, analysis, and interpretation of data and in writing, reviewing, and approval of the final publication. R.S.H., A.M., and A.Y.S. are employees of AbbVie and may own AbbVie stock.

This article is a PNAS Direct Submission. B.S. is a guest editor invited by the Editorial Board.

[1] To whom correspondence may be addressed. Email: Mark.Tuckerman@nyu.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2204414119/-/DCSupplemental.

Published October 17, 2022.

are accurately predicted, it remains nontrivial to deduce which of these frameworks are amenable to water uptake and how many water molecules may reside within the crystal lattice. This lack of understanding of the water distribution in crystal hydrates can only make it more difficult to predict and anticipate storage and downstream processing risks due to water uptake and could even result in patent litigation cases. As an example of the latter, consider the case of Paxil for which a class-action patent litigation ensued when a generic manufacturer marketing a hygroscopic anhydrous polymorph was accused of infringing on a patented hemihydrate polymorph due to a likely solid form conversion from exposure to humidity (18, 19). Retrospectively, detailed structural and energetic understanding of these solid forms would have provided clearer assertions in court and more distinctly defined the patent claims.

So far, no methods have been reported for predicting hydrates efficiently that do not rely on random packing searches with an *a priori* specification of a desired stoichiometry. While grand canonical Monte Carlo methods can be used to predict inclusion behavior through the insertion/removal/replacement of solvent in porous organic and inorganic solids (20, 21) and protein structures (22), these methods can have limitations in dense systems, such as molecular crystals, which exhibit highly directional interactions due to extremely low acceptance probabilities for both solvent insertion and removal (23, 24).

Inspired by computational methods used in drug design for the placement of water molecules within protein structures, including WaterMap, WScore, and other scoring or mapping approaches (25–27), we introduce a simple, high-throughput technique to map out likely locations of water molecules within these framework molecular crystals that does not rely on an energetic acceptance probability. Our approach, termed mapping approach for crystal hydrates (MACH), involves flooding the crystal lattice, followed by systematic removal of water molecules located at unfavorable locations through scoring approaches to determine favorable positions of water molecules within the crystal lattice while allowing their positions and interactions therein to relax. In this way, we can explicitly consider the chemical environments of potentially accessible voids within the crystal structure.

MACH allows access to hydrate crystal structures of varying and plausible stoichiometry simply from a framework structure without prior knowledge or specification of a stoichiometry of the hydrate structure. As such, this brings the predictions of partial stoichiometries and nonstoichiometric hydrates into scope for CSP. Crucially, our approach accurately differentiates between voids within the crystal structure that are likely to host water molecules from those that are not. In so doing, we show that favorable insertion of waters is dependent on the local chemical environments of a crystal structure's accessible voids. We demonstrate the ability of MACH to successfully map out waters and predict the stoichiometries and water positions of hydrates of brucine, paroxetine hydrochloride, and sitagliptin tartrate, the latter two being notable in regimens for the treatment of depression and diabetes, respectively. The dry frameworks of these pharmaceutically relevant hydrates contain void spaces of differing volumes and chemical environments, including open hydrophilic channels, large hydrophobic voids, and tight hydrophilic pockets. As such, these structures serve as excellent examples to test the ability of our approach to distinguish diverse chemical environments of each void space for water insertion. Last, we discuss how combining MACH with either standard full anhydrous CSP or evolutionary algorithm CSP constitutes a promising direction in CSP methodology.

Water Insertion Workflow Methodology

The approach for water insertion is shown in Fig. 1 and starts by overlaying an equilibrated water box with the crystal lattice to sample many possible water positions. Water molecules not within 3.2 Å of electronegative atoms (e.g., carbonyl oxygens) of the host molecule are then deleted, ensuring that water molecules are inserted in environments where at least one directional intermolecular interaction can be formed with the host molecule within the crystal lattice. The remaining water molecules that exhibit poor contacts (as determined through an atom-specific cutoff) or reside in hydrophobic regions (as determined through a radial symmetry function) are then deleted. These steps are repeated with random displacements of the original water box coordinates to sample other potential water insertion positions within the crystal lattice. Subsequently, the positions and geometry of the inserted water molecules are optimized. Due to the nature of this algorithm, it is possible for more than one water molecule to be inserted at a time if the cutoff criteria are met. These iterative loops of water insertion and optimization are repeated until no additional waters are accepted. Afterward, the hydrogen bonding and interaction network of the resulting water molecules are reassessed; waters that do not form at least two hydrogen bonds [as defined by Baker and Hubbard (28) and McGibbon et al. (29)] or an ion-coordinated interaction are deleted. A short 50-ps constant volume/constant temperature (canonical) molecular dynamics (NVTMD) run is carried out, and minimization of the inserted water molecules is performed to relax them into their optimal orientations and positions. Note that the MACH workflow could also be augmented with enhanced MD and Monte Carlo translations and rotations to sample favorable configurational spaces where large energetic barriers between water configurations may exist. These modifications will be considered in a future study. Last, the geometry and cell lattice are fully optimized to allow for unit cell expansions and conformational changes with the additional lattice waters. These steps can then be performed in additional iterative loops after structural optimization to determine potential higher hydrates if further accessible voids are generated through full structural optimization.

The atomic Van der Waals radii of the crystal host atoms and atom-specific probe radii for water molecules (*SI Appendix*, Table S1) are included to define generalizable atom-specific cutoff distances for step 3. This approach is analogous to those employed for mapping protein solvent accessible surfaces (30) and describes the directional interactions of water (i.e., hydrogen bonding) with certain atom types.

As a further refinement and confirmation of the validity of these cutoff distances, we analyzed over 10,000 organic hydrate crystal structures deposited in the Cambridge Crystallographic Database (CCDC) using criteria described in *Computational Methods*. In so doing, we created histograms of atomic distances between the water oxygens and the different atom types of the host crystal structure, showing that the cutoff distances appear at the tails of these distributions (*SI Appendix*, Fig. S1 and Table S2) and confirming that they both are physically realistic and allow us to capture most hydrate structures.

In addition, we hypothesized that water molecules are unlikely to reside in highly hydrophobic voids within the crystal structure. To study the chemical environments of experimental hydrates within the CCDC database, we have determined the number and distances of neighboring hydrophobic heavy atoms for each water molecule. These hydrophobic atoms are defined as nonionized heavy atoms excluding nitrogen and oxygen. Such atoms do

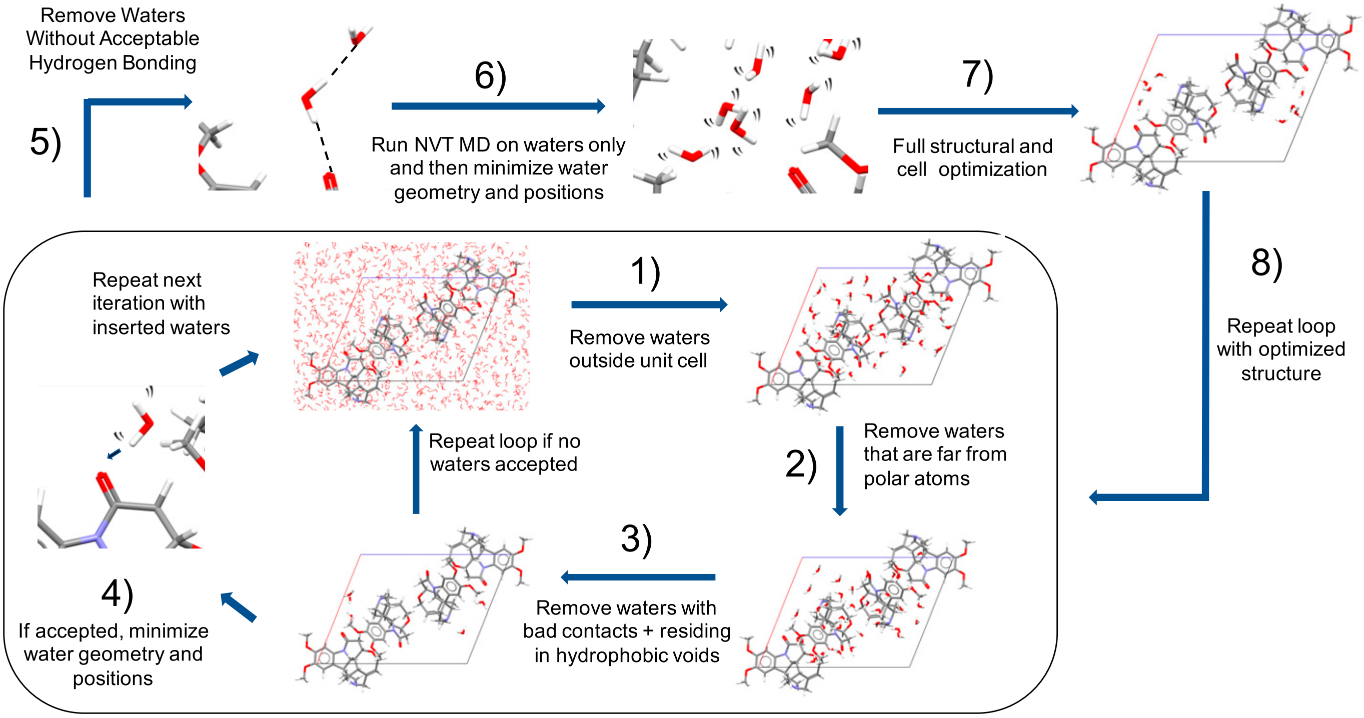
Remove Waters Without Acceptable Hydrogen Bonding
5)
6)
Run NVT MD on waters only and then minimize water geometry and positions
7)
Full structural and cell optimization
8)
Repeat loop with optimized structure
Repeat next iteration with inserted waters
1)
Remove waters outside unit cell
2)
Remove waters that are far from polar atoms
Repeat loop if no waters accepted
3)
Remove waters with bad contacts + residing in hydrophobic voids
4)
If accepted, minimize water geometry and positions

Fig. 1. Schematic representation and description for MACH, illustrating each of the main steps.

not typically participate in directional interactions with water. In order to quantify the local hydrophobic environment of each water molecule, we utilize a single radial symmetry function (31). The calculation is performed for each lattice water molecule, where the penalty score contribution, denoted $P(r_{ij})$, of each individual neighboring hydrophobic atom (Eq. **1**) is summed to determine an environmentally descriptive scoring function, $G_i(\boldsymbol{r})$, where $\boldsymbol{r}$ denotes the set of $N-1$ atomic coordinates, excluding those of atom *i*. The resulting $G_i(\boldsymbol{r})$ value of each individual water is then used to quantify the hydrophobicity of the water's surrounding environment.

$$
\begin{aligned}
G_i(\boldsymbol{r}) &= \sum_{j\neq i}^{N} P_{ij}(r_{ij}) \\
&= \sum_{j\neq i}^{N} e^{-\eta(|r_{ij}|-R_s)^2}\,\frac{1}{2}\left(\cos\left(\frac{|r_{ij}|-R_S}{r_{max}-R_S}\right)\pi+1\right). \qquad [\mathbf{1}]
\end{aligned}
$$

Here $G_i(\boldsymbol{r})$ is calculated for each water molecule *i*, r_{ij} denotes the distance between the center of mass of water molecule *i* and the neighboring atoms indexed as *j*, R_S denotes the atom-specific cutoff distances defined earlier, and r_{max} denotes the maximum distance considered for neighboring atoms. Lengths are input into Eq. **1** (assuming angstrom units) with η as a constant, set as 1 Å^{-2}. The function $G_i(\boldsymbol{r})$ assigns a higher penalty for hydrophobic atoms that are closer to the water molecule of interest. The Gaussian function provides a smooth decrease as a function of increasing interatomic distance, and the cosine function serves as a cutoff function (Fig. 2*A*).

The $G_i(\boldsymbol{r})$ values were calculated for each hydrate crystal structure within the CCDC database, with $r_{max} = R_S + 1.0$ Å, where 99% of the hydrate waters exhibit $G_i(\boldsymbol{r})$ of less than 4.0 (Fig. 2*B*). Thus, within our workflow, water molecules with $G_i(\boldsymbol{r})$ greater than 4.0 are considered unfavorable and not considered for insertion.

Results and Discussion

We first test MACH on the experimentally dehydrated brucine dihydrate structure (Fig. 3) to investigate how our approach performs on a channel hydrate that exhibits large, connected voids. The dihydrate crystal retains its framework structure once dehydrated (15, 32), and the energetic stability of the guest-free framework structure is confirmed through a CSP study by

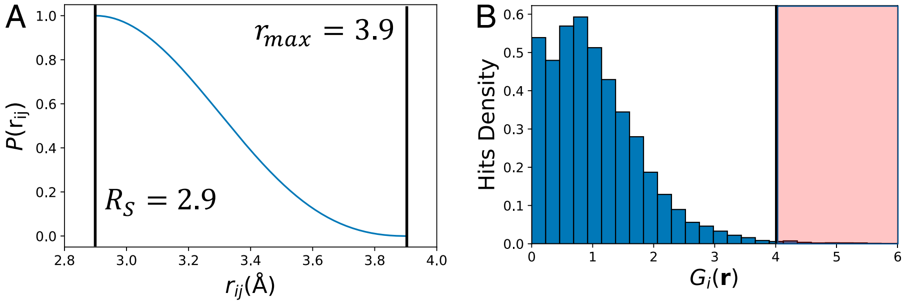

Fig. 2. (*A*) $P(r_{ij})$ for neighboring carbon atoms within 1.0 Å beyond the cutoff R_S. (*B*) Histogram of all $G_i(\boldsymbol{r})$ values from the CCDC hydrate structures. Water positions with $G_i(\boldsymbol{r})$ greater than 4.0 (as shown in red) are considered unfavorable for insertion in our approach.

Braun et al. (9), where this structure falls within 20 kJ/mol of the global minimum (*SI Appendix*, Fig. S2). In addition, upon dispersion-corrected density functional theory (DFT-D) minimization of the dehydrated structure, the framework is maintained, with an overall rmsd of 0.10 Å when compared to the fully hydrated framework (*SI Appendix*, Fig. S3). The abundance of open channels and the availablility of hydrogen-bond acceptor groups suggests this framework structure can clearly incorporate water. However, without experimental data, the water distribution within the lattice and its maximum water stoichiometry cannot be deduced.

Using the MACH protocol, we iteratively insert waters to completion, which correctly identifies the maximum stoichiometry and positions of water. The first water molecule inserted in the unit cell is placed within proximity of the exposed carbonyl oxygen, satisfying the criterion defined by step 2 of our workflow (Fig. 3*B*). The inserted water is then optimized, where it relaxes to its optimal position and orientation for forming a hydrogen bond to the host molecule. Through further iterations, subsequent water molecules are then inserted within proximity of the first water molecule as waters can also hydrogen bond to each other within the crystal structure. After 1,000 insertion attempts (steps 1 to 3) and their subsequent optimization and MD steps (steps 4 to 8 of Fig. 1), this approach was able to insert waters into the brucine structure closely matching their experimental positions, with an average deviation in the water positions of 0.34 Å from the experimental structure and 0.19 Å from the DFT-D optimized experimental structure (*SI Appendix*, Fig. S5). Ultimately capturing the brucine dihydrate structure (Fig. 3*A*) at its maximum water stoichiometry. The central processing unit (CPU) times of each of these steps are provided in *SI Appendix*, Table S3. However, it is important to note here that these timings will depend on the system being studied and the hardware, software, and algorithms utilized for the minimization and MD stages.

We note that due to the random nature of the positioning and displacements of the initial water box, variability may be observed with repeated runs in the final positions of the waters of up to 0.63 Å (*SI Appendix*, Fig. S6). However, this may be reflective of thermal variations in water positioning due to the dynamic nature of the water molecules as observed by the experimental thermal ellipsoids (*SI Appendix*, Fig. S4) and the fact that the observed experimental structures can often also be a dynamic ensemble of predicted structures (33, 34). To capture these potential variations, an ensemble average of the water positions generated from multiple MACH runs can be used. In

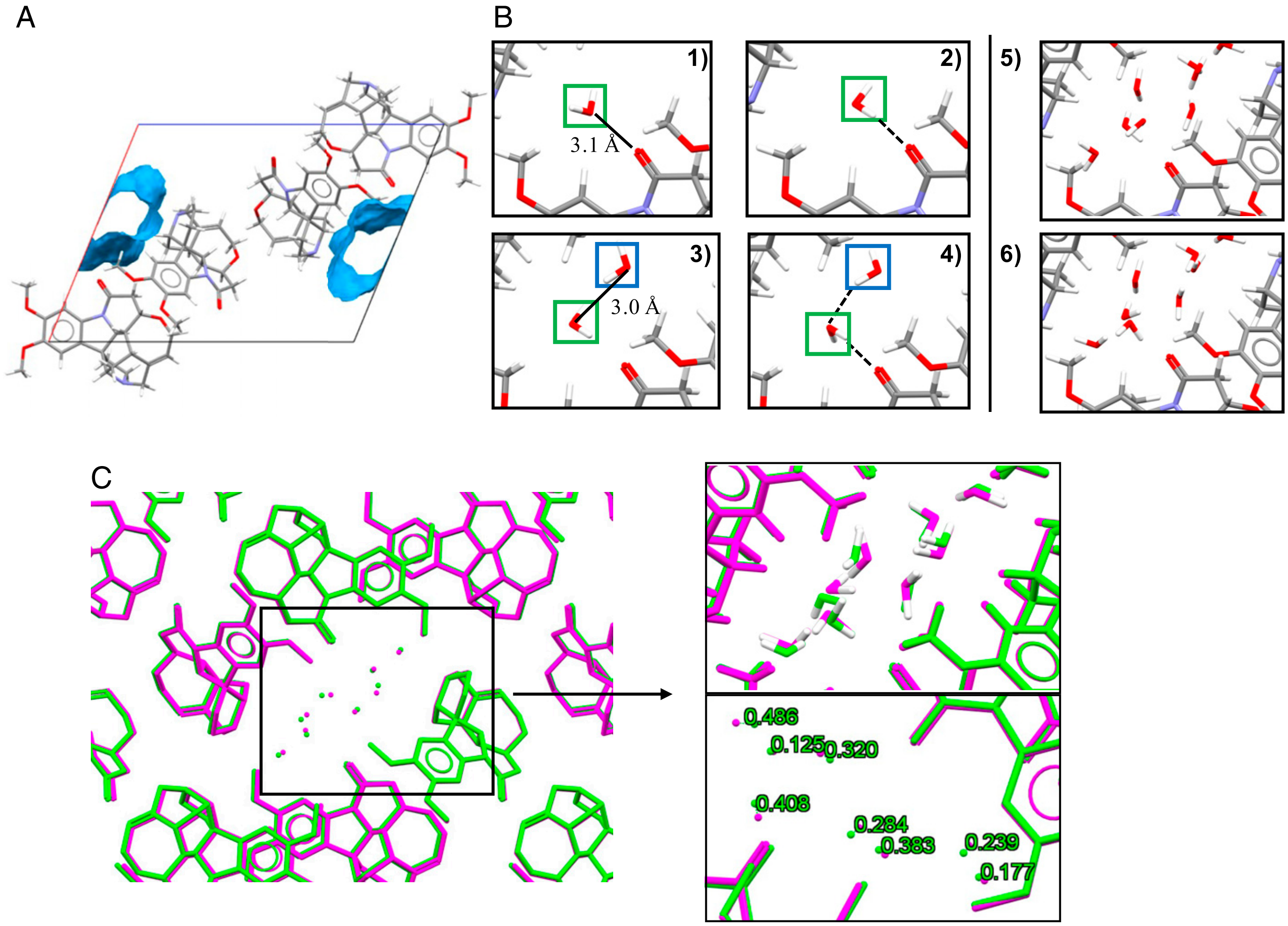

Fig. 3. (*A*) View of brucine dihydrate down the *b* axis showing channel-like voids. The blue contact surfaces represent accessible volume through a spherical probe radius of 1.2 Å. (*B*) Stepwise water insertion for brucine showing 1) the first water (green) being placed near a hydrogen bond acceptor, 2) geometry optimization of the inserted water, 3) insertion of the second water (blue) within proximity of the first water, 4) geometry optimization of both inserted waters, 5) total inserted waters from steps 1 to 4, and 6) final water positions and orientations after a short NVTMD and minimization. (*C*) Comparison of the final structure from this protocol (green) with the experimental crystal structure (magenta) showing the comparison of the water positions and orientations. (While the predicted water orientations do not exactly match the experimental orientations, it is likely that at finite temperatures, water molecules are mobile, and hydrogen atom positions are may not be exactly resolved experimentally.)

the case of brucine, we observed that the average positions of water centers from multiple MACH run repeats matched well with experimental positions, with an average deviation in water positions of 0.24 Å (*SI Appendix*, Fig. S6).

To predict the brucine dihydrate structure *de novo* through traditional hydrate CSP, extensive structure searches with three individual components (one brucine and two water), involving potentially hundreds to thousands of CPU hours (35), would be required. However, when provided a framework structure from an existing anhydrous CSP landscape, MACH can quickly access this structure in a matter of minutes or less. While traditional CSP structure generators may, in theory, be used to sample potential hydrate water positions when given an anhydrous framework, placement of waters into random, unfavorable positions will lead to overprediction of many energetically unfavorable hydrates, along with the additional computational overhead of minimizing each of these structures. In addition, such an approach would require *a priori* specification of the desired number of waters to incorporate within the crystal lattice.

In order to test MACH on fractional hydrates where the available void spaces are not as obvious, we have studied the stoichiometric hemihydrate form I of the paroxetine hydrochloride, the subject of notable class-action polymorph lawsuits (18, 19, 36–38). By simply visualizing the voids of the experimental dehydrated paroxetine hydrochloride form I, it is not obvious to conclude if this structure can accommodate water molecules. For example, by probing the void spaces in this structure with the default 1.2-Å probe radius in the CCDC's Mercury toolkit, as is typically done in structural void analysis (15, 39–41), there appear to be no apparent voids within the dehydrated paroxetine hydrochloride form I structure. However, using a smaller probe radius of 0.8 Å, there appear to be small, connected voids that occupy 5.8% of the unit cell volume (Fig. 4*A*). It is also unclear where the water molecules should be placed from structural inspection of these voids.

Using the MACH protocol, the positions around the chloride ion and the protonated amine are thoroughly sampled to determine the water positions within the tight voids that meet our cutoff criteria and scoring measures (Fig. 4*B*). Overall, these results show the applicability of MACH and its cutoff criteria for inserting water in nonobvious, tight void spaces of fractional hydrates, as well as for predicting the highly directional ion–water interactions within the crystal lattice. The hemihydrate of form I is also predicted when using the DFT-D optimized host framework of form I as an input to MACH, where its optimized framework is representative of a structure that is generated from CSP (*SI Appendix*, Figs. S7 and S8). In addition, MACH also correctly predicts the hydrate, form II, of paroxetine hydrochloride (*SI Appendix*, Fig. S10).

In a final example, we test MACH on the patented, phase 2 hemihydrate of sitagliptin L-tartrate (SLTPH2) as reported by Tieger et al. (42) and Richter et al. (43) to demonstrate how our approach differentiates between chemical environments of different structural voids. The dehydrated SLTPH2 framework structure exhibits a range of physically and chemically diverse void environments (Fig. 5*A*), as usually observed in complex pharmaceutical molecules (6, 44), and was used as a framework for water insertion.

Based on chemical intuition, the vast available void spaces and numerous polar atoms from both sitagliptin and tartrate counterion suggest that this structure is very amenable to hosting water molecules. However, contrary to these general observations and consistent with the experimental structure, only one water molecule is inserted within the small hydrophilic pockets (Fig. 5*B*). This is due to the exposed carboxylate group and hydroxyl substituents of the tartrate counterion, which allow for electrostatic and highly directional hydrogen bonding interactions with water. Concurrently, the absence of water molecules within the large void spaces suggests that no hydrogen bond acceptors/donors from the host molecule can be readily accessed here. Although hydrogen bond acceptors are present within these large void spaces and channels, they reside in hydrophobic environments or are shielded by surrounding hydrophobic atoms, hence rendering them inaccessible to water molecules. For example, within the extended channels, while the nitrogen of the triazole ring is partially exposed to the open void spaces, the surrounding hydrophobic functionalities, such as the trifluoromethyl and trifluorophenyl groups, prevent facile

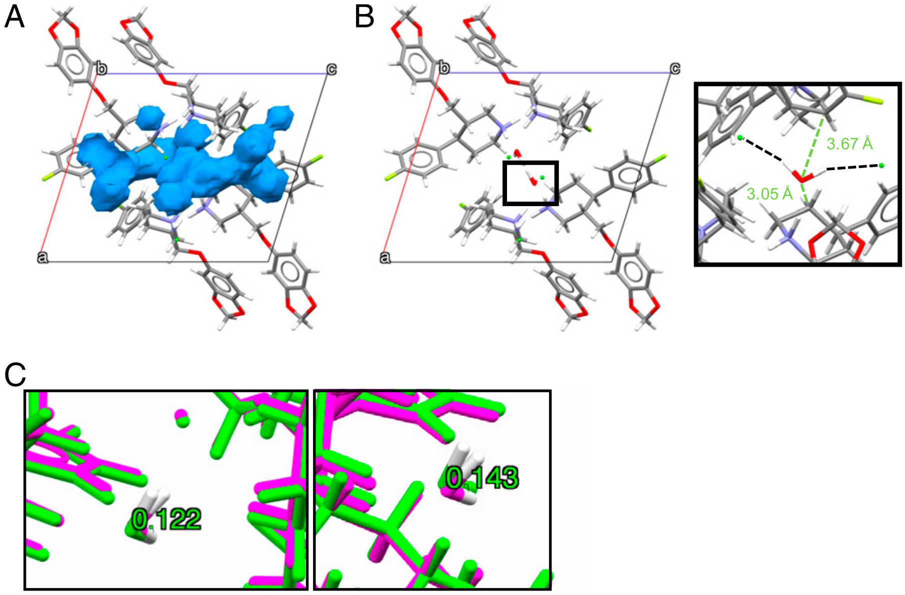

Fig. 4. Paroxetine hemihydrate form I (no water) showing (*A*) small, connected voids using a probe radius of 0.8 Å with a 0.3-Å contact spacing. (*B*) The hemihydrate structure and the chemical environment of the inserted water showing its directional interactions and distances from neighboring hydrophobic atoms. (*C*) Comparisons of water positions between the predicted and experimental structure show a difference of 0.12 and 0.14 Å in their relative water positions.

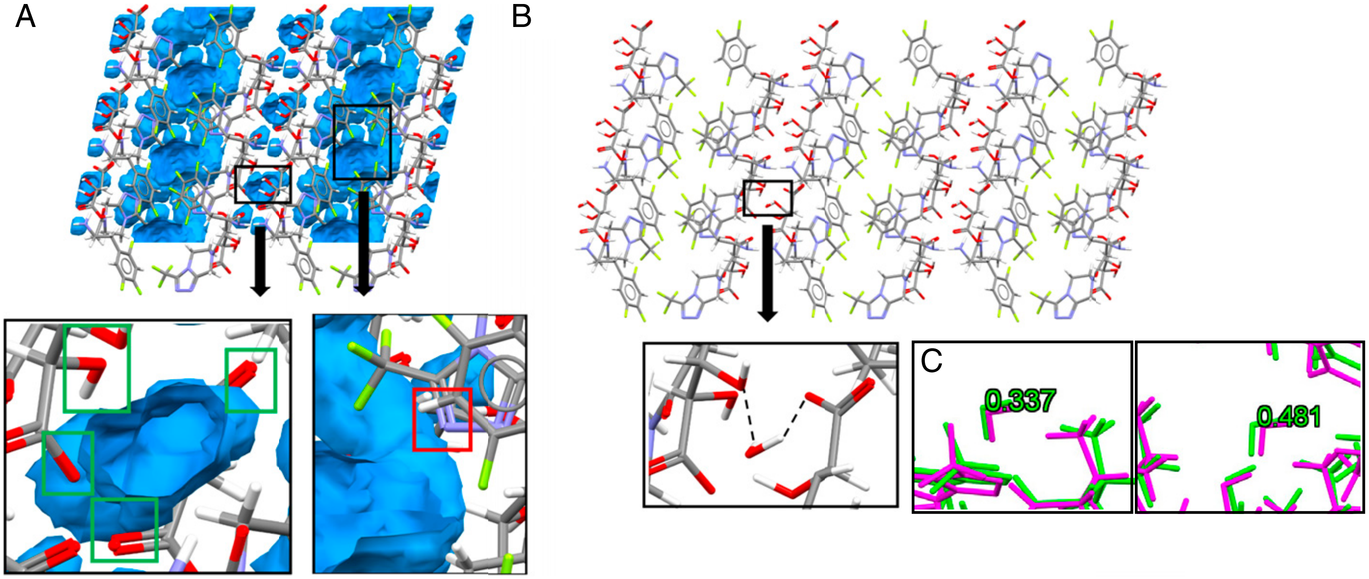

Fig. 5. (*A*) Void spaces of SLTPH2 hemihydrate and their respective chemical environments. Inaccessible hydrogen bond forming groups are shown in red, while accessible groups are shown in green. (*B*) Water positioning in the SLTPH2 hemihydrate within the tight, hydrophilic pocket voids, showing the hydrogen bonding interactions of the water molecule. (*C*) Comparison of the predicted water positions with the experimental water positions showing an average difference of 0.41 Å. However, when the experimental structure is DFT-D minimized, it matches exactly with the predicted hydrate structure (*SI Appendix*, Fig. S11), thus showing that the experimental structure and the MACH predicted hydrate structure fall into the same DFT-D minimum.

hydrogen bonding of water to the triazole nitrogen (Fig. 5*A*). MACH has also been tested on the DFT-D optimized dehydrated framework of SLTPH2, resulting in water insertions at the same positions and stoichiometries as the experimental structure (*SI Appendix*, Fig. S12).

General Applicability, Potential Limitations, and Future Work. In order to understand better the general applicability and potential limitations of MACH, we have also tested this approach on the DFT-D optimized frameworks of multiple other crystal structures (*SI Appendix*, Figs. S14–S18), including target XXIV from the sixth CCDC blind test, pharmaceutically relevant molecules, and an organic luminogen. From these examples, we note that MACH successfully predicts the correct hydrate structures, water positions, and stoichiometries in most of these examples.

In the test case for a different brucine hydrate, a nonstoichiometric 3.85 hydrate (*SI Appendix*, Fig. S17), we observed MACH generating multiple hydrate structures with varying fractional stoichiometries (4.0, 4.25, and 4.35) and water positions from different independent runs due to the random nature of the algorithm, the chemical environments of the notably vast void spaces present, and the dynamic nature of the inserted waters. Since these predicted structures differ in stoichiometry, an ensemble average of water positions was not taken for comparison. However, these structures likely represent an ensemble of plausible (fractional) water stoichiometries and configurations and can be used for further structural elucidation or refinement as the experimental water occupancies and positions are not well defined. Experimentally observed structures and stoichiometries for nonstoichiometric higher hydrates may likely constitute an ensemble of structures with varying water positions and stoichiometries (3). Yet, MACH is potentially able to distinguish these nonstoichiometric higher hydrates by generating an ensemble of different plausible hydrate structures. In its current form, MACH assumes full water occupancy and predicts the highest possible accessible water occupancy of a given framework structure. As an enhancement of the MACH workflow, grand canonical Monte Carlo deletion steps may be incorporated in the iterative loops to account for varying water levels as a function of chemical potential, as is commonly observed in nonstoichiometric hydrates. Ultimately, this can provide useful computational insights that may help in capturing the risk of observing such nonstoichiometric higher hydrates experimentally as these hydrates can bring significant challenges in pharmaceutical development (3, 6).

Lastly, we also observed that MACH was not able to insert any waters into the DFT-D minimized framework structure of paracetamol trihydrate as the DFT-D minimization results in formation of new hydrogen bonds that close off any accessible voids. Such behavior may occur more often in crystal hydrates of small, polar molecules where there are fewer overall interactions to stabilize the dehydrated framework. We note that when this approach is used with standard anhydrous CSP, it may not necessarily capture such labile hydrates where significant conformational changes or lattice collapse occur upon dehydration. However, our insertion scheme can be easily either augmented with evolutionary/genetic CSP algorithms that generate such porous crystal structures (45, 46) or even combined with a stoichiometric hydrate CSP applied to highly labile hydrates.

Due to the simplicity and speed of MACH, we believe its utilization introduces a promising direction for efficiently conducting hydrate CSP and could serve well an integral component in CSP workflows. In conjunction with standard anhydrous CSP (or even multicomponent CSP), MACH can be used to generate hydrate structures efficiently, wherein each generated structure from CSP can be used as a candidate for water insertion. This can be especially effective when augmented with newly developed CSP algorithms that efficiently generate accurate landscapes for structurally similar chemical entities (47). These results can guide further hydrate CSP or experimental studies by capturing the plausible hydrate stoichiometries, hence eliminating the need for brute force energy-based CSP searches of each plausible stoichiometry or partial stoichiometry.

In addition, these generated hydrate structures (or lack thereof) can provide unique insights to aid drug and materials design in identifying molecules or structural patterns that may be prone to hydrate formation during storage or downstream processing. In addition to its utilization in CSP, MACH also provides rich opportunities for hydrate structural elucidation, especially in cases where the electron density of waters within the crystal lattice cannot be easily resolved.

Finally, MACH provides the groundwork for the development of more advanced data/topologically driven insertion approaches where finer-grained cutoff and scoring criteria can be further refined through neural network approaches or explicit consideration of atomic polarization. These criteria can also be modified to capture solvate crystal structures or multicomponent mixed solvate/hydrate polymorphs, paving the path for potentially redefining multicomponent CSP.

Computational Methods

Water Position Optimization. Water optimization and MD runs for each iteration were done using the PINY_MD (48) package, where the General AMBER Force Field (49) intermolecular and intramolecular parameters were used with atomic point charges assigned from the restricted electrostatic potential (50) charge assignment scheme using RHF/6-31G*//MP2/6-31G* in Gaussian09 (51–55).

DFT Optimization. Periodic DFT-D optimizations and energy calculations for crystal structures were performed using the Perdew-Burke-Ernzerhof (PBE) functional (56) with the Neumann-Perrin (NP) dispersion correction (57), a plane wave basis set (520 eV, $2\ \pi \times 0.07\ \text{Å}^{-1}$ k-point grid), and default projected-augmented wave (PAW) pseudopotentials, as implemented in VASP 5.4.1 (58–60).

Hydrate Crystal Structure Database Analysis. The dataset for building the cutoff distance and hydrophobicity score histograms was obtained from a search of the Crystal Structure Database using the Conquest software (61). Structures included in this dataset had three-dimensional coordinates determined, have an R factor of ≤ 0.05, are nondisordered, include no errors, are not polymeric, are single crystal structures, and are not organometallics.

Void Analysis and Visualization. Voids within the crystal structure were visualized using CCDC Mercury (62), with contact surfaces as described by Barbour (63).

Structural Comparison. Comparisons between different entire hydrate crystal structures were performed using the COMPACK algorithm (64). The differences between the structures' water positions were then determined after the two crystal structures were overlaid.

Data, Materials, and Software Availability. All study data are included in the article and/or SI Appendix and will be made available by the authors upon request.

ACKNOWLEDGMENTS. M.E.T. acknowledges funding from the NSF, grant CHE-1955381.

Author affiliations: [a]Solid State Chemistry, Research & Development, AbbVie Inc., North Chicago, IL 60064; [b]Department of Chemistry, New York University, New York, NY 10003; [c]Courant Institute of Mathematical Sciences, New York University, New York, NY 10012; [d]New York University–East China Normal University Center for Computational Chemistry at New York University Shanghai, Shanghai 200062, China; and [e]Simons Center for Computational Physical Chemistry at New York University, New York, NY 10003

1. G. P. Stahly, Diversity in single- and multiple-component crystals. The search for and prevalence of polymorphs and cocrystals. *Cryst. Growth Des.* **7**, 1007–1026 (2007).
2. R. K. Khankari, D. J. W. Grant, Pharmaceutical hydrates. *Thermochim. Acta* **248**, 61–79 (1995).
3. R. S. Hong *et al.*, Distinct hybrid hydrates of paritaprevir: Combined experimental and computational assessment of their hydration–dehydration behavior and implications for regulatory controls. *Cryst. Growth Des.* **22**, 726–737 (2022).
4. F. Liu, D. E. Hooks, N. Li, N. A. Mara, J. A. Swift, Mechanical properties of anhydrous and hydrated uric acid crystals. *Chem. Mater.* **30**, 3798–3805 (2018).
5. G. Schneider-Rauber *et al.*, Understanding stress-induced disorder and breakage in organic crystals: Beyond crystal structure anisotropy. *Chem. Sci.* **12**, 14270–14280 (2021).
6. A. Y. Sheikh *et al.*, Implications of the conformationally flexible, macrocyclic structure of the first-generation, direct-acting anti-viral paritaprevir on its solid form complexity and chameleonic behavior. *J. Am. Chem. Soc.* **143**, 17479–17491 (2021).
7. R. D. Barnes *et al.*, Anti-depressant crystalline paroxetine hydrochloride hemihydrate. US4721723A (1988).
8. F. Zhou *et al.*, Unexpected organic hydrate luminogens in the solid state. *Nat. Commun.* **12**, 2339 (2021).
9. D. E. Braun, U. J. Griesser, Why do hydrates (solvates) form in small neutral organic molecules? Exploring the crystal form landscapes of the alkaloids brucine and strychnine. *Cryst. Growth Des.* **16**, 6405–6418 (2016).
10. A. T. Hulme, S. L. Price, Toward the prediction of organic hydrate crystal structures. *J. Chem. Theory Comput.* **3**, 1597–1608 (2007).
11. D. E. Braun, P. G. Karamertzanis, S. L. Price, Which, if any, hydrates will crystallise? Predicting hydrate formation of two dihydroxybenzoic acids. *Chem. Commun.* **47**, 5443–5445 (2011).
12. C. J. Tilbury, J. Chen, A. Mattei, S. Chen, A. Y. Sheikh, Combining theoretical and data-driven approaches to predict drug substance hydrate formation. *Cryst. Growth Des.* **18**, 57–67 (2018).
13. S. Datta, D. J. W. Grant, Crystal structures of drugs: Advances in determination, prediction and engineering. *Nat. Rev. Drug Discov.* **3**, 42–57 (2004).
14. A. J. Cruz-Cabeza, G. M. Day, W. Jones, Predicting inclusion behaviour and framework structures in organic crystals. *Chemistry* **15**, 13033–13040 (2009).
15. D. E. Braun, U. J. Griesser, Stoichiometric and non-stoichiometric hydrates of brucine. *Cryst. Growth Des.* **16**, 6111–6121 (2016).
16. A. J. Cruz Cabeza, G. M. Day, W. D. S. Motherwell, W. Jones, Prediction and observation of isostructurality induced by solvent incorporation in multicomponent crystals. *J. Am. Chem. Soc.* **128**, 14466–14467 (2006).
17. C. Zhao *et al.*, Digital navigation of energy-structure-function maps for hydrogen-bonded porous molecular crystals. *Nat. Commun.* **12**, 817 (2021).
18. Apotex wins latest round in generic Paxil litigation. *Nat. Rev. Drug Discov.* **3**, 468 (2004).
19. D.-K. Bučar, R. W. Lancaster, J. Bernstein, Disappearing polymorphs revisited. *Angew. Chem. Int. Ed. Engl.* **54**, 6972–6993 (2015).
20. D. P. McMahon *et al.*, Computational modelling of solvent effects in a prolific solvatomorphic porous organic cage. *Faraday Discuss.* **211**, 383–399 (2018).
21. B. Smit, Grand canonical Monte Carlo simulations of chain molecules: Adsorption isotherms of alkanes in zeolites. *Mol. Phys.* **85**, 153–172 (1995).
22. G. A. Ross, M. S. Bodnarchuk, J. W. Essex, Water sites, networks, and free energies with grand canonical Monte Carlo. *J. Am. Chem. Soc.* **137**, 14930–14943 (2015).
23. W. Shi, E. J. Maginn, Continuous fractional component Monte Carlo: An adaptive biasing method for open system atomistic simulations. *J. Chem. Theory Comput.* **3**, 1451–1463 (2007).
24. W. Shi, E. J. Maginn, Improvement in molecule exchange efficiency in Gibbs ensemble Monte Carlo: Development and implementation of the continuous fractional component move. *J. Comput. Chem.* **29**, 2520–2530 (2008).
25. E. Nittinger *et al.*, Placement of water molecules in protein structures: From large-scale evaluations to single-case examples. *J. Chem. Inf. Model.* **58**, 1625–1637 (2018).
26. R. Abel, T. Young, R. Farid, B. J. Berne, R. A. Friesner, Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *J. Am. Chem. Soc.* **130**, 2817–2831 (2008).
27. T. Young, R. Abel, B. Kim, B. J. Berne, R. A. Friesner, Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 808–813 (2007).
28. E. N. Baker, R. E. Hubbard, Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **44**, 97–179 (1984).
29. R. T. McGibbon *et al.*, MDTraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* **109**, 1528–1532 (2015).
30. S. Bhat, E. O. Purisima, Molecular surface generation using a variable-radius solvent probe. *Proteins* **62**, 244–261 (2006).
31. J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).
32. G. Smith, U. D. Wermuth, J. M. White, Pseudopolymorphism in brucine: Brucine-water (1/2), the third crystal hydrate of brucine. *Acta Crystallogr. C* **63**, 489–492 (2007).
33. R. S. Hong *et al.*, Insights into the polymorphic structures and enantiotropic layer-slip transition in paracetamol Form III from enhanced molecular dynamics. *Cryst. Growth Des.* **21**, 886–896 (2021).
34. E. C. Dybeck, D. P. McMahon, G. M. Day, M. R. Shirts, Exploring the multi-minima behavior of small molecule crystal polymorphs at finite temperature. *Cryst. Growth Des.* **19**, 5568–5580 (2019).
35. A. M. C. Reilly *et al.*, Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallogr. B* **72**, 439–459 (2016).
36. H. G. Brittain, Paroxetine hydrochloride: Polymorphs and solvatomorphs. *Profiles Drug Subst. Excip. Relat. Methodol.* **38**, 407–421 (2013).
37. J. A. Ibers, Paroxetine hydrochloride hemihydrate. *Acta Crystallogr. C* **55**, 432–434 (1999).
38. M. F. Pina *et al.*, An investigation into the dehydration behavior of paroxetine HCl Form I using a combination of thermal and diffraction methods: The identification and characterization of a new anhydrous form. *Cryst. Growth Des.* **14**, 3774–3782 (2014).
39. S. Sen *et al.*, Cooperative bond scission in a soft porous crystal enables discriminatory gate opening for ethylene over ethane. *J. Am. Chem. Soc.* **139**, 18313–18321 (2017).
40. M. J. Turner, J. J. McKinnon, D. Jayatilaka, M. A. Spackman, Visualisation and characterisation of voids in crystalline materials. *CrystEngComm* **13**, 1804–1813 (2011).

41. I. Brekalo *et al.*, Microporosity of a Guanidinium organodisulfonate hydrogen-bonded framework. *Angew. Chem. Int. Ed. Engl.* **59**, 1997–2002 (2020).
42. E. Tieger *et al.*, Studies on the crystal structure and arrangement of water in sitagliptin l-tartrate hydrates. *CrystEngComm* **18**, 3819–3831 (2016).
43. J. Richter *et al.*, K. modifikace 2 L-vínanu (3R)-3-amino-1-[3-(trifluormethyl)-6,8-dihydro-5H-[1,2,4,]triazolo[4,3-a]pyrazin-7-yl]-4-(2,4,5-trifluorfenyl) butan-1-onu. UV 27898, 01.13.2015 (2013).
44. K. Shubin, A. Bērziņš, S. Belyakov, Crystal structures of new ivermectin pseudopolymorphs. *Crystals* **11**, 172 (2021).
45. E. Berardo, L. Turcani, M. Miklitz, K. E. Jelfs, An evolutionary algorithm for the discovery of porous organic cages. *Chem. Sci.* **9**, 8513–8527 (2018).
46. F. Curtis *et al.*, GAtor: A first-principles genetic algorithm for molecular crystal structure prediction. *J. Chem. Theory Comput.* **14**, 2246–2264 (2018).
47. A. Mattei *et al.*, Efficient crystal structure prediction for structurally related molecules with accurate and transferable tailor-made force fields. *J. Chem. Theory Comput.***18**, 5725–5738 (2022).
48. M. E. Tuckerman, D. A. Yarne, S. O. Samuelson, A. L. Hughes, G. J. Martyna, Exploiting multiple levels of parallelism in molecular dynamics based calculations via modern techniques and software paradigms on distributed memory computers. *Comput. Phys. Commun.* **128**, 333–376 (2000).
49. J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
50. R. J. Woods, R. Chappelle, Restrained electrostatic potential atomic partial charges for condensed-phase simulations of carbohydrates. *J. Mol. Struct. Theochem.* **527**, 149–156 (2000).
51. D. R. Hartree, The wave mechanics of an atom with a non-Coulomb central field, Part I. Theory and methods. *Math. Proc. Camb. Philos. Soc.* **24**, 89–110 (1928).
52. M. J. Frisch, M. Head-Gordon, J. A. Pople, A direct MP2 gradient method. *Chem. Phys. Lett.* **166**, 275–280 (1990).
53. G. A. Petersson, M. A. Al-Laham, A complete basis set model chemistry. II. Open-shell systems and the total energies of the first-row atoms. *J. Chem. Phys.* **94**, 6081–6090 (1991).
54. J. S. Binkley, J. A. Pople, W. J. Hehre, Self-consistent molecular orbital methods. 21. Small split-valence basis sets for first-row elements. *J. Am. Chem. Soc.* **102**, 939–947 (1980).
55. M. J. Frisch et al., Gaussian 09, Revision A.02 (Gaussian, Inc., Wallingford, CT, 2016)
56. J. P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
57. M. A. Neumann, M.-A. Perrin, Energy ranking of molecular crystals using density functional theory calculations and an empirical van der Waals correction. *J. Phys. Chem. B* **109**, 15531–15541 (2005).
58. G. Kresse, J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B Condens. Matter* **54**, 11169–11186 (1996).
59. G. Kresse, J. Hafner, Ab initio molecular dynamics for liquid metals. *Phys. Rev. B Condens. Matter* **47**, 558–561 (1993).
60. G. Kresse, J. Hafner, Ab initio molecular-dynamics simulation of the liquid-metal-amorphous-semiconductor transition in germanium. *Phys. Rev. B Condens. Matter* **49**, 14251–14269 (1994).
61. I. J. Bruno *et al.*, New software for searching the Cambridge Structural Database and visualizing crystal structures. *Acta Crystallogr B* **58**, 389–397 (2002).
62. C. F. Macrae *et al.*, *Mercury 4.0*: From visualization to analysis, design and prediction. *J. Appl. Cryst.* **53**, 226–235 (2020).
63. L. J. Barbour, Crystal porosity and the burden of proof. *Chem. Commun.* **11**, 1163–1168 (2006).
64. J. A. Chisholm, S. Motherwell, COMPACK: A program for identifying crystal structure similarity using distances. *J. Appl. Cryst.* **38**, 228–231 (2005).