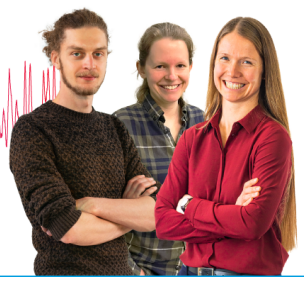
RESEARCH ARTICLE | FEBRUARY 08 2001

On the use of graph invariants for efficiently generating hydrogen bond topologies and predicting physical properties of water clusters and ice

Jer-Lai Kuo; James V. Coe; Sherwin J. Singer; Yehuda B. Band; Lars Ojamäe

J. Chem. Phys. 114, 2527–2540 (2001)
https://doi.org/10.1063/1.1336804



ARTICLES

On the use of graph invariants for efficiently generating hydrogen bond topologies and predicting physical properties of water clusters and ice

Jer-Lai Kuo, James V. Coe, and Sherwin J. Singer
Department of Chemistry, Ohio State University, Columbus, Ohio 43214

Yehuda B. Band
Departments of Chemistry and Physics, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Lars Ojamäe
Physical Chemistry, Arrhenius Laboratory, Stockholm University, S-106 91 Stockholm, Sweden

(Received 28 September 2000; accepted 7 November 2000)

Water clusters and some phases of ice are characterized by many isomers with similar oxygen positions, but which differ in direction of hydrogen bonds. A relationship between physical properties, like energy or magnitude of the dipole moment, and hydrogen bond arrangements has long been conjectured. The topology of the hydrogen bond network can be summarized by oriented graphs. Since scalar physical properties like the energy are invariant to symmetry operations, graphical invariants are the proper features of the hydrogen bond network which can be used to discover the correlation with physical properties. We demonstrate how graph invariants are generated and illustrate some of their formal properties. It is shown that invariants can be used to change the enumeration of symmetry-distinct hydrogen bond topologies, nominally a task whose computational cost scales like N^2, where N is the number of configurations, into an $N \ln N$ process. The utility of graph invariants is confirmed by considering two water clusters, the $(H_2O)_6$ cage and $(H_2O)_{20}$ dodecahedron, which, respectively, possess 27 and 30 026 symmetry-distinct hydrogen bond topologies associated with roughly the same oxygen atom arrangements. Physical properties of these clusters are successfully fit to a handful of graph invariants. Using a small number of isomers as a training set, the energy of other isomers of the $(H_2O)_{20}$ dodecahedron can even be estimated well enough to locate phase transitions. Some preliminary results for unit cells of ice-Ih are given to illustrate the application of our results to periodic systems. [DOI: 10.1063/1.1336804]

INTRODUCTION

Hydrogen bonds are long-lived structures in ice and cold water clusters, and, to a lesser extent, in liquid water. Thanks to the strong tendency of water to form hydrogen bonds in a tetrahedral arrangement, our understanding of the three-dimensional structure and dynamics of aqueous systems has long been couched in terms of a reduced description based on hydrogen bond connectivity.[1] Many phenomena illustrate how hydrogen bond topology serves as a critical structural descriptor: The zero-point entropy of ice-Ih, ''ordinary'' ice at atmospheric pressure, is thought to be a manifestation of frozen-in complete disorder among the possible hydrogen bond topologies of the ice lattice[2–5] (but see below). Transport properties of ice are understood in terms of defects in H-bond connectivity.[6] The language used to name the structural isomers of water clusters—''cage,''[7] for example, or ''cube''[8]—reflects the correspondence between H-bond topology and water cluster structure.[9,10] The strength of hydrogen bonds within the ice-Ih lattice has been conjectured to fall into strong or weak classes based on the local hydrogen bond topology,[11–15] although this distinction has recently been questioned.[16,17]

To determine the lowest energy hydrogen bonding arrangement or construct a statistical average requires, in general, a sampling of the hydrogen bond topologies. The number of available topologies grows exponentially with system size. In 1935 Pauling[2] estimated that the number of H-bond arrangements available to N water molecules in the ice-Ih crystal structure is $(3/2)^N$, an estimate that has been shown to be accurate within a few percent.[4,5] The number of structures in a simulation cell of even 100 water molecules is in the range of $\sim 10^{18}$, so it might appear that enumeration or sampling of topologies for a system of this size will fall exclusively in the province of either Monte Carlo[17,18] or more sophisticated variants of Monte Carlo.[19,20]

The purpose of this work is to provide analytic techniques for complicated hydrogen bonded systems, such as the myriad arrangements of ice-Ih, which would seem to

only be tractable by numerical simulations. The problem of H-bond structures in ice-Ih also happens to be a particularly fascinating one. The experimental residual entropy of ice-Ih at 0 K is close to Pauling's estimate,[2,3] leading to the conclusion that H-bond topological disorder becomes frozen into the ice-Ih lattice as temperature is lowered from the freezing point to absolute zero. Within the last decade, experiments have detected a phase transition in KOH-doped ice-Ih, weakly dependent on KOH concentration and tending toward 72 K in the limit of vanishing impurity concentration.[21–25] This seems to indicate that the KOH impurity catalyzes the rearrangement of hydrogen bonds, and that ordinary ice, if equilibrium could be attained, would undergo a proton ordering transition at 72 K. Neutron scattering[26–29] and thermal depolarization experiments[30,31] on KOH-doped ice-Ih suggest that the proton-ordered form of ice-Ih, known as ice-XI, is an orthorhombic ferroelectric crystal, although the interpretation of these experiments has been debated.[32–34] Most common potential models for water do not predict this structure as the ground state, which has forced a reappraisal of such models.[17] More recently, Antarctic ice cores have been investigated with Raman spectroscopy.[35] These samples are believed to have been equilibrated at temperatures controlled by their depth beneath the surface for tens of thousands of years. The Raman spectra indicate a phase transition at 237 K that has similar characteristics as, but lies far above, the phase transition in KOH-doped ice-Ih. Studies on Greenland ices failed to find similar evidence for a phase transition.[36] Therefore, the current understanding of ordinary ice is ripe for further experimental and theoretical insight.

Hydrogen bonds are directional, so hydrogen bond topologies are in one-to-one correspondence with oriented simple graphs, that is, collections of vertices connected by at most one directed edge. The direction of a hydrogen bond points from hydrogen donor to hydrogen acceptor. Enumeration of H-bond topologies becomes an exercise in graph theory, to list all possible graphs consistent with the so-called ''ice rules.''[37] These rules allow at most two edges emanating from a vertex because H_2O molecules can donate at most two hydrogens, and at most two edges incident upon a vertex since at most two hydrogen bonds can be accepted at the lone pairs. The ice rules are modified in an obvious way to accommodate the presence of species like OH^- and H^+.[9]

In anything but the smallest water clusters or unit cells of ice crystals, one is faced with huge numbers of configurations. In the more rigid ice clusters—the $(H_2O)_6$ cage, the $(H_2O)_8$ cube, the $(H_2O)_{20}$ dodecahedron, to name a few—and in ice-Ih, local minima of the potential energy surface are, to a good approximation, in one-to-one correspondence with oriented graphs. How does one find the ground state or construct a thermal average with anything but numerical sampling techniques? Perhaps the most useful result of the current work is a strategy by which the physical properties, including but not necessarily confined to the energy, of a large number of H-bonded structures can be summarized and predicted in terms a handful of parameters, each associated with a special linear combination of variables defined for oriented graphs called a *graph invariant*. The procedure by which physical properties are correlated with, and predicted by graph invariants is automatic and does not rely on special physical insight, although graph invariants will certainly facilitate deeper physical interpretation. The number of graph invariants used to fit and predict physical properties can be systematically enlarged, leading to a hierarchy of approximations. To determine the energy or find the ground state among a vast number of structures, we can therefore use graph invariants to avoid explicit and often costly calculations for all but a small training set of structures from which the parameters can be extracted. The procedure is tested in this work for two water clusters, the $(H_2O)_6$ cage and $(H_2O)_{20}$ dodecahedron, for which 27 and 30 026 hydrogen bonding arrangements are possible. We also illustrate the utility of graph invariants in the enumeration of the huge number of H-bond structures possible for a unit cell of ice-Ih, but leave further analysis of this challenging problem to a future work.

Computationally, explicit enumeration of allowed graphs for hydrogen-bonded systems is relatively straightforward, but eliminating structures that are related to each other by a symmetry operation is not. Nominally, elimination of symmetry-related graphs is an $\sim N^2/2$ process, where N is the number of graphs, because it involves comparison of pairs of graphs. Moreover, each of the $\sim N^2/2$ comparisons can be rather expensive when the symmetry group is large. In this work we show how use of graph invariants[38,39] can change the scaling of computational effort with N from N^2 to $N \ln N$. The desirability of eliminating symmetry-related structures is illustrated by our calculation for a 48-member hexagonal unit cell of an ice-Ih lattice: there are 2 404 144 962 graphs possible in total, but only 8 360 361 symmetry-distinct structures. With such large numbers of configurations, $N \ln N$ scaling is an enormous improvement. Elimination of symmetry-related configurations has been attempted previously ''by hand'' for small ice unit cells.[40] The results of this effort are in apparent conflict with attempts to group symmetry-related structures based on energetic criteria.[17,41] Therefore, efficient and reliable computational methods of generating hydrogen bond graphs are needed.

Counting *total* numbers of allowed graphs on an infinite periodic lattice (or regular finite structures) can be addressed by series expansion methods.[4,5] Even though series expansions have only been used, to our knowledge, for counting total numbers of hydrogen bond arrangements in regular structures, these methods can presumably be extended to calculate certain averages with generating function techniques.[42] However, the graphs themselves, not just total numbers or averages, are needed to construct explicit molecular structures for further study, as would be needed as input for an *ab initio* or empirical potential calculation.

The notion of a graph invariant is introduced in Sec. I. Graph invariants often have a simple physical interpretation, as we illustrate in this section. The use of graph invariants to change the enumeration of symmetry-distinct graphs from an N^2 to $N \ln N$ process is described in Sec. II. This section may be omitted by readers who are not interested in the numerical problems associated with enumeration of H-bonded structures. When the energy and other physical properties can be

correlated with topological properties of the hydrogen bonded network, as first noted for water clusters by Radhakrishnan and Herndon[43] and also by McDonald, Ojamäe, and Singer,[9] then the physical properties, themselves being invariant with respect to symmetry operations, should be expressible in terms of the values of graph invariants.

The link between physical properties and H-bond topology suggests the following strategy for global optimization, put forward in Sec. III, which is useful even when the number of such isomers is too large to enumerate or analyze: A Monte Carlo procedure can be used to generate a training set of structures.[17,18] The training set can be used to establish a relationship between physical properties and the values of graph invariants, either by least-squares fitting or more sophisticated methods. The tentative relationship between physical properties and topology can be refined by further sampling. In this way, low-entropy structures, which may tend to be overlooked in Monte Carlo methods, can be identified by selective enumeration. Invariants can concisely parametrize the energy of a large number of H-bond arrangements, thereby permitting the calculation of physical properties that involve the entire ensemble of H-bond topologies, like phase transitions, with a minimum of input. We illustrate this capability with a model calculation of a cluster phase transition for the $(H_2O)_{20}$ dodecahedron in Sec. III.

In Sec. IV we show how invariants may be applied to periodic systems by giving results for hexagonal unit cells of ice-Ih. We conclude with a brief discussion in Sec. V.

I. GRAPH INVARIANTS

Hydrogen bonds between water molecules are directional. One water ''donates'' its covalently bonded hydrogen to the bond while the second water ''accepts'' that hydrogen in the vicinity of the lone pair electronic cloud. The hydrogen bonded network within a water cluster or ice crystal is summarized by graphs in which the vertices represent oxygen atoms and directed bonds (or edges) connecting vertices represent hydrogen bonds. By convention, the directed bonds point from donor to acceptor. The so-called ''ice rules'' stipulate that each neutral water vertex has a maximum of four neighbors. At each vertex there are a maximum of two

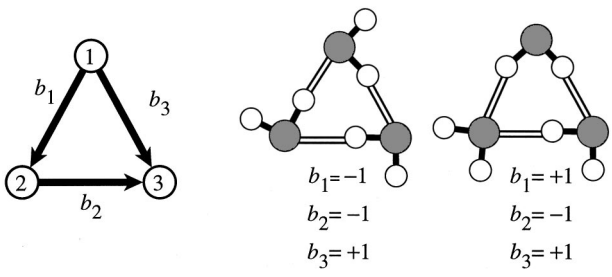


FIG. 1. A simple example of an oriented graph, which might represent the configuration of a $(H_2O)_3$ cluster, is shown on the left. The direction shown on the edges indicates the orientation of the edges if all the bond variables b_r were taken equal to $+1$, canonical orientations chosen arbitrarily for each bond. Two different hydrogen bond topologies for $(H_2O)_3$ are shown on the right, along with the value of the bond variables, as referenced to the canonical orientations of the graph on the left.

TABLE I. Group of vertex permutations for the triangle graph shown in Fig. 1, and the induced group of signed permutations on the bonds of the triangle graph. For the vertex permutations, we denote the permutation taking vertices 1, 2, and 3 to *i*, *j*, and *k* as (*ijk*). Often this permutation is indicated by $\binom{1\ 2\ 3}{i\ j\ k}$. It is also common to indicate permutations in terms of independent cycles, in terms of which, for example, the σ_v permutation would be written as (1)(23). Our notation for signed bond permutations follows that for vertices.

g_α	Vertex permutation	Signed bond permutation
E	(123)	(b_1,b_2,b_3)
C_3^{-1}	(231)	$(b_2,-b_3,-b_1)$
C_3	(312)	$(-b_3,b_1,-b_2)$
σ_v	(132)	$(b_3,-b_2,b_1)$
$\sigma_{v'}$	(213)	$(-b_1,b_3,b_2)$
$\sigma_{v''}$	(321)	$(-b_2,-b_1,-b_3)$

outgoing edges, the covalently bonded hydrogens, and two incoming edges, where the two lone pairs can accept an H-bond. The formalism can easily accommodate excess protons or hydroxide by allowing the appropriate local deviations from these rules.[9]

The hydrogen bond topology of a finite or infinite system of water molecules is summarized by a collection of variables b_{ij}, one for each vertex pair *ij*, whose value is

$$b_{ij}=\begin{cases} 1 & \text{if water } i \text{ donates to water } j \\ -1 & \text{if water } j \text{ donates to water } i \\ 0 & \text{if there is no hydrogen bond between } i \text{ and } j. \end{cases} \tag{1}$$

The order of the indices on b_{ij} is meaningful, so to describe the same physical configuration, $b_{ji}=-b_{ij}$, it is convenient to let a single index *r* replace the dual index for bond pairs *ij*. It is arbitrary whether *r* stands for *ij* or *ji*, but some canonical ordering of the bond pairs must be specified. For example, in the triangle graph of Fig. 1 the canonical direction chosen for b_1 is from vertex 1 to vertex 2. With this convention, $b_1=1$ indicates that molecule 1 donates to molecule 2, while $b_1=-1$ specifies that molecule 2 accepts a hydrogen bond from molecule 1. In Fig. 1 we give some examples of hydrogen bond arrangements in a triangular cluster of three water molecules, and the value of bond variables that specify these physical configurations *relative to the canonical orientations* given to the left in Fig. 1. In general, once the canonical orientation for each directed edge is specified, the physical meaning of b_r is clear: If *r* stands for *ij*, then $b_r=1$ indicates that water *i* donates to water *j*, while if *r* stands for *ji*, then $b_r=-1$ must be used to indicate the same arrangement.

Symmetry properties of a cluster or crystal are manifested by a group of permutation operations mapping the set of vertices onto themselves. The list of *adjacent* vertices (vertices connected by a bond, irrespective of the bond's direction) is preserved by each of the symmetry operations. It is important to note that the symmetry group pertains to the oxygen atom ''scaffold,'' and is not dependent on particular orientations of hydrogen bonds. The group of symmetry operations for the vertices of the triangular graph shown in Fig. 1 are the 3! vertex permutations given in the second

column of Table I. The group members are labeled by corresponding real-space point group operations in Table I to provide a useful mnemonic, but fundamentally our formalism deals with connectivity, or topology, and not geometry.

We stress that the utility of the graph formalism is not at all dependent on the physical isomers having the full symmetry of the vertex and bond permutation groups. Consider two H-bond topologies that are symmetry related. They each correspond to distorted local minima, and possibly have low or no symmetry. As long as the topologies are symmetry related, then the energy and any other physical property of the distorted local minima corresponding to those topologies will be identical. Therefore, parametrization of energy of distorted structures in terms of invariants based on higher symmetry of the oxygen scaffold is appropriate. While the values of bond lengths and angles are irrelevant in the graph theoretical formalism, it is not true that physical geometry is entirely irrelevant. Firstly, adjacent bonds reflect physical proximity. Secondly, the symmetry group may be chosen to reflect expected physical geometry, which may lower the symmetry from that based solely on connectivity.

As a hypothetical example, consider a ring of five vertices which represent a geometry in which four vertices are coplanar and the fifth lies far outside the plane. The symmetry group determined by vertex adjacency is D_{5h}, but one may choose a smaller symmetry group, perhaps C_s, to reflect the nonplanarity of the ring. However, use of the higher symmetry group may still be appropriate. Consider geometry optimizations initiated from a planar starting structure. If two planar initial structures which are equivalent within D_{5h} symmetry optimize to the same distorted structure, then the energy of the isomers will be described by the more compact invariants of the D_{5h} symmetry group. The C_s symmetry group will generate a larger set of initial structures and (if they exist on the potential energy surface) will enumerate more physical isomers. In this situation, the symmetry group should be chosen to suit the goals of the calculation and the properties of the potential energy surface.

The symmetry group on vertices induces a group of signed permutations on the bonds. We signify the image of group operation α on bond b_r with the notation $g_\alpha(b_r)$. For example, the C_3 operation on the triangle graph in Fig. 1 brings vertices 1 and 2 to vertices 3 and 1, respectively (also see Table I). Therefore, bond b_1 is moved to the location of bond b_3. By the convention chosen in Fig. 1, $b_1=+1$ indicates an H-bond from 1 to 2, and $b_3=+1$ indicates an H-bond from 1 to 3. However, the C_3 operation takes a bond from vertex 1 to vertex 2 to another bond from vertex 3 to vertex 1, not from 1 to 3. Therefore, the image of b_1 under the C_3 operation is $-b_3$, not b_3, and $g_{C_3}(b_1)=-b_3$.

A. Generation of graph invariants

We seek functions of the bond variables $\mathbf{b}=\{b_1,b_2,\ldots,b_r,\ldots\}$ that, like physical properties, are unchanged by application of symmetry operations. These special functions, the *invariants*, have the property

$$g_\alpha(I(\mathbf{b}))=I(g_\alpha(b_1),g_\alpha(b_2),\ldots)=I(\mathbf{b}). \tag{2}$$

The functions $I(\mathbf{b})$ transform as the totally symmetric representation of the induced group on bonds. According to standard group theory, invariants can be constructed using a projection operator, which takes a particularly simple form because the characters for the totally symmetric representation are all unity. For example, the application of a projection operator to a single bond variable takes the form

$$I_r=C_r\sum_{\alpha=1}^{G} g_\alpha(b_r), \tag{3}$$

where C_r is a normalization constant chosen for convenience. Now consider the application of a group operation g_β to I_r:

$$g_\beta(I_r)=C_r\sum_{\alpha=1}^{G} g_\beta(g_\alpha(b_r)). \tag{4}$$

According to the requirements for group operations the composition $g_\beta(g_\alpha(\cdots))$, $\alpha=1,2,\ldots,G$, generates each of group operations once and only once. [If $g_\beta(g_\alpha(\cdots))$ and $g_\beta(g_{\alpha'}(\cdots))$] gave the same resultant operation, then g_β would fail to have a unique inverse.] Therefore $g_\beta(I_r)=I_r$.

By precisely the same reasoning, $I_{rs},I_{rst},\ldots$, as defined below, are also invariants:

$$I_{rs}=C_{rs}\sum_{\alpha=1}^{G} g_\alpha(b_rb_s), \tag{5}$$

$$I_{rst}=C_{rst}\sum_{\alpha=1}^{G} g_\alpha(b_rb_sb_t),\ldots. \tag{6}$$

We refer to I_r as a first-order invariant, I_{rs} as a second order invariant, and so on. It is obvious that the order of the indices of bond generators does not change the invariants (i.e., $I_{rs}=I_{sr}$).

The number of first order invariants is generally less than the number of bonds for several reasons. Many invariants may turn out to be identically zero. The necessary and sufficient condition for any invariant $I_{rst\ldots}$ to vanish identically is the existence of a group operation that takes the product of bond variables $b_rb_sb_t\ldots$ into minus itself, as shown in the Appendix. In many cases, it is possible to find an operation which takes a single bond b_r into $-b_r$ and therefore many first-order bond invariants vanish identically. The number of first-order invariants may be less than the number of bonds for another reason. If $g_\alpha(b_r)=\pm b_s$, application of the projection operator in Eq. (3) to b_r and b_s yields the same result within an overall constant, in which case I_r and I_s are equivalent. The number of different first-order invariants depends on the number of independent orbits of the bond group, which in turn depends on the structure of the induced group on bonds. Similar considerations apply to the higher-order invariants. When all bonds are filled, and therefore all bond variables $b_r=\pm 1$, $I_{rrstu\ldots}=I_{stu\ldots}$. There can be no linearly independent invariants of order greater than the number of bonds when all bonds are filled. Given a group of symmetry operations expressed as permutations of vertices, we have found symbolic algebra

programs convenient for computationally generating the induced group on bonds, and then a table of independent invariants.

Let us use the triangle graph of Fig. 1 as an example to illustrate the properties of invariants mentioned above. Directly from Table I, it can be seen that all the first-order invariants are identically zero. As for second-order invariants, after convenient normalization we have that

$$I_{11}=I_{22}=I_{33}=b_1^2+b_2^2+b_3^2, \tag{7}$$

$$I_{12}=I_{23}=I_{13}=b_1b_2-b_1b_3-b_2b_3. \tag{8}$$

Since the image of $b_1b_2b_3$ is $+b_1b_2b_3$ under the first three group operations in Table I and $-b_1b_2b_3$ under the last three operations, $I_{123}=0$. Other third-order invariants, like

$$I_{122}=-(b_2+b_3)b_1^2+(b_1+b_3)b_2^2-(b_1-b_2)b_3^2, \tag{9}$$

may take nonzero values if one of the bonds is empty.

Products of invariants are also invariant. Therefore, products of two first-order invariants can be expanded as a linear combination of second-order invariants, products of first and second are a linear combination of third-order invariants, and so on:

$$I_rI_s=\sum_{tu} c_{tu}^{r,s}I_{tu}, \tag{10}$$

$$I_rI_{st}=\sum_{uvw} c_{uvw}^{r,st}I_{uvw},\ldots. \tag{11}$$

Standard group representation theory governs the resolution of invariant products.

B. Physical interpretation of graph invariants

The invariants can be interpreted in terms of physical quantities, often in terms of several such quantities. For example, second-order invariants of the triangle graph (Fig. 1) with identical indices like I_{11} in Eq. (7) simply count the number of nonempty bonds within a group theoretical orbit. If we would write the dipole moment of water molecules in terms of bond dipoles of magnitude μ_{bond}, the total dipole moment due to ring dipoles would be

$$\mu_x=\tfrac{1}{2}\mu_{\text{bond}}(-b_1+2b_2+b_3), \tag{12}$$

$$\mu_y=-\frac{\sqrt{3}}{2}\mu_{\text{bond}}(b_1+b_3). \tag{13}$$

The squared magnitude of the dipole moment is

$$\begin{aligned}\mu_x^2+\mu_y^2&=\mu_{\text{bond}}^2[\tfrac{1}{4}(b_1^2+4b_2^2+b_3^2-2b_1b_3-4b_1b_2\\&\quad+4b_2b_3)+\tfrac{3}{4}(b_1^2+b_3^2+2b_1b_3)]\\&=\mu_{\text{bond}}^2[(b_1^2+b_2^2+b_3^3)-(b_1b_2-b_1b_3-b_2b_3)]\\&=\mu_{\text{bond}}^2[I_{11}-I_{12}],\end{aligned}$$

a quantity invariant to symmetry operations. It therefore comes as no surprise that second-order invariants nicely capture the dependence of the squared magnitude of the dipole moment on hydrogen bonded configurations.

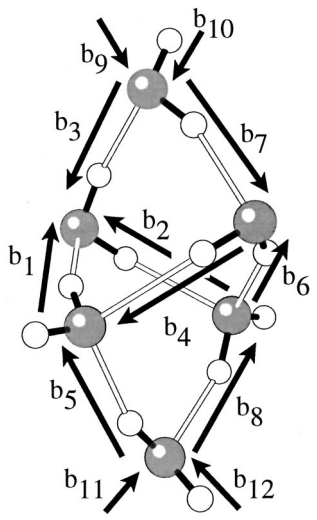


FIG. 2. The cage structure of $(H_2O)_6$. One of the 27 possible symmetry-distinct hydrogen bonding arrangements for the cage structure is shown. The arrows and bond labels indicate the directions of the bonds when the bond variables are equal to $+1$.

Second-order invariants can always be understood in terms of counting nonempty bonds and the magnitude of the dipole moment. However, these are not the only possible interpretations of the second-order invariants. For example, the invariants $I_{3,7}=I_{5,8}$ of the $(H_2O)_6$ cage structure shown in Fig. 2 are connected to the number of single-donor/single-acceptor molecules, a feature which strongly affects the total energy of the cage.[10] The precise relation is

TABLE II. First- and second-order invariants for the $(H_2O)_6$ cage structure shown in Fig. 2. The invariants are calculated using a permutation symmetry group on the vertices isomorphic to the D_{2d} point group. In the text we refer to the invariants by any of the bond products that generate the invariant by application of a projection operator. For example, $I_3=I_5=\cdots$ is at the top of the list of first-order invariants, while $I_{1,1}=I_{2,2}\cdots$ heads the list of second-order invariants and $I_{3,9}=I_{7,9}=\cdots$ is at the bottom of the list.

First-order invariants
$b_3+b_5+b_7+b_8$
$b_9+b_0+b_{11}+b_{12}$
Second-order invariants
$b_1^2+b_2^2+b_4^2+b_6^2$
$b_3^2+b_5^2+b_7^2+b_8^2$
$b_9^2+b_{10}^2+b_{11}^2+b_{12}^2$
$b_2b_4-b_1b_6$
$b_3b_7+b_5b_8$
$b_9b_{10}+b_{11}b_{12}$
$b_1b_2-b_1b_4+b_2b_6-b_4b_6$
$b_3b_5+b_5b_7+b_3b_8+b_7b_8$
$b_5b_9+b_8b_{10}+b_3b_{11}+b_7b_{12}$
$b_8b_9+b_5b_{10}+b_7b_{11}+b_3b_{12}$
$b_9b_{11}+b_{10}b_{11}+b_9b_{12}+b_{10}b_{12}$
$b_1b_3+b_2b_3-b_1b_5+b_4b_5-b_4b_7+b_6b_7-b_2b_8-b_6b_8$
$b_1b_9-b_4b_9+b_2b_{10}+b_6b_{10}-b_1b_{11}-b_2b_{11}+b_4b_{12}-b_6b_{12}$
$b_2b_9+b_6b_9+b_1b_{10}-b_4b_{10}+b_4b_{11}-b_6b_{11}-b_1b_{12}-b_2b_{12}$
$b_3b_4+b_2b_5-b_3b_6+b_5b_6-b_1b_7-b_2b_7+b_1b_8-b_4b_8$
$b_3b_9+b_7b_9+b_3b_{10}+b_7b_{10}+b_5b_{11}+b_8b_{11}+b_5b_{12}+b_8b_{12}$

$$(\text{number of single-donor/single-acceptors}) = 1 - \tfrac{1}{2} I_{3,7}. \quad (14)$$

Incidentally, the $(H_2O)_6$ cage is an example of a structure which has nonzero first-order invariants. The complete list of invariants for the $(H_2O)_6$ cage is given in Table II. The first-order invariant $I_3 = b_3 + b_5 + b_7 + b_8$ counts the number of H-bonds among bonds 3, 5, 7, and 8 pointing away from the center of the cage, while the invariant $I_9 = b_9 + b_{10} + b_{11} + b_{12}$ has a similar interpretation for bonds 9, 10, 11, and 12. (See Fig. 2 for the definition of the bond variables. Later we explain the rationale for including four bonds incident on the apical vertices, even though the apical waters participate in only two hydrogen bonds and have only one dangling hydrogen.)

While we have stressed that invariants possess physical interpretations, their power lies in the fact that their generation and use can be automated. Trends can be deduced without reliance on physical insight or tedious trial-and-error process. Better yet, the use of invariants can guide the discovery of physical interpretation, or show that several interpretations are equivalent. We will show that certain critical physical properties are captured by surprisingly low order invariants because the constraints of the ice rules link properties in nonobvious ways.

II. GRAPH INVARIANTS AS A TOOL FOR ENUMERATING H-BOND TOPOLOGIES

Graph invariants can be used to change the computation of all symmetry-distinct hydrogen bond topologies for a cluster or crystal unit cell from a process scaling as N^2, where N is the total number of topologies, to an $N \ln N$ process. Readers who are not interested in the computational aspects of enumerating hydrogen bond arrangements can proceed to the next section.

Enumerating all possible hydrogen bond arrangements for a cluster is accomplished by considering each of the n_{bond} hydrogen bonds in turn. After completing the assignment of i_{bond} out of the total n_{bond} hydrogen bonds, a list of all hydrogen bond topologies for the first i_{bond} bonds consistent with the ice rules is in hand. Each entry of this list is a sequence of 1s, -1s, or 0s of length i_{bond} for each configuration. The ± 1s stand for the two orientations of an H-bond, and the 0 means the bond is left empty. Usually there are just the two orientations signified by ± 1 but in certain cases, discussed below, we will see that also allowing a bond to be empty is useful. Instead of using bond variables, the H-bond configuration is sometimes parametrized by the arrangement of edges incident at each vertex[44,41] (e.g., the six bond arrangements at a vertex of a four-coordinate water). The advantage furnished by invariants does not depend on how the H-bond configurations are parametrized.

The addition of bond $(i_{\text{bond}}+1)$ means making a new list of H-bond topologies by attempting to fit in all orientations of the $(i_{\text{bond}}+1)$th H-bond with each member of the old list. When an orientation of the new bond is allowed by the ice rules, that configuration of $i_{\text{bond}}+1$ bonds is added to the new list. Eventually a new list containing all possible orientations of H-bonds among $i_{\text{bond}}+1$ bonds is completed, and the process of adding bond $i_{\text{bond}}+2$ is started. After some or all of bonds are added, the list of configurations is checked to eliminate configurations that are related by a symmetry operation. Symmetry-related configurations can be safely eliminated from the list when only a fraction of the bonds have been added, for if two partial (that is, containing i_{bond} out of the full n_{bond} bonds) configurations are symmetry-related, then the configurations grown from these two by filling in the rest of the bonds will also be symmetry-related. Eliminating symmetry-related graphs is the difficult step in enumerating H-bond topologies. It involves comparing pairs of graphs and is therefore nominally an N^2 process, where N is the number of graphs before symmetry comparison. Symmetry comparison involves applying each group operation in turn and checking if a match is found. Two graphs are determined to be symmetry-distinct only after all group operations are applied. Therefore the pairwise comparisons are a lengthy process, more so when the symmetry group is large and symmetry reduction offers the most benefit. By comparison, checking the ice rules after adding an H-bond, a process that scales as N, is far less costly. The cost of the symmetry comparisons can be reduced by only performing this check after certain of the H-bonds have been added. It is best to order the bonds into cycles or orbits, groups of bonds that are related by a symmetry operation, and check for symmetry only after all bonds of an orbit are added. However, even with this strategy, the computational cost of checking for symmetry grows rapidly with system size and would quickly render most large calculations infeasible if there was not some way around the N^2 scaling.

In this section we demonstrate that, using graph invariants, the nominally N^2 process of symmetry comparison can be turned into a calculation that scales like $N \ln N$. The strategy is one of divide and conquer. The set of N graphs is sorted into groups of target size n, such that graphs in different groups must be symmetry-distinct. Within each group a conventional (n^2-scaling) symmetry comparison method is used. Since there are N/n such groups, the total work associated with conventional symmetry comparison scales like $N/n \times n^2 = Nn$, linear with the number of graphs. If sorting can be made to scale more efficiently than N^2, the overall efficiency of symmetry comparisons will be improved.

A. Sorting strategy

The value of all invariants of two symmetry-related graphs must be identical. Therefore, if we divide N graphs into P groups, each one with a different value of a particular invariant, symmetry comparisons need only be done within each group. The work of calculating the value of that particular invariant for all N graphs and sorting them into P groups scales like N, while the work of symmetry comparisons now scales like $(N/P)^2$ within each group. After separation into P groups, the work of symmetry comparison scales like $P(N/P)^2 = N^2/P$, an improvement by a factor P in efficiency.

Instead of using just one invariant to sort the graphs into smaller subsets, imagine using m different invariants to sort the graphs m times. For simplicity, we will assume that the graphs are sorted into P equal piles according to each new

invariant. After employing the first invariant the graphs are divided into P groups. Then the next invariant divides each of those groups into P subgroups, making a total of P^2 groups. Finally, after m such sorts, the graphs are partitioned into P^m groups of size

$$n = \frac{N}{P^m}. \tag{15}$$

The goal is to reduce the groups to a target size n which is small enough to employ a conventional symmetry comparison method. From Eq. (15), the number of sorts required to reach a target size n is

$$m = \frac{\ln(N/n)}{\ln P}. \tag{16}$$

Each time the graphs are sorted, an invariant is calculated for each of the graphs, and that graph is either labeled or moved to another location in memory or on disk. The cost of each sort is proportional to N, the total number of graphs. The computational cost of m sorts is proportional to

$$mN = N\frac{\ln(N/n)}{\ln P}. \tag{17}$$

Associating a coefficient A with the computational cost of sorting the graphs into groups of target size n, and another coefficient B associated with the conventional symmetry comparison within each group, the total work of eliminating symmetry-related graphs scales like

$$AN\frac{\ln(N/n)}{\ln P} + B\frac{N}{n}n^2 = \frac{A}{\ln P}N\ln N + \left(Bn - A\frac{\ln n}{\ln P}\right)N. \tag{18}$$

The total work contains components that scale as $N\ln N$ and as N, far more efficient than conventional N^2-scaling symmetry comparison.

We arrived at $N\ln N$ scaling by assuming that each sort breaks the graphs into P groups of equal size. Actual computations are more complicated. The number of groups into which the graphs are sorted is the number of different values an invariant takes over the set of graphs. This varies from invariant to invariant, so invariants differ in their ability to resolve the graphs into smaller groups. Moreover, in each sort the graphs are, in general, broken into groups of unequal size. Therefore, the parameter P used in Eqs. (15)–(18) must be taken as an average or effective number of groups. The basic idea is confirmed, and evidence presented below shows $N\ln N$ scaling in realistic calculations.

B. Performation of sorting algorithm for realistic calculations

We have previously enumerated the 30 026 symmetry distinct H-bond topologies of the $(H_2O)_{20}$ dodecahedral clathrate.[9] After the addition of each hydrogen bond to the structure, we determined which graphs were related to others by a symmetry operation and eliminated all but one representative from each set of symmetry related structures. If symmetry-related graphs can be eliminated from smaller substructures, then we avoid the work of adding and testing redundant new structures built from the symmetry-related

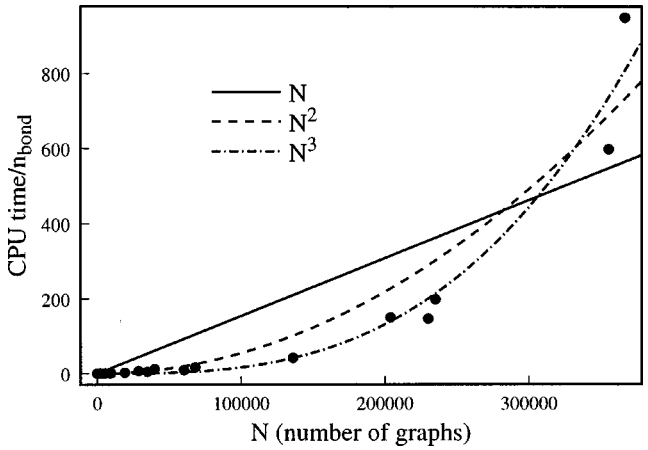


FIG. 3. Enumeration of all symmetry-distinct H-bond topologies for a $(H_2O)_{20}$ dodecahedral clathrate was performed by considering a sequence of structures containing fewer bonds than the full dodecahedron. Additional H-bonds were added to the structures after all symmetry-related duplicates were eliminated. This process furnishes data on the computational cost of eliminating symmetry-related structures as a function of the number of graphs. This data shown is for the calculation as performed in Ref. 9, without the use of the sorting method introduced in this work. The computational cost per graph edge is plotted as a function of the number of graphs N before symmetry comparisons were made. Least-squares fits of CPU time to N, N^2 and N^3 clearly show that the computational cost scales as N^2 or worse without the sorting method.

structures. The process of adding occupied bonds one at a time furnishes a data set on which we can compare different methods of graph enumeration. Our previous calculation[9] did use a crude version of what we now call invariants to speed up symmetry comparisons, but ultimately it was an N^2-scaling method because the graphs were not sorted as described in the previous section. The CPU time needed to eliminate symmetry-related duplicate structures is plotted against the number of graphs after each of the 30 bond additions in Fig. 3. The CPU time clearly increases faster than linear with the number of graphs, and the old method even appears to scale with N more steeply than N^2.

The same $(H_2O)_{20}$ clathrate calculation was repeated using graph invariants to sort the graphs until each group contained no more that $n=500$ structures. The CPU time for sorting is plotted against the number of graphs in the top panel of Fig. 4. It is difficult to distinguish whether the computational cost is actually scaling as N or $N\ln N$ (actually a happy state of affairs!), and clearly the calculation no longer scales as N^2 or worse. The CPU time for symmetry comparisons within the group of size $n\leqslant 500$ is shown in the bottom panel of Fig. 4, clearly compatible with linear scaling.

At the time we first published the enumeration of the 30 026 H-bond topologies of the $(H_2O)_{20}$ dodecahedron,[9] it was a challenging calculation. With the help of graphical invariants, this calculation is rather quick. The most challenging example we have tackled to date is the enumeration of the 8 360 361 symmetry-distinct H-bond topologies of a 48-water hexagonal unit cell from the ice-Ih lattice. There are 2 404 144 962 H-bond topologies of this system, including symmetry-related configurations, indicating how valuable the symmetry reduction can be.

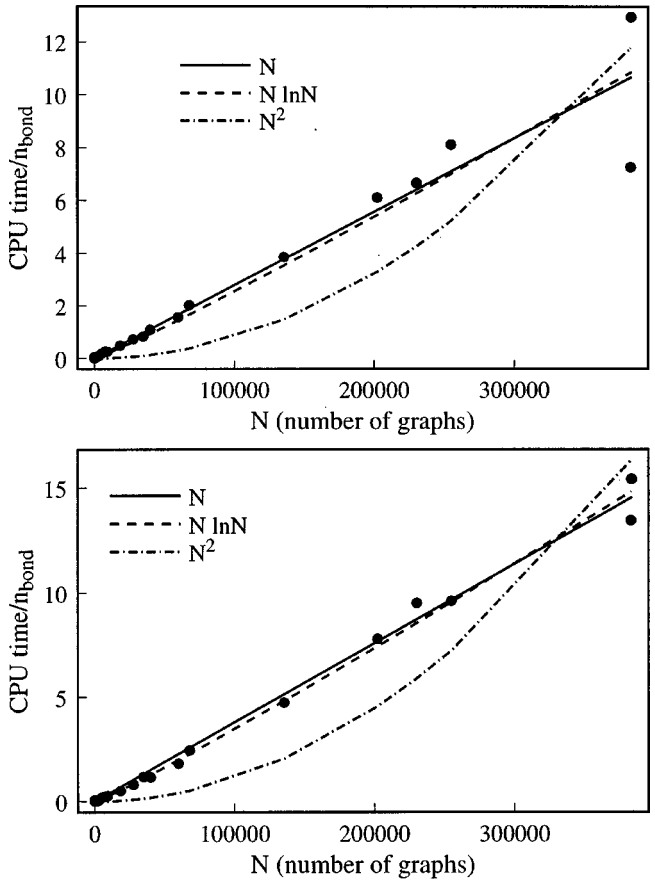


FIG. 4. Data for the same calculation as in Fig. 3, this time employing the sorting method introduced in this work. CPU time per graph edge for sorting the graphs is plotted against the number of graphs N in the top panel. The total CPU time for symmetry comparisons within groups of size $n \leqslant 500$ is shown in the bottom panel. Least-square fits clearly show that the computational cost scales as either N or $N \ln N$ in each case, and definitely not like N^2 as in Fig. 3. On the basis of arguments presented in Sec. II we expect $N \ln N$ scaling in the bottom panel and linear scaling in the top panel.

III. CORRELATION AND PREDICTION OF PHYSICAL PROPERTIES FROM H-BOND TOPOLOGY USING GRAPHICAL INVARIANTS

The number of hydrogen-bonded arrangements for a given water framework, increasing exponentially with system size, quickly grows beyond the point where it is practical to calculate the energy, or other properties, of each structure by *ab initio*, or even semiempirical or empirical potential methods. If physical properties correlate with hydrogen bonding topology, then connecting physical properties with features of the hydrogen bonding network provides a new means of understanding hydrogen bonded structures, and a new route to predicting their properties based on limited input. Physical properties are themselves invariant to symmetry operations. If a correlation exists between physical properties and hydrogen bonding topology, the graph invariants furnish the required connection.

A. $(H_2O)_6$ and $(H_2O)_{20}$

Our first test of the correlation between H-bond topology and physical properties is the $(H_2O)_6$ cage. Using the semiempirical PM3 method,[45] Tissandier *et al.* recently determined optimized geometries of isomers corresponding to each of the 27 hydrogen bond topologies possible for the $(H_2O)_6$ cage structure (Fig. 2), and calculated the energy and dipole moment at each optimized geometry. Where comparison could be made, they found the semiempirical energies consistent with previous *ab initio* results.[46–50] Here we take those results and test whether these properties correlate with hydrogen bond topology, and, if so, how effective invariants are in fitting those properties. Of course, what we are testing is not dependent on the absolute accuracy of the PM3 method. As long as both PM3 properties and those from more accurate methods exhibit a similar level of dependence on H-bond topology, then PM3 properties can be used to gauge the effectiveness of our graph invariant technique.

Before discussing how graph invariants apply to the $(H_2O)_6$ cage, we pause to discuss the use of both filled and empty bonds in this case. The canonical orientations we arbitrarily choose for the cage structure are shown in Fig. 2. The actual bond orientations are specified relative to the canonical orientations. For example, for the physical configuration of H-bonds shown in Fig. 2, $b_3 = -1$ while $b_7 = +1$. The apical water molecules (top and bottom of the cage structure in Fig. 2) are only two-coordinate. When these molecules are single-donor/single-acceptors, *ab initio* calculations on $(H_2O)_6$ have shown there are two minimum energy positions of the apical waters that arise from these molecules accepting a hydrogen bond at either of their two lone pairs. The isomers that arise in this case are conveniently enumerated by adding two ''ghost'' atoms for each of the two-coordinate waters, and to which the two-coordinate waters can be treated as donating a hydrogen bond.[10] Of course, we only allow the ghost atoms to be a hydrogen bond acceptor, not donor. Bonds 9, 10, 11, and 12 in Fig. 2 involve ghost atoms. The variables b_9, b_{10}, b_{11}, and b_{12} can sometimes take the value 0, while $b_1 - b_8$ only assume the values ± 1.

The vertex and bond permutation group of the $(H_2O)_6$ cage was taken to be isomorphic to the D_{2d} point group, although each of the 27 isomers is distorted from perfect D_{2d} symmetry. As discussed earlier in Sec. I, use of graph theory and invariants for the $(H_2O)_6$ cage in no way requires that the isomers have D_{2d} symmetry. We regard the value of the energy or other physical properties for each of the isomers as a 27-dimensional vector that we wish to express, in a least-squared sense, as a linear combination of several 27-dimensional vectors which contain the value of one of the invariants of Table II for each of the 27 isomers. There are a total of 18 first- and second-order invariants, 2 first and 16 second-order, for the $(H_2O)_6$ cage structure. However, the number of independent vectors from among the first-and second-order invariants available to fit physical properties turned out to be much less than 18. Because of constraints among the bonds imposed by the ice rules, some invariants evaluate to be linearly dependent on others. Second-order invariants with repeated indices (I_{ii}) merely count the number of filled bonds ($b_r \neq 0$) within a group theoretical orbit or cycle. Some cycles contain bonds which are always filled bonds, and therefore they give rise to an invariant that is represented by a 27-dimensional vector all of whose components are identical. The existence of more than one such

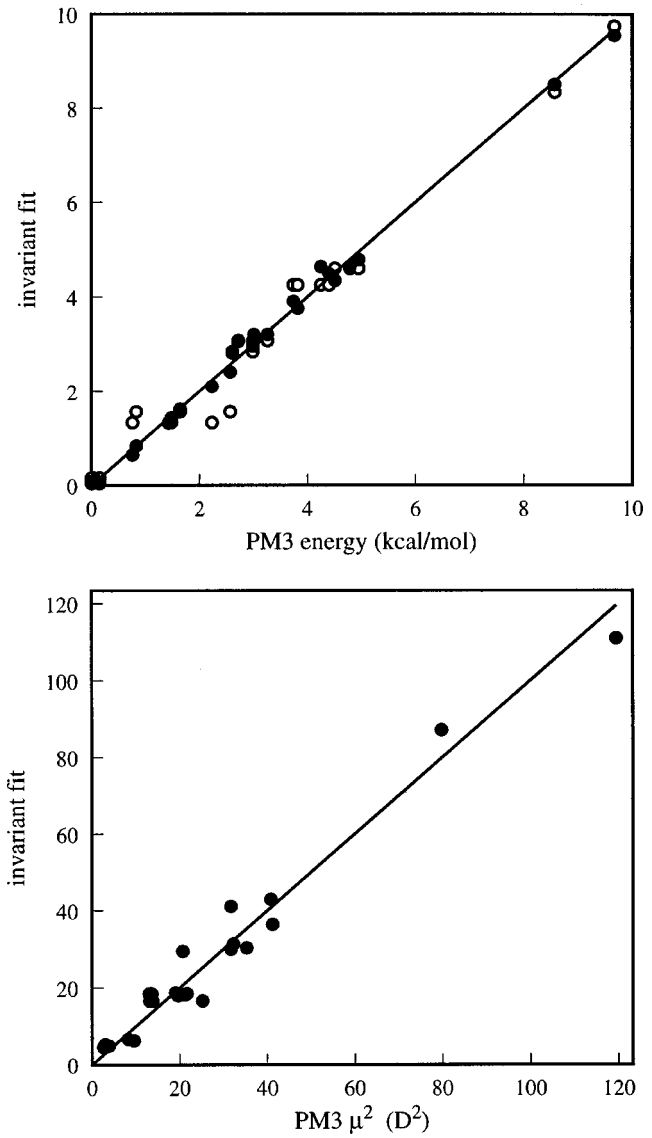


FIG. 5. In the upper panel, we test the degree to which the energies of the 27 isomers of the $(H_2O)_6$ cage are correlated with hydrogen bond topology, and the effectiveness of graphical invariants in capturing that trend. The *x*-coordinate is the energy of the isomers using the PM3 semiempirical theory. The *y*-coordinate is the result of a least-squares fit to these energies using all nine linearly independent first and second graph invariants (filled symbols), or just four out of those nine that proved to be most important (open symbols). If the fit was perfect all points would lie along the straight line. The lower panel exhibits the fit of the squared dipole moment to nine linearly independent first- and second-order invariants.

''bond-counting'' invariant is another source of linear dependence. For the $(H_2O)_6$ cage structure, the sources of linear dependence just mentioned conspire to limit the number of independent first- and second-order invariants available to fit physical properties to 9.

The ability of the first- and second-order invariants to capture the trends in energy and squared total dipole for the $(H_2O)_6$ cage is confirmed in Fig. 5. The root-mean-squared deviation of the invariant fit from the actual energy is 0.15 kcal/mol and the maximum deviation is .38 kcal/mol, compared to a range of 9.7 kcal/mol between least and most stable isomers. Specifically, the nine-invariant fit portrayed in Fig. 5 is given as

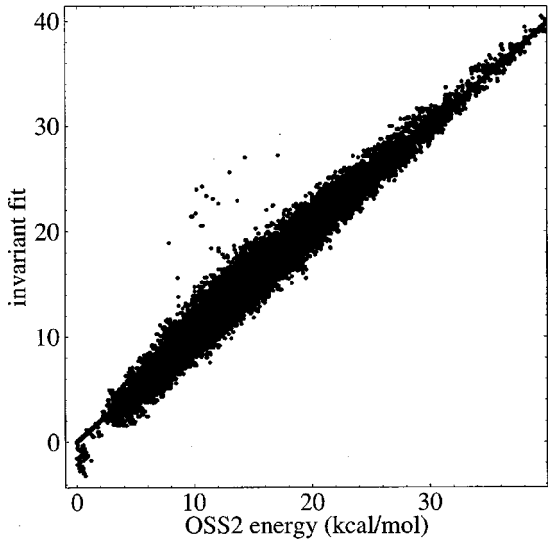


FIG. 6. This plot evaluates the degree to which the energies of the 30 026 isomers of the dodecahedral $(H_2O)_{20}$ cage are correlated with hydrogen bond topology, and the effectiveness of graphical invariants in capturing that trend. The *x*-coordinate is the energy of the isomers using the OSS2 empirical potential (Refs. 51, 53). The *y*-coordinate is the result of a least-squares fit to these energies using the seven linearly independent second graph invariants. If the fit was perfect all points would like along the straight line. A training set of only 20 randomly selected configurations was used to parametrize the energy as a linear combination of invariants.

$$E(\text{kcal/mol}) \approx 0.956 I_3 - 2.235 I_9 + 0.668 I_{2,4} + 0.755 I_{3,7} + 0.158 I_{3,5} - 0.393 I_{5,9} + 0.159 I_{1,3} - 0.364 I_{2,9} + 0.571 I_{3,4}, \quad (19)$$

where the invariants I_r, I_{rs} for the $(H_2O)_6$ cage are defined in Table II. Since there are linear dependencies among the invariants, the linear fit in Eq. (19) could be written in many different ways, although the fit itself is uniquely defined. The root-mean-squared and maximum deviations for the squared dipole moment are $4.7D^2$ and $9.4D^2$, respectively, compared to a range of $119D^2$ between least and greatest squared dipole moment. Further invariants—third, fourth, and even higher-order—could have been used to fit physical properties. It is enlightening that physical properties of the $(H_2O)_6$ cage do correlate well with the hydrogen bond topology, and encouraging that only first- and second-order invariants of the H-bond topology can capture the trends in physical properties. Actually, in Fig. 5 we show that only four out of the nine available first- and second-order invariants are really important in describing the cluster energy.

Similar encouraging results are obtained upon examination of a dodecahedral cage of 20 water molecules. We have calculated the optimized geometry, and energy at that geometry, for each of the 30 026 symmetry-distinct isomers of the $(H_2O)_{20}$ dodecahedron using the OSS2 empirical potential.[51–53] The data in Fig. 6 confirms that the energies of $(H_2O)_{20}$ isomers are correlated with H-bond topology, and that the trend is captured well by the seven linearly-independent second-order graph invariants for the dodecahedron. [All first-order invariants for the dodecahedron are identically zero. There are eight second-order invariants, but one linear dependence among this set caused by factors men-

tioned above in the discussion of the $(H_2O)_6$ cage.] The energy as a linear combination of the seven independent graph invariants was determined by least-squares fit using 20 isomers randomly selected from the full set of 30 026 as a training set. Figure 6 shows that the second-order graph invariants can be effectively employed to predict the energy of the remaining isomers based on data from a small training set. The fit could be further improved by using still more isomers in the training set, or by including third- and higher-order invariants. We did not pursue these refinements since even the lowest level calculation seemed quite adequate for estimating thermal properties or selecting candidates for the lowest energy isomer.

In our original work on the $(H_2O)_{20}$ dodecahedron,[9] we noted that the energy of various isomers largely depended on a single topological feature, the number of nearest-neighbor pairs of double acceptor waters in the clathrate structure. One hydrogen of each of the ten double acceptor waters is a ''dangling'' hydrogen and does not participate in a hydrogen bond. Therefore, the energy of the dodecahedral clathrate is determined by the number of nearest-neighbor dangling hydrogen pairs. The number of such pairs is indeed a topological feature captured by graph invariants of the dodecahedron at second-order. Therefore, the success of the second-order graph invariants in Fig. 6 is not surprising. However, the procedure used in this work is entirely automatic. It can be applied in situations (we have ice-Ih in mind) where a serendipitous discovery of an important topological feature, like the dangling hydrogen pairs of the dodecahedron, is lacking. The graph invariants provide a *non*serendipitous route for discovery of such important topological features.

B. Using invariants to calculate phase transitions: Dodecahedral $(H_2O)_{20}$ as a dry run

The potential energy surface for ice or cold water clusters consists of a number of deep minima, each corresponding to a different hydrogen bond topology. Working within the framework of classical statistical mechanics, the classical configuration integral for these systems can be written as a sum of contributions from each of the M symmetry-distinct local minima of the potential energy surface:[54–62]

$$Z_N = \int d\mathbf{r}^{\mathbf{N}}\, e^{-\beta V(\mathbf{r}^{\mathbf{N}})} = \sum_{i=1}^{M} f_i e^{-\beta E_i} \int_{\mathcal{D}_i} d\mathbf{r}^{\mathbf{N}}\, e^{-\beta[V(\mathbf{r}^{\mathbf{N}}) - V(\mathbf{r}_i^{\mathbf{N}})]}. \tag{20}$$

We use $\mathbf{N}$ to stand for (N_O, N_H), the number of hydrogen and oxygen atoms. The position of the atoms at the ith local minimum is denoted as $\mathbf{r}_i$, $\mathcal{D}_i$ is an $\mathbf{N}$-dimensional integration domain about the ith minimum, $E_i \equiv V(\mathbf{r}_i^N)$ is the potential energy at the ith minimum, and f_i is the number of symmetry-related configurations which are represented by one symmetry-distinct configuration. The canonical partition function of the system is given as

$$Q_N = \sum_{i=1}^{M} \frac{f_i}{\Lambda^{3N}\mathbf{N}!} e^{-\beta E_i} \int_{\mathcal{D}_i} d\mathbf{r}^{\mathbf{N}}\, e^{-\beta[V(\mathbf{r}^{\mathbf{N}}) - V(\mathbf{r}_i^{\mathbf{N}})]} \tag{21a}$$

$$\equiv \sum_{i=1}^{M} f_i e^{-\beta(E_i + A_{\mathrm{vib},i})}. \tag{21b}$$

In keeping with the notation of Eq. (20), $\Lambda^{3\mathbf{N}}\mathbf{N}!$ stands for $(\Lambda_O^{3N_O}\Lambda_H^{3N_H} N_O! N_H!)$.

The contribution of each isomer to the partition function is determined by the potential energy E_i of the isomer, and an integral over ''vibrational'' or ''phonon'' fluctuations about the ith local minimum of the potential energy surface whose contribution we call the vibrational free energy $A_{\mathrm{vib},i}$. In a harmonic approximation, $V(\mathbf{r}^{\mathbf{N}}) - V(\mathbf{r}_i^{\mathbf{N}})$ would be taken as a quadratic function in deviations from $\mathbf{r}_i^{\mathbf{N}}$ and the range of integration over $\mathcal{D}_i$ could be safely extended to all space. It also might be a reasonable assumption to replace $A_{\mathrm{vib},i}$ by an average value $\overline{A}_{\mathrm{vib}}$ for each of the isomers, in which case $Q_{\mathbf{N}} \approx e^{-\beta \overline{A}_{\mathrm{vib}}} \sum_{i=1}^{M} f_i e^{-\beta E_i}$. Consider now the calculation of the average energy as a function of temperature. Under the framework of classical statistical mechanics, by equipartition the vibrational contribution to the average configurational energy is $1/2 k_B T$ for each degree of freedom, regardless of how the spectrum of harmonic frequencies may change among the isomers. At this level of approximation, the average energy $\langle E \rangle$ and the heat capacity C_V are unaffected by the details of vibrational structure. Classical statistical mechanics may be acceptable, even for the ice-Ih→ice-XI transition: the heat capacity peak assigned to this transition only shifts from 72 to 76 K upon replacing H_2O by D_2O.[24] If the vibrational contribution to the heat capacity does not vary widely among the isomers, then an expression in the same form as Eq. (21b) applies even in the quantum regime.

Calculating the E_i and $A_{\mathrm{vib},i}$ for all possible hydrogen bond topologies is a daunting task, yet it is what would be needed to, say, predict proton ordering phase transitions in ice-Ih. The graph invariants we introduce in this work provide a way to circumvent the need to calculate all the E_i and $A_{\mathrm{vib},i}$. Provided these quantities can be fit, using a relatively small training set, as a linear combination of graph invariants, we would have all information in hand to calculate thermodynamic properties and phase transitions. This idea is tested here for a hypothetical, unphysical situation: We calculate proton ordering phase transitions for the $(H_2O)_{20}$ dodecahedron. This calculation is not experimentally relevant because it may never be possible to study the $(H_2O)_{20}$ dodecahedral cluster, and even more unlikely to obtain properties as a function of temperature. At higher energies the dodecahedron will transform into a more stable isomer or dissociate into smaller fragments. The calculation of proton ordering transitions in the $(H_2O)_{20}$ dodecahedron is presented here as a dry run for an especially physically relevant situation, proton ordering transitions in ice.

The top panel of Fig. 7 shows the heat capacity of the $(H_2O)_{20}$ dodecahedron calculated using the energy of all 30 026 isomers in Eqs. (21a)–(21b), which were optimized using the OSS2 potential energy surface.[51] These curves are labeled as ''exact'' in Fig. 7, but they are only exact under the reasonable assumptions (harmonic fluctuations about local minima of the potential energy surface or similar $A_{\mathrm{vib},i}$ among the many isomers) and unreasonable assumptions

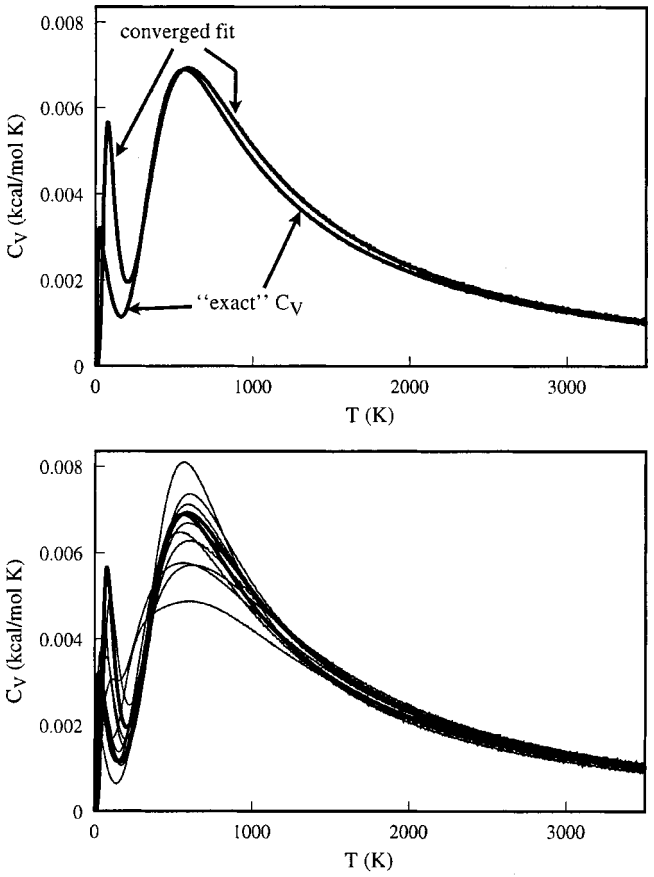


FIG. 7. Configurational energy and heat capacity of a model $(H_2O)_{20}$ dodecahedral cluster. The ''exact'' curve in the top panel is calculated from the partition function in Eqs. (21a), (21b) using the energy of all 30 026 isomers of the $(H_2O)_{20}$ dodecahedron. The vibrational contribution to the energy, which would only add a linear term to the average energy and a constant to the heat capacity under the assumption of harmonic fluctuations about each local minimum, is not included. The curve labeled ''converged fit'' is obtained using an arbitrarily large number of isomers in the training set. It represents the best fit possible using only second-order invariants. The curves with thin lines in the bottom panel give the results of the invariant fitting procedure, as fully explained in the text. The heat capacity was calculated using only 30 of the 30 026 isomers as input for a fit to invariants, after which the energies E_i in Eqs. (21a), (21b) were calculated from the fit. To portray the variability arising from fitting to randomly selected points, we give results for nine independent trials.

[that the $(H_2O)_{20}$ dodecahedron will reach several hundred degrees Kelvin without transforming into another structure or dissociating into smaller fragments] that underly the application of Eqs. (21a)–(21b) to dodecahedral $(H_2O)_{20}$. All these assumption *would* be reasonable for ice-Ih. The fit labeled ''converged fit'' in Fig. 7 is the heat capacity derived from cluster energy expressed as a linear combination of second-order invariants with the linear coefficients determined by least-squared fit to all 30 026 isomers of the dodecahedron.

The curves in Fig. 7 indicate there are the cluster analogs of two phase transitions for this model system, one near 50 K and another near 600 K. The fit using only seven second-order invariants (recall that all first-order invariants of the dodecahedron are zero) accurately describes the main heat capacity peak. The smaller peak at low temperature is slightly displaced. This peak is more difficult to reproduce because it involves relatively few configurations. Using higher-order invariants would narrow the gap between the ''exact'' and ''converged fit.'' Of course, even if an ensemble of $(H_2O)_{20}$ dodecahedra could be prepared, the high temperature transition would certainly never be observed since these clusters would fragment long before the temperature reached this value.

Next, just 20 randomly selected isomers from the full set of 30 026 were used as an initial training set to parametrize the energies as a function of graph invariants. The resulting fit gave a good representation of the overall behavior of the energy and heat capacity, including the heat capacity peak around 600 K, and indicated some structure in the low ($<$250 K) energy range. However, it gave a poor representation of the low energy peak in the heat capacity. This is not surprising, since the isomers range in energy up to $\sim$50 kcal/mol, or $\sim$25 000 K, about the lowest energy isomer. Therefore, the structure at low temperature involves an extremely small fraction of the isomers. We added ten more randomly selected isomers to the fitting set, this time chosen from isomers that, according to the initial fit, were within 350 K of the lowest energy configuration. The resulting fit, obtained by randomly sampling from 0.1% of the total number of isomers, reproduces the qualitative features of the heat capacity function. We repeated the entire procedure nine times and show the results as the thin lines in the bottom panel of Fig. 7 to illustrate the errors incurred at this level. It is very encouraging that using the lowest-order nonvanishing invariants and a sparse sample of configurations gives a qualitatively correct description of the 30 026 H-bond isomers of the $(H_2O)_{20}$ dodecahedron.

IV. GRAPH INVARIANTS IN PERIODIC SYSTEMS

The same formalism employed for clusters is applicable to periodic systems. Invariants are constructed by applying the crystal space group to the bonds of the crystal lattice. The invariants are therefore infinite sums of bonds. When the hydrogen bond arrangement repeats according to a periodic unit cell, then the invariants can be expressed as finite sums over bonds from one unit cell. A less rigorous, more intuitive approach (but with some pitfalls) is to picture the crystal unit cell as a finite cluster in which hydrogen bonds to atoms outside the unit cell wrap around to the corresponding atoms in the original unit cell. The formalism governing invariants in hydrogen-bonded crystals will be presented in a separate publication.[63] As a demonstration of the capabilities of our formalism, we present the results of enumerating hydrogen bond topologies for several unit cells from the ice-Ih lattice, which are the most challenging systems we have tackled to date.

Under normal condition, the space group of the oxygen lattice in Ice-Ih belongs to the $P6_3/mmc$ space group. In the ice-Ih$\rightarrow$XI transition, neutron scattering experiments indicate that the $P6_3/mmc$ symmetry is broken by ordering of the hydrogen bonds, and the lattice symmetry falls to $Cmc2_1$.[26–29] The distortion of the lattice vectors from $P6_3/mmc$ symmetry is measured to be 2% in the *ab*-plane and 0.5% along the *c*-axis.[27] According to the discussion in Sec. I, enumerated H-bond arrangement and invariants for

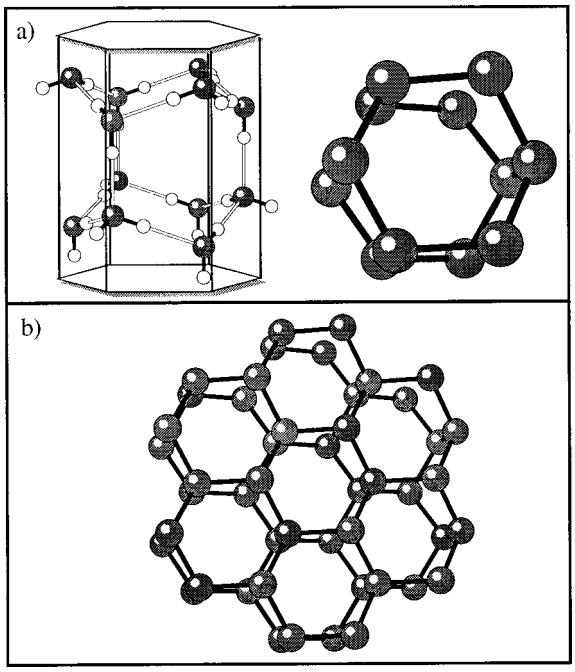


FIG. 8. (a) Hexagonal $H_{1\times1}$ unit cell for ice-Ih containing 12 water molecules. An all-atom depiction looking parallel to the *ab*-plane is shown on the left. On the right, only the oxygen atoms are shown looking along the *c*-direction. (b) The oxygen atoms of the 48-water $H_{2\times1}$ cell are shown.

the high-symmetry lattice will even describe distorted structures, provided there is some correspondence (such as a decreasing energy path) between undistorted and distorted structures. We have enumerated H-bond configurations using the hexagonal $P6_3/mmc$ lattice of ice-Ih, but enumerations using the $Cmc2_1$ lattice are certainly also possible.

Our smallest unit cell is a hexagonal prism of 12 water molecules (Fig. 8). This unit cell can be extended in both *ab*-plane and *c*-axis. We use $H_{m\times n}$ to designate these hexagonal unit cells where *m* stands for the extension in *ab*-plane and *n* stands for the extension in *c*-axis. By this convention the smallest hexagonal unit cell (with 12 molecules) is named $H_{1\times1}$. $H_{1\times2}$ has 24 molecules while $H_{2\times1}$ has 48 molecules.

The result of graph enumeration for different unit cells in summarized in Table III. The rapid growth of the total number of configurations and symmetry-distinct configurations with unit cell size is apparent. Isolating symmetry-distinct configurations reduces the total number by a factor of 37 for the $H_{1\times1}$ cell and 288 for the larger $H_{2\times1}$ cell. Calculating the energy of all configurations of the large cell is certainly not feasible. However, invariants can be transferred from small cells to larger cells[63] and it is very feasible to estimate the energy of all 8 360 361 configurations for the larger cell using invariants.

TABLE III. Results of enumerating hydrogen bond arrangements for various units cells of the ice-Ih lattice. The unit cell designations are explained in Sec. IV. n_{mol} is the number of water molecules in each unit cell, n_{Gv} is the order of the permutation group, n_{Inv} is the number of second-order invariants. (All first-order invariants for the ice-Ih lattice are identically zero.) N_{dist} is the number of symmetrically distinct structures, N_{total} is the number of structures satisfying the ice rule. Let m_i be the number of permutation symmetry elements of the *i*th graph. According to group theory, there are $n_{Gv}/m_i=f_i$ graphs which are generated from the *i*th graph by a symmetry operation. Therefore, we have the relation $N_{\mathrm{total}}=\Sigma_{k=1}^{N_{\mathrm{dist}}}(n_{Gv}/m_k)$.

Unit cell	n_{mol}	n_{Gv}	n_{Inv}	N_{dist}	N_{total}
$H_{1\times1}$	12	72	13	14	522
$H_{2\times1}$	48	288	36	8 360 361	2 404 144 962

V. DISCUSSION

In the past, it has seemed natural to assume some correlation between the hydrogen bond topology and physical properties like energy and dipole moment.[11–15] Oriented graph invariants provide a means to quantify the correlation between hydrogen bond arrangement and physical properties, and systematize the process determining the features of the hydrogen bond network which most affects the energy, and other physical properties. Such correlations would prove exceedingly useful in resolving outstanding questions regarding thermal properties and possible phase transitions in common ice-Ih. The verdict on the degree to which H-bond topology predicts energy in ice must await further investigation. This article furnishes the tools needed to analyze such putative relationships. We also provide a highly efficient method for either exhaustive or selective enumeration of hydrogen bond topologies of clusters and crystals.

The utility of graph invariants rests on the extent to which physical properties depend on hydrogen bond topology. The basic idea underlying the formalism is that scalar physical properties, like energy or magnitude of the dipole moment, are invariant to symmetry transformations. If two hydrogen bonding arrangements are symmetry related, then physical properties associated with these arrangements must also be identical. For example, the association might be pictured as arising from geometry optimization from a high-symmetry initial structure. The symmetry properties of the oriented graph and the geometry-optimized physical isomers may be different. Remarkably, as long as symmetry-equivalent initial structures optimize to the same distorted structure, invariants based on the high-symmetry structure are still appropriate for describing the relationship between hydrogen bonding topology and physical properties of the distorted final structures.

ACKNOWLEDGMENTS

S.J.S. gratefully acknowledges receipt of a Dozor Fellowship enabling him to visit to Ben-Gurion University of the Negev where this work was initiated. L.P.R. acknowledges support from the Swedish Natural Science Research Council (NFR). Some of the computations reported here were made possible by a resource grant from the Ohio Supercomputer Center.

APPENDIX

Here we demonstrate a necessary and sufficient condition for a graphical invariant to be identically zero. We first show that a sufficient condition for the first-order invariant,

$$I_r = \sum_{\alpha=1}^{G} g_\alpha(b_r), \tag{A1}$$

to vanish is the existence of a group element that takes the bond b_r into minus itself,

$$g_\alpha(b_r) = -b_r. \tag{A2}$$

Apply the projection operator for the totally symmetric representation as in Eq. (3) to both sides of the above equation,

$$\sum_{\beta=1}^{G} g_\beta(g_\alpha(b_r)) = -\sum_{\beta=1}^{G} g_\beta(b_r) = -I_r. \tag{A3}$$

Let $g_\beta(g_\alpha(b_r)) = g_\gamma(b_r)$ and realize that from the group requirement of a unique inverse that no two elements g_β can give rise to the same g_γ. It follows that

$$I_r = \sum_{\gamma=1}^{G} g_\gamma(b_r) = -\sum_{\beta=1}^{G} g_\beta(b_r) = -I_r, \tag{A4}$$

and therefore the existence of a group element that takes b_r into minus itself is a sufficient condition for $I_r = 0$. To show that this is a necessary condition note that if I_r vanishes identically, then the result of applying the projector of Eq. (3) onto b_r can be grouped into pairs of terms that cancel each other,

$$\begin{aligned} I_r &= \sum_{\alpha=1}^{G} g_\alpha(b_r) \propto \cdots (g_\beta(b_r) + g_\gamma(b_r)) + \cdots \\ &= \cdots (b_s - b_s) + \cdots, \end{aligned} \tag{A5}$$

for if $g_\beta(b_r) = b_s$ there must be another group operation g_γ such that $g_\gamma(b_r) = -b_s$ if I_r vanishes identically. However, if $g_\beta(b_r) = -g_\gamma(b_r)$, then there must be a group element g_α with the property that

$$g_\alpha(b_r) = g_\gamma^{-1}(g_\beta(b_r)) = -b_r. \tag{A6}$$

This shows that condition (A2) is both necessary and sufficient for I_r to vanish identically.

The same reasoning can be applied to show that any invariant $I_{rst\cdots}$ vanishes identically if and only if there exists a group elements such that

$$g_\alpha(b_r b_s b_t \cdots) = -b_r b_s b_t \cdots. \tag{A7}$$

Condition (A7) is sufficient because applying the projector of the totally symmetric representation to both sides of Eq. (A7) leads to

$$\begin{aligned} I_{rst\cdots} &= \sum_{\beta=1}^{G} g_\beta(g_\alpha(b_r b_s b_t \cdots)) \\ &= -\sum_{\beta=1}^{G} g_\beta(b_r b_s b_t \cdots) = -I_{rst\cdots}. \end{aligned} \tag{A8}$$

Condition (A7) is necessary because if $I_{rst\cdots}$ vanishes it must be possible to group its terms into pairs that cancel each other:

$$\begin{aligned} I_{rst\cdots} &= \sum_{\alpha=1}^{G} g_\alpha(b_r b_s b_t \cdots) \\ &\propto \cdots (g_\beta(b_r b_s b_t \cdots) + g_\gamma(b_r b_s b_t \cdots)) + \cdots \\ &= \cdots (b_{r'} b_{s'} b_{t'} \cdots - b_{r'} b_{s'} b_{t'} \cdots) + \cdots. \end{aligned} \tag{A9}$$

Therefore there must exist a group element g_α that satisfies condition (A7).

[1] D. Eisenberg and W. Kauzmann, *The Structure and Properties of Water* (Oxford, New York, 1969).
[2] L. Pauling, J. Am. Chem. Soc. **57**, 2680 (1935).
[3] W. F. Giauque and J. W. Stout, J. Am. Chem. Soc. **58**, 1144 (1936).
[4] E. A. DiMarzio and F. H. Stillinger, J. Chem. Phys. **40**, 1577 (1964).
[5] J. F. Nagle, J. Math. Phys. **7**, 1484 (1966).
[6] P. V. Hobbs, *Ice Physics* (Oxford, New York, 1974).
[7] J. K. Gregory, D. C. Clary, K. Liu, M. G. Brown, and R. J. Saykally, Science **275**, 814 (1997).
[8] C. J. Gruenloh, J. R. Carney, C. A. Arrington, T. S. Zwier, S. Y. Fredericks, and K. D. Jordan, Science **276**, 1678 (1997).
[9] S. McDonald, L. Ojamäe, and S. J. Singer, J. Phys. Chem. A **102**, 2824 (1998).
[10] M. D. Tissandier, S. J. Singer, and J. V. Coe, J. Phys. Chem. A **104**, 752 (2000).
[11] N. Bjerrum, Science **115**, 386 (1952).
[12] K. S. Pitzer and J. Polissar, J. Am. Chem. Soc. **60**, 1140 (1956).
[13] E. R. Davidson and K. Morokuma, J. Chem. Phys. **81**, 3741 (1984).
[14] B. J. Yoon, K. Morokuma, and E. R. Davidson, J. Chem. Phys. **83**, 1223 (1985).
[15] J. C. Li and D. K. Ross, Nature (London) **365**, 327 (1993).
[16] J. S. Tse and D. D. Klug, Phys. Lett. A **198**, 464 (1995).
[17] V. Buch, P. Sandler, and J. Sadlej, J. Phys. Chem. B **102**, 8641 (1998).
[18] A. Rahman and F. H. Stillinger, J. Chem. Phys. **57**, 4009 (1972).
[19] W. J. Pullan, Comput. Phys. Commun. **107**, 137 (1997).
[20] R. Judson, Rev. Comput. Chem. **10**, 1 (1997).
[21] S. Kawada, J. Phys. Soc. Jpn. **32**, (1972).
[22] Y. Tajima, T. Matsuo, and H. Suga, Nature (London) **299**, 810 (1982).
[23] Y. Tajima, T. Matsuo, and H. Suga, J. Phys. Chem. Solids **45**, 1135 (1984).
[24] T. Matsuo, Y. Tajima, and H. Suga, J. Phys. Chem. Solids **47**, 165 (1986).
[25] T. Matsuo and H. Suga, J. Phys. Colloq. **C1**, 477 (1987).
[26] R. Howe and R. W. Whitworth, J. Chem. Phys. **90**, 4450 (1989).
[27] S. M. Jackson, V. M. Nield, R. W. Whitworth, M. Oguro, and C. C. Wilson, J. Phys. Chem. B **101**, 6142 (1997).
[28] A. J. Leadbetter, R. C. Ward, J. W. Clark, P. A. Tucker, T. Matsuo, and H. Suga, J. Chem. Phys. **82**, 424 (1985).
[29] C. M. B. Line and R. W. Whitworth, J. Chem. Phys. **104**, 10008 (1996).
[30] S. M. Jackson and R. W. Whitworth, J. Chem. Phys. **103**, 7647 (1995).
[31] S. M. Jackson and R. W. Whitworth, J. Phys. Chem. **101**, 6177 (1997).
[32] M. J. Iedema, M. J. Dresser, D. L. Doering, J. B. Rowland, W. P. Hess, A. A. Tsekouras, and J. P. Cowin, J. Phys. Chem. B **102**, 9203 (1998).
[33] R. W. Whitworth, J. Phys. Chem. **103**, 8192 (1999).
[34] J. P. Cowin and M. J. Iedema, J. Phys. Chem. **103**, 8194 (1999).
[35] H. Fukazawa, S. Mae, S. Ikeda, and O. Watanabe, Chem. Phys. Lett. **294**, 554 (1998).
[36] Y. Wang, J. C. Li, A. I. Kolesnikov, S. Parker, and S. J. Johnsen, Physica B **276–278**, 282 (2000).
[37] J. D. Bernal and R. H. Fowler, J. Chem. Phys. **1**, 515 (1933).
[38] F. Harary, *Graph Theory* (Addison-Wesley, Reading, MA, 1969).
[39] F. Harary and E. M. Palmer, *Graphical Enumeration*, (Academic, New York, 1973).
[40] R. Howe, J. Phys. (Paris) Colloq. **48**, C1-599 (1987).
[41] J. Lekner, Physica B **252**, 149 (1998).
[42] C. Domb, *Phase Transitions and Critical Phenomena* (Academic, New York, 1974), Vol. 3, p. 1.
[43] T. P. Radhakrishnan and W. C. Herndon, J. Phys. Chem. **95**, 10609 (1991).
[44] J. A. Hayward and J. R. Reimers, J. Chem. Phys. **106**, 1518 (1997).
[45] J. J. P. Stewart, J. Comput. Chem. **10**, 209 (1989), **10**, 221 (1989).
[46] C. J. Tsai and K. D. Jordan, Chem. Phys. Lett. **213**, 181 (1993).
[47] C. Lee, H. Chen, and G. Fitzgerald, J. Chem. Phys. **101**, 4472 (1994).

[48] K. Kim, K. D. Jordan, and T. S. Zwier, J. Am. Chem. Soc. **116**, 11568 (1994).
[49] J. M. Pedulla, K. Kim, and K. D. Jordan, Chem. Phys. Lett. **291**, 78 (1998).
[50] J. Kim and K. S. Kim, J. Chem. Phys. **109**, 5886 (1998).
[51] L. Ojamäe, I. Shavitt, and S. J. Singer, J. Chem. Phys. **109**, 5547 (1998).
[52] C. V. Ciobanu, L. P. Ojamäe, I. Shavitt, and S. J. Singer (unpublished). Preliminary density functional theory and MP2 *ab initio* results confirm the trends of the OSS2 potential for the $(H_2O)_{20}$ dodecahedron. In particular the isomers become less stable as the number of nearest-neighbor dangling hydrogens increase.
[53] Using the OSS2 potential, roughly 10% of the 30 026 isomers failed to optimize to a structure with the same hydrogen bond topology as the starting structure. Including or deleting these exceptions makes no perceptible difference in Figs. 6 and 7. The exceptions are nevertheless quite interesting. They are not an artifact of the OSS2 potential, and the effects are confirmed in DFT/B3LYP and MP2 level *ab initio* studies. They will be described in a future report.
[54] D. J. McGinty, J. Chem. Phys. **55**, 580 (1971).
[55] J. J. Burton, J. Chem. Phys. **56**, 3133 (1972).
[56] M. R. Hoare, Adv. Chem. Phys. **40**, 49 (1979).
[57] F. H. Stillinger and T. A. Weber, Phys. Rev. A **25**, 978 (1982).
[58] F. H. Stillinger and T. A. Weber, Phys. Rev. **28**, 2408 (1983).
[59] P. G. Mezey, in *Potential Energy Hypersurfaces*, Studies in Physical and Theoretical Chemistry (Elsevier, New York, 1987), Vol. 53.
[60] G. Franke, E. R. Hilf, and P. Borrmann, J. Chem. Phys. **98**, 3496 (1993).
[61] D. J. Wales, Mol. Phys. **78**, 151 (1993).
[62] J. P. K. Doye and D. J. Wales, J. Chem. Phys. **102**, 9659 (1995).
[63] J.-L. Kuo and S. J. Singer (unpublished).