

# Benchmarks of graph invariants for hydrogen-bond networks in water clusters of different topology

Andrey M. Tokmachev<sup>1</sup> · Andrei L. Tchougréeff<sup>2,3,4</sup> · Richard Dronskowski<sup>4</sup>

Received: 18 July 2015 / Accepted: 1 September 2015  
© Springer-Verlag Berlin Heidelberg 2015

**Abstract** The diversity of the various forms of water stems from systems of hydrogen bonds. Cooperative behaviour of hydrogen-bond networks gives rise to unique properties of water systems. A number of approaches to understand and model the collective behaviour of hydrogen bonds and predict their properties on the basis of a small number of calculations have been put forward. Among them, the concept of graph invariants provides most general descriptors for hydrogen-bond networks, which are routinely used to predict properties of water systems. In the present work, we examine the formalism of graph invariants and propose its modification which may be beneficial for water structures with defects. To benchmark graph invariants, we carried out quantum-chemical calculations of more than  $10^7$  water clusters with different hydrogen-bond configurations. The quality of the approximation is studied as a function of the type of graph invariant and its order. The results demonstrate that the method is applicable only to cage-like structures without significant strains.

**Keywords** Water clusters · Hydrogen bonds · Graphs · Invariants · APSG

## 1 Introduction

Water in all its forms constitutes the basis of our existence. The properties of water are unique and its beauty is limitless. Aggregates of water molecules are among the most studied chemical systems, but despite the intensive concerted efforts, water is still far from being understood. The reason behind is the set of hydrogen (H-) bonds keeping water molecules together. Networks of hydrogen bonds exhibit complex behaviour manifested as cooperative contributions to properties of water systems. The enormity of these contributions gives impetus to studies of the collective structure of H-bond networks. Among the most recent successes in electronic-structure theory devoted to H-bonded systems, we mention density-functional studies on carefully quantifying H-bonds in molecular crystals [1, 2] and in assessing the amount of covalency involved [3].

Water systems allow for numerous H-bond configurations even when the morphology of the system (adjacency matrix) is fixed and the restrictions of the ice rules (basically requiring the absence of ionized individual water molecules in a water system) [4] are imposed. The number of these configurations grows exponentially with the size of the system. Simple gas hydrate shells are a perfect illustration: the  $5^{12}$  D-cage ( $\text{H}_2\text{O}$ )<sub>20</sub> allows for 30,026 symmetry independent configurations, the numbers for the larger  $5^{12}6^2$  T-cage ( $\text{H}_2\text{O}$ )<sub>24</sub> and the  $5^{12}6^4$  H-cage ( $\text{H}_2\text{O}$ )<sub>28</sub> are 3,043,836 and 61,753,344, respectively [5]. Clearly, not all of them are important, but information extracted from studies of the minimal energy configuration (which is commonly the only subject of the investigation) is often

Published as part of the special collection of articles “Festschrift in honour of P. R. Surjan”.

✉ Andrei L. Tchougréeff  
andrei.tchougreff@ac.rwth-aachen.de

<sup>1</sup> NBICS Centre, NRC Kurchatov Institute, Kurchatov Sq. 1, Moscow 123182, Russia

<sup>2</sup> Moscow Center for Continuous Mathematical Education, Bol'shoy Vlas'evskiy 11, Moscow 119002, Russia

<sup>3</sup> Department of Chemistry, Moscow State University, Vorob'evy Gory 1, Build. 3, Moscow 119992, Russia

<sup>4</sup> Institut für Anorganische Chemie, RWTH Aachen University, Landoltweg 1, 52056 Aachen, Germany

insufficient. Any reasonable prediction of properties of such systems would require a thorough analysis of H-bond configurations because a significant part of them can be thermally populated.

Construction of a predictive model correlating the energy and other properties with the H-bond network structure is a long-standing problem. A number of approaches have been put forward, mostly for water clusters. A most straightforward approach is based on classification of all H-bonds in a cluster according to their local environment with subsequent counting the numbers of bonds of different types. The required classification can rely upon notions of weak and strong H-bonds (which are often associated with cis and trans H-bonds)—now so easily quantified by DFT [3]—or it can take into account the first neighbours of H-bonds leading to two types of trans-bonds and three types of cis-bonds [6]. Similar ideas underline the so-called Strong–Weak–Effective-Bond model for polyhedral water clusters which is based on estimating the effective pair interactions between nearest neighbours [7]. An alternative model, especially suited to water cages, classifies bonds according to the state of oxygen atoms (having a dangling O–H bond or a lone pair) [8], but it leads to the same five types of H-bonds. Our own calculations of all H-bond isomers of dodecahedral water cluster (H<sub>2</sub>O)<sub>20</sub> indicate that there exists a correlation between the cluster energy and the number of trans or cis H-bonds or the number of pairs of dangling O–H bonds in vicinal positions, but it is not strong enough to build a predictive model [9].

A viable alternative to this approach is given by a general set of entities called graph invariants [10]. Those are automatically generated descriptors of the H-bond structure taking into account the symmetry of the system. The energy and other properties can be approximated as linear combinations of graph invariants. This approach seems very promising because it employs only the adjacency matrix, does not rely on special physical insight, and encompasses a hierarchy of approximations. Incidentally, other common descriptors of H-bond networks (such as numbers of H-bonds of different types, see above) emerge as particular

cases of graph invariants. Moreover, the concept is highly useful for generation or enumeration of symmetry-distinct H-bond networks reducing  $N^2$  scaling of these procedures to  $N \log N$ , where  $N$  is the total number of configurations.

The method has originally been developed for finite clusters, but its generalization to periodic systems [11] has been no less successful. Graph invariant descriptors provide a simple algorithm to extrapolate from a handful of expensive quantum-mechanical calculations to a larger set of H-bond configurations. Similarly, they can be used to extrapolate from calculations for smaller unit cells to statistical mechanical simulations employing larger unit cells. The method is routinely applied in studies of H-bond topology and proton-ordering transitions for different phases of ice [12–16]. At the same time, its applicability is not thoroughly tested in different aspects. First, only second-order invariants are used in all numerical applications. Second, only water systems without significant strains and any defects are studied. Third, all the applications correspond to a very small number of actually calculated H-bond isomers with the only exception provided by the 30,026 isomers of the dodecahedral water cluster (H<sub>2</sub>O)<sub>20</sub>, where the graph invariant approximation is compared with the exhaustive force field OSS2 calculations [10].

In the present paper, graph invariants are benchmarked against results of quantum-chemical calculations. Exhaustive calculations of all H-bond networks for water clusters of 24 different morphologies provide reference values of the energy. A series of tests attempting to approximate the energy by a linear combination of graph invariant descriptors reveals limits for applicability of the method.

## 2 Computational details

Benchmarking requires calculations for a number of water clusters, each providing a set of H-bond networks, large enough to be studied with graph invariants. To be representative, the set of water clusters should include clusters of different size, symmetry and strains. Most of the

**Fig. 1** Structures of ice nano-tubes formed by rings of **a** 4; **b** 5; **c** 6 and **d** 7 water molecules

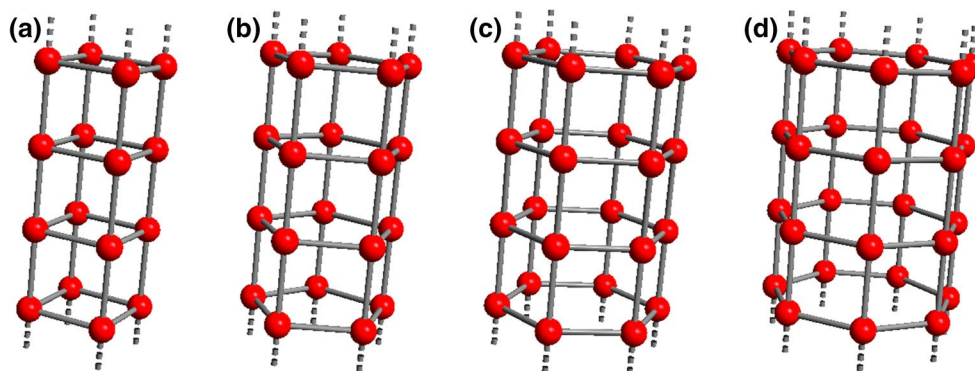
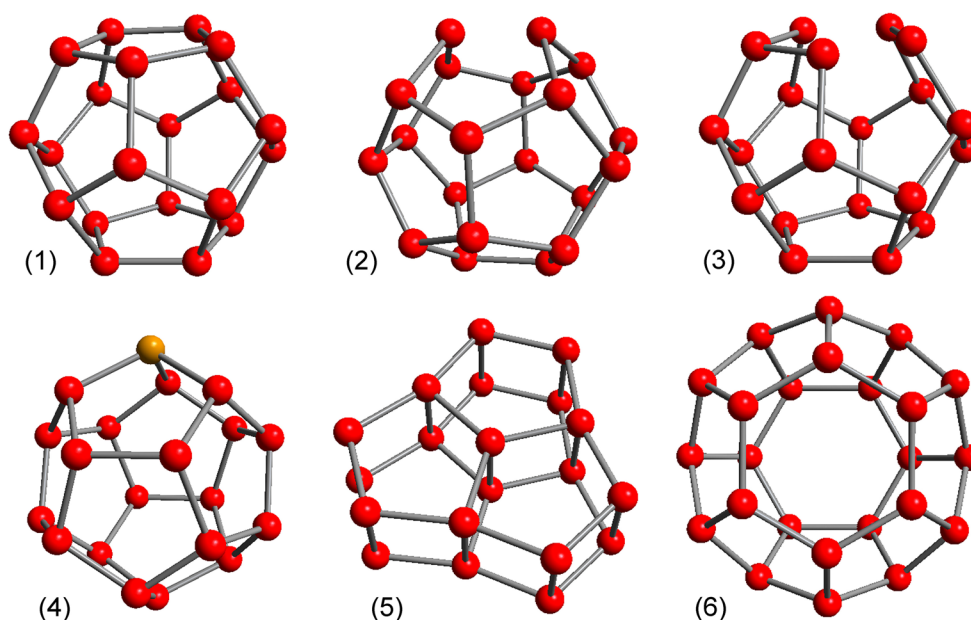


**Fig. 2** Water clusters: **1** dodecahedral cage; **2** dodecahedral cage with one H-bond broken; **3** dodecahedral cage with two H-bonds broken; **4** dodecahedral cage with one water molecule replaced by HF; **5** edge-sharing pentagonal prisms; **6** tetrakaidecahedral cage



structures considered in the paper are finite fragments of experimentally known ice nanotubes (see Fig. 1), which can be viewed as stacks of  $n$ -membered rings (or rolled square-net sheets) with  $n$  being from 4 to 7. The fragments are denoted as  $\text{INT}_n^m$  where  $m$  is the number of  $n$ -membered layers. We studied structures with sufficient but not exceedingly large number of H-bond configurations, namely  $\text{INT}_4^m$  with  $m = 3 \div 8$ ,  $\text{INT}_5^m$  with  $m = 2 \div 6$ ,  $\text{INT}_6^m$  with  $m = 2 \div 5$ , and  $\text{INT}_7^m$  with  $m = 2 \div 4$  [17]. The other clusters studied are shown in Fig. 2. They are the  $5^{12}$  D-cage ( $\text{H}_2\text{O}$ )<sub>20</sub> (1); the same cage but having a single defect due to one H-bond broken (2) or two defects due to a couple of H-bonds broken (3); the same cage with one water molecule replaced by hydrogen fluoride HF (4); the isomer of the D-cage with the structure of edge-sharing pentagonal prisms (5); and the  $5^{12}6^2$  T-cage ( $\text{H}_2\text{O}$ )<sub>24</sub> (6). The total number of configurations in the water clusters mentioned above is 10,366,824.

Quantum-chemical calculations on such a grand scale pose special requirements to the computational scheme. Clearly, the task is far beyond the scope of sophisticated ab initio methods of quantum chemistry. To solve the problem, we devised a specialized semiempirical method dubbed strictly local geminals (SLG) [18–21]. The method is based on geminals [22]—two-electron wave functions enabling construction of efficient methods for large molecular systems [23]. The water molecule is represented by an antisymmetrized product of four geminals—two describing the chemical bonds O–H and two describing the electron lone pairs. Interaction between molecules represented by geminal wave functions can be accounted for perturbatively [24], but we employ a different scheme. The

H-bonds are represented by three-orbital four-electron wave functions. The resulting wave function combines geminals with other strictly local electron groups [25]. An important characteristic of the method is that it is ultra-fast but still highly reliable with respect to energy calculations for many classes of molecules. A modified PM3 parameterization is specially devised to describe water systems with this method [21]. The resulting binding energies are similar to MP2 and DFT benchmarks. The method is successfully applied to a number of water clusters [9, 17].

The present analysis requires the energy for each of H-bond networks under consideration. The simplest way to proceed is to build an idealized O frame for each water cluster morphology, fixed H atom positions being determined separately for each H-bond network. Energy calculations for such idealized structures may already give some impression of the usefulness of graph invariants. However, many water clusters are strained and the calculations without optimization of atomic positions would lead to a distorted picture. Therefore, we optimized the spatial structures of clusters for each H-bond network. It is important that the optimization does not change the topology of the H-bond network because the latter is explicitly incorporated into the structure of the wave function used. It should be stressed that the difference between the optimized and the idealized structures strongly depends on the H-bond pattern. In particular, the H-bond network structure determines the H–O–H angle in the idealized structure and thus determines the local distortion of the structure after optimization.

### 3 Results and discussion

Graph invariants are described in detail in Ref. [10], but it is instructive to recall their definition. H-bonds between water molecules are directional: the H atom is closer to one of O atoms participating in the bond. Therefore, any H-bond network can be represented as an oriented graph. Each water system (such as 24 structures mentioned above) defines the adjacency graph but not orientations of individual edges. In general, the graph can generate  $2^l$  oriented graphs, where  $l$  is the number of edges, but only a small part of them satisfies the ice rules. In addition, the system may have a non-trivial symmetry and some of the oriented graphs become identical. It further reduces the number of the oriented graphs allowed. Thus, each water cluster (morphology)  $M$  generates a set of  $N_{\text{conf}}^M$  H-bond networks (oriented graphs)  $\{X_k^M\}$ , where  $k = 1 \div N_{\text{conf}}^M$ . The aim is to find descriptors distinguishing between networks  $X_k^M$  for different  $k$  and allowing for predictions of physical properties.

First, one introduces a set of functions  $b_{ij}^M$  with  $i$  and  $j$  denoting vertices of the adjacency graph for morphology  $M$  (it is assumed that  $i < j$  to avoid duplication):  $b_{ij}^M = 1$  if the edge between  $i$  and  $j$  is oriented towards  $j$ ;  $b_{ij}^M = -1$  if the same edge is oriented towards  $i$ ;  $b_{ij}^M = 0$  if  $i$  and  $j$  are not connected. The arguments of functions  $b_r^M$  ( $r$  here is a complex index denoting a pair of vertices) are H-bond networks  $X_k^M$ . Values of  $b_r^M(X_k^M)$ ,  $[b_r b_s](X_k^M)$ ,  $[b_r b_s b_t](X_k^M)$ , etc. are general descriptors of H-bond networks determining the direction of H-bonds, their pair correlations, triple correlations etc. Nevertheless, direct application of such descriptors is not generally recommended because most of the known water systems have that or another symmetry: physical entities such as energy are invariant with respect to symmetry operations, and this fundamental property should be reflected in the definition of H-bond topology descriptors.

The symmetry meant is not the spatial symmetry of an actual water cluster. Instead, symmetry operations are permutations of graph vertices preserving the adjacency matrix. The resulting symmetry group  $G$  often corresponds to the spatial symmetry group of the ideal framework of O atoms. Permutations of vertices induce transformations of bond functions  $b_r$ . Symmetry-invariant descriptors for H-bond networks are generated by projecting on the totally symmetric representation of group  $G$ . The first-order invariants (up to normalization constant) are given as:

$$I_r^M = \sum_{\alpha=1}^{n_G} g_{\alpha} \left( b_r^M \right), \quad (1)$$

where  $g_{\alpha}$  are symmetry operations and the summation is over all  $n_G$  elements of the symmetry group. In full analogy, the second-order invariants are defined as:

$$I_{rs}^M = \sum_{\alpha=1}^{n_G} g_{\alpha} \left( b_r^M b_s^M \right), \quad (2)$$

where the group elements act upon products of bond variables. The generalization to higher orders is straightforward. Not all generated invariants are independent: some invariants are zero, while some other coincide.

To illustrate the definitions, we consider a model case of four-membered cyclic clusters with low symmetry group  $C_4$ . The idealized O framework is formed by a square of atoms  $A$ ,  $B$ ,  $C$ , and  $D$ . H-bonds 1–4 are between  $A$  and  $B$ ,  $B$  and  $C$ ,  $C$  and  $D$ ,  $D$  and  $A$ , respectively. The only first-order invariant for such a system is simply the sum of all bond functions:

$$I_1^M = b_1 + b_2 + b_3 + b_4. \quad (3)$$

Three independent second-order invariants can be constructed:

$$\begin{aligned} I_{11}^M &= b_1^2 + b_2^2 + b_3^2 + b_4^2; \\ I_{12}^M &= b_1 b_2 + b_2 b_3 + b_3 b_4 + b_4 b_1; \\ I_{13}^M &= 2(b_1 b_3 + b_2 b_4). \end{aligned} \quad (4)$$

If one considers the network of H-bonds directed from  $A$  to  $B$ , from  $B$  to  $C$ , from  $C$  to  $D$ , and from  $D$  to  $A$ , then all variables  $b$  are equal to 1 and all the above invariants are equal to 4. Another network of H-bonds directed from  $A$  to  $B$ , from  $C$  to  $B$ , from  $C$  to  $D$ , and from  $A$  to  $D$  is characterized by  $b_1 = b_3 = 1$  and  $b_2 = b_4 = -1$  leading to the following values for the invariants:  $I_1^M = 0$ ,  $I_{11}^M = I_{13}^M = 4$ ,  $I_{12}^M = -4$ .

One can notice that the definition of graph invariants strongly depends on the definition of bond functions. We employ this degree of freedom and propose an alternative definition of graph invariants based on the occupation numbers. Instead of functions  $b_{ij}^M$  with  $i < j$ , we introduce twice as many functions  $d_{ij}^M$  defined for  $i \neq j$ :  $d_{ij}^M = 1$  if the edge between  $i$  and  $j$  is oriented towards  $j$  and  $d_{ij}^M = 0$  otherwise. That means, for example, that each existing H-bond is represented by two bond functions ( $d_{ij}^M$  and  $d_{ji}^M$ ), one of them produces 1, while the other 0. The obvious advantage of this alternative scheme is that dangling O–H bonds can be naturally incorporated into the definition and used for generating the graph invariants. Graph invariants from functions  $d$  are formed using the same projection operation as before (Eqs. 1, 2). At the same time, the set of graph invariants for systems of H-bonds without defects remains unchanged: both schemes lead to the same linearly independent graph invariants because for each existing H-bond  $b_{ij}^M = 2d_{ij}^M - 1$  and the other bond function  $d_{ji}^M = 1 - d_{ij}^M$ .

**Table 1** Number of invariants of different orders for H-bond networks in ice nanotubes

| System                        | 1-st order | 2-nd order | 3-rd order | 4-th order | 5-th order |
|-------------------------------|------------|------------|------------|------------|------------|
| INT <sub>3</sub> <sup>2</sup> | 1          | 7          | 10         | 52         | 94         |
| INT <sub>3</sub> <sup>3</sup> | 1          | 15         | 56         | 279        | 963        |
| INT <sub>3</sub> <sup>4</sup> | 1          | 26         | 140        | 928        | 4374       |
| INT <sub>3</sub> <sup>5</sup> | 2          | 40         | 303        | 2350       | 14,138     |
| INT <sub>3</sub> <sup>6</sup> | 2          | 57         | 534        | 5004       | 36,204     |
| INT <sub>3</sub> <sup>7</sup> | 3          | 77         | 886        | 9457       | 80,171     |
| INT <sub>3</sub> <sup>8</sup> | 3          | 100        | 1336       | 16,384     | 158,692    |
| INT <sub>3</sub> <sup>9</sup> | 4          | 126        | 1949       | 26,568     | 289,818    |
| INT <sub>4</sub> <sup>3</sup> | 1          | 24         | 91         | 644        | 2606       |
| INT <sub>4</sub> <sup>4</sup> | 1          | 42         | 236        | 2172       | 12,372     |
| INT <sub>4</sub> <sup>5</sup> | 2          | 65         | 513        | 5540       | 40,828     |
| INT <sub>4</sub> <sup>6</sup> | 2          | 93         | 916        | 11,848     | 106,300    |
| INT <sub>4</sub> <sup>7</sup> | 3          | 126        | 1523       | 22,456     | 237,790    |
| INT <sub>4</sub> <sup>8</sup> | 3          | 164        | 2312       | 38,984     | 474,744    |
| INT <sub>5</sub> <sup>3</sup> | 1          | 24         | 144        | 1083       | 5,931      |
| INT <sub>5</sub> <sup>4</sup> | 1          | 42         | 374        | 3797       | 28,641     |
| INT <sub>5</sub> <sup>5</sup> | 2          | 65         | 806        | 9896       | 95,302     |
| INT <sub>5</sub> <sup>6</sup> | 2          | 93         | 1441       | 21,455     | 249,987    |
| INT <sub>5</sub> <sup>7</sup> | 3          | 126        | 2388       | 41,049     | 561,823    |
| INT <sub>5</sub> <sup>8</sup> | 3          | 164        | 3628       | 71,753     | 1,126,283  |
| INT <sub>6</sub> <sup>3</sup> | 1          | 33         | 202        | 1857       | 11,529     |
| INT <sub>6</sub> <sup>4</sup> | 1          | 58         | 529        | 6540       | 56,735     |
| INT <sub>6</sub> <sup>5</sup> | 2          | 90         | 1141       | 17,082     | 190,510    |
| INT <sub>6</sub> <sup>6</sup> | 2          | 129        | 2048       | 37,081     | 503,133    |
| INT <sub>6</sub> <sup>7</sup> | 3          | 175        | 3396       | 71,003     | 1,135,599  |

becomes redundant. The schemes are different only when the system of H-bonds allows for defects like broken H-bonds.

Before we start our analysis of the quality of the energy approximation by graph invariants, it is reasonable to look at their total numbers for different systems. Table 1 shows these numbers for some ice nanotubes INT<sub>n</sub><sup>m</sup>. In particular, the dependence of the numbers of distinct graph invariants on the order of the invariant and the indices *m* and *n* is revealed. The values in the table can be found numerically by applying the graph invariant definition to a specific water cluster. Alternatively, one can determine these numbers analytically on the basis of symmetry considerations. Transformation of bond variables defines a (reducible) representation *T* of the symmetry group *G* of the cluster. The number of the *k*th order invariants is simply the number of the totally symmetric representations in the *k*th symmetric power of *T*. Thus, only characters of *T* are necessary to determine numbers in Table 1.

One can notice that the numbers of graph invariants grow very fast with their order. The number of fifth-order invariants is comparable and sometimes even exceeds the total number of H-bond configurations making their use unjustified. Even in the case of the highly symmetric 5<sup>12</sup> D-cage, the total number of fifth-order invariants is 2286, which is only one order of magnitude smaller than the total number of configurations. At the same time, such unfavourable ratio of the numbers of H-bond configurations and high-order graph invariants is caused by the exclusion of configurations with defects. Moreover, many of the graph invariants are linearly dependent when they are computed for the given set of H-bond networks.

We carried out calculation of standard second-order graph invariants for all the systems mentioned above. The calculations correspond to two cases: one takes into account only H-bonds, while the other is based on dangling O–H bonds. Table 2 presents the results of our calculations for all 24 water clusters. The first column shows numbers of H-bond configurations *N*<sub>conf</sub><sup>*M*</sup> satisfying the ice rules. When a graph invariant *I*<sub>α</sub><sup>*M*</sup> is applied to the set of H-bond configurations {*X*<sub>*k*</sub><sup>*M*</sup>}, it produces a vector of dimension *N*<sub>conf</sub><sup>*M*</sup> comprising numbers *I*<sub>α</sub><sup>*M*</sup>(*X*<sub>*k*</sub><sup>*M*</sup>). Some of these vectors are linearly dependent. The next two columns of Table 2 present numbers of linearly independent second-order graph invariants (determined for both systems of bonds mentioned above) with respect to H-bond configurations satisfying the ice rules.

The calculated energies of H-bond networks also constitute vectors *E*<sub>*k*</sub><sup>*M*</sup> of dimension *N*<sub>conf</sub><sup>*M*</sup>. The linear predictive model approximates the energies by linear combinations of graph invariant descriptors:

$$E_k^M = \sum_{\alpha} C_{\alpha}^M I_{\alpha}^M(X_k^M). \quad (5)$$

The coefficients *C*<sub>α</sub><sup>*M*</sup> are determined by the method of least squares for all 24 clusters. The quality of the approximation is characterized by *R*<sup>2</sup>, a coefficient of determination routinely used in statistics. The remaining two columns of Table 2 show *R*<sup>2</sup> as determined for the same two systems of graph invariants (H-bonds and O–H bonds). The quality of the approximation turns out to be strongly dependent on the type of the cluster. In particular, energies of clusters with a cage structure are well described by linear combinations of graph invariant descriptors. The invariants defined for H-bonds provide slightly better results than those defined for O–H bonds, but both approximations are highly accurate. Quite expectedly, the system of O–H bonds is not sufficient to describe a defect in a system of H-bonds. Therefore, energies of clusters with defects (2) and (3) are reasonably well described only with graph invariants for systems of H-bonds.

**Table 2** Description of H-bond networks in water clusters by second-order invariants: number of H-bond configurations  $N_{\text{conf}}^M$ ; number of linearly independent second-order invariants for dangling O–H bonds and H-bonds ( $N_{\text{inv}}^M(\text{O–H})$  and  $N_{\text{inv}}^M(\text{H-bond})$ ); the quality of the approximations ( $R^2$ -squared)

| System                        | $N_{\text{conf}}^M$ | $N_{\text{inv}}^M(\text{O–H})$ | $N_{\text{inv}}^M(\text{H-bond})$ | $R^2(\text{O–H})$ | $R^2(\text{H-bond})$ |
|-------------------------------|---------------------|--------------------------------|-----------------------------------|-------------------|----------------------|
| INT <sub>4</sub> <sup>3</sup> | 178                 | 5                              | 14                                | 0.15332           | 0.47787              |
| INT <sub>4</sub> <sup>4</sup> | 978                 | 5                              | 21                                | 0.26268           | 0.71627              |
| INT <sub>4</sub> <sup>5</sup> | 5588                | 5                              | 29                                | 0.04810           | 0.44210              |
| INT <sub>4</sub> <sup>6</sup> | 33,073              | 5                              | 39                                | 0.07379           | 0.49441              |
| INT <sub>4</sub> <sup>7</sup> | 198,706             | 5                              | 50                                | 0.06196           | 0.46588              |
| INT <sub>4</sub> <sup>8</sup> | 1,210,106           | 5                              | 63                                | 0.06761           | 0.54263              |
| INT <sub>5</sub> <sup>2</sup> | 102                 | 5                              | 9                                 | 0.98234           | 0.99138              |
| INT <sub>5</sub> <sup>3</sup> | 860                 | 5                              | 14                                | 0.19914           | 0.57754              |
| INT <sub>5</sub> <sup>4</sup> | 7480                | 5                              | 21                                | 0.39873           | 0.83217              |
| INT <sub>5</sub> <sup>5</sup> | 66,160              | 5                              | 29                                | 0.10688           | 0.52272              |
| INT <sub>5</sub> <sup>6</sup> | 591,328             | 5                              | 39                                | 0.12638           | 0.56749              |
| INT <sub>6</sub> <sup>2</sup> | 408                 | 7                              | 13                                | 0.96178           | 0.98239              |
| INT <sub>6</sub> <sup>3</sup> | 4962                | 7                              | 20                                | 0.19364           | 0.48253              |
| INT <sub>6</sub> <sup>4</sup> | 64,835              | 7                              | 30                                | 0.46010           | 0.85003              |
| INT <sub>6</sub> <sup>5</sup> | 865,243             | 7                              | 41                                | 0.11721           | 0.55329              |
| INT <sub>7</sub> <sup>2</sup> | 1474                | 7                              | 13                                | 0.94966           | 0.96614              |
| INT <sub>7</sub> <sup>3</sup> | 28,896              | 7                              | 20                                | 0.35136           | 0.38212              |
| INT <sub>7</sub> <sup>4</sup> | 580,200             | 7                              | 30                                | 0.35224           | 0.66997              |
| (1)                           | 30,026              | 5                              | 7                                 | 0.96378           | 0.96403              |
| (2)                           | 443,112             | 5                              | 8                                 | 0.46732           | 0.97290              |
| (3)                           | 2,772,313           | 5                              | 8                                 | 0.24939           | 0.93086              |
| (4)                           | 394,000             | 39                             | 78                                | 0.99290           | 0.99345              |
| (5)                           | 22,960              | 9                              | 36                                | 0.41787           | 0.63397              |
| (6)                           | 3,043,836           | 19                             | 35                                | 0.98046           | 0.98140              |

The quality of the approximation deteriorates when significant strains are present in the structure (like those in ice nanotubes). Figure 3 shows the correlation between the calculated energy and its graph invariant approximation for clusters (1) and (5). The former is not strained, while the latter has mild strains. In line with this observation, the quality of the approximation is better for stable INT<sub>5</sub><sup>m</sup> and INT<sub>6</sub><sup>m</sup> than for strained INT<sub>4</sub><sup>m</sup> and INT<sub>7</sub><sup>m</sup>. As for the indices  $m$ , the value of  $R^2$  depends on its parity:  $R^2$  is larger for even  $m$  than for odd  $m$ . Taking into account that the number of dangling O–H bonds does not depend on the size parameter  $m$  of INT<sub>n</sub><sup>m</sup>, it is not surprising that the quality of the energy approximation by graph invariants for O–H bonds is inadequate for large  $m$ . All these conclusions are valid only for calculations with the optimized spatial structure. If we consider idealized structures of ice nanotubes, the value of  $R^2$  becomes close to 1 because the energy of H-bond configurations becomes largely determined by the number of strained water molecules with both O–H bonds placed along the nanotube axis, and this parameter is expressible via second-order graph invariant descriptors.

There are different possibilities to improve the description based on increasing the number of graph invariants in play. One of them is to take a larger set of variables combining both H-bonds and dangling O–H bonds. This approach slightly increases the value of  $R^2$ : for example, in the case of cluster (2) it increases from 0.97290 to 0.97716, while in the case of cluster (3) it changes from 0.93086 to 0.93906. In the case of ice nanotubes, the effect of adding dangling O–H bonds to the set of variables is very small. Another obvious option is to increase the order of the invariants and this, indeed, increases the quality of the approximation, but

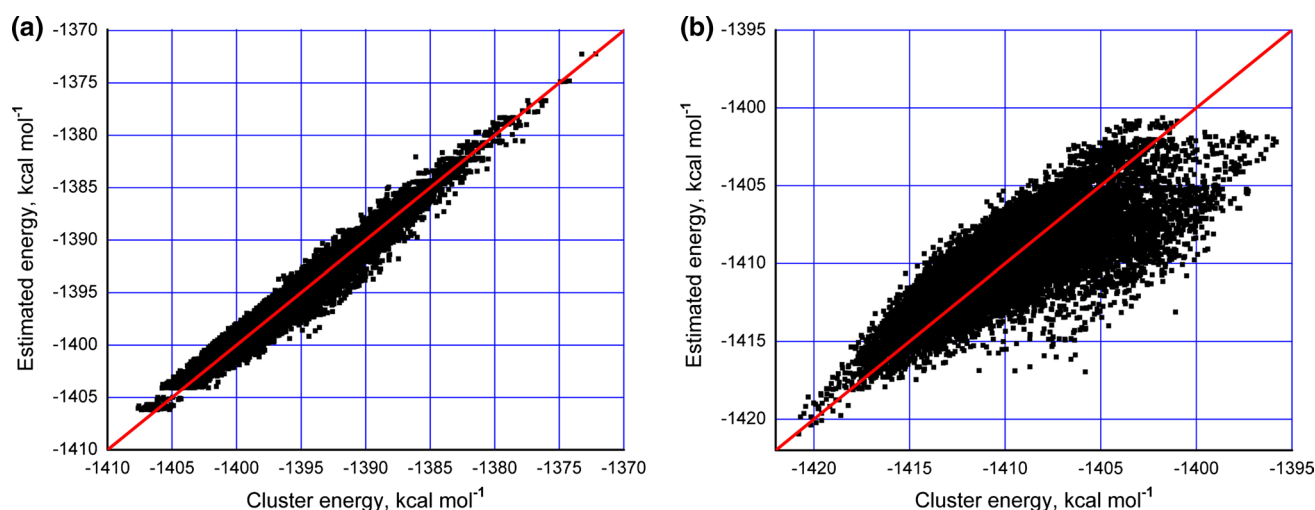


**Fig. 3** Correlation between the calculated energy and its graph invariant approximation for clusters (1) and (5)

it is necessary to bear in mind that it comes with a significant increase in the number of graph invariants. For example, in the case of  $\text{INT}_4^4$  the number of independent invariants increases from 21 to 57 and  $R^2$  changes from 0.71627 to 0.74785 when the second-order graph invariants for H-bonds are replaced by third-order ones. The increase in  $R^2$  is smaller when invariants for dangling O–H bonds are considered (for example,  $R^2$  changes from 0.26268 to 0.27331 when the second-order invariants are replaced by fourth-order invariants for the same ice nanotube  $\text{INT}_4^4$ ). Even a significant increase in the number of graph invariants doesn't necessarily bring serious improvement of the approximation. For example, in the case of graph invariants defined for dangling O–H bonds of water cage (1), the number of invariants increases from 5 to 58 when the order of invariants changes from 2 to 4, but the value of  $R^2$  is improved only marginally (from 0.96378 to 0.96907).

Some systems allow for specific approaches. In the case of mixed cluster (4), it is possible to separately consider the configurations with a dangling F–H bond and those with the H-bond  $\text{F–H}\cdots\text{O}$ . It turns out that the value of  $R^2$  is smaller for the first set (0.98550) than for the other (0.99435). Approaches employing graph invariants based on bond functions  $b$  and  $d$  are different in the case of clusters with broken H-bonds. Our calculations show that the change of the definition of the graph invariants does not bring significant improvements in terms of the quality of the approximation: the values of  $R^2$  for two schemes are relatively close (larger for that based on bond functions  $d$ ) and are equal to 0.97290 and 0.97763 for cluster (2) and 0.93086 and 0.94077 for cluster (3), respectively.

Finally, we provide a few thoughts about the general idea of graph invariants as given by totally symmetric projectors of the products of bond variables. Sometimes other representations of the symmetry group may be useful. An example is provided by ice nanotubes. They are thought to possess ferroelectric order with a net polarization when  $n$  is odd and to be “anti-ferroelectric” when  $n$  is even [26, 27]. This prediction is based on a simple model discarding most of the H-bond configurations. Alternatively, we can consider an idealized structure of ice nanotubes. In this case, the polarization parallel to the INT axis is proportional to the difference of the numbers of dangling O–H bonds on the INT ends, thereby confirming the original predictions. This parameter should also play a significant role for relaxed structures, but it corresponds to the representation  $A_1'$  for  $\text{INT}_n^m$  with an odd  $n$  and to the representation  $A_{1u}$  for  $\text{INT}_n^m$  with an even  $n$ . Therefore, analysis of properties like polarization may require specialized graph descriptors based on the symmetry of the system.

## 4 Conclusion

In summary, a very large database of the energies of H-bond networks in water clusters is constructed with the help of an ultra-fast semiempirical method based on strictly local electron groups. It includes water clusters of different symmetry and morphology, clusters with ideal and defected structures, systems where one water molecule is replaced with hydrogen fluoride. Two varieties of graph invariant descriptors are benchmarked against quantum-chemical calculations, one of them being proposed in the present paper. Both types of descriptors lead to identical results when applied to ideal networks satisfying ice rules but differ for structures with defects. We established that the use of high-order invariants is not justified unless H-bond networks significantly deviating from ice rules are taken into consideration. Approximation of the water cluster energies by linear combinations of second-order graph invariants shows that this approach gives surprisingly good results for clusters without significant strains (including those with defects) but fails when four-membered water rings are present in the structure.

**Acknowledgments** This work is partially supported by National Research Center “Kurchatov Institute”, Russian Science Foundation (Grant 14-19-00662), and Russian Foundation for Basic Research (Grants 13-07-00095 and 14-03-00867).

## References

1. Hoepfner V, Deringer VL, Dronskowski R (2012) Accurate hydrogen positions in organic crystals: assessing a quantum-chemical aide. *Cryst Growth Des* 12:1014–1021
2. Hoepfner V, Deringer VL, Dronskowski R (2012) Hydrogen-bonding networks from first principles: exploring the guanidine crystal. *J Phys Chem A* 116:4551–4559
3. Deringer VL, Englert U, Dronskowski R (2014) Covalency of hydrogen bonds in solids revisited. *Chem Commun* 50:11547–11549
4. Bernal JD, Fowler RH (1933) A theory of water and ionic solution, with particular reference to hydrogen and hydroxyl ions. *J Chem Phys* 1:515–548
5. Yoo S, Kirov MV, Xantheas SS (2009) Low-energy networks of the T-cage ( $\text{H}_2\text{O}$ )<sub>24</sub> cluster and their use in constructing periodic unit cells of the structure I (sI) hydrate lattice. *J Am Chem Soc* 131:7564–7566
6. Chihai V, Adams S, Kuhs WF (2004) Influence of water molecules arrangement on structure and stability of  $5^{12}$  and  $5^{12}6^2$  buckyball water clusters. A theoretical study. *Chem Phys* 297:271–287
7. Kirov MV, Fanourgakis GS, Xantheas SS (2008) Identifying the most stable networks in polyhedral water clusters. *Chem Phys Lett* 461:180–188
8. Anick DJ (2003) Application of database methods to the prediction of B3LYP-optimized polyhedral water cluster geometries and electronic energies. *J Chem Phys* 119:12442–12456

9. Tokmachev AM, Tchougréeff AL, Dronskowski R (2010) Hydrogen-bond networks in water clusters ( $\text{H}_2\text{O}$ )<sub>20</sub>: an exhaustive quantum-chemical analysis. *ChemPhysChem* 11:384–388
10. Kuo J-L, Coe JV, Singer SJ, Band YB, Ojamäe L (2001) On the use of graph invariants for efficiently generating hydrogen bond topologies and predicting physical properties of water clusters and ice. *J Chem Phys* 114:2527–2540
11. Kuo J-L, Singer SJ (2003) Graph invariants for periodic systems: towards predicting physical properties from the hydrogen bond topology of ice. *Phys Rev E* 67:016114
12. Singer SJ, Kuo J-L, Hirsch TK, Knight C, Ojamäe L, Klein ML (2005) Hydrogen-bond topology and the ice VII/VIII and ice Ih/XI proton-ordering phase transitions. *Phys Rev Lett* 94:135701
13. Knight C, Singer SJ, Kuo J-L, Hirsch TK, Ojamäe L, Klein ML (2006) Hydrogen bond topology and the ice VII/VIII and ice Ih/XI proton ordering phase transitions. *Phys Rev E* 73:056113
14. Knight C, Singer SJ (2006) A reexamination of the ice III/IX hydrogen bond ordering phase transition. *J Chem Phys* 125:064506
15. Knight C, Singer SJ (2008) Hydrogen bond ordering in ice V and the transition to ice XIII. *J Chem Phys* 129:164513
16. Knight C, Singer SJ (2009) Site disorder in ice VII arising from hydrogen bond fluctuations. *J Phys Chem A* 113:12433–12438
17. Tokmachev AM, Dronskowski R (2011) Hydrogen-bond networks in finite ice nanotubes. *J Comp Chem* 32:99–105
18. Tokmachev AM, Tchougréeff AL (2001) Semiempirical implementation of strictly localized geminals for analysis of molecular electronic structure. *J Comp Chem* 22:752–764
19. Tokmachev AM, Tchougréeff AL (2003) Fast NDDO method for molecular structure calculations based on strictly localized geminals. *J Phys Chem A* 107:358–365
20. Tokmachev AM, Tchougréeff AL (2005) Efficient multipole model and linear scaling of NDDO-based methods. *J Phys Chem A* 109:7613–7620
21. Tchougréeff AL, Tokmachev AM, Dronskowski R (2009) Multipole model for the electron group function method. *J Phys Chem A* 113:11406–11415
22. Surján PR (1999) An introduction to the theory of geminals. *Top Curr Chem* 203:63–88
23. Tokmachev AM (2015) Perspectives of geminal methods for large molecular systems. *Int J Quantum Chem*. doi:[10.1002/qua.24963](https://doi.org/10.1002/qua.24963)
24. Jeszenszki P, Nagy PR, Zoboki T, Szabados Á, Surján PR (2014) Perspectives of APSG-based multireference perturbation theories. *Int J Quantum Chem* 114:1048–1052
25. Tokmachev AM, Tchougréeff AL (2006) Group functions approach based on the combination of strictly local geminals and molecular orbitals. *Int J Quantum Chem* 106:571–587
26. Luo C, Fa W, Zhou J, Dong J, Zeng XC (2008) Ferroelectric ordering in ice nanotubes confined in carbon nanotubes. *Nano Lett* 8:2607–2612
27. Mikami F, Matsuda K, Kataura H, Maniwa Y (2009) Dielectric properties of water inside single-walled carbon nanotubes. *ACS Nano* 3:1279–1287