JCTC Journal of Chemical Theory and Computation

Article

Cite This: J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

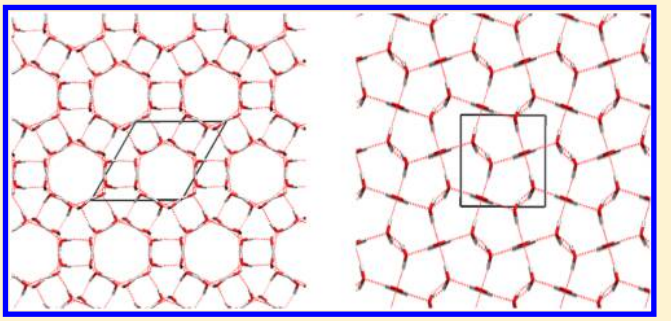
pubs.acs.org/JCTC

Crystal Structure Prediction via Basin-Hopping Global Optimization Employing Tiny Periodic Simulation Cells, with Application to Water−Ice

Christian J. Burnham and Niall J. English*

School of Chemical and Bioprocess Engineering, University College Dublin, Belfield, Dublin 4, Ireland

ABSTRACT: A crystal structure prediction algorithm for use in periodic boundary conditions with empirical rigid models is presented, which employs (i) unrestricted cutoff radii for the real-space interactions, thus allowing the treatment of even very small unit cells, and (ii) a global-optimization algorithm based on the basin-hopping method of Wales et al. (D. J. Wales and J. P. K. Doye, *J. Phys. Chem. A* **1997**, *101*, 5111). The algorithm is then applied to the TIP4P model of water (W. L. Jorgensen et al., *J. Chem. Phys.* **1983**, *79*, 926.) in order to find the lowest enthalpy water−ice crystalline structures in the pressure region 0−8000 bar, in unit cells holding in the range of 1−16 molecules, and a database of the 10 lowest enthalpy structures found at pressures 0, 4000, and 8000 bar is presented. The algorithm finds many of the ice polymorphs and, in particular, finds that the lowest energy structure at zero pressure is almost exactly tied between an ice Ic (cubic ice) and ice Ih (hexagonal ice) structure, having near-identical energies.



INTRODUCTION

Algorithmic crystal structure prediction is one of the most important problems in physical chemistry. In a nutshell, given a potential energy surface, we would like to be able to predict which crystalline structures are favored under different thermodynamic conditions.

Crystal structure prediction can be thought of as an optimization problem, where the free-energy is the quantity to be optimized, the most probable crystalline structure being that with the minimum free-energy. A commonly used approximation is to neglect temperature and attempt to optimize the enthalpy of the system, essentially searching for the most stable structures at 0 K. This essentially turns the crystal structure prediction problem into that of global optimization, for which, over the past few decades, many good algorithms have been developed.

A real crystal comprises too many molecules to realistically simulate. Fortunately, however, we can employ the well-established simulation method of periodic boundary conditions, so that we only need search for the enthalpy minimum with respect to the coordinates in one simulation cell, which it is assumed are then replicated to describe an infinite crystal. Thus, the general approach is to use optimization algorithms which search the configurational space of nuclear coordinates inside one simulation cell, trying to locate structures with the lowest enthalpy.

Recent insightful discussion/review papers by Price[1−3] provide a good overview of the problems and encouraging state-of-the-art progress in crystal structure prediction. In particular, these papers focus on polymorphism and kinetics, i.e., when a variety of different crystal structures are thermodynamically possible, which can often to be a common occurrence for organic molecules. This presents challenges for crystal structure prediction, because it is not as simple as just finding one lowest-energy structure; rather a search is required for all thermodynamically feasible structures. Furthermore, not all the low-lying energy minima will correspond to distinct polymorphs, as they may be unstable with respect to the free-energy surface. The other problem, discussed in detail by Price,[1−3] is that polymorphism requires a potential-energy surface which is accurate enough not only to correctly rank the global minimum but to do a reasonable job of describing the relative energies between minima.

Thus, there is, naturally, a great deal more to crystal structure prediction than just the search algorithm. There also needs to be some way of generating suitable potential-energy surfaces to describe the interaction of the molecules, which are accurate enough to rank the various polymorphs. As Price explains,[1−3] there are three common approaches in the literature:

(i) The use of empirical (generally pairwise additive) literature force fields. These have the advantage that the energy/force evaluations are usually very quick compared to electronic-structure methods, but they are also likely to be considerably less accurate and less transferable.

(ii) The use of density functional theory (DFT)-based methods. These have the advantage that they are potentially considerably more accurate than empirical models, but at the price of being spectacularly slower to run. It is also worth mentioning that there are many levels

Received: January 25, 2019

DOI: 10.1021/acs.jctc.9b00073
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

of electronic structure theory available, of different accuracies and speeds (and not merely DFT), so as usual, there is generally a trade-off between empirical-potential and ab initio methods.

(iii) The use of multipole-based models, such as those used by Price and co-workers,[4] which are based on the work of Stone.[5] These are expected to be considerably more accurate than relatively simple pairwise additive force fields but are much more difficult to implement.

This present study is focused on the "search-algorithm" part of the crystal structure prediction problem, and as such, we are here content to simply make use of literature empirical force fields in order to test the search algorithm. However, in future work, we plan to address the issue of parametrizing force fields from electronic-structure data and their use in crystal structure prediction.

Here, it is also worth mentioning the "Blind Tests" organized by the Cambridge Crystallographic Data Centre, in which different groups from academia and industry compete to see if their crystal structure prediction approaches can predict the experimentally observed crystallographic structure for a given molecule, where the structure is unknown at the time of prediction. A report on the sixth, and most recent, Blind Test[6] describes the considerable advances in recent years, particularly in the use of more advanced molecular-modeling approaches, going beyond the use of simple point charges, and this work also discusses advancements in going beyond the 0 K approximation. Particularly striking and encouraging is the impressive performance recently of industry participants, such as Avant-Garde Materials Simulation, which emphasizes the key industrial relevance of crystal structure prediction across a range of commercial activity, particularly in the pharmaceuticals sector.

Elking et al.[7] have performed local optimizations of molecules interacting under empirical potentials under different space-group symmetries, where the system is constrained to the desired symmetry throughout the optimization. The initial coordinates are randomized, but the symmetry constraints are enough to locate interesting minima, and they have managed to reproduce the experimental structures for a variety of molecular crystals. These optimizations are performed with their own empirical models, which include site−site multipolar interactions (but not polarizability).

Local optimization methods can sometimes produce good results, but because the number of minima increases dramatically with the system size, for even moderately sized systems, there may be too many minima to have a realistic chance of locating the lowest ones, which are the ones we are really interested in. Thus, there has been much study of more "global" optimization approaches, which attempt to use strategies to not just locate local minima but to more intelligently search the space for the lowest lying minima, and in the best case, to locate the global minimum.

Wang et al.[8] have used a particle swarm global optimization algorithm to find crystal structures under an ab initio potential energy surface. The particle swarm algorithm, invented by Kennedy and Eberhart,[9] works on a population of systems, which together form an interacting "swarm" in configurational space, where there is communication between the swarm members, such that the trajectories eventually converge on the best solution.

Oganov and co-workers have employed an evolutionary algorithm to search for low-lying and global minimum structures.[10−13] Their method works by representing the state of the system in terms of real numbers representing fractional coordinates of the atomic nuclei and six real numbers for the cell vectors, all of which can then be "mutated" (altered), before rerelaxing to a possibly new local minimum. This is done for a population of trial solutions, over successive generations, providing a diversity of solution candidates. Crucially, the genetic algorithm also includes a "heredity" stage, in which information from (two or more) parent structures combine to produce offspring structures for the next generation. In this stage, slices are taken from the unit cells of the parent structures and then combined in the child structures to produce offspring combining features of the parents. And along with heredity, there is also "death" in which only the best performing structures are allowed to survive to the next generation.

Oganov and co-workers have interfaced their algorithm with DFT code to achieve impressive results, notably finding novel structures for crystals of high-pressure phases of boron,[14] sodium,[15] superconducting oxygen,[16] carbon,[17] hydrogen-rich hydrides,[18,19] and polymorphs of calcium carbonate.[20]

Using their evolutionary algorithm, Oganov and Glass[10] have also managed to locate both the ice Ih and the ice Ic crystal structures for water, using a 12 molecule unit cell on a DFT potential energy surface.

In this work, we shall describe a crystal structure prediction algorithm which uses a (modified) basin-hopping algorithm.

The basin-hopping algorithm of Wales and Doye[21] is a global optimization approach which works by performing a walk on a *transformed* potential-energy surface, where the transformed, or "plateaued", potential energy surface is defined by the energy of the local minimum at each point, which is obtained by performing a numerical optimization starting from the coordinates on that step. The transformed surface eliminates the downhill barriers and reduces the uphill barriers between minima, and it has been shown[21−24] that walks on this surface have a much greater chance of sampling low-lying minima, including the global minimum.

Middleton and Wales et al.[25−27] have used this algorithm to investigate glassy solids, in particular, the crystal structures of binary Lennard-Jones particles, and silicon, with the latter using the three-body empirical model of Stillinger and Weber.[28] The main aim of their studies was not crystal structure prediction as such, but to map the potential energy landscape by examining the connectivity of minima and the barriers between them. However, their work did necessitate a search for low-lying minima, which were obtained via the basin-hopping approach.

These authors' work on constant pressure calculations has particular relevance to this current study. In ref 27, Middleton and Wales investigate glassy solids at constant pressure, which necessitates adding a PV term to the energy term, to give the enthalpy, and finding the derivatives with respect to cell vectors, and a similar approach will be used in this work.

Also worth noting is that Wales et al. have also recently proposed[29] a modification to their basin-hopping approach for crystal structure prediction at nonzero temperatures, in which they incorporate vibrational contributions to the free energy obtained through a harmonic analysis. Wales et al. have also explored combining the basin-hopping algorithm with a parallel-tempering-type approach, in order to greatly speed up calculations.[30]

Goedecker's "minima-hopping" algorithm[31] is an ingenious molecular-dynamics based variant on the basin-hopping type method, in which new minima are found through following short

DOI: 10.1021/acs.jctc.9b00073
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

molecular dynamics "escape trajectories" in an attempt to cross the barriers separating the minima. When a new candidate structure has been found, it is then relaxed using standard local minimum optimization algorithms, and the minimum is accepted depending on its relative energy with respect to the last accepted minimum. This algorithm dynamically adjusts the temperature of the escape trajectories, such that it favors low-energy barriers, which, according to the Bell-Evans Polyani principle, is associated with a greater likelihood of connecting with lower energy minima on the other side of the barrier.

Amsler and Goedecker[32] used the minima hopping algorithm for crystal structure prediction of silicon clathrates and a binary Lennard-Jones solid, both using empirical models, and were able to locate (record low-energy) putative global minimum structures of the binary Lennard-Jones solid, and a clathrate silicon structure, an impressive feat, given that this structure requires a 46 particle unit cell.

Advantages of the basin-hopping type approaches include that the algorithms involve few parameters, are reasonably easy to implement, and provide competitive performance with respect to other approaches for locating low-lying and global minima. However, there are also drawbacks to this type of approach. A particular problem is that if multiple walks are performed, then there can be no communication between members of the search population, so that if one member finds a route to low-lying minima, it has no way of communicating that information to the others. This can be contrasted with evolutionary type approaches, which allow successful solutions to spread throughout the population. Thus, it might seem that evolutionary approaches have the advantage, but a comparison for cluster global optimization showed that,[33] in fact, the minima-hopping algorithm outperforms an evolutionary approach, at least for the systems studied.

The TIP4P model is a common choice for benchmarking global optimization algorithms. It was chosen as a benchmark by Wales and Hodges in their study[22] of water-cluster global minima, n = 2..21, using Wales' basin-hopping algorithm. And the TIP4P phase diagram for various ices has, in a tour de force calculation, been mapped out by Abascal et al.,[34] through integration of the Clapeyron equation over coexisting ice phases.

It is in particular worth mentioning the work of Buch et al.[35] who used a molecular dynamics based method based on the inherent structures analysis of Stillinger and Weber[36] for searching for the crystal structures of TIP4P ice. The idea is to run molecular dynamics trajectories to sample the potential energy surface, from which local conjugate gradient optimizations are spawned (at regular time intervals), with the goal of finding low energy minima.

Buch et al. begin their study by fixing the simulation cell dimensions to the experimental values appropriate for different ice structures and then ran molecular dynamics trajectories at fixed volume, from which local optimizations were spawned (relaxing the coordinates and the cell vectors). Using this approach they were able to locate several different polymorphs of ice, although it should be emphasized that the approach did require some information from experiment, in the form of experimentally derived unit-cell sizes. They then tried extending the above approach by performing a more unconstrained search in which the density is fixed to a reasonable value for ice, and the cell vector ratios are stepped over, in the hope that they could find reasonable structures without the aid of experimentally derived lattice vectors (although, experimentally derived densities were still used, meaning that their method was not completely free of input from experiment). With this second approach and using unit cells in the range of 8−16 molecules, they were able to identify several known polymorphs of ice, without the need for experimentally derived lattice vectors.

We admit from the outset that this study will not do dramatically better than Buch et al. in locating the known ice polymorphs using the TIP4P model, Buch et al. having already found most of them. However, we can make improvements in some respects. First, we will be using a more advanced algorithm for locating low-lying minima, and our access to better computational resources than was presumably available to Buch et al. will allow us to perform a more thorough search. Second, we will aim to perform an entirely unconstrained search, in which no experimental data is used as input. In particular, our algorithm does not require the need for experimentally realistic densities or cell vectors, which are allowed to vary during the course of the search as the algorithm sees fit. And, last, we will be using a full Ewald sum for the electrostatic interactions, whereas Buch's study appears to employ a purely real-space sum, out to a large cutoff, which is likely to be less accurate than the complete Ewald sum approach.

An important part of Buch's study was the use of very small unit cells, containing only a few molecules, in the range n = 8..24, for which their crystal structure prediction algorithm had a realistic chance of finding low lying minima, if not the global minimum for that size range. Fortunately, the minimal unit cells of ice polymorphs, as well as many other crystal structures, are in this size range, and so the use of small unit cells seems ideal for many crystal structure prediction purposes.

There is, however, a problem with using very small unit cells with conventional simulation methods. For technical reasons we will explain in this work, the cutoff sphere is very often restricted to be small enough to fit inside the simulation cell, making it difficult to converge the nonelectrostatic part of the interactions.

As Buch et al. noted, these difficulties can be essentially solved through modifying the standard algorithm to allow the use of unrestricted cutoff spheres, allowing the interaction energy to be converged, no matter how small the unit cell.

In this work, we will describe one approach to implementing unrestricted cutoff spheres, through use of a supercell method. The central idea is to replicate the unit cell to create a larger supercell, capacious enough to hold a cutoff sphere capable of converging the energy sum.

The supercell method, in conjunction with our modified basin-hopping algorithm, will then be used to perform a systematic search for crystalline ice structures using the TIP4P (4-site transferable interaction potential) model of Jorgensen et al.,[37] a popular empirical model for water, which has often been used as a benchmark system for optimization algorithms. Our aim will be to perform a systematic search of low energy structures over the pressure range 0−8000 bar, and over cell sizes of 1−16 molecules, where we do not just look for known experimental structures but allow the algorithms to rank all the best structures over the search range, to see which structures this water model predicts.

ENERGY SUM FOR TRICLINIC CELLS

Suppose that we are simulating a system in periodic boundary conditions with N particles per unit cell interacting under a (nonelectrostatic) pair potential, where the energy of a pair of particles with separation r is given by $u(r)$, with $u(r) \to 0$ as $r \to \infty$. The true energy of the periodic system includes interactions between every particle image with every other particle image

DOI: 10.1021/acs.jctc.9b00073
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

between all replicas, but the configurational energy *per simulation cell* is commonly taken to be approximated by

$$U = \sum_{i}^{N} \sum_{j>i}^{N} u(\tilde{r}_{ji}) + U^{LR} \qquad \tilde{r}_{ji} < r_c \tag{1}$$

where $\tilde{\tilde{r}}_{ji} = |\tilde{\tilde{\boldsymbol{r}}}_{ji}|$, with $\tilde{\tilde{\boldsymbol{r}}}_{ji}$ being the so-called *minimum image* of $\boldsymbol{r}_{ji} = \boldsymbol{r}_j - \boldsymbol{r}_i$ (see Appendix I), and the sum is assumed to only count pair energy contributions for which $\tilde{r}_{ji} < r_c$, where r_c is the cutoff radius. $U^{LR}(r_c)$, to be discussed presently, is a long-range correction, which approximates interactions outside of the cutoff sphere.

Also note that, in the case of a nonzero pressure, the above is to be supplemented by a PV term, to give the *enthalpy* per simulation cell: $H = U + PV$, where P is the externally applied pressure and V is the cell volume.

The minimum image (see Appendix I) and the cutoff radius work in tandem to restrict the energy sum to include only those translationally distinct pair interactions in the periodic system which lie inside the cutoff sphere. In standard approaches, the size of the cutoff radius is restricted such that the associated cutoff sphere fits inside a unit cell. This is required to be consistent with the minimum image process, which always returns vectors inside one unit cell.

The maximum value of the cutoff radius then is determined by the largest sphere that can fit inside a triclinic cell, and it is not too difficult to show that, for a unit cell of lattice vectors $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, the radius of such a sphere must satisfy

$$r_c \leq \frac{1}{2} \min\left(\frac{1}{a^*}, \frac{1}{b^*}, \frac{1}{c^*}\right) \tag{2}$$

where $a^* = |\boldsymbol{a}^*|$ are magnitudes of the reciprocal cell vectors, $\boldsymbol{a}^* = \boldsymbol{b} \times \boldsymbol{c}/V$, with $V = \boldsymbol{a}(\boldsymbol{b} \times \boldsymbol{c})$ being the cell volume and similarly for $\boldsymbol{b}^*$ and $\boldsymbol{c}^*$. (A result which has been derived in a slightly different form by Smith.[38])

The aforementioned long-range correction approximates the contribution to the energy from interactions outside the cutoff sphere. Approximating the density of particles at $r > r_c$ by a uniform distribution, this long-range correction takes the form

$$U^{LR} = 2\pi \frac{N^2}{V} \int_{r_c}^{\infty} u(r) r^2 \, \mathrm{d}r \tag{3}$$

For sufficiently fast converging interactions, the energy sum, including the long-range correction, is often quite a good approximation to the actual energy per unit cell, i.e., when the entire periodic system is taken into account, and, in the limit, it ought to converge to the ideal result. However, as the energy sum stands, the cutoff radius is always limited by the size of the simulation cell. Thus, obtaining converged results requires that the simulation cell be large enough to contain a sufficiently large cutoff sphere.

THINKING OUTSIDE THE BOX: THE SUPERCELL METHOD

As mentioned above, the maximum radius of the cutoff sphere is limited to be no larger than what can fit inside one simulation cell. This can become a serious problem for small simulation cells, as the cutoff radius may be too small to properly converge the energy. It could also be a problem if the simulation cell is allowed to change size during the course of a simulation, as the simulation cell could shrink such that it is no longer large enough to support the cutoff sphere.

One solution is to use an Ewald summation technique, in which the dispersion interactions are partly summed in reciprocal space,[39] an approach that has been implemented by Elking et al.[7] in their crystal structure prediction work. However, in this work, we will try a supercell approach, in which the energy sum is extended to range over a supercell consisting of replicas of the original simulation cell, the idea being that this supercell can always be made large enough to support any cutoff radius.

The general supercell approach is not original to us, as both GMIN, the basin-hopping code developed by Wales et al., and the CHARMM molecular simulation software package, for instance, appear to implement a similar scheme,[40] and Buch et al.[41] also briefly make mention of a presumably similar approach in passing, but neither is this method as well-known or as well-documented as it ought to be, and so, in this section, we think it worth outlining our implementation of the supercell approach in some detail.

Let the (triclinic) supercell be such that it has cell vectors $M_a\boldsymbol{a}$, $M_b\boldsymbol{b}$, $M_c\boldsymbol{c}$, where $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$ are the cell vectors of the original simulation cell and M_a, M_b, M_c are integers and where the supercell comprises $M = M_aM_bM_c$ replicas, each of which are identical to the original simulation cell.

The coordinates of the ith particle in the mth replica are given by $\boldsymbol{r}_i^m = \boldsymbol{r}_i + \boldsymbol{R}_m$, where $\boldsymbol{R}_m = m_a\boldsymbol{a} + m_b\boldsymbol{b} + m_c\boldsymbol{c}$ is the lattice generated by integer multiples of the unit cell vectors.

Also, if $\boldsymbol{r}_{ji} = \boldsymbol{r}_j - \boldsymbol{r}_i$ is the pair vector between two particles in the simulation cell, we will write $\boldsymbol{r}_{ji}^{n,m} = \boldsymbol{r}_j^n - \boldsymbol{r}_i^m = \boldsymbol{r}_{ji} + \boldsymbol{R}_{n-m}$ for the pair vector between particles across replicas.

The standard minimum image process (see Appendix I) can be applied to the supercell. Defining $\bar{\boldsymbol{r}}$ to be the supercell minimum image of $\boldsymbol{r}$, the supercell minimum image of $\boldsymbol{r}_{ji}^{n,m}$ is given by

$$\bar{\boldsymbol{r}}_{ji}^{n,m} = \boldsymbol{r}_{ji}^{n,m} - \boldsymbol{C}^{super} \, \mathrm{nint}(\boldsymbol{C}^{super-1} \boldsymbol{r}_{ji}^{n,m}) \tag{4}$$

where $\mathrm{nint}(x)$ returns the nearest integer to x (or nearest even integer if x is exactly halfway between two consecutive integers), and

$$\boldsymbol{C}^{super} = \begin{bmatrix} M_a a_x & M_b b_x & M_c c_x \\ 0 & M_b b_y & M_c c_y \\ 0 & 0 & M_c c_z \end{bmatrix} \tag{5}$$

is the matrix of *supercell* cell vectors (here presented in upper-triangular form; c.f. the treatment for a standard unit cell in Appendix I).

The cutoff radius now has to fit inside the supercell and so must satisfy (c.f. eq 2)

$$r_c \leq \frac{1}{2} \min\left(\frac{M_a}{a^*}, \frac{M_b}{b^*}, \frac{M_c}{c^*}\right) \tag{6}$$

Suppose that the simulation cell vectors are allowed to change during the course of a simulation. The cutoff radius, as usual, is held fixed during the course of the simulation, but the size of the supercell is *not* fixed, and the M_a, M_b, M_c integers controlling its dimensions can be adjusted on the fly such that the supercell is always just large enough to hold its cutoff sphere. These values can be calculated by inverting eq 6, to obtain

$$M_a = \mathrm{int}(2r_ca^*) + 1 \tag{7}$$

DOI: 10.1021/acs.jctc.9b00073
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

and similarly for M_b and M_c, where the int(x) function (for $x > 0$) rounds down to the nearest integer.

The energy *per replica*, i.e., per simulation cell, is given by summing over every pair interaction in the supercell smaller than the cutoff radius and dividing by the number of replicas, M. Excepting the long-range correction, this energy is given by

$$U = \frac{1}{2}\frac{1}{M}\sum_{i}^{N}\sum_{j}^{N}\sum_{\mathbf{m}}^{\mathbf{M}}\sum_{\mathbf{m}'}^{\mathbf{M}} u(\bar{r}_{ji}^{\boldsymbol{m}',\boldsymbol{m}}) \tag{8}$$

where the sums over $\boldsymbol{m}$ and $\boldsymbol{m}'$ are over all M replicas in the supercell and where it is assumed that the sum excludes the nonphysical $\tilde{r}_{ii}^{\boldsymbol{m},\boldsymbol{m}} = 0$ terms describing the interaction of each particle with itself. As usual, only interactions inside the cutoff radius are counted.

In principle, the above approach should work, but it is very wasteful, as it involves counting translationally equivalent pair interactions multiple times. In fact, as we shall presently show, the above sum sums over every distinct ij interaction M times, once for each replica in the supercell.

To remove the overcounting, first consider the following identity

$$\sum_{\boldsymbol{m}}^{\mathbf{M}} u(\bar{r}_{ji}^{\boldsymbol{m},\boldsymbol{k}}) = \sum_{\boldsymbol{m}}^{\mathbf{M}} u(\bar{r}_{ji}^{\boldsymbol{m},\boldsymbol{l}}) \tag{9}$$

which holds because the LHS sums the interaction of $\boldsymbol{r}_i^k$ with every replica of $\boldsymbol{r}_j$ in the supercell and the RHS sums the interaction of $\boldsymbol{r}_i^l$ with every replica of $\boldsymbol{r}_j$ in the supercell. But, in periodic boundary conditions, $\boldsymbol{r}_i^k$ and $\boldsymbol{r}_i^l$ describe the same particle, just in different replicas, so the two sums must be equal.

Using this result allows us to write

$$\sum_{\boldsymbol{m}}^{\mathbf{M}}\sum_{\boldsymbol{m}'}^{\mathbf{M}} u(\bar{r}_{ji}^{\boldsymbol{m}',\boldsymbol{m}}) = \sum_{\boldsymbol{m}}^{\mathbf{M}}\sum_{\boldsymbol{m}'}^{\mathbf{M}} u(\bar{r}_{ji}^{\boldsymbol{m}',\mathbf{0}}) = M\sum_{\boldsymbol{m}}^{\mathbf{M}} u(\bar{r}_{ji}^{\boldsymbol{m},\mathbf{0}}) \tag{10}$$

which, when applied to eq 8, gives the energy sum as a single sum over replicas

$$U = \frac{1}{2}\sum_{i}^{N}\sum_{j}^{N}\sum_{\boldsymbol{m}}^{\mathbf{M}} u(\bar{r}_{ji}^{\boldsymbol{m},\mathbf{0}}) \tag{11}$$

We still have an overcounting problem because both ij and ji are summed over. We can try to avoid this by restricting the sum to $j > i$, which works fine for the $i \neq j$ interactions, but the $i = j$ interactions cannot be summed this way. Considering these interactions separately and using the identity $r_{ii}^{\boldsymbol{m},\mathbf{0}} = \boldsymbol{R}_{\boldsymbol{m}}$, we have

$$U = \sum_{i}^{N}\sum_{j>i}^{N}\sum_{\boldsymbol{m}}^{\mathbf{M}} u(\bar{r}_{ji}^{\boldsymbol{m},\mathbf{0}}) + \frac{N}{2}\sum_{\boldsymbol{m}\neq\mathbf{0}}^{\mathbf{M}} u(\bar{R}_{\boldsymbol{m}}) \tag{12}$$

This, then is our final form for the supercell version of the energy sum, a sum in which the cutoff radius is no longer restricted to fit inside the simulation cell but which can now be made arbitrarily large. And as usual, it is possible to supplement the above with the long-range correction of eq 3, to take into account interactions outside the cutoff sphere.

The above is very similar to the standard energy sum of eq 1, except that (i) it involves a sum over replicas, (ii) the minimum images are now applied with respect to the supercell cell vectors, and (iii) the interactions of particles with their replicas now also have to be summed over. Thus, we expect that it should not be too difficult to modify standard molecular simulation codes to implement this method.

The force-gradients $\boldsymbol{F}_k = -\nabla_k\{U\}$ for the above energy sum can be analytically found in the usual manner, but note, the second part of the sum involving interactions between particles and their images makes no contribution to the forces, as it has no dependence on the particle coordinates.

Finally, it is worth briefly comparing the above algorithm to that used by the GMIN basin-hopping code of Wales et al. GMIN also allows for a real-space sum over cells, in order that it can accommodate cutoff spheres of arbitrary size. However, GMIN's implementation makes use of a minimum image with respect to the *unit cell* (as opposed to the supercell, as done in this work). GMIN's summation method is essentially equivalent to

$$U = \sum_{i}^{N}\sum_{j>i}^{N}\sum_{\boldsymbol{m}=-\boldsymbol{K}}^{\boldsymbol{K}} u(\tilde{\boldsymbol{r}}_{ji} + \boldsymbol{R}_{\boldsymbol{m}}) + \frac{N}{2}\sum_{\boldsymbol{m}=-\boldsymbol{K}\neq\mathbf{0}}^{\boldsymbol{K}} u(R_{\boldsymbol{m}})$$

where $\tilde{\tilde{\boldsymbol{r}}}_{ji}$ is the minimum image in fractional coordinates with respect to the unit cell (see Appendix I), the sum over cells is between $-K_a \leq m_a \leq K_a$ (and similarly for m_b, m_c), where $K_a = \text{int}(r_c a^* + 0.5)$ (again, similarly for K_b, K_c), and where, as usual, it is assumed that only the interactions within the cutoff are counted.

The above should give the same answer as our approach. However, it is constrained to sum over a supercell of size $M = (2K_a + 1)(2K_b + 1)(2K_c + 1)$ unit cells, which is expected to make it more inefficient than our approach, which allows for supercells having any number of unit cells along each side. This can make a large difference in cases where the cutoff radius is only just bigger than the unit cell.

In terms of the Supporting Information, a Fortran code for calculating the radial distribution function is available on a web repository,[42] which gives an illustrative example implementation of the supercell method as described in this section.

BASIN-HOPPING

Now that we have a method for handling small unit cells, we can proceed to searching for global minima of the periodic system.

In the basin-hopping method, a simulation proceeds on a *transformed* potential energy surface, U', which is given by

$$U'(\boldsymbol{X}) = U^{\min}(\boldsymbol{X}) \tag{13}$$

where $\boldsymbol{X} = X_1, X_2, ..., X_N$ are N coordinates specifying the system, and $U^{min}(\boldsymbol{X})$ is the local minimum, starting from $\boldsymbol{X}$, under the potential energy surface U. That is, $U'(\boldsymbol{X})$ is obtained by using a local-optimization algorithm, which starts from initial coordinates $\boldsymbol{X}$, and then travels downhill until it finds a local minimum. Also, note that in the case of an externally applied pressure, a transformed *enthalpy* surface, $H'(\boldsymbol{X}) = H^{min}(\boldsymbol{X})$, will be used.

In this method, a Metropolis Monte Carlo algorithm is commonly used to walk around the transformed potential energy surface. This will result in the system visiting a succession of different local minima, with the hope that, if the simulation is run long enough, it may eventually find its way to the global minimum.

Most local-optimization algorithms require both the energy and the derivatives of the energy with respect to each coordinate; i.e., they need to be supplied with a subroutine which returns the N derivatives $\partial U(\boldsymbol{X})/\partial X_i$ for any value of $\boldsymbol{X}$.

DOI: 10.1021/acs.jctc.9b00073
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

And to make the code efficient, these derivatives usually need to be implemented analytically.

In our case, a convenient set of coordinates is given by (i) the $3N_{mol}$ molecular center of mass displacements $f_{ia}^{com}, f_{ib}^{com}, f_{ic}^{com}$, for $i = 1 \ldots N_{mol}$, (ii) the $3N_{mol}$ Euler angles ϕ_i, θ_i, ψ_i specifying the rotation of each molecule about its center of mass (see, e.g., the treatment of Allen and Tildesley[43]), and (iii) the 6 independent lattice parameters, a_x, b_x, b_y, c_x, c_y, c_z. Thus, the code needs to be furnished with analytic derivatives: $\partial H/\partial r_{ix}^{com}$, $\partial H/\partial\theta_i$, and $\partial H/\partial a_x$ (and like components). (In the coding stage, these can and should be checked against numerical derivatives.)

A problem we often encountered in using the Monte Carlo algorithm on the transformed surface is that we quite often saw situations where there would be a large number of consecutive rejections, with the Monte Carlo walk effectively getting stuck in regions where it is surrounded by multiple higher minima. However, this problem can be largely solved by resetting to the local minimum on every accept. This method, which has been used with success by Wales et al.,[44−46] has been found by White and Mayne[47] to be substantially more efficient; this is a claim which has been confirmed by our own internal tests.

Basin-hopping approaches typically use quite large Monte Carlo step sizes, and another problem we encountered was that sometimes the trial step could move a molecule unphysically close to another molecule, resulting in huge forces, which, in turn, could cause problems for the local optimization algorithm, making it hard to find the local minimum.

To solve this problem, we restricted the possible space of trial moves that the Monte Carlo walk was allowed to take, such that valid trial moves are to configurations where every intermolecular particle pair is larger than a specified minimum distance. (If a trial move does not satisfy this criterion, then a new trial move is generated, until the algorithm finds a move which does.)

Explicitly, our implementation of the basin-hopping algorithm takes the form

(i) Begin the algorithm at a local minimum (obtained through local optimization) with coordinates $\boldsymbol{X}$ and enthalpy H.
(ii) Generate a random displacement vector $\boldsymbol{\Delta X}$.
(iii) Attempt a trial move by displacing the coordinates to $\boldsymbol{X}^{trial} = \boldsymbol{X} + \boldsymbol{\Delta X}$. If the trial coordinates have two particles with an intermolecular separation closer than r^{min}, then return to (iii) and pick a new random displacement vector.
(iv) Using the local optimization algorithm, find the relaxed enthalpy of the trial coordinates: $H^{trial} = H^{min}(\boldsymbol{X}^{trial})$.
(v) Decide whether to accept or reject the move based on the standard Metropolis Monte Carlo criterion: Accept if $R < \min[1, \exp(-\beta(H^{trial} - H))]$, where R is a uniform random number between 0 and 1 and β is the inverse temperature, $\beta = 1/k_BT$, where k_B is Boltzmann's constant.
(vi) If the step is accepted, set $\boldsymbol{X} = \boldsymbol{X}^{min}(\boldsymbol{X}^{trial})$ and $H = H(\boldsymbol{X}^{min})$. Otherwise, return $\boldsymbol{X}$ to the value of the last accepted local minimum before rejection.
(vii) Return to (ii).

The $\boldsymbol{\Delta X}_i$ random displacements are chosen from distributions ρ_i. In our case, we have chosen to use Gaussian distributions (although uniform distributions between $\pm S/2$ are also common). Explicitly, we used the following: the displacements in each fractional coordinate are taken from a Gaussian distribution $\rho_f = \exp(-X_f^2/2\sigma_f^2)$, with sigma values σ_f; the displacements in the Euler angles are taken from a Gaussian distribution with sigma values σ_e; and the displacements in each of the six cell vector components are taken from a Gaussian distribution with sigma values σ_c. The values of σ_f, σ_e, σ_c are adjustable parameters, which affect the efficiency of the basin-hopping. The inverse temperature, β, is also an adjustable parameter, and should generally be chosen such that the system has a reasonable probability of sometimes accepting a move to higher energy/enthalpy, so that it can overcome barriers. But the temperature should not be made so large that the walk spends all its time exploring high energy configurations.

CASE STUDY: WATER−ICE

In this section, we will attempt to use our crystal structure prediction algorithm to search for the lowest enthalpy structures of the 4-site TIP4P empirical molecular model of water of Jorgensen et al.[37] in the range 4000−8000 bar.

Technical Details. The energy for this model is given by summing over all interatomic pairs, where the energy for an ij pair is given by

$$u(r_{ij}) = \left(\frac{A_{12}}{r_{ij}^{12}} + \frac{A_6}{r_{ij}^{6}}\right) + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{14}$$

TIP4P is a rigid water model, having a fixed geometry determined by a HOH angle, θ_{HOH}, and OH bond distance, r_{OH}. It has a single Lennard-Jones interaction site on the oxygen site of each water, charges q_H on the hydrogen atoms, and a charge of $-2q_H$ on a massless "M-site", which is placed on the molecular bisector, at a distance of r_{OM} from the oxygen nucleus toward the H nuclei. The model parameters are given in Table 1.

Table 1. Model Parameters of the TIP4P Model of Water[a]

q_H (lel)	0.52
q_M (lel)	−1.04
A_{12} (kJ/mol Å^{12})	2 510 400.00
A_6 (kJ/mol Å^6)	−2552.24
θ_{HOH} (deg)	104.52
r_{OH} (Å)	0.9572
r_{OM} (Å)	0.15

[a]Note the reported densities were calculated using the (slightly idealized) atomic masses $m_H = m_P$, $m_O = 16.0m_P$, and $m_M = 0$, where m_P is the proton mass: $m_P = 1.67262178 \times 10^{-27}$ kg.

The electrostatic interactions were handled by implementing an Ewald sum for triclinic periodic cells, where the real space part of the sum together with the Lennard-Jones interactions were summed using the supercell technique, which was found to be good enough to converge the energy per simulation cell to about 1/10th of a kJ/mol. The reciprocal space part of the sum is unaffected by the supercell method and was implemented in the standard manner.

The Ewald method was used, as opposed to particle-mesh variants, owing to the Ewald approach being faster for up to ∼1000 point charges,[42] which is very effective when using the tiny periodic cells made possible by our innovative supercell approach. Naturally, for larger systems, approaching 1000 point charges in size, particle-mesh Ewald approaches would then be recommended.

To prevent discontinuities at the cutoff, the Lennard-Jones pair interactions were multiplied by a sigmoidal cubic smooth step function, $S(x)$, which smoothly interpolates from $S(x) = 1$, $x < 0.9r_c$, to $S(x) = 0$, $x > r_c$. With this function implemented, it is

DOI: 10.1021/acs.jctc.9b00073
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

appropriate to modify the expression for U^{LR}, the long-range energy, in eq 3 such that the integral is from 0.95 r_c to ∞.

The real space part of the Ewald sum involves summing over Gaussian charge distributions of the form $G(k) = \exp(-k^2/4\zeta^2)$, and the reciprocal space part of the sum involves summing over its Fourier transform pair: $G(k) = \exp(-k^2/4\zeta^2)$, where ϵ is a freely chosen parameter, which determines the convergence of the Ewald sum. To ensure good convergence, we want $g(r)$ to be very small at the real space cutoff, and we will also choose a cutoff in reciprocal space, k_c, such that $G(k)$ is similarly small at k_c. To this end, we choose ζ and k_c such that

$$\epsilon = g(r_c) = G(k_c) \tag{15}$$

where δ is a very small number. (For this work, we have used $\delta = 10^{-7}$.)

Inverting the above gives

$$\zeta = -\sqrt{\log(\epsilon)}/r_c \tag{16}$$

and

$$k_c = 2\zeta^2 r_c \tag{17}$$

The calculations were all performed using our own in-house code, but the energies were checked against a standard molecular dynamics package, to make sure that the results are correct and reproducible.

Local optimizations were done using the FRPRMN subroutine of Numerical Recipes,[48] which is an implementation of the Fletcher-Reeves conjugate-gradient algorithm.

The triclinic simulation cell vectors are not unique, and in order to avoid long, thin unit cells, after each local optimization is performed, the minimized cell vectors were checked to see if an equivalent set with shorter cell vectors could be found (see Appendix II).

The basin-hopping implementation used steps drawn from Gaussian distributions with sigma values of $\sigma_f = 0.08$ for each fractional center of mass coordinate, $\sigma_e = 1$ radians for each Euler angle, and $\sigma_c = 0.05$ Å for each of the six independent cell vector components. Each Monte Carlo step consisted of first choosing a molecule at random and then performing a random translation and rotation for that molecule, coupled with a random displacement in all six cell vector components. A temperature of 250 K was found to be near optimal for the Metropolis Monte Carlo acceptance criterion.

Locating global minima is a hard problem, even with a good algorithm, and so to stand the best chance of actually finding the global minimum, for each case, we used multiple independent basin-hopping simulations, each starting from different random initial configurations, and using different random-number seeds for the Monte Carlo displacements. In total, a population of 24 independent basin-hopping trajectories was used for each structure, which were each run on separate cores. This acts to effectively parallelize the problem and also to increase diversity in the search, diversity being a key concern given that individual walks can be highly correlated over long times, as they can get trapped in funnels of low-lying minima.

Another advantage of using independent trajectories is that it gives some assurance that the putative global minima are likely correct, if the same lowest minimum structure is found from multiple trajectories, begun from different initial conditions.

In this work, we will be not just interested in finding the global minimum structures; we will also be concerned with finding a range of the best (i.e., lowest enthalpy) structures, where the best candidates hopefully include the true global minimum structure. To do this, we will store not just the best structure found but every minimum accepted during the course of each quasi-Monte Carlo walk, making a list of candidate minima, which can later be sorted and ranked.

As mentioned above, the initial structures were generated by random placement of the molecules (random center of mass, random orientations). In practice, we chose a cubic box, of size large enough to accommodate the required number of molecules without any molecules having to be unphysical close, and then attempted to place each molecule at random, rejecting and trying again if a site was chosen too close to one already placed.

From the perspective of finding the best crystal structure, a random placement method is likely suboptimal. For instance, Wang et al.[8] generate structures constrained to one of 230 space groups. However, random structure placement does have the advantage of (i) being easy to implement and (ii) being completely unbiased.

Hydrogen-Bond Ring Analysis and Structure-Naming Scheme. Many of the ice polymorphs are proton disordered. That is, for a given hydrogen-bond configuration, the position of the protons can adopt any number of possible configurations, while still obeying the ice rules. With this in mind, we were interested in finding the best (i.e., lowest enthalpy) structures only up to a proton ordering, such that we have chosen to only count the lowest enthalpy structure found for each different type of hydrogen-bond network.

For each minimum-energy structure, we first enumerate the list of H-bonds, where two water molecules are considered hydrogen-bonded if their OO distance is less than 3.5 Å, and the H−O---H angle is less than 30°.

We then want to know whether two structures share the same hydrogen-bond network, which is a decidedly nontrivial task, essentially equivalent to the graph-theory problem of deciding the isomorphism of two graphs. Our (partial) solution was to perform a ring analysis of each minimum energy structure, in which we enumerate the number of water tetramer, pentamer, hexamer, etc. rings present in the structure. Each ring is a hydrogen-bonded circuit of water molecules, where only those rings which cannot be further decomposed into smaller rings are listed.

Rings are found through use of a recursive algorithm, which begins on a graph node and then takes steps until it finds a path which connects back to the original node. The rings are then tested to see if they contain shorter rings by using another recursive algorithm, which attempts to enumerate all the connecting paths between every two nodes on the ring. If a connecting path is found that is shorter than the separation of the two nodes on the ring, then the ring is rejected.

The ring-decomposition problem is complicated by the use of periodic boundary conditions, for which it is no longer true that the topology of the hydrogen-bond network can be unambiguously determined from its associated graph. This is particularly a problem for small unit cells, for which rings can "interfere" with themselves across replicas. Our solution to this problem was to have the algorithm replicate small unit cells using the supercell approach, where the size of the supercell is constructed to be large enough to incorporate a cutoff sphere of a user-defined radius, such that the supercell is large enough to avoid the rings from interfering with themselves. The graph is then constructed from the H-bond network of the supercell, and

DOI: 10.1021/acs.jctc.9b00073
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

the resulting ring counts are divided by the number of unit cells in the supercell, to give the number of rings per unit cell.

In terms of providing supporting code material, our FORTRAN code for ring decomposition is available online on a web repository.[49]

Note that our ring-decomposition approach is not a full solution, because it is possible that two structures could have the same number of molecules per unit cell and the same ring decomposition and yet have different hydrogen-bonding networks. Nevertheless, this is the method that we shall use in this work.

This ring-decomposition approach allows us to develop a naming scheme, in which the structures are labeled by their number of rings and the number of molecules per unit cell. In this scheme, the structure $S12/5^8 6^4 7^8 8^8$, for example, has 12 (the number following the "S") molecules per unit cell. The numbers following the forward slash give the ring count, and this structure has eight five-membered rings, four six-membered rings, eight seven-membered rings, and eight eight-membered rings.

In this work, we will calculate ring decompositions up to 10-membered rings.

Calculations and Results. We performed a systematic search in the range 1−14 TIP4P water molecules in periodic boundary conditions and at pressures 0 bar, 4000 bar, and 8000 bar.

Our search located hundreds of structures, but Tables 2, 3, and 4 list the 10 best (lowest-enthalpy) structures found at each pressure; a selection of the best structures is shown in Figure 1, and the coordinates for all structures are supplied in Supporting Information.

Table 2. Ten Lowest Enthalpy per Molecule Structures at 0 bar, Together with Their Enthalpies and Densities

structure	enthalpy (kJ/mol)	density (g/cm^3)
$S4/6^8$ (ice Ic)	−57.104	0.9851
$S8/6^{16}$ (ice Ih)	−57.104	0.9835
$S12/5^8 7^8 8^8$ (ice III)	−56.793	1.2508
$S12/4^1 6^{20} 8^{10}$	−56.647	0.9660
$S14/5^2 6^{22} 7^2 8^6$	−56.593	0.9840
$S14/6^{28} 8^4$	−56.581	1.0046
$S12/5^8 6^2 7^4 8^6$	−56.580	1.2366
$S12/5^8 6^4 7^8 8^8$	−56.578	0.9535
$S12/4^2 5^4 6^4 7^8 8^6$	−56.577	1.2362
$S12/5^2 6^{16} 7^8$	−56.557	0.9924

Table 3. Ten Lowest Enthalpy per Molecule Structures at 4000 bar[a]

structure	enthalpy (kJ/mol)	density (g/cm^3)
$S12/5^8 7^8 8^8$ (ice III)	−51.086	1.2895
$S12/4^2 5^4 6^4 7^8 8^6$	−50.818	1.2800
$S12/5^8 6^2 7^4 8^6$	−50.802	1.2725
$S12/4^2 6^8 8^{22} 10^{30}$	−50.672	1.3586
$S12/6^{14} 8^{18} 10^{30}$ (ice II)	−50.609	1.2969
$S12/5^6 6^7 7^4 8^8 9^4$	−50.548	1.3300
$S6/7^8 8^{12}$	−50.394	1.4020
$S14/4^4 5^6 6^4 8^6 9^4 10^{24}$	−50.380	1.3727
$S14/5^7 6^1 7^9 8^{10} 9^6 10^1$	−50.362	1.3402
$S10/5^4 6^2 7^4 8^{16}$	−50.342	1.3758
...	...	...
$S10/4^{10} 8^{18}$ (ice VI)	−49.989	1.4529

[a]Also presented is an ice VI structure, $S10/4^{10}8^{18}$.

Table 4. Ten Lowest Enthalpy per Molecule Structures at 8000 bar[a]

structure	enthalpy (kJ/mol)	density (g/cm^3)
$S12/5^8 7^8 8^8$ (ice III)	−45.533	1.3220
$S12/4^2 6^8 8^{22} 10^{30}$	−45.380	1.3817
$S6/7^8 8^{12}$ (ice XII)	−45.261	1.4234
$S14/6^{10} 7^{16} 8^{20} 9^8$	−45.256	1.4434
$S12/4^2 5^4 6^4 7^8 8^6$	−45.227	1.3132
$S12/5^6 6^6 7^4 8^8 9^4$	−45.172	1.3671
$S12/5^8 6^2 7^4 8^6$	−45.169	1.3018
$S14/4^4 5^6 6^4 8^6 9^4 10^{24}$	−45.143	1.3970
$S10/5^4 6^2 7^4 8^{16}$	−45.114	1.3983
$S8/4^2 7^4 8^{20} 9^4$	−45.112	1.4383
...	...	...
$S2/6^4$ (ice VII)	−39.691	1.6157

[a]Also presented is an ice VII structure, $S2/6^4$.

Turning to Table 2, we see that, at zero pressure, the best structure is tied between an ice Ic and an ice Ih structure (structures $S4/6^8$ and $S8/6^{16}$), which have (to this precision) identical energies of −57.104 kJ/mol (per molecule).

Of interest, we did find ice lattices with even smaller unit cells. Our crystal structure prediction code did locate an ice Ic lattice using just a two-molecule simulation cell and an ice Ih lattice with just a four-molecule simulation cell, but both of these were higher in energy than could be found with the larger cells, presumably due to the better proton ordering afforded by the larger unit cells.

The third best structure we found is an ice III structure, $S12/5^8 7^8 8^8$, which is just 0.3 kJ/mol higher in energy than the best ice Ih and Ice Ic structures.

The ordering of the structures completely changes at 4000 bar, and looking at Table 3, we see that, at this higher pressure, the ice Ih and ice Ic structures are no longer even in the best 10, and an ice III structure is now the one with the lowest enthalpy. Outside of the top 10, we also located an ice VI structure, $S10/4^{10}8^{18}$.

Finally, at 8000 bar, the ice III structure is still in first place, but our crystal structure prediction algorithm also locates, in third place, an ice XII structure, $S6/7^8 8^{12}$, just behind $S12/4^2 6^8 8^{22} 10^{30}$, which does not appear to be one of the known polymorphs. Outside of the top 10, we also located an ice VII structure, $S2/6^4$, which is a high pressure phase consisting of two interpenetrating ice Ic lattices.

In the spirit of providing further supporting material to the community, the *xyz* coordinates of all structures listed in Tables 2, 3, and 4 have been made available at an online depository.[50]

Analysis. Two zero-pressure crystalline structures were found possessing the lowest energy per molecule: an ice Ic structure, with a smallest unit cell of four molecules, and an ice Ih structure, with a smallest unit cell of eight molecules, with both structures having near identical energies (to five significant figures).

That the energies of ice Ic and ice Ih are very close is not too much of a surprise, given that they are both ice structures with a practically identical tetrahedrally coordinated first-neighbor shell around each molecule. But it is surprising that the best structures we found appear to have practically identical energies. However, the structures do differ (however slightly) in their

DOI: 10.1021/acs.jctc.9b00073
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

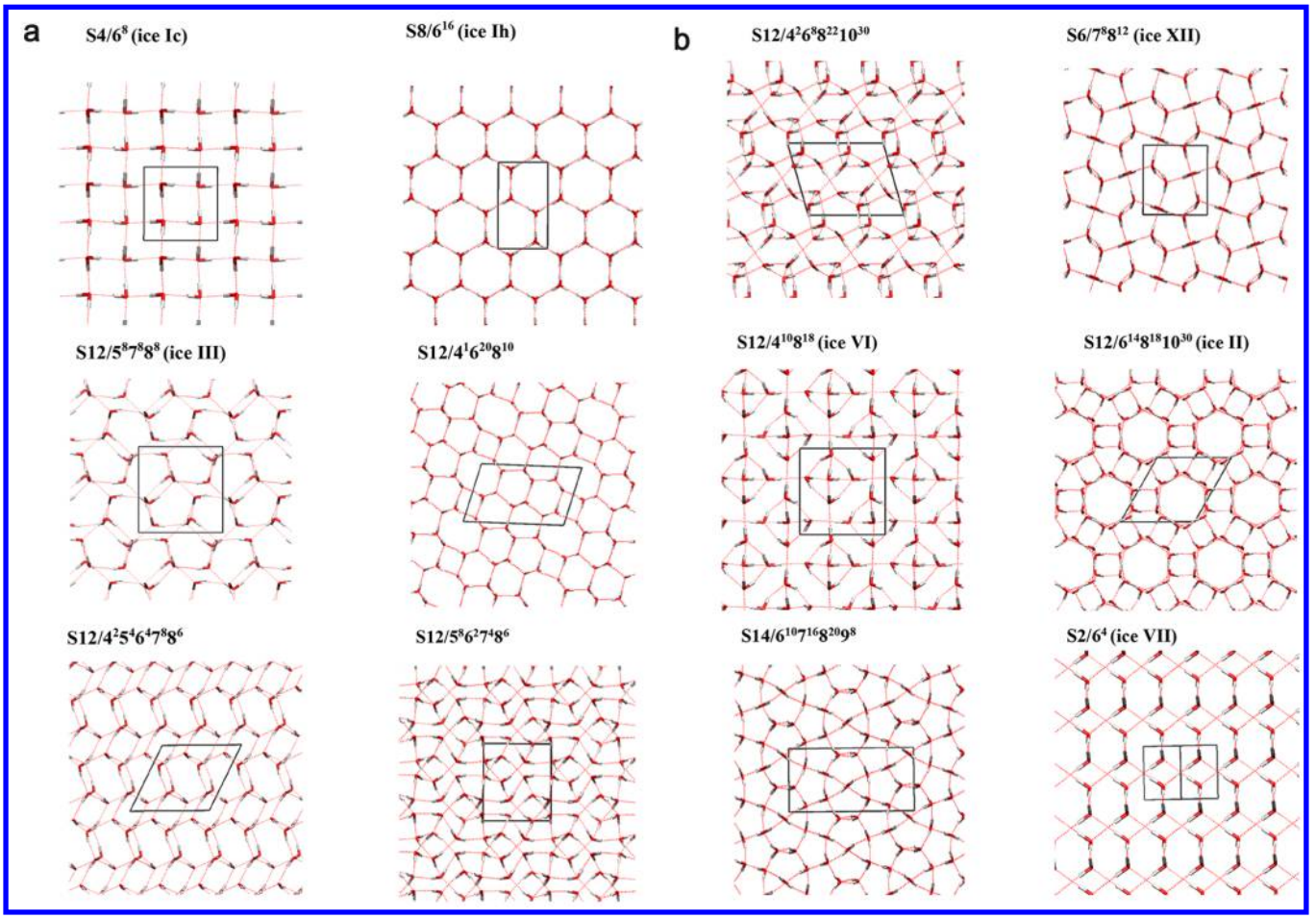


Figure 1. Crystal structures and unit cells for a selection of structures in Tables 2–4.

densities, so we think the agreement in their energies is the result of a numerical coincidence and not for any deeper physical reason. Indeed, we find that slightly varying the model charges or applying an external pressure is enough to break the agreement.

Of the known polymorphs, we did miss a couple. Our algorithm was unable to find any ice IV structures, which have a 16 molecule unit cell, and we did not get ice V, which requires a 28-molecule unit cell and is thus out of our search range. We were also unable to find the very high-pressure ice X structure, which has shared protons, and so cannot be described, ipso facto, by the TIP4P rigid water model.

As previously mentioned, Abascal et al.[34] have calculated the phase diagram for TIP4P, and in their diagram it was shown that TIP4P undergoes a transformation to ice II at ∼2000 bar, at temperatures below ∼230 K, whereas, at this pressure ice III is predicted in the temperature region 230–250 K. Thus, it was surprising to us that our crystal structure prediction algorithm predicts that our best ice II structure is *higher* in enthalpy than our best ice III structure at all the pressures considered. This may be simply because our algorithm failed to find a better ice II structure which is present on the surface but which was not visited during any of the walks. Or it may be because, at nonzero temperatures, entropic effects will change the ordering. Still, we would like to understand this better and think it worth investigating further.

CONCLUSIONS

One of the key messages of this work is that, contrary to much received wisdom, the cutoff sphere need not be restricted to fit inside the simulation cell and it is quite possible to write a molecular simulation program, or modify an existing one, such that an unrestricted cutoff radius of arbitrary length can be used.

Furthermore, in this work, we described a remarkably simple supercell summation method for unrestricted cutoff spheres that can be implemented in just about any code for molecular simulation with empirical models in periodic boundary conditions.

The supercell approach is particularly useful for crystal structure prediction, which often involves very small simulation cells. Indeed, it would have been very hard to have found many of the structures in this study using the conventional restrictions on the cutoff sphere size.

Given the inherent algorithmic complexity of the problem, with the number of possible minima increasing exponentially with system size, all global minimum optimization routines will eventually struggle for large enough systems, and indeed, we found that, past 10 or so water molecules, it became much harder to locate global minima with any certainty using a basin-hopping approach. Nevertheless, many important crystal unit cells are composed of just a few molecules, for which basin-hopping type approaches should be very useful.

This work concluded with a case study of the crystal structure prediction for the TIP4P water model, in which we found that, at 0 bar, the lowest energy structure, i.e., the putative global

DOI: 10.1021/acs.jctc.9b00073
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

minimum for TIP4P water, is a tie between an ice Ic structure with a four-molecule unit cell, having an energy of −57.104 kJ/mol per molecule, and a density of 0.9851 g/cm^3, and an ice Ih structure with an eight molecule unit cell, also having an energy of −57.104 kJ/mol per molecule, having a density of 0.9834 g/cm^3. That the TIP4P model can crystallize as ice Ih and ice Ic with almost indistinguishable energies is already known in the literature,[51] but our approach adds extra confirmation that these are indeed likely to be the true global minimum structures at 0 bar.

Searching over pressures 0, 4000, and 8000 bar, and in the range 1−16 molecules per unit cell, our algorithm successfully located seven polymorphs of ice: ice Ih, ice Ic, ice II, ice III, ice VI, ice VII, and ice XII. Given that TIP4P is such a benchmark system, it is perhaps not surprising that all of these experimentally observed polymorphs have been found previously for this model.[35] However, our present approach differs from previous searches in that (i) it uses a full Ewald sum, (ii) our search results return the best proton ordering for each structure, and (iii) we have performed a more systematic search in which we do not just look for the known ice polymorphs but build up a database of all the structures encountered, which are then ranked by their enthalpies. Of course, we cannot be absolutely sure that there are structures which were missed by our crystal structure prediction algorithm, but we did run the searches long enough such that most of the putative global minima for each unit-cell size were converged upon by more than one walk, meaning that we performed a fairly comprehensive survey.

It is a bit disturbing that several of the ice polymorphs found by our algorithm were not ranked particularly highly under the TIP4P water model in the pressure range considered, with other (experimentally unknown) structures having lower enthalpies. Thus, although the crystal structure prediction algorithm did find them, it does not predict that these particular crystal structures are favored. It may be that this is due to the inaccuracy of the model, TIP4P being quite a simple nonpolarizable water model, but our neglect of temperature effects is likely also a problem. Still, we have to start somewhere, and even locating the lowest enthalpy crystal structures at 0 K on an empirical potential energy surface is a formidable task.

In conclusion, we have presented an algorithm that shows real promise for the general crystal structure prediction problem using empirical models. Our algorithm gave very convincing results for an empirical water model, successfully finding several ice polymorphs, and we hope to apply it to other systems in the future. Also, as mentioned previously in the Introduction, in future work, we plan to address the issue of parametrizing force fields from electronic-structure data and leveraging their use in crystal structure prediction.

APPENDIX I: FRACTIONAL COORDINATES AND THE MINIMUM IMAGE

Consider a standard molecular dynamics or Monte Carlo simulation in periodic boundary conditions with a triclinic simulation cell of cell vectors $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$. The simulation cell is a unit cell of the periodic system and is replicated at positions

$$\boldsymbol{R}_{\boldsymbol{m}} = m_a\boldsymbol{a} + m_b\boldsymbol{b} + m_c\boldsymbol{c} \tag{A1}$$

where $\boldsymbol{m} = (m_a, m_b, m_c)$ is an integer triplet labelling the $\boldsymbol{m}$th replica.

If the Cartesian coordinate of the ith particle is given by $\boldsymbol{r}_i$, then the coordinates of the same particle in the $\boldsymbol{m}$th replica is given by

$$\boldsymbol{r}_i^{\boldsymbol{m}} = \boldsymbol{r}_i + \boldsymbol{R}_{\boldsymbol{m}} \tag{A2}$$

It is convenient to define fractional coordinates, $\boldsymbol{f} = (f_a, f_b, f_c)$, which label positions as fractional displacements along the unit cell vectors. The conversion between Cartesian and fractional coordinates is given by $\boldsymbol{r} = \boldsymbol{C}\boldsymbol{f}$ and $\boldsymbol{f} = \boldsymbol{C}^{-1}\boldsymbol{r}$, where

$$\mathbf{C} = \begin{bmatrix} a_x & b_x & c_x \\ 0 & b_y & c_y \\ 0 & 0 & c_z \end{bmatrix} \tag{A3}$$

is a matrix of cell vectors in upper triangular form.

The zero elements in the above are the result of choosing a form of the cell vector matrix with the rotational degrees of the cell removed. This works because the x axis can always be chosen to be parallel to the $\boldsymbol{a}$ cell vector, and the y axis is such that the $\boldsymbol{b}$ cell vector falls in the x−y plane.

Furthermore, it can be shown that if $\boldsymbol{C}$ is upper triangular, then so is $\boldsymbol{C}^{-1}$, its inverse, which is given by

$$\mathbf{C}^{-1} = \begin{bmatrix} a_x^* & a_y^* & a_z^* \\ 0 & b_y^* & b_z^* \\ 0 & 0 & c_z^* \end{bmatrix} \tag{A4}$$

where $\boldsymbol{a}^* = \boldsymbol{b}\times\boldsymbol{c}/V$ is a reciprocal lattice vector (and similarly for $\boldsymbol{b}^*$ and $\boldsymbol{c}^*$), with V being the cell volume (which, when the upper-triangular form is used, is simply given by $V = a_x b_y c_z$).

Now, let the triplet $\tilde{\boldsymbol{f}}_i = (\tilde{f}_{ia}, \tilde{f}_{ib}, \tilde{f}_{ic})$ be the *minimum image* of $\boldsymbol{f}_i$, which is defined according to

$$\tilde{\boldsymbol{f}}_i = \boldsymbol{f}_i - \mathrm{nint}(\boldsymbol{f}_i) \tag{A5}$$

where $\mathrm{nint}(x)$ is the *nearest integer* function, returning the nearest integer to x.

It is easy to see that each component of $\tilde{\boldsymbol{f}}_i$ is folded into the range $-0.5 < \tilde{f}_{ia} < 0.5$ (and similarly for $\tilde{f}_{ib}$ and $\tilde{f}_{ic}$), which encompasses one unit cell, centred at the origin, in fractional coordinates.

But, given the periodicity of the system, if $\boldsymbol{f}_i$ is the fractional coordinate of particle i, then $\tilde{\boldsymbol{f}}_i$ must be one of its images, in another (or possibly the same) replica. Thus, the minimum image process returns, out of all the periodic images, the unique image residing in an origin-centred unit cell.

The equivalent expression in Cartesians is obtained by premultiplying both sides of the above by $\boldsymbol{C}$, the matrix of cell vectors, which gives

$$\tilde{\boldsymbol{r}}_i = \boldsymbol{r}_i - \boldsymbol{C}\,\mathrm{nint}(\boldsymbol{C}^{-1}\boldsymbol{r}_i) \tag{A6}$$

where, once again, the minimum image process returns the unique image of $\boldsymbol{r}_i$ inside the origin-centered unit cell. (And it follows that the minimum image of $\boldsymbol{r}_i^{\boldsymbol{m}}$ is also equal to $\tilde{\boldsymbol{r}}_i$.)

The minimum image process can also be applied to pair vectors $\boldsymbol{r}_{ij}$. And doing so returns the unique pair vector, out of all the pair vectors in the periodic system, that lies within the origin-centered unit cell.

DOI: 10.1021/acs.jctc.9b00073
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

APPENDIX II: EQUIVALENT LATTICE VECTORS

For a given periodic system, there is more than one way to choose the lattice vectors of the unit cell; i.e., the lattice vectors are not unique. This doesn not matter so much as far as the crystal structure prediction algorithm goes, but it can result in unit cells which would be unlikely to be chosen by a crystallographer to represent a given periodic structure.

To make this more concrete, suppose the lattice vectors $\boldsymbol{L} = \{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}\}$ describe the unit cell of a periodic system. But, the set, $\boldsymbol{L}' = \{\boldsymbol{a} + \boldsymbol{b}, \boldsymbol{b}, \boldsymbol{c}\}$ has the same cell volume as $\boldsymbol{L}$, and also like $\boldsymbol{L}$, it contains just one lattice point. And since both $\boldsymbol{L}$ and $\boldsymbol{L}'$ tile the whole space, we conclude that both are equally good choices for the unit cell. Thus, we are always free to make the substitution $\boldsymbol{a}^{new} = \boldsymbol{a} \pm \boldsymbol{b}$ (or $\boldsymbol{a}^{new} = \boldsymbol{a} \pm \boldsymbol{c}$) and still have a valid cell (and similarly for the other two lattice vectors).

We can freely choose between any of the equivalent lattice vectors, and in this work, we have chosen to use the set which minimizes the lengths of each lattice vector. We do this by making the substitutions: $\boldsymbol{a}^{new} = \boldsymbol{a} \pm \boldsymbol{b}$, if $|\boldsymbol{a} \pm \boldsymbol{b}| < |\boldsymbol{a}|$ (and similarly for $\boldsymbol{a}^{new} = \boldsymbol{a} \pm \boldsymbol{c}$, and for the other two vectors). This process is then iterated over all lattice vectors until no more reductions can be found. (It is here worth mentioning that Oganov and Glass have employed[11] a closed-form non-iterative approach to solve this problem, which may be slightly more elegant and which presumably gives the same results.)

One technical issue with changing the lattice vectors is that our code uses an upper-diagonal representation of the lattice vectors throughout, and it is possible that the new lattice vectors aren not in this form; i.e., the new $\boldsymbol{a}$ axis is not parallel to $\boldsymbol{x}$ or the new $\boldsymbol{b}$ axis is no longer in the $x-y$ plane. This was remedied by following the generation of new lattice vectors by a rotation to a new $x-y-z$ axis, in which both conditions are satisfied.

AUTHOR INFORMATION

Corresponding Author

*(N.J.E.) E-mail: niall.english@ucd.ie.

ORCID

Christian J. Burnham: 0000-0001-5574-4339

Niall J. English: 0000-0002-8460-3540

Funding

Both authors thank Enterprise Ireland under Grant CF 2017 0777-P

Notes

The authors declare no competing financial interest.

As mentioned in the main text, we have included a database of the *xyz* coordinates of all structures listed in Tables [2], [3], and [4], in an online web repository.[50] Also available in online repositories are FORTRAN codes for ring decomposition[49] and for calculating the radial distribution function with the supercell approach,[42] alongside associated "README" files, instructions/advice for usage, etc.

REFERENCES

(1) Price, S. L. From Crystal Structure Prediction to Polymorph Prediction: Interpreting the Crystal Energy Landscape. *Phys. Chem. Chem. Phys.* **2008**, *10*, 1996.

(2) Price, S. L. Control and Prediction of the Organic Solid State: A Challenge to Theory and Experiment. *Proc. R. Soc. A* **2018**, *474*, 20180351.

(3) Price, S. L. Is Zeroth Order Crystal Structure Prediction (Csp_0) Coming to Maturity? What Should We Aim for in an Ideal Crystal Structure Prediction Code? *Faraday Discuss.* **2018**, *211*, 9.

(4) Price, S. L.; Leslie, M.; Welch, G. W. A.; Habgood, M.; Price, L. S.; Karamertzanis, G.; Day, G. M. Modelling Organic Crystal Structures Using Distributed Multipole and Polarizability-Based Model Intermolecular Potentials. *Phys. Chem. Chem. Phys.* **2010**, *12*, 8478.

(5) Stone, A. *The Theory of Intermolecular Forces*; Oxford University Press: 2016.

(6) Reilly, A. M.; et al. Report on the Sixth Blind Test of Organic Crystal Structure Prediction Methods. *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2016**, *B72*, 439.

(7) Elking, D. M.; Fusti-Molnar, L.; Nichols, A. *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2016**, *72*, 488.

(8) Wang, Y.; Lv, J.; Zhu, L.; Ma, Y. Calypso: A Method for Crystal Structure Prediction. *Comput. Phys. Commun.* **2012**, *183*, 2063.

(9) Kennedy, J.; Eberhart, R. C. A Discrete Binary Version of the Particle Swarm Algorithm. *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*; IEEE: 1997; p 4104.

(10) Oganov, A. R.; Glass, C. W. Crystal Structure Prediction Using Ab Initio Evolutionary Techniques: Principles and Applications. *J. Chem. Phys.* **2006**, *124*, 244704.

(11) Oganov, A. R.; Glass, C. W. Evolutionary Crystal Structure Prediction as a Tool in Materials Design. *J. Phys.: Condens. Matter* **2008**, *20*, 064210.

(12) Lyakhov, A. O.; Oganov, A. R.; Valle, M. Crystal Structure Prediction Using Evolutionary Approach. In *Modern Methods of Crystal Structure Prediction*; Oganov, A. R., Ed.; Wiley-VCH Verlag GMbH & Co. KGaA: 2010.

(13) Oganov, A. R.; Ma, Y.; Lyakhov, A. O.; Valle, M.; Gatti, C. Evolutionary Crystal Structure Prediction as a Method for the Discovery of Minerals and Materials. *Rev. Mineral. Geochem.* **2010**, *71*, 271.

(14) Oganov, A. R.; Chen, J.; Gatti, C.; Ma, Y.; Ma, Y.; Glass, C. W.; Liu, Z.; Yu, T.; Kurakevych, O. O.; Solozhenko, V. L. Ionic High-Pressure Form of Elemental Boron. *Nature* **2009**, *457*, 863−867.

(15) Ma, Y.; Eremets, M.; Oganov, A. R.; Xie, Y.; Trojan, I.; Medvedev, S.; Lyakhov, A. O.; Valle, M.; Prakapenka, V. Transparent Dense Sodium. *Nature* **2009**, *458*, 182−5.

(16) Ma, Y.; Oganov, A. R.; Glass, C. W. Structure of the Metallic Zeta-Phase of Oygen and Isosymmetric Nature of the Epsilon-Eta Phase Transition: Ab Initio Simulations. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2007**, *76*, 064101.

(17) Gao, G.; Oganov, A. R.; Bergara, A.; Martinez-Canales, M.; Cui, T.; Iitaka, T.; Ma, Y.; Zou, G. Superconducting High Pressure Phase of Germane. *Phys. Rev. Lett.* **2008**, *101*, 107002.

(18) Zurek, E.; Hoffmann, R.; Ashcroft, N. W.; Oganov, A. R.; Lyakhov, A. O. A Little Bit of Lithium Does a Lot for Hydrogen. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 17640.

(19) Martinez-Canales, M.; Oganov, A. R.; Ma, Y.; Yan, Y.; Lyakhov, A. O.; Bergara, A. Novel Structures and Superconductivity of Silane under Pressure. *Phys. Rev. Lett.* **2009**, *102*, 087005.

(20) Oganov, A. R.; Glass, C. W.; Ono, S. High-Pressure Phases of Caco3: Crystal Structure Prediction and Experiment. *Earth Planet. Sci. Lett.* **2006**, *241*, 95.

(21) Wales, D. J.; Doye, J. P. K. 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.

(22) Wales, D. J.; Hodges, M. P. Global Minima of Water Clusters (H2o)N, N ≤ 21, Described by an Empirical Potential. *Chem. Phys. Lett.* **1998**, *286*, 65.

(23) Wales, D. J. Energy Landscapes and Structure Prediction Using Basin-Hopping. In *Modern Methods of Crystal Structure Prediction*; Oganov, A. R., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: 2006; p 29.

(24) Wales, D. J. *Energy Landscapes*; Cambridge University Press: 2013.

(25) Middleton, T. F.; Wales, D. J. Energy Landscapes of Some Model Glass Fomers. *Phys. Rev. B* **2001**, *64*, 024205.

DOI: 10.1021/acs.jctc.9b00073
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

(26) Middleton, T. F.; Hernández-Rojas, J.; Mortenson, P. N.; Wales, D. J. Crystals of Binary Lennard-Jones Solids. *Phys. Rev. B* **2001**, *64*, 184201.
(27) Middleton, T. F.; Wales, D. J. Energy Landscapes of Model Glasses. Ii. Results for Constant Pressure. *J. Chem. Phys.* **2003**, *118*, 4583.
(28) Stillinger, F. H.; Weber, T. A. Computer Simulation of Local Order in Condensed Phases of Silicon. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1985**, *31*, 5262.
(29) Sutherland-Cash, K. H.; Wales, D. J.; Chakrabarti, D. Free Energy Basin-Hopping. *Chem. Phys. Lett.* **2015**, *625*, 1.
(30) Strodel, B.; Lee, J. W. L.; Whittleston, C. S.; Wales, D. J. Transmembrane Structures for Alzheimer's A. *J. Am. Chem. Soc.* **2010**, *132*, 13300.
(31) Goedecker, S. Minima Hopping: An Efficient Search Method for the Global Minimum of the Potential Energy Surface of Complex Molecular Systems. *J. Chem. Phys.* **2004**, *120*, 9911.
(32) Amsler, M.; Goedecker, S. Crystal Structure Prediction Using the Minima Hopping Method. *J. Chem. Phys.* **2010**, *133*, 224104.
(33) Schönborn, S. E.; Goedecker, S.; Roy, S.; Oganov, A. R. The Performance of Minima Hopping and Evolutionary Algorithms for Cluster Structure Prediction. *J. Chem. Phys.* **2009**, *130*, 144108.
(34) Abascal, J. L. F.; Vega, C. A General Purpose Model for the Condensed Phases of Water: Tip4p/2005. *J. Chem. Phys.* **2005**, *123*, 234505.
(35) Buch, V.; Martoñak, R.; Parrinello, M. A New Molecular-Dynamics Based Approach for Molecular Crystal Structure Search. *J. Chem. Phys.* **2005**, *123*, 051108.
(36) Stillinger, F. H.; Weber, T. A. Packing Structures and Transitions in Liquids and Solids. *Science* **1984**, *225*, 983.
(37) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926.
(38) Smith, W. The Minimum Image Convention in Non-Cubic Md Cells. *CCP5 Newsletter*, 1989; Vol. *30*.
(39) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577.
(40) Brooks, B. R.; et al. Charmm: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30*, 1545.
(41) Buch, V.; Sandler, P.; Sadlej, J. Simulations of H2o Solid, Liquid, and Clusters, with an Emphasis on Ferroelectric Ordering Transition in Hexagonal Ice. *J. Phys. Chem. B* **1998**, *102*, 8641.
(42) Burnham, C. J. Super_Rdf Code Repository. https://github.com/christianjburnham/super_rdf.
(43) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford Science Publications: 1987.
(44) Wales, D. J.; Scheraga, H. A. Global Optimization of Clusters, Crystals, and Biomolecules. *Science* **1999**, *285*, 1368.
(45) Wales, D. J.; Hodges, M. P. Global Minima of Water Clusters Ž H 2 O/ N, N F 21, Described by an Empirical Potential. *Chem. Phys. Lett.* **1998**, *286*, 65.
(46) Doye, J. P. K.; Wales, D. J. Global Minima for Transition Metal Clusters Described by Sutton−Chen Potentials. *New J. Chem.* **1998**, *22*, 733.
(47) White, R. P.; Mayne, H. R. An Investigation of Two Approaches to Basin Hopping Minimization for Atomic and Molecular Clusters. *Chem. Phys. Lett.* **1998**, *289*, 463.
(48) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P.; Metcalf, M. *Numerical Recipes in Fortran 77: The Art of Scientific Computing*, 2nd ed.; Cambridge University Press: 1992.
(49) Burnham, C. J. Findrings Code Repository. https://github.com/christianjburnham/findrings/tree/master.
(50) Burnham, C. J. Ice Structure Database. https://github.com/christianjburnham/ice_structures.
(51) Zaragoza, A.; Conde, M. M.; Espinosa, J. R.; Valeriani, C.; Vega, C.; Sanz, E. Competition between Ices Ih and Ic in Homogeneous Water Freezing. *J. Chem. Phys.* **2015**, *143*, 134504.

DOI: 10.1021/acs.jctc.9b00073
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX