Topology-based Phase Identification of Bulk, Interface, and Confined Water using an Edge-Conditioned Convolutional Graph Neural Network

A. Moradzadeh[1], H. Oliaei[1], and N. R. Aluru[2*]

***Corresponding Author, e-mail: aluru@utexas.edu**

[1]Department of Mechanical Science and Engineering, University of Illinois at Urbana−Champaign, Urbana, IL, 61801 United States, [2]Oden Institute for Computational Engineering and Sciences, Walker Department of Mechanical Engineering, The University of Texas at Austin, Austin, TX, 78712 United States

Abstract

Water plays a significant role in various physicochemical and biological processes. Understanding and identifying water phases in various systems such as bulk, interface, and confined water is crucial in improving and engineering state-of-the-art nanodevices. Various order parameters have been developed to distinguish water phases, including bond-order parameters, local structure index, and tetrahedral order parameters. These order parameters are often developed with the assumption of homogenous bulk systems, while most applications involve heterogeneous and non-bulk systems, thus limiting their generalizability. Our study develops a methodology based on a graph neural network to distinguish water phases directly from data and to learn features instead of predefining them. We provide comparisons between baseline methods trained using conventional order parameters as features and a graph neural network model trained using radial distance and hydrogen-bonding information to study phase classification and continuous and discontinuous phase transitions of bulk, interface, and confined water.

I. Introduction

Water is an indispensable part of many physicochemical and biological phenomena, and its phase can significantly alter physicochemical and biological phenomena such as CO_2 reduction,[1] proton transport,[2] power generation[3], and water desalination[4]. Furthermore, properties of water such as the diffusion coefficient, dielectric permittivity, and density, as well as structural properties[5], depend on its phase. Therefore, it is of great importance to accurately identify different water phases. Similar to most other liquids, a significant understanding of water is developed through computational studies, where atomistic level data about the positions and velocities are available.[6,7] Therefore, the prediction of water phases from water molecule topology, *i.e.,* the configuration of other molecules around a tagged water molecule, is a task worth studying and understanding, especially for confined systems such as carbon nanotubes (CNTs) due to the technological applications of CNTs.

Due to the high dimensionality and uninterpretable nature of atomistic simulation data, researchers have developed a wide variety of order parameters to reduce dimensionality and predict the phase of a system from reduced dimensions. Motivated by the importance of water in various areas, water is studied through multiple order parameters (OPs) such as the bond-order parameter (BOP),[8,9] tetrahedral order parameter[10], and local-structure index.[11,12] Even though these order parameters are widely adopted in various studies about ice nucleation[13], phase discrimination/identification,[14] liquid–liquid transitions,[15,16] and free energy calculation[17], they are far from ideal. In many cases, it requires considerable domain expertise and effort to combine multiple order parameters to reach conclusive findings or even define new order parameters.[17,18] The problem is particularly pronounced for confined systems, as OPs are usually defined for homogenous and bulk systems, which is not the case for confined water. Furthermore, due to the interplay between fluid-fluid and fluid-wall interactions in a confined system,[19] confined systems have richer physics accompanied by anomalous behavior in the phase transition region, where both continuous and discontinuous phase transitions can occur (the discontinuous phase transition is characterized by a sharp change in the potential energy, enthalpy, or OP of the system, while a continuous phase transition shows only a critical point).[20,21]

Various computational and experimental studies have been performed to investigate and identify the phase behavior of confined water.[22,23] However, many challenges remain, one of which is to predict the phase of water directly from positional information, especially for confined systems.

Similar to the order parameter design in the phase identification task is the design of kernel and feature engineering in image, speech, and text processing applications, which require considerable domain expertise and human time.[24–26] During the last several decades, however, the process of kernel and feature engineering has been revolutionized by deep learning-based methods, which are adapted for a wide variety of applications in physics, chemistry, and biology. Water, as one of the most complex and important liquids, has been successfully studied using various deep learning methods in applications such as force field development and phase identification.[27–35] However, recent deep-learning methods still try to use traditional OPs as features for the phase identification of water, which does not address the issue of order parameter definition for nonhomogeneous systems.

The main bottleneck of phase identification stems from the nature of the data obtained from MD simulations. The data used for training the machine learning models should ensure that the input features are permutation, rotation, and translation invariant.[24] The atomic coordinates obtained from MD simulation do not possess these properties, which hinders the application of many conventional deep learning algorithms unless some sort of transformation is applied. Initial attempts to classify the water phase using deep learning-based methods started with a study where multiple features requiring multiple transformations are fed into a multilayer perceptron. However, the method requires arbitrary rotational transformations of the dataset to enforce the rotational invariance. Recent progress in graph neural networks (GNNs) provides a suitable tool to deal with atomistic data, as they are best described in a non-Elucidation space.[36] In addition to addressing permutation, rotation, and translation invariance, GNN addresses the variable size of the data, which is the case for confined water with different numbers of neighbors depending on the water phase and its distance from the wall. GCIceNet[37] was developed to solve the problem of rotational and permutational invariance using a GNN. Even though GCIceNet is successful, GCIceNet

constructs node features using OPs, which is an edge feature (it depends on the distance between atoms). Additionally, OPs used as node features are not well defined for confined systems.

In this study, we use the latest advances in GNNs, particularly edge-conditioned convolutional (ECC) graph neural networks, to address the problem of phase identification of water in bulk, interface, and confined systems in an end-to-end fashion.[38] In short, ECC is successfully applied to the point cloud dataset, which mimics the problem of phase identification in many ways. We formulate the phase-identification problem as a graph classification task and use the ECC layers to remove the need for human-engineered order parameters. To do so, we construct our graphs $G = \{V, E\}$ by collecting the oxygen atoms within the cutoff distance of a tagged oxygen atom. Based on the performance and computational cost, we keep all the oxygen atoms or several closest oxygen atoms. The oxygen atoms form the nodes (V) of the graph, and the pairwise distance between all oxygen atoms (nodes) is the edge feature. The node feature ($X \in \mathbb{R}^{|V|\times 2}$) is the one-hot encoded vector $\{0,1\}$, where the tagged oxygen atom, *i.e.*, the water molecule whose phase we want to predict, has a different node feature compared to its neighboring oxygen atoms. The graphs are fully-connected, i.e., every two nodes are connected together with an edge. We collect all the pairwise distances between all nodes as the edge feature of our graphs. Additionally, we collect information regarding the hydrogen bonds between water molecules by determining whether an edge corresponds to a donor-acceptor or acceptor-donor hydrogen bond as well as a no-hydrogen bond. Hydrogen bonding is incorporated as an edge feature, as hydrogens are omitted in the graph representation. Hydrogen bond information is extracted based on the methodology developed by Wernet *et al.*[39] as implemented in the MDTraj package.[40] In Wernet et al. methodology, donor (D) and acceptor (A) are both oxygen atoms, and one of the hydrogen atoms covalently bonded to donor oxygen forms hydrogen bond with the acceptor oxygen atom. The geometric criterion for hydrogen bond depends on the distance of donor and acceptor oxygen atoms (r_{DA}) and angle between donor, hydrogen, and acceptor atoms (δ_{HDA}). Mathematically, if $r_{DA} + 0.00044\,{\delta^2}_{HDA} < 0.33\,nm$ is true, hydrogen bond exists, otherwise there is no hydrogen bond between the donor oxygen, hydrogen covalently boned to donor, and acceptor oxygen. Therefore, any triplet

formed by an oxygen atom and its two covalently bonded hydrogen atoms with any other oxygen atoms and its hydrogens can participate in a hydrogen bond as a donor-acceptor or acceptor-donor hydrogen bond, as well as not forming any hydrogen bond

. In short, there are three possible conditions, which add three dimensions to each edge. The dimension of an edge, therefore, is a vector of size 4 ($E \in \mathbb{R}^{|V|\times|V|\times 4}$), three of which represent hydrogen bond as a one-hot encoding and one of which represents radial distance. The output of the graph classification task for bulk water is a vector of dimension n_c, where n_c is the number of different water phases in the dataset. We study 9 different phases of water (Ih, Ic, II, III, VI, VII, VIII, and IX Ices) as well as liquid water. For interface and confined systems, we use a similar input as for a bulk system, but the output is a binary value indicating whether water is liquid or solid. GNNs are trained to predict whether a particular configuration of atoms is liquid-like or solid-like inside CNTs or at the interface for various temperatures. We study CNT (10,10), inside which both continuous and discontinuous phase transitions can occur. Our reference solid and liquid systems used for training are picked from temperatures away from the phase transition temperature. The model successfully shows both sharp and smooth changes in the fraction of liquid-like molecules near the phase transition, allowing us to predict the phase transition temperature faster than normal methods. This is one of the significant advantages of our method, as previous studies need averaging over many trajectories to calculate the OP or thermodynamic properties to obtain phase transition temperatures. We refer to GNN model trained in this work as Top2Phase.

The rest of the paper is organized as follows. First, we describe the details of MD simulation and calculation of order parameters, followed by training of graph neural network and random forest models and comparison between their performance. Finally, we summarize the findings of our study.

II. Methods

MD simulations:

Molecular dynamics (MD) simulations of water in bulk and confinement are performed using the GROMACS package.[41] Water is modeled using the TIP4P/Ice model, as it performs better for phase transitions.[42] For water at the interface, we study the ice h/vapor interface, where a quasi-liquid layer[43] can form at the interface of solid and vapor due to the missing hydrogen bonding in the interface of solid and vapor. For confined cases, carbon-water interactions are modeled using the parameters from reference[44]. The temperature and pressure of the systems are controlled using the Nosè-Hoover thermostat with a time constant of 0.2 ps and the Parrinello-Rahman barostat with a time constant of 2.0 ps, respectively.[45] Initial configurations of the bulk system are generated using the GenIce package.[46] After energy minimization steps on the initial bulk configurations, MD simulation is performed for 25 ns at the corresponding temperature and pressure of the phase (see SI for the temperature and pressure of each phase). The data for machine learning model training as well as OPs are obtained from the last 10 ns of simulation. For the confined systems, we fill CNTs using a reservoir. Once filled, the isolated periodic CNT mimicking an infinite CNT is simulated at different temperatures by gradually decreasing the temperature from 390 *K* to 10 *K* at a rate of 1 K/ns . For every 10-*K* decrease, we simulate the system for 20 ns; again, the last 10 ns are used for postprocessing.

To compare the performance of the GNN with conventional machine learning methods as a baseline, we calculate OPs including the local-structure index (LSI), BOP, and tetrahedral OP. The LSI indicates the translational order of the system, and it considers $\left|\mathcal{N}\left(i, r_{cf} = 0.37\ nm\right)\right|$ neighboring water molecules by ordering them in ascending pairwise distances ($r_{j+1} > r_j \ \forall\, j \in \mathcal{N}\left(i, r_{cf} = 0.37\ nm\right)$). Mathematically, it is defined as

$$LSI = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} [\Delta(j) - \overline{\Delta}]^2 \tag{1}$$

where $\Delta(j)$ is the difference between the pairwise distance of two neighboring water molecules, *i.e.,* ($\Delta(j) = r_{j+1} - r_j$), and $\overline{\Delta}$ is the average value of $\Delta(j)$.

The BOP of order l (q_l) is the other OP used in the baseline machine learning method, where it is a coarse-grained representation of Steinhardt parameter q_{lm},[9,14] which can be expressed as follows:

$$q_{lm}(i) = \frac{1}{\left|\mathcal{N}\left(i,\ r_{cf} = r_{cf,6}\right)\right|} \sum_{j \in \mathcal{N}(i)} Y_{lm}(\theta_{ij}, \phi_{ij}) \tag{2}$$

where Y_{lm} is the spherical harmonic function of degree l and order m. θ_{ij} and ϕ_{ij} are polar angles. The cutoff distance ($r_{cf} = r_{cf,6}$) of the neighbor list is chosen such that $\left|\mathcal{N}\left(i,\ r_{cf} = r_{cf,6}\right)\right|$ equals 6. The BOP of order l and degree m is defined as

$$Q_{lm}(i) = \frac{1}{\left|\mathcal{N}\left(i,\ r_{cf} = r_{cf,6}\right)\right| + 1}\left(q_{lm}(i) + \sum_{j \in \mathcal{N}(i)} q_{lm}\left(\theta_{ij}, \phi_{ij}\right)\right) \tag{3}$$

which is coarse-grained by averaging over degree through the following expression:

$$q_l(i) = \sqrt{\frac{4\pi}{2l+1} \sum_{m=-l}^{l} |Q_{lm}|^2} \tag{4}$$

The tetrahedral OP is defined based on the four nearest molecules and takes a value between 0 and 1, where 0 and 1 correspond to an ideal gas and perfect tetrahedron, respectively. It can be expressed as

$$q_{te} = 1 - \frac{3}{8} \sum_{j=1}^{3} \sum_{k=j+1}^{4} \left(\cos \psi_{jk} + \frac{1}{3}\right) \tag{5}$$

where ψ_{jk} is the angle formed from the tagged molecule and two of the four closest water molecules. BOPs and tetrahedral OPs calculated using the PyBoo package are used to train the baseline machine learning models.[47] In this study, we use the random forest as our baseline.[48] We also store the pairwise distance between atoms within a cutoff distance from a tagged water molecule as an edge feature for Top2Phase training along with the presence and type of hydrogen bonding as a one-hot-encoded vector.

Machine Learning

To investigate the performance of Top2Phase over other machine learning algorithms, we train a baseline machine learning method, *i.e.,* random forest (RF). Before going into the details of our training, we describe the task and procedure we have taken. We treat the problem as a classification problem, and for a bulk system, our output is a one-hot encoded vector of 9 different phases. The task in the confined and interface systems, however, is simplified to the identification of solid and liquid phases, where the output is binary, *i.e.,* 0 or 1. The data for classification are selected from multiple temperatures, both above and below the melting temperature of the TIP4P/Ice models, exposing the model to both liquid and solid phases. The performance of the method is of special interest for confined systems, where determination of the phase transition temperature from the conventional method (first-order change in order parameters) is not straightforward and requires a long simulation. It can also suffer from significant noise, as the timescale of phase transition can be large for continuous phase transition. However, the model can still predict phase behavior well (more quantitative analyses are provided later). We also compare GCIceNet with Top2Phase model in the supporting information to show generalizability of current framework in challenging cases.

Random Forest

The random forest algorithm (RF), as an ensemble learning method, selects a subset of features (in our cases, a vector of dimension 7, composed of LSI and tetrahedral order parameters as well as BOPs with degrees of $\{4,6,8,10,12\}$. For each selected subset, a decision tree is trained by randomly selecting a subset of features and dataset, followed by the construction of decision trees. Once the training of various trees is done, a majority vote is taken to determine the class for a given data. We apply grid search with 5-fold cross-validation to obtain the optimal depth and number of trees for RF.

Graph Neural Network

The graphs in this study are denoted by $G = \{V, E\}$, where V is set of nodes (oxygen atoms of water molecules) and each node has a feature vector of two. We construct for each timestep of MD trajectory as

many graphs as the number of water molecules, the central water molecule tagged for phase identification has node feature different from its neighboring water molecules, *i.e.,* node feature has a dimension of 2 corresponding to one-hot-encoding based on whether the node is a tagged oxygen atom or a neighboring oxygen atom $x_i \in \{(1,0),(0,1)\}$ (one might choose an embedding layer prior to graph convolution layer, but here we use a simple encoding with 2 dimensions instead, conveying the same information for tagged and its neighbors as a {0, 1} labeling leads to a very poor performance and slow training). The graph in this study is fully connected, *i.e.*, every node has an edge with other nodes in the graph. Every edge has four dimensions built by concatenation of pairwise distance between every pair of nodes and one-hot-encoding of hydrogen bonding of every edges. Edges set $E = \{e_{ij} \in R^4 | i,j \in V\}$ is the set of edges with 4 attributes, i.e., the pairwise distance and the one-hot-encoding of hydrogen bonding corresponding to donor-acceptor, acceptor-donor, or no-hydrogen-bond cases. Note that depending on the computational cost and classification performance, we use different numbers of water molecules to form the graph (see Figure 1 for schematic representations of graph construction).

In general, most of the GNN methods belong to the message-passing networks, which utilize combinations of message, aggregation, and update.[36] In this study, we use the ECC layer to build the Top2Phase model.[38] The hidden representation of nodes h^l at layer l is equal to a weighted sum of hidden representation h^{l-1} in its neighborhood. The weights in the ECC are generated by another network, also known as filter generating networks, which is usually modeled with an MLP with trainable parameters. Mathematically, the following operations are performed in the *l*-th layer:

$$h^l(i) = h^{l-1}(i)W^l_{root} + \sum_{j \in N(i)} h^{l-1}(j)\mathrm{MLP}\left(e_{ji}, w^l\right) + b^l \tag{6}$$

where b^l is the *l*-th layer bias, and $N(i)$ is the neighborhood of node i ($N(i) = \{j; (j,i) \in E\}$). h^l are embeddings of the *l*-th layer. Note that h^0 corresponds to the input feature x, and w^l are learnable parameters of the multilayer perceptron, *i.e.,* weight generating function (MLP above). W^l_{root} are learnable weights corresponding to the contribution from the hidden representation of the *i*-th node itself. After 3 or

4 ECC layers, we use a pooling function (sum pooling) to find a representation for each graph. The role of pooling is to reduce node embeddings of the whole graph into a single vector. Additionally, the pooling layer should be invariant to the permutation of nodes, and we use the sum function as our pooling layer. The pooled representation is fed to a multilayer perceptron with 1 hidden layer. All the layers, except the last multilayer perceptron, use the ReLU activation function. Further details regarding the structure of the layers are given in the SI.

To train the parameters of the Top2Phase model, we use either binary- or categorical-cross-entropy losses defined as

$$\mathcal{L} = -\sum_{i \in n_c} p_i \log v_i \tag{7}$$

where p_i and $\log v_i$ are the i-th element of vectors with dimensions equal to the number of classes (n_c) representing one-hot-encoding and Top2Phase predictions, respectively. The Adam optimizer is used to train the model with a learning rate of 0.00005 for 100000 epochs with early stopping if the loss is saturated for 5 consecutive epochs on the validation dataset (0.2 of the dataset). A batch size of 16, 32, or 64 is used depending on the computational cost. The Spektral and TensorFlow packages[49,50] are used to build and train Top2Phase model, and the Top2Phase package is developed as a Python package for broader usage (see the GitHub link).

III. Results and Discussions

We first perform data-exploratory analysis to examine the sufficiency of BOPs to separate different bulk water phases (bulk systems are shown in Figure 2a). To do so, we obtain 2D scatter plots of BOPs for two sets of degrees, namely, (q_6, q_8) and (q_{10}, q_{12}). In Figure 2b, we show the results of the analysis, where we observe a large overlap between any selected BOPs. Following this step, we train RF methods with different numbers of trees and depths and find the optimal RF parameters. The model reaches an average accuracy of 89.2 percent. Training the Top2Phase model with the same dataset leads to an average accuracy of 99.9 percent. Figure 3 shows a more quantitative analysis of the accuracy of both the Top2Phase and RF

models based on the confusion matrix. The confusion matrix shows the percentage of dataset misclassified for off-diagonal elements, and diagonal elements show the percentage of correctly classified samples per class. The Top2Phase confusion matrix (Figure 3b) shows far superior behavior, as the off-diagonal elements are far less than their counterparts in the RF confusion matrix (Figure 3b).

Along the same lines, we study the ice h/vapor system, where we simulate the system for the temperature range of 10 K to 300 K with a 10-K step (see Figure 4 for exploratory data analysis as well as the schematic representation of water at different temperatures before and during phase transition). The experimental and computational investigations show the formation of a quasi-liquid layer at the ice h/vapor interface.[43,51–53] The predicted melting temperature from the experiment and simulation is approximately 270 ± 5 K. We select temperatures of [10,140] K and [290, 300] K as our reference solid and liquid systems for ML training. After training the model, which shows an accuracy of 99% for the Top2Phase and 97% for the RF, we feed the data from other temperatures to predict the melting temperature and compare the classification results obtained from the RF and Top2Phase. To do so, we compare the liquid fraction at each temperature using both the RF and Top2Phase model in Figure 5. Additionally, we show the potential energy of the system, which shows a sharp change near the melting temperature. The Top2Phase model predicts a melting temperature of 275 K, RF predicts a melting temperature of 275 K, and the potential energy indicates a melting temperature of 275.0 ± 2.5 K. The behavior of the Top2Phase prediction is monotonic within the temperature range of our study, showing an increasing number of liquid-like molecules, while RF shows a non-monotonic and inconsistent behavior with increasing temperature.

Next, we study confined systems, where we simulate water confined inside a (10,10) CNT for the temperature range of 10 K to 390 K with a 10-K step (see Figure 6 for exploratory data analysis as well as the schematic representation of waters in liquid and solid phases of water with average densities of 16.75 nm^{-3} and 19.14 nm^{-3}). Confined water in general shows more complex behavior compared to bulk water; for example, with an increase in density, the phase transition becomes continuous, especially for CNTs with a smaller diameter. This phenomenon is usually attributed to the interplay between interface-water and

water–water interactions. As shown in Figure 6a-d, the solid phase of water inside the (10,10) CNT shows heptagons and heptagons with single-file water, respectively, at densities of 16.75 nm^{-3} (low) and 19.14 nm^{-3} (high), while the liquid phases of both densities inside the CNT look like each other. The larger overlap of the scatter plots of BOPs in high-density cases shown in Figure 6e-h indicates difficulty in using BOPs. Note that the high-density case corresponds to a continuous phase transition. The predicted melting temperatures for low- and high-density cases are approximately 270 ± 10 K and 290 ± 10, respectively. This trend is consistent with previous computational studies. Similar to the ice h/vapor case, we select two representative temperatures for both liquid and solid; in this case, we use [10,150] K and [310, 380] K as our reference solid and liquid systems, respectively. After training the model, the Top2Phase model achieves 0.994 and 0.949 accuracy, respectively, for low- and high-density cases. The accuracies of RF are 0.997 and 0.809 for low and high densities, respectively, lower than that of GNN accuracies (see SI for confusion matrix and more details on model performance). The Top2Phase model outperforms RF model in the high-density case, thereby demonstrating the capability of the GNN in complex cases. The lower accuracy of RF is attributed to the large overlap of BOPs, as shown in Figure 6 e-h. Top2Phase, however, learns its featurization based on data and does not face many difficulties in distinguishing solid and liquid phases. Once the models are trained, we feed the data from other temperatures to predict melting temperature and compare the results obtained from RF and Top2Phase. In Figure 7a-b, we compare the liquid fraction at each temperature using both RF and Top2Phase models for low- and high-density cases (we also compare the performance of Top2Phase and GCIceNet in the supporting information). Additionally, we show the potential energy and axial diffusion coefficient, which exhibit a sharp change near the melting temperature for low density and a smooth transition for high density, signatures of discontinuous and continuous phase transitions. The predicted melting temperatures are close to the MD simulation results using both the RF and Top2Phase models. However, the behavior of the Top2Phase is more monotonic with temperature change, as shown in Figure 7a-b. For example, the fractions of liquid-like molecules are a non-decreasing function of temperature, while RF shows a nonmonotonic and inconsistent behavior with increasing temperature. We also note that the larger deviation of both Top2Phase

and RF for the high-density case can be attributed to the difficulty in reaching complete equilibrium in the high density as the timescale of relaxation of simulation is large. Overall, the results of the high-density case prove the abilities of the Top2Phase in more complex environments.

IV. Conclusions

In this study, we trained a graph neural network model to classify different phases of water in bulk, interfacial, and confined environments. To address the issue with the definition of order parameters in the confined environments, we trained the model to learn features from the positional data, *i.e.,* the distance between the oxygen atom of tagged molecules and all other water molecules oxygen atoms within a cutoff distance. We augmented the edge features with hydrogen-boding information (acceptor-donor, donor-acceptor, or lack of hydrogen-bonding), as hydrogen atoms are coarse-grained in the graph representation. The results showed successful employment of the model in bulk, interfacial, and confined water inside a carbon nanotube, especially in terms of its generalization compared to the baseline method trained using the classical order parameters model. Furthermore, the predicted melting temperature and behavior of the model in both continuous and discontinuous phase transitions inside carbon nanotubes were in good agreement with the change in the potential energy and dynamics of waters. In summary, the methodology presented here provides a robust data-driven tool to classify and study the phase behavior of complex systems. Code and data are available for practitioners at https://github.com/moradza/Top2Phase.

Supporting Information Available

The Supporting Information is available free of charge on the ACS Publications website at DOI: XXXXXX and includes the reference Systems with thermodynamic conditions of water phases, confusion matrix of the CNT, comparison of Top2Phase with GCIceNet, and embedding of different phases of water (t-SNE plots) (PDF).

Acknowledgments

This work was supported by the National Science Foundation under Grants 2140225 and 2137157. The authors acknowledge the use of Blue Waters supercomputing resources at the University of Illinois at Urbana-Champaign. Furthermore, this work partially used the Extreme Science and Engineering Discovery Environment (XSEDE) Stampede2 at the Texas Advanced Computing Center through allocation TG-CDA100010. This work also utilized resources supported by the National Science Foundation's Major Research Instrumentation program, grant #1725729, as well as by the University of Illinois at Urbana-Champaign.

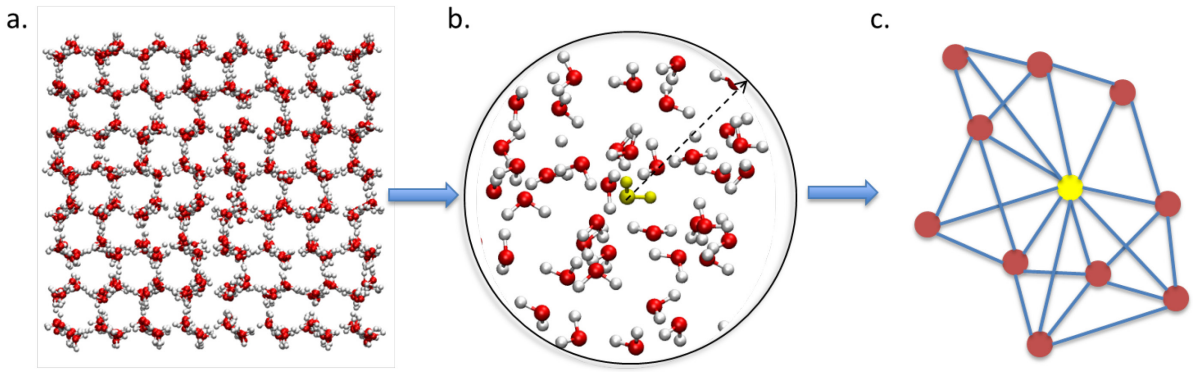


Figure 1. Schematic representation of water and the corresponding graph representing the water structure. a. atomistic configuration b. neighbor list formation based on a tagged water molecular c. graph representation with nodes as oxygen atoms, and edges representing connection with blue color. Each edge has four dimensions, representing distance and H-bond. Node color represents whether it is the tagged molecule or not.

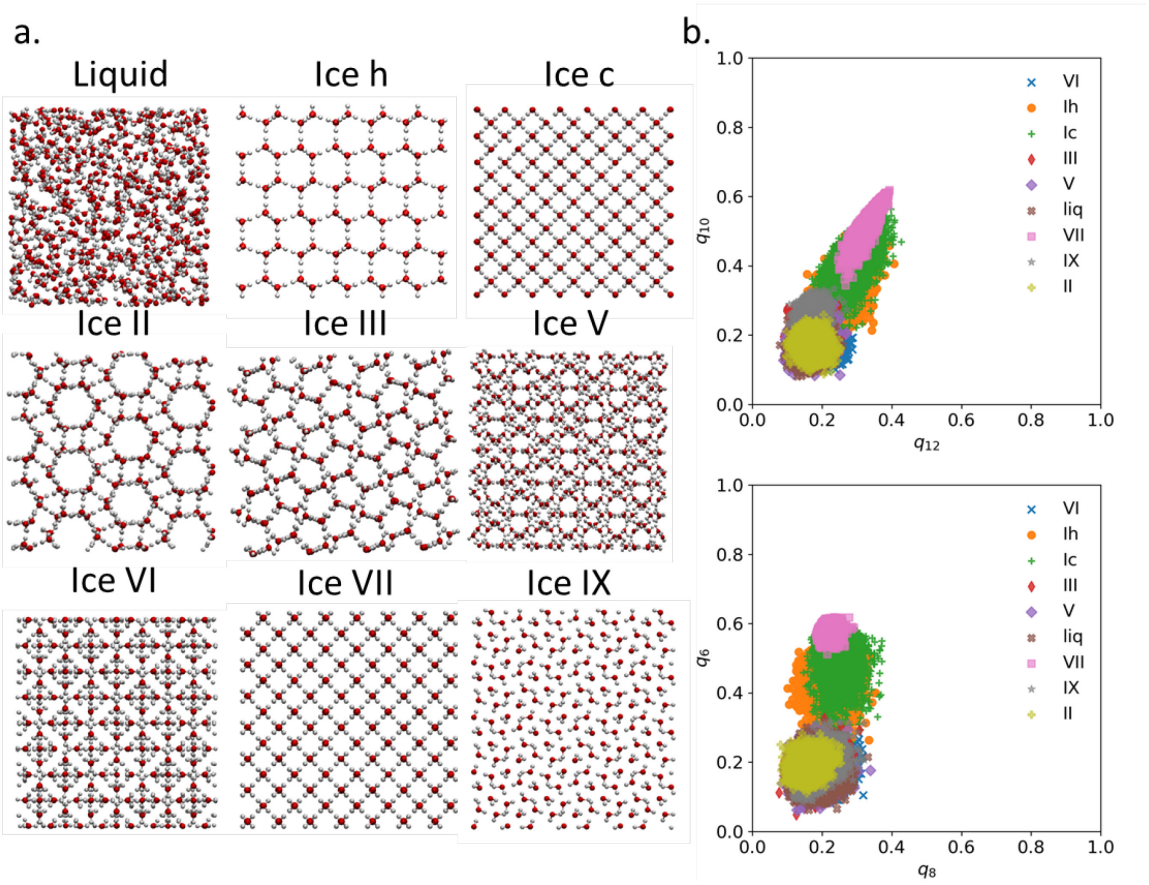


Figure 2. Exploratory data analysis of bulk water phases. a. Schematic representation of different bulk water phases b. scatter plot of (q_{10}, q_{12}) and (q_6, q_8) of bulk phases of water.

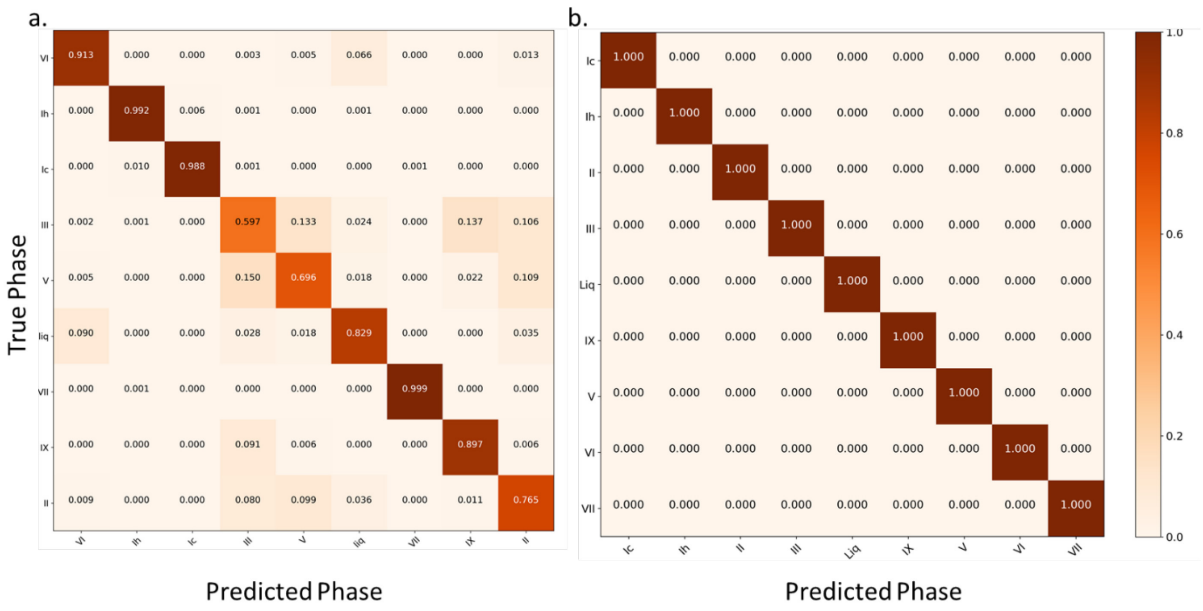


Figure 3. Confusion matrix for classification of bulk water. a. using RF. b. using Top2Phase.

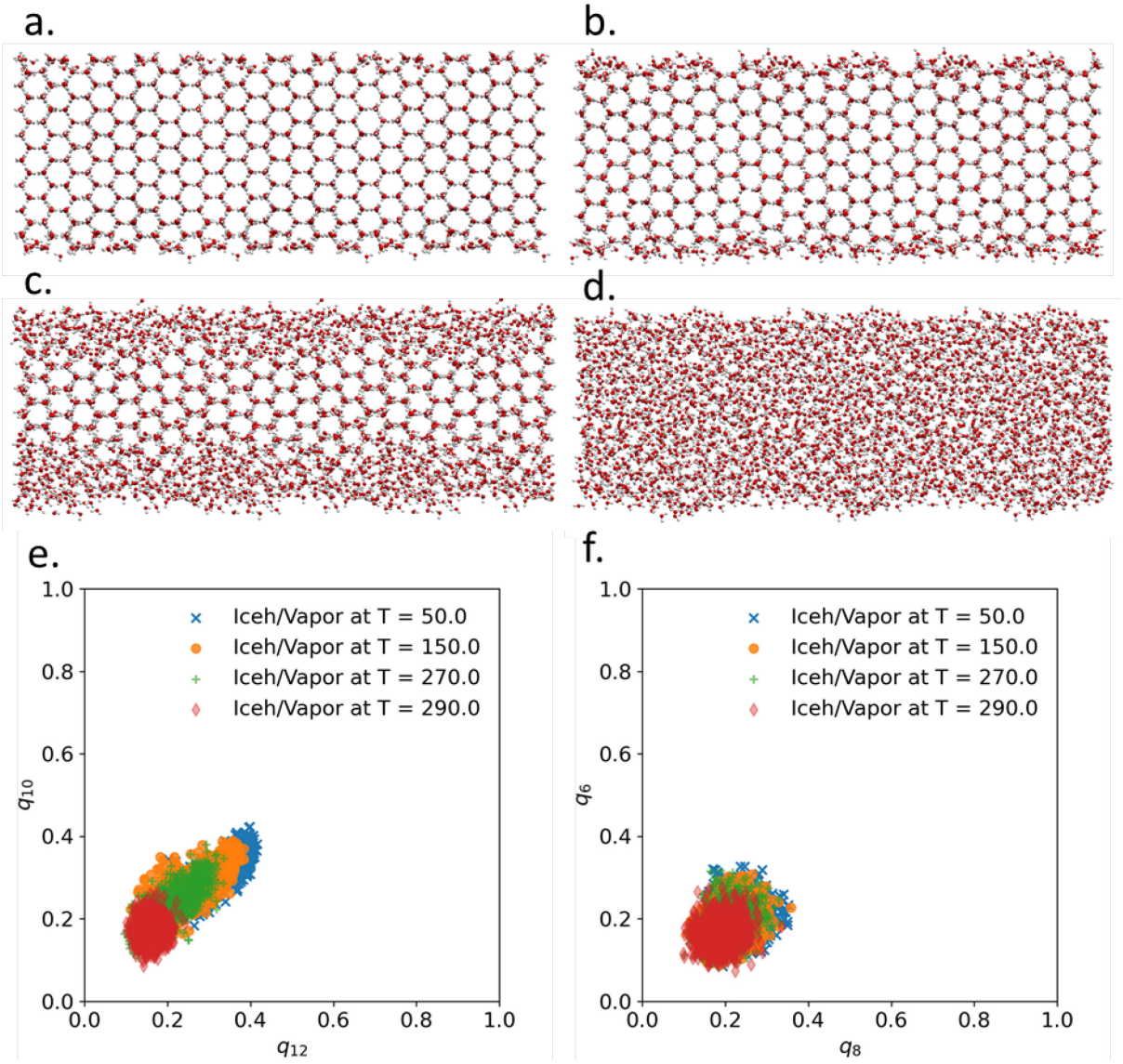


Figure 4. Schematic representation of Ice h/vapor interface at different temperatures along with exploratory data of analysis. a. configuration at 50 K b. configuration at 150 K c. configuration at 270 K d. configuration at 290 K e. scatter plot of (q_{10}, q_{12}) at different temperatures f. scatter plot of (q_6, q_8) at different temperatures. Distribution of order parameter pairs shows significant overlap at different temperatures.

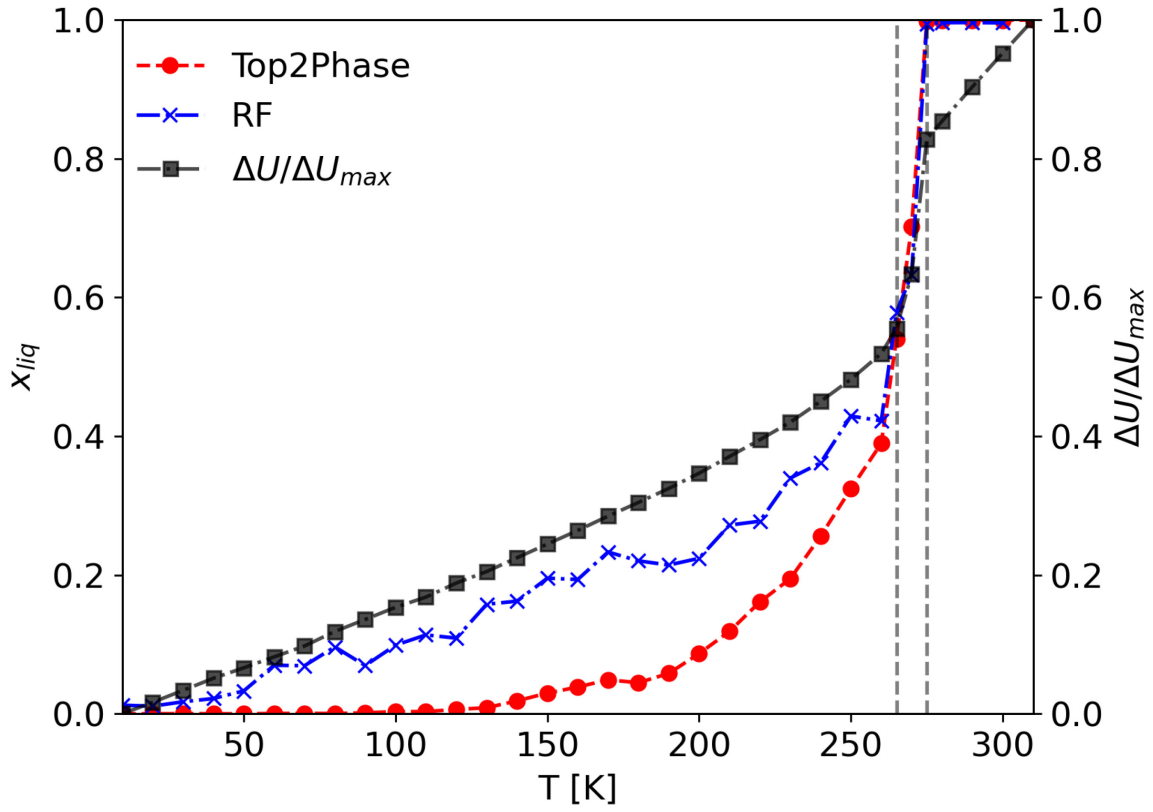


Figure 5. Phase transition of Ice h/vapor system. The black line with squares shows scaled potential energy of water at different temperatures. The red circle and blue cross represent fraction of liquid-like molecules at different temperatures obtained using RF and Top2Phase, respectively. Dashed vertical lines indicate temperature range at which all Ice-like molecules disappear due to temperature increase.

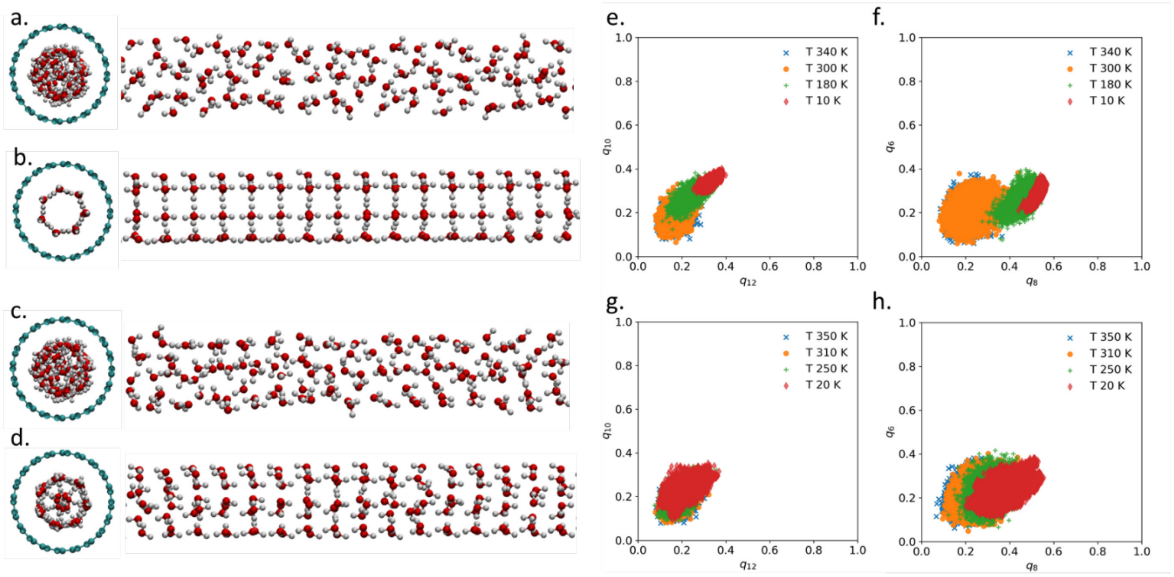


Figure 6. Schematic representation of confined water inside a (10,10) CNT at different temperatures and densities along with exploratory data analysis. a. liquid water configuration at 350 K and 16.75 nm^{-3}. b. solid water configuration at 20 K and 16.75 nm^{-3}. c. liquid water configuration at 350 K and 19.14 nm^{-3}. d. solid water configuration at 20 K and 19.14 nm^{-3}. e. scatter plot of (q_{10}, q_{12}) at different temperatures and density of 16.75 nm^{-3}. f. scatter plot of (q_6, q_8) at different temperatures and density of 16.75 nm^{-3}. g. scatter plot of (q_{10}, q_{12}) at different temperatures and density of 19.14 nm^{-3}. h. scatter plot of (q_6, q_8) at different temperatures and density of 19.14 nm^{-3}. Distribution of order parameter pairs show significant overlap for different phases.

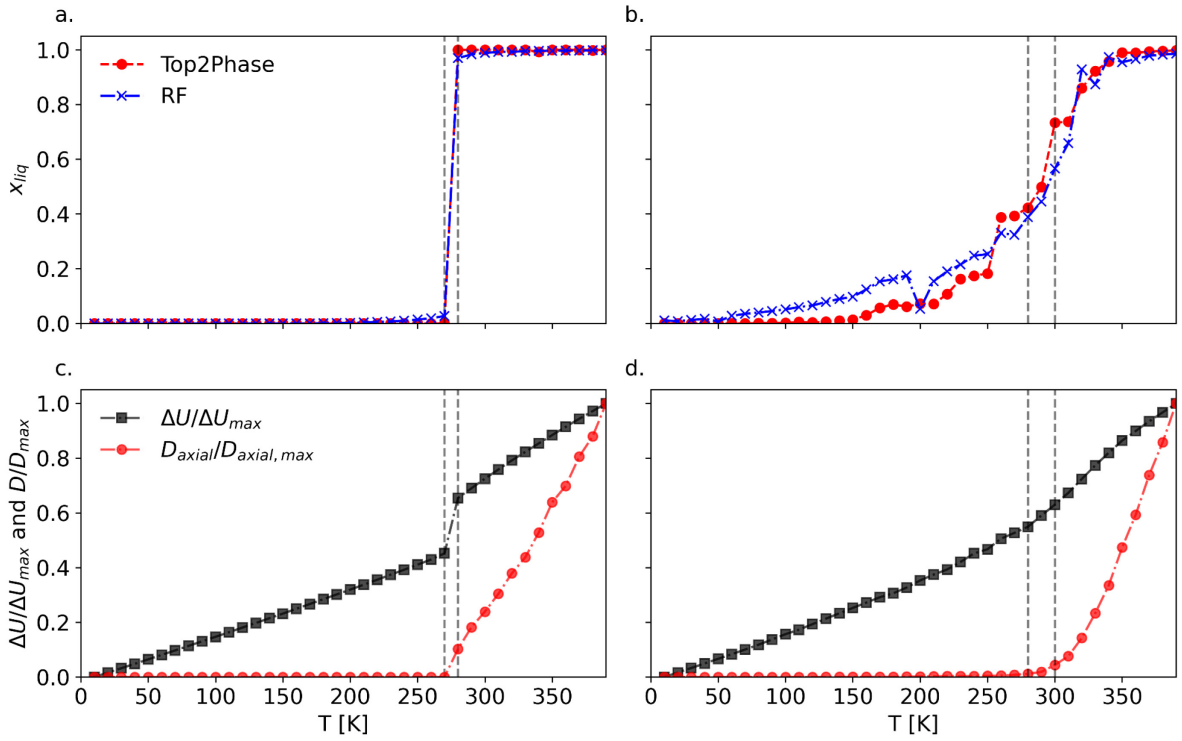


Figure 7. Phase transition of confined water with discontinuous and continuous phase transition. Comparison between the normalized potential energy and diffusion coefficient change at different temperatures and two densities. a. fraction of liquid-like molecules for CNT with density of 16.75 nm^{-3} predicted using RF and Top2Phase models. b. fraction of liquid-like molecules for CNT with density of 19.14 nm^{-3} predicted using RF and Top2Phase models. c. potential energy and diffusion coefficient at different temperatures for CNT with density of 16.75 nm^{-3}. d. potential energy and diffusion coefficient at different temperatures for CNT with density of 19.14 nm^{-3}. vertical lines show phase transition region. In a-b, red circles and blue crosses show results of RF and Top2Phase models, respectively. In c-d, potential energy and diffusion coefficients are show with black squares and red circles, respectively.

References

(1) Asadi, M.; Kim, K.; Liu, C.; Addepalli, A. V.; Abbasi, P.; Yasaei, P.; Phillips, P.; Behranginia, A.; Cerrato, J. M.; Haasch, R.; Zapol, P.; Kumar, B.; Klie, R. F.; Abiade, J.; Curtiss, L. A.; Salehi-Khojin, A. Nanostructured Transition Metal Dichalcogenide Electrocatalysts for CO2 Reduction in Ionic Liquid. *Science* **2016**, *353* (6298), 467–470. https://doi.org/10.1126/science.aaf4767.

(2) Janssen, G. J. M. A Phenomenological Model of Water Transport in a Proton Exchange Membrane Fuel Cell. *J Electrochem Soc* **2001**, *148* (12), A1313. https://doi.org/10.1149/1.1415031.

(3) Feng, J.; Graf, M.; Liu, K.; Ovchinnikov, D.; Dumcenco, D.; Heiranian, M.; Nandigana, V.; Aluru, N. R.; Kis, A.; Radenovic, A. Single-Layer MoS2 Nanopores as Nanopower Generators. *Nature* **2016**, *536* (7615), 197–200. https://doi.org/10.1038/nature18593.

(4) Heiranian, M.; Farimani, A. B.; Aluru, N. R. Water Desalination with a Single-Layer MoS2 Nanopore. *Nat Commun* **2015**, *6*, 8616. https://doi.org/10.1038/ncomms9616.

(5) Errington, J. R.; Debenedetti, P. G. Relationship between Structural Order and the Anomalies of Liquid Water. *Nature* **2001**, *409* (6818), 318–321. https://doi.org/10.1038/35053024.

(6) P, S.; J, K. Computer Simulation of Liquids. *J Mol Liq* **1988**, *38* (3–4), 267. https://doi.org/10.1016/0167-7322(88)80022-9.

(7) Bacher, A. K.; Schrøder, T. B.; Dyre, J. C. Explaining Why Simple Liquids Are Quasi-Universal. *Nat Commun* **2014**, *5* (1), 5424. https://doi.org/10.1038/ncomms6424.

(8) Lechner, W.; Dellago, C. Accurate Determination of Crystal Structures Based on Averaged Local Bond Order Parameters. *Journal of Chemical Physics* **2008**, *129* (11), 114707. https://doi.org/10.1063/1.2977970.

(9) Steinhardt, P. J.; Nelson, D. R.; Ronchetti, M. Bond-Orientational Order in Liquids and Glasses. *Phys Rev B* **1983**, *28* (2), 784–805. https://doi.org/10.1103/PhysRevB.28.784.

(10) Monroe, J. I.; Shell, M. S. Decoding Signatures of Structure, Bulk Thermodynamics, and Solvation in Three-Body Angle Distributions of Rigid Water Models. *Journal of Chemical Physics* **2019**, *151* (9), 094501. https://doi.org/10.1063/1.5111545.

(11) Truskett, T. M.; Torquato, S.; Debenedetti, P. G. Towards a Quantification of Disorder in Materials: Distinguishing Equilibrium and Glassy Sphere Packings. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* **2000**, *62* (1 B), 993–1001. https://doi.org/10.1103/PhysRevE.62.993.

(12) Duboué-, E.; Dijon, D.-; Laage, D. Characterization of the Local Structure in Liquid Water by Various Order Parameters. *J. Phys. Chem. B* **2015**, *119*, 47. https://doi.org/10.1021/acs.jpcb.5b02936.

(13) Haji-Akbari, A.; Debenedetti, P. G. Computational Investigation of Surface Freezing in a Molecular Model of Water. *Proc Natl Acad Sci U S A* **2017**, *114* (13), 3316–3321. https://doi.org/10.1073/pnas.1620999114.

(14) Lechner, W.; Dellago, C. Accurate Determination of Crystal Structures Based on Averaged Local Bond Order Parameters. *Journal of Chemical Physics* **2008**, *129* (11), 114707. https://doi.org/10.1063/1.2977970.

(15) Hamm, P. Markov State Model of the Two-State Behaviour of Water. *Journal of Chemical Physics* **2016**, *145* (13), 134501. https://doi.org/10.1063/1.4963305.

(16) Singh, R. S.; Biddle, J. W.; Debenedetti, P. G.; Anisimov, M. A. Two-State Thermodynamics and the Possibility of a Liquid-Liquid Phase Transition in Supercooled TIP4P/2005 Water. *Journal of Chemical Physics* **2016**, *144* (14), 144504. https://doi.org/10.1063/1.4944986.

(17) Niu, H.; Yang, Y. I.; Parrinello, M. Temperature Dependence of Homogeneous Nucleation in Ice. *Phys Rev Lett* **2019**, *122* (24), 245501. https://doi.org/10.1103/PhysRevLett.122.245501.

(18) Piaggi, P. M.; Parrinello, M. Calculation of Phase Diagrams in the Multithermal-Multibaric Ensemble. *Journal of Chemical Physics* **2019**, *150* (24), 244119. https://doi.org/10.1063/1.5102104.

(19) Aydin, F.; Moradzadeh, A.; Bilodeau, C. L.; Lau, E. Y.; Schwegler, E.; Aluru, N. R.; Pham, T. A. Ion Solvation and Transport in Narrow Carbon Nanotubes: Effects of Polarizability, Cation-ΠInteraction, and Confinement. *J Chem Theory Comput* **2021**. https://doi.org/10.1021/acs.jctc.0c00827.

(20) Raju, M.; Van Duin, A.; Ihme, M. Phase Transitions of Ordered Ice in Graphene Nanocapillaries and Carbon Nanotubes. *Sci Rep* **2018**, *8* (1), 1–11. https://doi.org/10.1038/s41598-018-22201-3.

(21) Takaiwa, D.; Hatano, I.; Koga, K.; Tanaka, H. Phase Diagram of Water in Carbon Nanotubes. *Proc Natl Acad Sci U S A* **2008**, *105* (1), 39–43. https://doi.org/10.1073/pnas.0707917105.

(22) Algara-Siller, G.; Lehtinen, O.; Wang, F. C.; Nair, R. R.; Kaiser, U.; Wu, H. A.; Geim, A. K.; Grigorieva, I. V. Square Ice in Graphene Nanocapillaries. *Nature* **2015**, *519* (7544), 443–445. https://doi.org/10.1038/nature14295.

(23) Pugliese, P.; Conde, M. M.; Rovere, M.; Gallo, P. Freezing Temperatures, Ice Nanotubes Structures, and Proton Ordering of TIP4P/ICE Water inside Single Wall Carbon Nanotubes. *Journal of Physical Chemistry B* **2017**, *121* (45), 10371–10381. https://doi.org/10.1021/acs.jpcb.7b06306.

(24) Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep Learning for Computational Chemistry. *J Comput Chem* **2017**, *38* (16), 1291–1307. https://doi.org/10.1002/jcc.24764.

(25) Vincent, P.; Larochelle, H.; Bengio, Y.; Manzagol, P.-A. Extracting and Composing Robust Features with Denoising Autoencoders. In *Proceedings of the 25th international conference on Machine learning - ICML '08*; ACM Press: New York, New York, USA, 2008; pp 1096–1103. https://doi.org/10.1145/1390156.1390294.

(26) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521* (7553), 436–444. https://doi.org/10.1038/nature14539.

(27) Wang, H.; Zhang, L.; Han, J.; E, W. DeePMD-Kit: A Deep Learning Package for Many-Body Potential Energy Representation and Molecular Dynamics. *Comput Phys Commun* **2018**, *228*, 178–184. https://doi.org/10.1016/j.cpc.2018.03.016.

(28) Moradzadeh, A.; Aluru, N. R. Transfer-Learning-Based Coarse-Graining Method for Simple Fluids: Toward Deep Inverse Liquid-State Theory. *Journal of Physical Chemistry Letters* **2019**, *10* (6), 1242–1250. https://doi.org/10.1021/acs.jpclett.8b03872.

(29) Moradzadeh, A.; Aluru, N. R. Understanding Simple Liquids through Statistical and Deep Learning Approaches. *J Chem Phys* **2021**, *154* (20), 204503. https://doi.org/10.1063/5.0046226.

(30) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys Rev Lett* **2007**, *98* (14), 146401. https://doi.org/10.1103/PhysRevLett.98.146401.

(31) Nguyen, T. T.; Székely, E.; Imbalzano, G.; Behler, J.; Csányi, G.; Ceriotti, M.; Götz, A. W.; Paesani, F. Comparison of Permutationally Invariant Polynomials, Neural Networks, and Gaussian Approximation Potentials in Representing Water Interactions through Many-Body Expansions. *Journal of Chemical Physics* **2018**, *148* (24), 241725. https://doi.org/10.1063/1.5024577.

(32) Moradzadeh, A.; Aluru, N. R. Molecular Dynamics Properties without the Full Trajectory: A Denoising Autoencoder Network for Properties of Simple Liquids. *J Phys Chem Lett* **2019**, 7568–7576. https://doi.org/10.1021/acs.jpclett.9b02820.

(33) Moradzadeh, A.; Aluru, N. R. Many-Body Neural Network-Based Force Field for Structure-Based Coarse-Graining of Water. *J Phys Chem A* **2022**, acs.jpca.1c09786. https://doi.org/10.1021/ACS.JPCA.1C09786.

(34) Schütt, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Müller, K. R. SchNet – A Deep Learning Architecture for Molecules and Materials. *J Chem Phys* **2018**, *148* (24), 241722. https://doi.org/10.1063/1.5019779.

(35) Jeong, J.; Moradzadeh, A.; Aluru, N. R. Extended DeepILST for Various Thermodynamic States and Applications in Coarse-Graining. *J Phys Chem A* **2022**, acs.jpca.1c10865. https://doi.org/10.1021/ACS.JPCA.1C10865.

(36) Xu, K.; Jegelka, S.; Hu, W.; Leskovec, J. How Powerful Are Graph Neural Networks? In *7th International Conference on Learning Representations, ICLR 2019*; International Conference on Learning Representations, ICLR, 2019.

(37) Kim, Qh.; Ko, J. H.; Kim, S.; Jhe, W. GCIceNet: A Graph Convolutional Network for Accurate Classification of Water Phases. *Physical Chemistry Chemical Physics* **2020**, *22* (45), 26340–26350. https://doi.org/10.1039/d0cp03456h.

(38) Simonovsky, M.; Komodakis, N. Dynamic Edge-Conditioned Filters in Convolutional Neural Networks on Graphs. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* **2017**, *2017-January*, 29–38.

(39) Wernet, P.; Nordlund, D.; Bergmann, U.; Cavalleri, M.; Odelius, N.; Ogasawara, H.; Näslund, L. Å.; Hirsch, T. K.; Ojamäe, L.; Glatzel, P.; Pettersson, L. G. M.; Nilsson, A. The Structure of the First Coordination Shell in Liquid Water. *Science (1979)* **2004**, *304* (5673), 995–999. https://doi.org/10.1126/SCIENCE.1096205/SUPPL_FILE/WERNET.SOM.PDF.

(40) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L. P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the

Analysis of Molecular Dynamics Trajectories. *Biophys J* **2015**, *109* (8), 1528. https://doi.org/10.1016/J.BPJ.2015.08.015.

(41) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, Flexible, and Free. *J Comput Chem* **2005**, *26* (16), 1701–1718. https://doi.org/10.1002/jcc.20291.

(42) Abascal, J. L. F.; Sanz, E.; Fernández, R. G.; Vega, C. A Potential Model for the Study of Ices and Amorphous Water: TIP4P/Ice. *Journal of Chemical Physics* **2005**, *122* (23), 234511. https://doi.org/10.1063/1.1931662.

(43) Asakawa, H.; Sazaki, G.; Nagashima, K.; Nakatsubo, S.; Furukawa, Y. Two Types of Quasi-Liquid Layers on Ice Crystals Are Formed Kinetically. *Proc Natl Acad Sci U S A* **2016**, *113* (7), 1749–1753. https://doi.org/10.1073/PNAS.1521607113/SUPPL_FILE/PNAS.1521607113.SM03.AVI.

(44) Wu, Y.; Aluru, N. R. Graphitic Carbon-Water Nonbonded Interaction Parameters. *Journal of Physical Chemistry B* **2013**, *117* (29), 8802–8813. https://doi.org/10.1021/jp402051t.

(45) Nosé, S. A Unified Formulation of the Constant Temperature Molecular Dynamics Methods. *J Chem Phys* **1984**, *81* (1), 511–519. https://doi.org/10.1063/1.447334.

(46) Matsumoto, M.; Yagasaki, T.; Tanaka, H. GenIce: Hydrogen-Disordered Ice Generator. *J Comput Chem* **2018**, *39* (1), 61–64. https://doi.org/10.1002/jcc.25077.

(47) Leocmach, M. Pyboo. November 26, 2017. https://doi.org/10.5281/zenodo.1066568.

(48) Zdeborová, L. Machine Learning: New Tool in the Box. *Nat Phys* **2017**, *13* (5), 420–421. https://doi.org/10.1038/nphys4053.

(49) Grattarola, D.; Alippi, C. Graph Neural Networks in TensorFlow and Keras with Spektral. **2020**. https://doi.org/10.48550/arxiv.2006.12138.

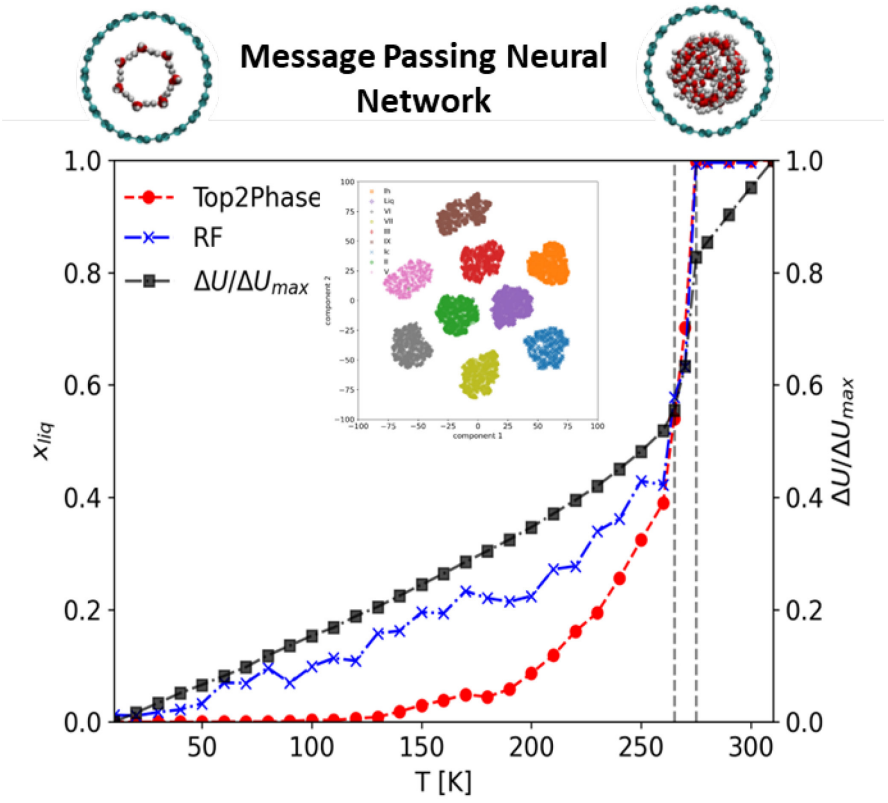
(50) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mane, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viegas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. **2016**.

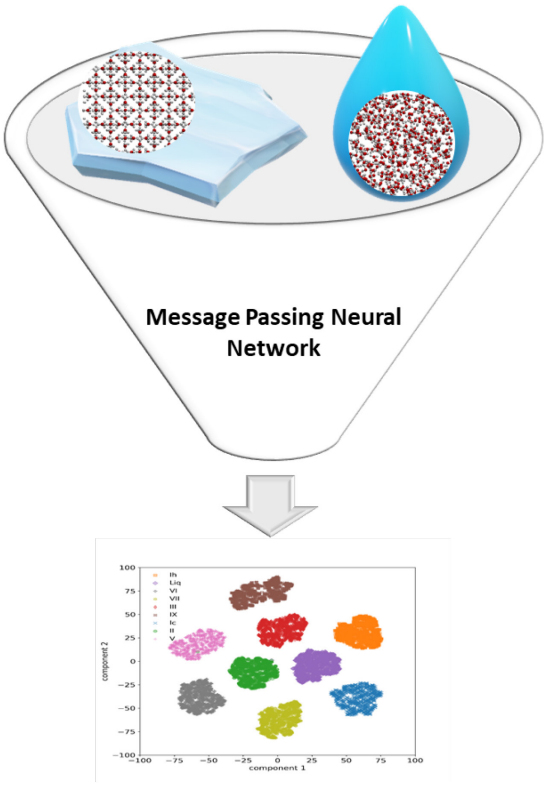
(51) Slater, B.; Michaelides, A. Surface Premelting of Water Ice. *Nature Reviews Chemistry 2019 3:3* **2019**, *3* (3), 172–188. https://doi.org/10.1038/s41570-019-0080-8.

(52) Kling, T.; Kling, F.; Donadio, D. Structure and Dynamics of the Quasi-Liquid Layer at the Surface of Ice from Molecular Simulations. *Journal of Physical Chemistry C* **2018**, *122* (43), 24780–24787. https://doi.org/10.1021/ACS.JPCC.8B07724/SUPPL_FILE/JP8B07724_SI_001.PDF.

(53) Constantin, J. G.; Gianetti, M. M.; Longinotti, M. P.; Corti, H. R. The Quasi-Liquid Layer of Ice Revisited: The Role of Temperature Gradients and Tip Chemistry in AFM Studies. *Atmos Chem Phys* **2018**, *18* (20), 14965–14978. https://doi.org/10.5194/ACP-18-14965-2018.

TOC Graphic:


Message Passing Neural Network
Top2Phase
RF
$\Delta U/\Delta U_{max}$
x_{liq}
T [K]


Message Passing Neural Network