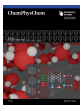
ChemPhysChem

Articles
doi.org/10.1002/cphc.202100599

Chemistry Europe
European Chemical Societies Publishing



A Machine Learning Model to Classify Dynamic Processes in Liquid Water**

Jie Huang,[a] Gang Huang,*[b] and Shiben Li*[a]

The dynamics of water molecules plays a vital role in understanding water. We combined computer simulation and deep learning to study the dynamics of H-bonds between water molecules. Based on *ab initio* molecular dynamics simulations and a newly defined directed Hydrogen (H-) bond population operator, we studied a typical dynamic process in bulk water: interchange, in which the H-bond donor reverses roles with the acceptor. By designing a recurrent neural network-based model, we have successfully classified the interchange and breakage processes in water. We have found that the ratio between them is approximately 1:4, and it hardly depends on temperatures from 280 to 360 K. This work implies that deep learning has the great potential to help distinguish complex dynamic processes containing H-bonds in other systems.

1. Introduction

As one of the big questions in the 21st century,[1] the structure of water is essential for understanding cells, biological processes, and ecosystems.[2–5] Water's surprising properties,[6–8] such as increased density on melting, high surface tension, maximum density at 4 °C, are closely related to the H-bonds.[9–11] Despite the fact that it is tough to capture the ultrafast motion of atoms during dynamic processes,[12] watching water molecules as they dance is the key to understand the dynamic properties of water[13] from the molecular level. Many methods over the past three decades were used to study water molecules' motion, such as scanning tunneling microscopy (STM),[14,15] femtosecond pump-probe,[16] infrared (IR) spectroscopy,[12,17–19] X-rays,[13,20,21] neutron scattering,[22] and computer simulations.[9,23–25]

In this work, we focus on one specific dynamic process *in bulk water*: interchange,[26] in which the H-bond donor reverses roles with the acceptor in the same H-bond. This process was observed in the gas-phase water dimer by Saykally and coworkers. The interchange process, which involves the quantum tunneling effect,[15,23,25–27] is essential for understanding water molecules' dynamics. Also, since interchange processes are closely related to the H-bond network dynamics, it is likely to play a critical role in biological processes, like proton transfer.[28–32] So far, interchange processes have been found in water dimer adsorbed on metal surfaces.[15] Using *ab initio* molecular dynamics (AIMD) simulations,[33] Ranea et al.[23] found that the interchange process can be used to explain the rapid diffusion behavior of water dimer on the Pd(111) surface. Fang et al.[25] found that interchange process is a mechanism of the rapid movement of water dimers on metal surfaces. As for bulk water, Lagge and Hynes found that the redirection of water molecules involves large-angle jumps,[24] which involves the redirection of *multiple* water molecules referred to as *H-bond exchange*, and it is supported by the subsequent experiments.[34,35]

The interchange process involves the concerted rotation of both water molecules engaged in a H-bonded pair. This mechanism is important in small clusters where the future hydrogen-bond donor OH group is typically initially dangling. There are some simulation studies on the interchange process in water clusters,[26,27,36–38] as far as we know, the question of the ratio of interchange to other dynamic processes related to H-bonds *in bulk water* has not been discussed. To determine the proportion of interchange processes, we simulated bulk water in a canonical (NVT) ensemble using a specific AIMD simulation method: the density functional molecular dynamics (DFTMD) simulation. We observed interchange processes in bulk water by analyzing the dynamic trajectory.

As it's tough to quantify interchange processes in a large number of ultrafast dynamic processes in liquid water, we have designed a recurrent neural network (RNN)-based model to classify the H-bond dynamic processes. Unlike general classification methods, this model has the capability of classifying the *dynamic processes* related to H-bonds in bulk water. Using this model, we have obtained the relative ratio of interchange and breakage processes in bulk water and explored the effect of temperature on this ratio. Our work presents the great capacity to use the RNN-based deep learning method to study the dynamic properties of liquid water.

The aims of this work is to provide a machine learning-based model to classify dynamic processes and to determine the proportion of interchange processes in water as one useage of the model. The organization of the paper is the following. We present the results and discussion in Sec. 2. At first, the dynamic graph representation of H-bond networks is intro-

[a] J. Huang, Prof. Dr. S. Li
Department of Physics,
Wenzhou University, Wenzhou,
Zhejiang 325035, China
E-mail: shibenli@wzu.edu.cn
[b] Dr. G. Huang
Institute of Theoretical Physics,
Chinese Academy of Sciences,
Beijing 100190, China
E-mail: hg08@lzu.edu.cn

[**] *A previous version of this manuscript has been deposited on a preprint server (https://arxiv.org/abs/2104.07965)*

Supporting information for this article is available on the WWW under https://doi.org/10.1002/cphc.202100599

duced in 2.1 and the main characteristics of interchange processes are obtained in 2.2. Then we implement the RNN-based classifier for different types of dynamic processes in liquid water in 2.3 and explore the temperature dependence of the relative ratios of interchange and breakage processes in 2.4; The discussion of two factors, the mean number of H-bonds and the rate of breakage and formation of H-bonds in liquid water, related closely to the temperature dependence are discussed in 2.5. Finally, we present the methods details and conclusions of our study in Sec. 3 and 4, respectively.

2. Results and discussion

2.1. Dynamic graph representation of H-bond networks

As shown in Figure 1, a directed *dynamic graph* is used to describe the N water molecules within the cubic simulation box in bulk water. Each water molecule may form an H-bond with any of the remaining N-1 molecules. For convenience, we call any pair of water molecules (i,j) a *quasi*-hydrogen bond (Q-bond), denoted as b_{ij} and represented as a dashed line in Figure 1.

Inspired by Luzar and Chandler's H-bond population operator,[39] we define a *directed* H-bond population operator $\tilde{h}_{ij}$ for b_{ij} $(i<j)$ at time t as Eq. 1.

$$\tilde{h}_{ij}(t)=\begin{cases}1 & \text{H-bonded, i is the donor}\\ 0 & \text{Not H-bonded}\\ -1 & \text{H-bonded, j is the donor}\end{cases} \quad (1)$$

We know from $\tilde{h}_{ij}$ whether an H-bond exists in b_{ij} and the donor-acceptor pair of the formed H-bond. At the bottom of Figure 1, we demonstrate four typical H-bond configuration change processes by using the sequences of $\tilde{h}$: interchange, breakage, formation, and no change. Besides, the Q-bonds likely to form H-bonds are the most relevant water molecule pairs to the breakage and reforming of H-bond networks. The following geometric criteria[40–42] of an H-bond is used: O–O distance $R_{OO'}<R_{cutoff}=3.5$ Å and angle O–H···O$>\theta_{cutoff}=120^{\circ}$. As shown in Figure 2, $R_{OO'}$, θ_a, θ_b, θ_c, and θ_d are monitored for Q-bonds to study the reorientation and breakage mechanism of H-bonds.

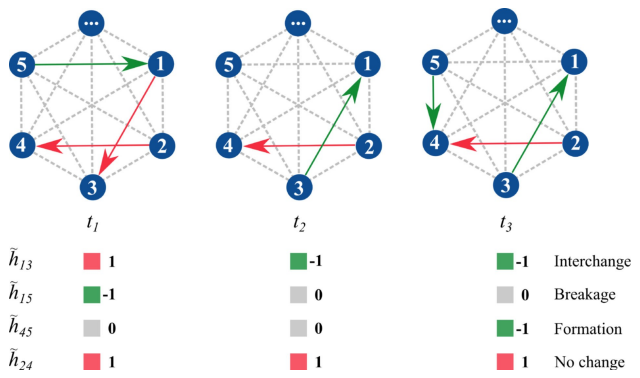


Figure 1. Dynamic graph representation of the H-bond network in simulated bulk water. Nodes represent water molecules; solid red or green arrows represent H-bonds; and dashed grey lines represent Q-bonds. The colors red, grey, and green indicate $\tilde{h}_{ij}=1$, $\tilde{h}_{ij}=0$, and $\tilde{h}_{ij}=-1$, respectively. From the time sequence of $\tilde{h}_{ij}$, we know how the H-bond configuration of b_{ij} changes over time. Four typical H-bond configuration change processes are illustrated for b_{13}, b_{15}, b_{45}, and b_{24}, corresponding to interchange, breakage, formation, and no change, respectively.

2.2. Interchange process

The AIMD simulation trajectory allows us to observe the details of the H-bond dynamics. Figure 3 demonstrates the dynamics of the distance, angles, and directed H-bond population for $b_{OO'}$. Intervals I_1, I_2, and I_3 correspond to three typical H-bond dynamic processes. *Interchange* (I_1): We notice $\theta_b>\theta_{cutoff}$ in the first half and $\theta_d>\theta_{cutoff}$ in the second half. Besides, $h_{OO'}$ changes from 1 to −1, indicating that the donor and acceptor have exchanged. *Breakage* (I_2): $\tilde{h}_{OO'}=-1$ in the first half of I_2, and $\tilde{h}_{OO'}=-0$ for most of the second half. There is no H-bond in the second half because $R_{OO'}>R_{cutoff}$, i.e., the increase of distance $R_{OO'}$ causes the H-bond to break. *Bifurcation rearrangement* motion (I_3):[26,43] At first, the hydrogen atom H′1 is donated to form an H-bond as $\theta_c>\theta_{cutoff}$. Then θ_c decreases and θ_d increases until $\theta_d>\theta_{cutoff}$, i.e., the other hydrogen atom H′2 of the donor is donated to form the H-bond. Therefore, the hydrogen atom contributed by the donor is changed. Because of the identity of hydrogen atoms, it is impossible to distinguish the configuration of water molecules before and after the process. However, during the interchange process, the direction of the water molecules' dipole moment will change, indicating that the water molecules' microscopic configuration will change. So in the rest of the article, we focus on the interchange process.

Figure 4 shows a typical interchange process in water (see Figure S4 in SI Sec. 3 and movies in supplementary material for more H-bond configuration change processes). A dashed line

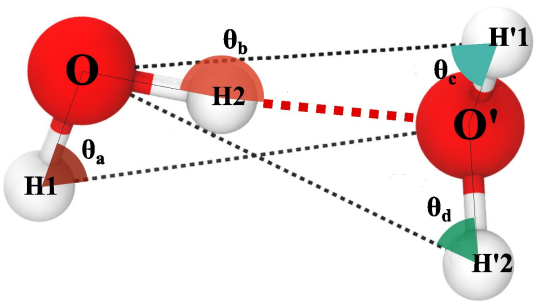


Figure 2. Scheme of the geometric coordinates. $R_{OO'}$ is the O–O distance. Four angles $\widehat{OH1O'}$, $\widehat{OH2O'}$, $\widehat{O'H'1O}$, and $\widehat{O'H'2O}$ are represented as θ_a, θ_b, θ_c, and θ_d, respectively. If $R_{OO'}<3.5$ Å, and any angle $\theta>120^{\circ}$ ($\theta\in\{\theta_a,\theta_b,\theta_c,\theta_d\}$), then an H-bond exists in this Q-bond. Here, the oxygen atom O as a donor donates the hydrogen atom H2 to the acceptor O′. Since $R_{OO'}<3.5$ Å and $\theta_b>120^{\circ}$, we describe this state of $b_{OO'}$ at this time t by $\tilde{h}_{OO'}(t)=1$.

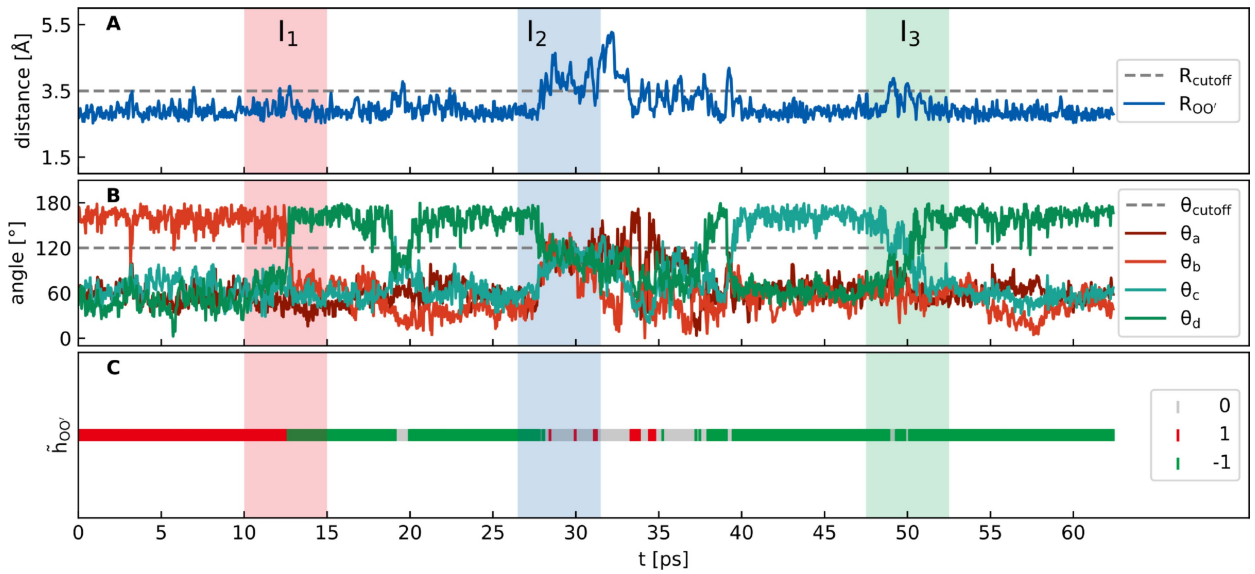


Figure 3. Interchange (I_1), breakage (I_2), and bifurcation rearrangement (I_3) process for one typical Q-bond in bulk water. When an H-bond exists in a Q-bond if $\theta_a > \theta_{cutoff}$ or $\theta_b > \theta_{cutoff}$, then the oxygen atom O is the donor; else, if $\theta_c > \theta_{cutoff}$ or $\theta_d > \theta_{cutoff}$, then the oxygen atom O′ is the donor. Three typical processes are interchange, where the water molecule pairs exchange their roles as H-bond donor and acceptor; breakage, where the H-bond is breaking as the distance increase of this water molecule pair; and bifurcation rearrangement, where the donated hydrogen atom of the H-bond donor exchanged. Through $\tilde{h}$, we can see whether an H-bond exists between a Q-bond, also know the donor and acceptor if an H-bond exists. In panel (C), the grey, red, and green lines indicate the $\tilde{h}_{OO'}$ states.

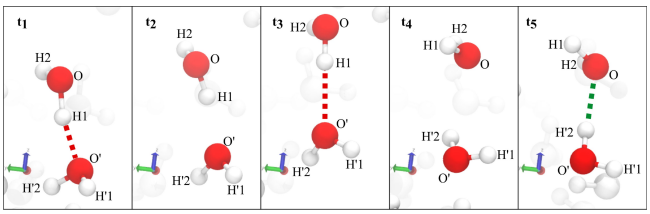


Figure 4. A typical interchange process, where two water molecules exchange their roles as H-bond donor and acceptor via water molecules' reorientation in an concerted manner. The donor oxygen atom has changed from the original O to O′ (color of dashed line changed from red to green). Besides, we have also noticed that the H-bond briefly breaks during the interchange process, causing the fluctuation of the $\tilde{h}$ sequence.

represents an H-bond, and its color (red or green) indicates its direction. Using $\tilde{h}$, we can describe the H-bond configuration change progress without paying attention to the distance and angles. Therefore, $\tilde{h}$ dramatically simplifies the description for the H-bond configuration change process. Nevertheless, during dynamic processes, the fluctuations of $\tilde{h}$ that can result from the vibration of water molecules will bring a huge challenge for the classification of H-bond configuration change processes. In addition, due to a large number of Q-bonds in the simulated bulk water, finding a specific H-bond configuration change process in 60 ps is like finding a needle in a haystack. Therefore, we design an RNN-based model that recognizes the dynamic processes related to H-bonds and uses it to determine various processes in water, thereby determining the ratio of interchanges.

2.3. RNN-based classifier for H-bond configuration change process

We can see the interchange and breakage processes intuitively from $\tilde{h}$. Specifically, in the interchange process, $\tilde{h}$ changes from ± 1 to ∓ 1; in the breakage process, $\tilde{h}$ changes from ± 1 to 0. Therefore, in principle, by observing the sequence of $\tilde{h}$ within a time window, we can classify the H-bond configuration changes during this period. Although we can see some change patterns in interchange and breakage processes, it is still challenging to distinguish different $\tilde{h}$ sequences due to the fluctuation. Therefore, we have designed a processing flow to classify the H-bond configuration change process based on RNN, as shown in Figure 5.

In the preprocessor, we use a low-pass filter to filter out the high-frequency fluctuations of $\tilde{h}$ sequences. As we focus on the configuration change processes of H-bonds, we exclude sequences without H-bonds at the beginning (T_1) and the sequences whose H-bond configuration are unchanged (T_2, T_3) according to the initial value and the variance of $\tilde{h}$ sequences (see Methods section). After preprocessing, the task we need to deal with is a time series classification problem: In addition to interchange and breakage processes, there are also many irregular and complicated processes. We call the sequences of interchange and breakage *positive* and all sequences other than these two types *negative* for convenience. Negative sequences (T_4) do not have any particular pattern. We do not expect that general supervised learning can be used to distinguish them. Nevertheless, we can teach a machine to learn to *recognize* positive sequences. Due to the need to classify time series, we use a

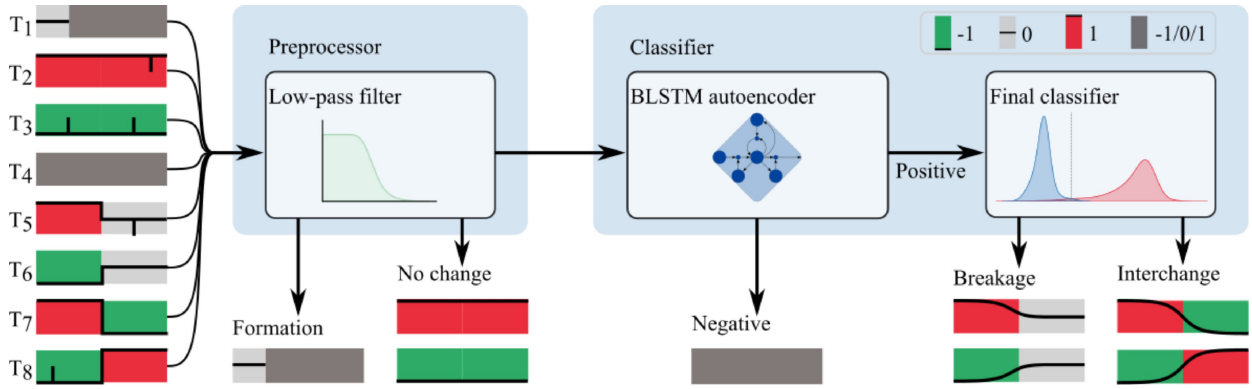


Figure 5. The processing flow of the H-bond configuration change classifier based on RNN. (i). Different types of $\tilde{h}$ sequences: T_1: Formation or no H-bond; T_2, T_3: No change; T_4: Negative sequence; T_5, T_6: Diffusion; T_7, T_8: Interchange. We refer to the sequences of breakage and interchange as positive sequences. (ii). The preprocessor filters out the high-frequency components of $\tilde{h}$ and excludes T_1, T_2, and T_3. (iii). The classifier consists of a BLSTM AE to separate the positive and negative sequences and a final classifier to distinguish breakage and interchange sequences.

typical method for modeling ordered data,[44–46] recurrent neural network (RNN).[47,48] Specifically, we have designed a bidirectional long short-term memory (BLSTM) autoencoder (AE), whose goal is to reconstruct the input sequences as much as possible. We have trained this AE using positive sequences only and evaluated how well the AE reconstructs for an input sequence using reconstruction error $\mathcal{L}(\mathbf{x})$ (see Methods Section 4.3., Eq. 5). After training, the autoencoder can reconstruct positive processes very well. However, when we input negative sequences into the AE, likely, it would not be able to reconstruct them well, leading to the reconstruction errors of these negative sequences greater than that of the positive sequences. Through the reconstruction error, we can determine whether a $\tilde{h}$ sequence is positive or negative. Finally, we use a *final classifier* to distinguish sequences between interchange and breakage processes from positive sequences. We use the range of a positive sequence ($\mathbf{x}$) to determine whether it is interchange or breakage, which is defined as $\delta(\mathbf{x}) = \max \mathbf{x} - \min \mathbf{x}$.

Figure 6(A) shows the densities of the reconstruction errors for interchange, breakage, and negative sequences. Since BLSTM AE can reconstruct positive sequences well, the reconstruction errors of interchange and breakage sequences are small, most of which are smaller than the reconstruction error threshold $\mathcal{L}_T$ ($\mathcal{L}_T$ determination and corresponding accuracy analysis are described in SI Sec. 3, Figure S3). Negative sequences are not used to train the autoencoder, so it is much more difficult to reconstruct them. Hence, the reconstruction errors are relatively large, most of which are greater than $\mathcal{L}_T$. As long as we find a suitable reconstruction error threshold, we can get a classifier for positive and negative sequences. Figure 6(B) shows the densities for the range of normalized interchange and breakage sequences. The two distributions are significantly different from each other. Therefore, the final classifier can distinguish interchange and breakage sequences very well via $\delta_T = 0.66$, as shown in the dashed line (see the classification process in SI Sec. 3, Figure S4–S5). Therefore, we have obtained an H-bond configuration change classifier based on an RNN autoencoder.

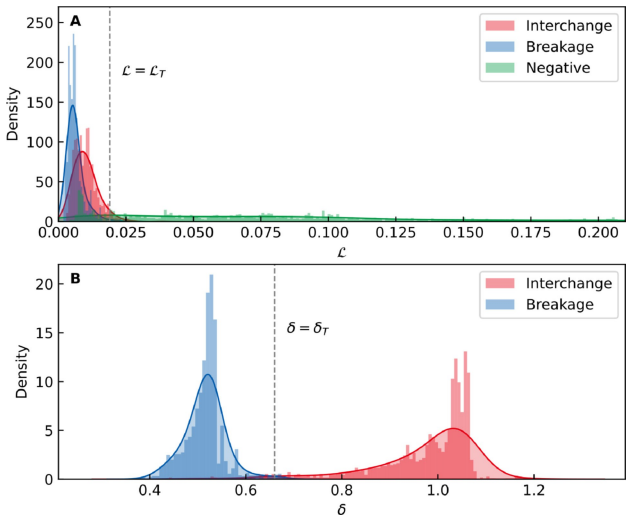


Figure 6. (A) Densities of reconstruction error $\mathcal{L}$ for interchange, breakage, and negative sequences. (i) BLSTM AE can reconstruct positive sequences well. Hence, the reconstruction errors for interchange and breakage sequences are relatively small, mainly less than $\mathcal{L}_T$. (ii) Since negative sequences are not used to train BLSTM AE, it is much more difficult for the autoencoder to reconstruct them. Therefore, the reconstruction errors are relatively large, mainly greater than $\mathcal{L}_T$. (iii) Once $\mathcal{L}_T$ is determined, we use it as the threshold to distinguish positive and negative sequences. (B) Densities of the range δ for interchange and breakage sequences. The two densities are significantly different from each other.

2.4. Proportions of interchange at different temperatures

To explore the effect of temperature on the H-bond configuration change process, we have simulated nine bulk water systems containing $N = 64$ water molecules. The temperature ranges from 280 to 360 K every 10 K. Using the RNN-based model, we classify $\tilde{h}$ sequence, count the number of interchange and breakage sequences at each temperature.

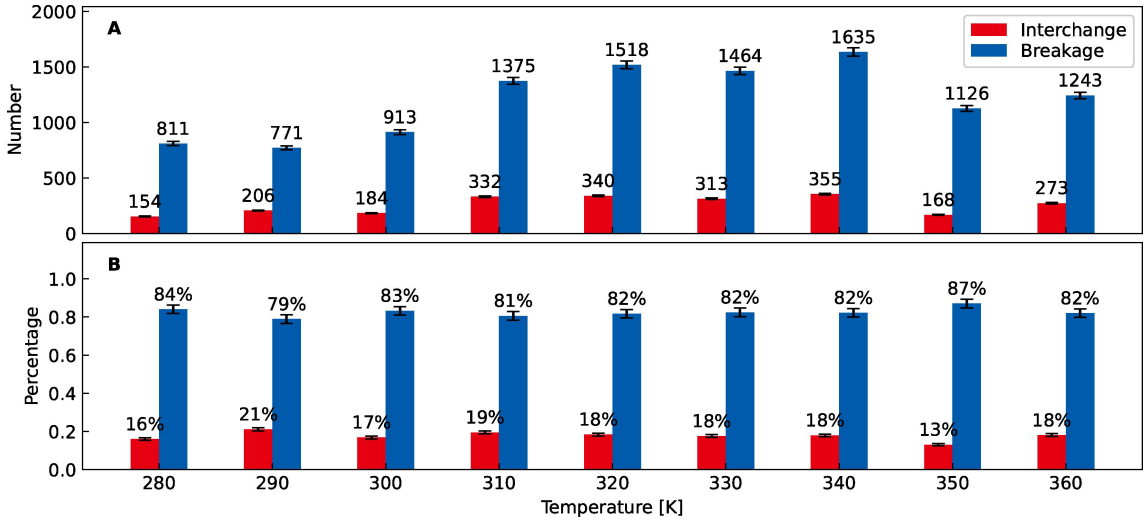


Figure 7. The number (A) and proportion (B) of interchange and breakage processes determined by the RNN-based classifier at different temperatures. (i) With the temperature increasing, the number of interchange and breakage processes increases first and then decreases on the whole. (ii) The relative ratio of interchange to breakage basically does not depend on temperature.

As shown in Figure 7, the number of interchange and breakage processes shows a "rising first, then decreasing" trend as the temperature increases. In other words, there is an overall upward trend from 280 to 330 K. However, as the temperature continues to rise, the number of detected interchange and breakage processes tends to decrease. As we use the method of width-fixed sliding window, the absolute number of interchanges and breakages would change along the step size of sliding window. These numbers would increase as we decrease the step size. Therefore, we focused on the trend of the detected processes over the temperatures. On the other hand, although the number of interchange and breakage processes vary at different temperatures, the relative ratios between the two are almost unchanged, which is still about 1 : 4 (see SI Sec. 4 for the step size effect of the sliding window). This result indicates that the relative ratio is almost not dependent on temperature, and the interchange process is another important mechanism in bulk water besides the breakage process. Next, we will explain this trend of the number of interchange and breakage processes from the following two aspects: the number of H-bond per molecule and the change rate of the coarse-grained H-bond network configuration.

2.5. The trend of interchange and breakage process number

To understand the trend in Figure 7(A), we first calculate the number of H-bonds per molecule (n_{HB}) in the simulated system. At time t, n_{HB} can be expressed as Eq. 2,

$$n_{HB}(t) = \frac{2}{N}\sum_{i=1}^{N}\sum_{j>i}^{N}|\tilde{h}_{ij}(t)| \quad (2)$$

where $N = 64$ is the number of water molecules in bulk water systems, and $|\tilde{h}_{ij}(t)|$ is the absolute value of $\tilde{h}_{ij}(t)$, i.e., the H-bond direction is ignored. The factor 2 is derived from the fact that one H-bond in water is shared by two water molecules. For a certain trajectory at one temperature, by counting n_{HB} at each time t, we get the distribution of n_{HB} (one density plot in Figure 8 A). Then we use an L-dimensional vector $\tilde{\mathbf{h}}$ to represent the coarse-grained H-bond network *configuration* for the simulated bulk water system at time t in Eq. 3,

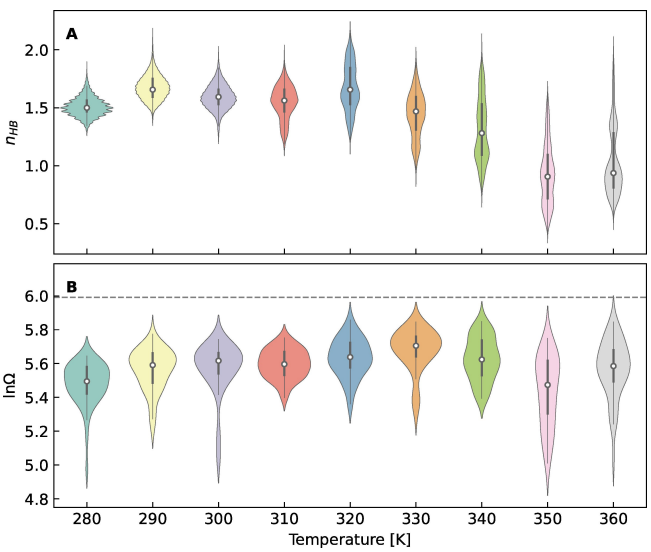


Figure 8. The temperature dependence of (A) The distributions of the number n_{HB} of H-bonds per molecule. (B) The distributions of $\ln\Omega$ characterizing the rate of H-bond breakage and reforming. The dashed line denotes the upper bound of $\ln\Omega$ in the unit time of 1 ps.

$$\tilde{\mathbf{h}}(t) = \left(\tilde{h}_{12}(t), \tilde{h}_{13}(t), ..., \tilde{h}_{ij}(t), ..., \tilde{h}_{N-1,N}(t)\right) \quad (3)$$

where $L = N(N-1)/2$ is the number of Q-bonds in the system. So in a unit time, we get a set H of $\tilde{\mathbf{h}}(t)$ in Eq. 4,

$$H = \left\{\tilde{h}(t) | t = t_0 + k\Delta t, k = 0, 1, ..., M\right\} \quad (4)$$

where t_0 represents the start time of the unit time window, Δt is the time interval between two adjacent frames, and M is the length of the unit time window. In a unit time $t_W = M\Delta t$, the number of graph configuration can be expressed as $\Omega = |H|$, where $|H|$ is the size of the set H, i.e., the number of different $\tilde{\mathbf{h}}$ vectors in this unit time. The number Ω of graph configuration per unit time characterizes the *rate* of breakage and reforming of the H-bonds in bulk water. The theoretical upper bound of Ω in t_w is $M+1$; in this case, all $\tilde{\mathbf{h}}$ vectors are different. By changing the start point t_0, we get the distribution of Ω.

Figure 8 shows the temperature dependence of the distributions of n_{HB} and $\ln\Omega$. The width of a density plot indicates the probability of n_{HB} or $\ln\Omega$ at the corresponding temperature. From the medians (white dots) of violin plots in Figure 8(B), we see Ω is relatively smaller at lower temperatures, indicating fewer changes of H-bond configuration in the unit time. This result explains why the number of interchange and breakage processes at lower temperatures in Figure 7(A) are smaller. Besides, the direct reason for the decrease in the number of interchange and breakage processes at higher temperatures is that thermal motions tend to break H-bonds (thus reducing n_{HB}). Therefore, the number of interchange and breakage processes in Figure 7(A) is determined by n_{HB} and Ω together.

Methods

AIMD simulations

AIMD simulations were carried out for bulk water of 64 water molecules within the canonical NVT ensemble using CP2K/QUICKSTEP (v7.1).[49] The number N of water molecules was 64 for all bulk water systems at different temperatures from 280 to 360 K. The length of the periodic cubic box was 12.4295 Å. The discretized integration time step Δt was set to 0.5 fs. The simulation time was 60 ps. The BLYP functional, which consists of Becke non-local exchange[50] and Lee-Yang-Parr correlation,[51] was used; Interactions between the valence electrons and the ionic cores were described by GTH pseudopotentials;[52,53] Valence electrons were expanded in a basis set consisting of double-zeta Gaussian functions[54] and plane waves with a cutoff energy of 280 Ry.[49] The Nosé-Hoover chain thermostat[55] was used to conserve temperature. DFT–D3 correction[56] for the dispersion interaction was used to obtain a more accurate description of the vibrational properties. It is worth mentioning that the analysis method we proposed can be used on various simulation data, and the AIMD simulation used here is one of the options. The graph-based analysis method is independent of simulation data, so this method can be used to analyze more accurate simulation data in the future.

Sequence collection and preprocessing

The sequence length of $\tilde{h}$ was 200 corresponding to 8 ps simulation time. Positive sequences in which only one interchange or breakage process occurred were collected. Negative $\tilde{h}$ sequences used to evaluate the BLSTM AE classifier were also collected. BLSTM AE was trained by 6786 positive sequences, of which the interchange and breakage processes each accounted for half (754 positive sequences at each temperature). There were 18,931 negative sequences for evaluating the BLSTM AE classifier. The filtered sequence $\tilde{h}_f[n]$ was obtained by second-order Butterworth filter implemented by Scipy.[57] In addition, if $\tilde{h}_f[0] - 0.5 < 0.15$, indicating no H-bond at the beginning (T_1). If the standard deviation σ of $\tilde{h}_f[n]$ satisfy $\sigma < 0.1$, then we consider the H-bond configuration in the Q-bond has not changed (T_2, T_3).

Bidirectional LSTM autoencoder classifier

The encoder and the decoder of BLSTM AE can be expressed as two transformations, $\phi : \mathcal{X} \to \mathcal{F}$ and $\psi : \mathcal{F} \to \mathcal{X}$, where $\mathcal{X}$ and $\mathcal{F}$ are the input space and the feature space, respectively. The dimension of $\mathcal{F}$ is smaller than that of $\mathcal{X}$, and the feature vector $\phi(\mathbf{x})$ is the compressed representation of input $\mathbf{x}$. The input $\mathbf{x}$ of BLSTM AE is the normalized and filtered directed H-bond population operator sequence $\tilde{h}_f[n]$. The reconstruction error of BLSTM AE for a sequence $\mathbf{x} = \tilde{h}_f[n]$ is defined as

$$\mathcal{L}_{\omega,\omega'}(\mathbf{x}) = \mathbf{x} - \psi_{\omega'}(\phi_{\omega}(\mathbf{x}))^2 \quad (5)$$

where ω, ω' represent the parameters of the encoder and decoder respectively. The purpose of training is to obtain the optimal ω, ω',

$$\omega^*, \omega'^* = \arg\min_{\omega,\omega'} \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}_{\omega,\omega'}(\mathbf{x}^i) \quad (6)$$

where $\mathbf{x}^i$ represents the i-th sequence (SI, Figure S1–S2).

3. Summary

In summary, we have designed and trained a deep learning-based model to recognize different types of processes related to H-bonds. The priority of this model are its remarkable ability to classify different dynamic processes of water molecules and its wide range of applications to different kinds of simulation methods. The model can be transfered to other dynamic systems containing H-bonds with the form of O–H···O. As a feasible example, combined with AIMD simulations, we have found that the relative ratio of interchange and breakage processes in bulk water is approximately 1:4, and this ratio hardly depends on temperature.

Moreover, the key concepts used in this work are the dynamic graph and the newly defined directed H-bond population. This reasonable coarse-grained description of the H-bond network simplifies the analysis of H-bond dynamics dramatically. This work demonstrates that the semi-supervised RNN-based model has an outstanding capability of

classifying the dynamic processes related to H-bonds in bulk water, which implies the great potential to extend our present scheme to distinguish more complex dynamic processes in other systems like the water-vapor interface and electrolyte solutions.

In this work, we monitor one variable, the directed H-bond population, for water molecule pairs along the trajectory, which gives a new perspective to detect interchange processes and diffusions in liquid water. The graph representation for bulk water used in this work serves as a platform that allows us to use graph-based approaches[58–60] to explore more complex properties of H-bond networks. And the inspiration of viewing water as a network composed of many directed rings makes it possible to study the complex H-bonded superstructures characteristic of liquid water.[32,61,62] Therefore, extending our framework from H-bonded water pairs to larger H-bonded water rings is promising if reasonable variables to represent the properties of H-bonded rings are used, which we would like to include in the forthcoming study.

Acknowledgment

This research was supported by the National Natural Science Foundation of China (NSFC) (Grant No. 21973070) and the Graduate Scientific Research Foundation of Wenzhou University. The simulations were performed on the cluster in the College of Mathematics and Physics at Wenzhou University.

Conflict of Interest

The authors declare no conflict of interest.

Keywords: AIMD · deep learning · dynamic process classification · hydrogen bond dynamics · LSTM

[1] D. Kennedy, *Science* **2005**, *309*, 75–75.
[2] F. Franks, Water: a matrix of life, Royal Society of Chemistry, 2 edition, **2000**.
[3] S. K. Pal, A. H. Zewail, *Chem. Rev.* **2004**, *104*, 2099–2124.
[4] M. Chaplin, *Nat. Rev. Mol. Cell Biol.* **2006**, *7*, 861–866.
[5] P. Ball, *Proc. Nat. Acad. Sci.* **2017**, *114*, 13327–13335.
[6] F. H. Stillinger, *Science* **1980**, *209*, 451–457.
[7] J. R. Errington, P. G. Debenedetti, *Nature* **2001**, *409*, 318–321.
[8] I. Dumé, *Phys. World* **2020**, *33*, 7i–7i.
[9] R. Kumar, J. R. Schmidt, J. L. Skinner, *J. Chem. Phys.* **2007**, *126*, 204107.
[10] A. Nilsson, L. G. M. Pettersson, *Nat. Commun.* **2015**, *6*.
[11] U. Wilhelmsen, D. Deutsches Elektronen-Synchrotron, The strangest liquid in the world: water amazes scientists time and again, volume 20, Deutsches Elektronen Synchrotron, DESY, Hamburg, **2020**.
[12] E. T. Karamatskos, S. Raabe, T. Mullins, A. Trabattoni, P. Stammer, G. Goldsztejn, R. R. Johansen, K. Długołecki, H. Stapelfeldt, M. J. J. Vrakking, S. Trippel, A. Rouzée, J. Küpper, *Nat. Commun.* **2019**, *10*.
[13] F. Perakis, G. Camisasca, T. J. Lane, A. Späh, K. T. Wikfeldt, J. A. Sellberg, F. Lehmkühler, H. Pathak, K. H. Kim, K. Amann-Winkel, S. Schreck, S. Song, T. Sato, M. Sikorski, A. Eilert, T. McQueen, H. Ogasawara, D. Nordlund, W. Roseker, J. Koralek, S. Nelson, P. Hart, R. Alonso-Mori, Y. Feng, D. Zhu, A. Robert, G. Grübel, L. G. M. Pettersson, A. Nilsson, *Nat. Commun.* **2018**, *9*.
[14] T. Mitsui, *Science* **2002**, *297*, 1850–1852.
[15] T. Kumagai, M. Kaizu, S. Hatta, H. Okuyama, T. Aruga, I. Hamada, Y. Morikawa, *Phys. Rev. Lett.* **2008**, *100*.
[16] S. Woutersen, U. Emmerichs, H. J. Bakker, *Science* **1997**, *278*, 658–660.
[17] H. J. Bakker, H.-K. Nienhuys, *Science* **2002**, *297*, 587–590.
[18] C. J. Fecko, *Science* **2003**, *301*, 1698–1702.
[19] K. Ichi Inoue, M. Ahmed, S. Nihonyanagi, T. Tahara, *Nat. Commun.* **2020**, *11*.
[20] T. Iwashita, B. Wu, W.-R. Chen, S. Tsutsui, A. Q. R. Baron, T. Egami, *Sci. Adv.* **2017**, *3*, e1603079.
[21] Z.-H. Loh, G. Doumy, C. Arnold, L. Kjellsson, S. H. Southworth, A. A. Haddad, Y. Kumagai, M.-F. Tu, P. J. Ho, A. M. March, R. D. Schaller, M. S. B. M. Yusof, T. Debnath, M. Simon, R. Welsch, L. Inhester, K. Khalili, K. Nanda, A. I. Krylov, S. Moeller, G. Coslovich, J. Koralek, M. P. Minitti, W. F. Schlotter, J.-E. Rubensson, R. Santra, L. Young, *Science* **2020**, *367*, 179–182.
[22] T. Head-Gordon, G. Hura, *Chem. Rev.* **2002**, *102*, 2651–2670.
[23] V. A. Ranea, A. Michaelides, R. Ramírez, P. L. de Andres, J. A. Vergés, D. A. King, *Phys. Rev. Lett.* **2004**, *92*.
[24] D. Laage, J. T. Hynes, *Science* **2006**, *311*, 832–835.
[25] W. Fang, J. Chen, P. Pedevilla, X.-Z. Li, J. O. Richardson, A. Michaelides, *Nat. Commun.* **2020**, *11*.
[26] F. N. Keutsch, R. J. Saykally, *Proc. Nat. Acad. Sci.* **2001**, *98*, 10533–10540.
[27] R. S. Fellers, C. Leforestier, L. B. Braly, M. G. Brown, R. J. Saykally, *Science* **1999**, *284*, 945–948.
[28] N. Agmon, *Chem. Phys. Lett.* **1995**, *244*, 456–462.
[29] A. Hassanali, M. K. Prakash, H. Eshet, M. Parrinello, *Proc. Nat. Acad. Sci.* **2011**.
[30] J. L. Thomaston, R. A. Woldeyes, T. Nakane, A. Yamashita, T. Tanaka, K. Koiwai, A. S. Brewster, B. A. Barad, Y. Chen, T. Lemmin, M. Uervirojnangkoorn, T. Arima, J. Kobayashi, T. Masuda, M. Suzuki, M. Sugahara, N. K. Sauter, R. Tanaka, O. Nureki, K. Tono, Y. Joti, E. Nango, S. Iwata, F. Yumoto, J. S. Fraser, W. F. DeGrado, *Proc. Nat. Acad. Sci.* **2017**, *114*, 13357–13362.
[31] M. D. Gelenter, V. S. Mandala, M. J. M. Niesen, D. A. Sharon, A. J. Dregni, A. P. Willard, M. Hong, *Commun. Biol.* **2021**, *4*, 338.
[32] A. Hassanali, F. Giberti, J. Cuny, T. D. Kühne, M. Parrinello, *Proc. Nat. Acad. Sci.* **2013**, *110*, 13723–13728.
[33] T. D. Kühne, M. Iannuzzi, M. D. Ben, V. V. Rybkin, P. Seewald, F. Stein, T. Laino, R. Z. Khaliullin, O. Schütt, F. Schiffmann, D. Golze, J. Wilhelm, S. Chulkov, M. H. Bani-Hashemian, V. Weber, U. Borštnik, M. Taillefumier, A. S. Jakobovits, A. Lazzaro, H. Pabst, T. Müller, R. Schade, M. Guidon, S. Andermatt, N. Holmberg, G. K. Schenter, A. Hehn, A. Bussy, F. Belleflamme, G. Tabacchi, A. Glöß, M. Lass, I. Bethune, C. J. Mundy, C. Plessl, M. Watkins, J. VandeVondele, M. Krack, J. Hutter, *J. Chem. Phys.* **2020**, *152*, 194103.
[34] D. E. Moilanen, D. Wong, D. E. Rosenfeld, E. E. Fenn, M. D. Fayer, *Proc. Nat. Acad. Sci.* **2009**, *106*, 375–380.
[35] M. Ji, M. Odelius, K. J. Gaffney, *Science* **2010**, *328*, 1003–1005.
[36] R. Schulz, Y. von Hansen, J. O. Daldrop, J. Kappler, F. Noé, R. R. Netz, *J. Chem. Phys.* **2018**, *149*, 244504.
[37] N. R. Samala, N. Agmon, *ACS Omega* **2019**, *4*, 22581–22590.
[38] E. Méndez, D. Laria, *J. Chem. Phys.* **2020**, *153*, 054302.
[39] A. Luzar, D. Chandler, *Nature* **1996**, *379*, 55–57.
[40] F. Sciortino, S. L. Fornili, *J. Chem. Phys.* **1989**, *90*, 2786–2792.
[41] S. Balasubramanian, S. Pal, B. Bagchi, *Phys. Rev. Lett.* **2002**, *89*.
[42] N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, O. Beckstein, *J. Comput. Chem.* **2011**, *32*, 2319–2327.
[43] M. G. Brown, F. N. Keutsch, R. J. Saykally, *J. Chem. Phys.* **1998**, *109*, 9645–9647.
[44] T. W. Hughes, I. A. D. Williamson, M. Minkov, S. Fan, *Sci. Adv.* **2019**, *5*.
[45] N. Rank, B. Pfahringer, J. Kempfert, C. Stamm, T. Kühne, F. Schoenrath, V. Falk, C. Eickhoff, A. Meyer, *npj Digital Medicine* **2020**, *3*.
[46] S.-T. Tsai, E.-J. Kuo, P. Tiwary, *Nat. Commun.* **2020**, *11*.
[47] J. J. Hopfield, *Proc. Nat. Acad. Sci.* **1982**, *79*, 2554–2558.
[48] S. Hochreiter, J. Schmidhuber, *Neural Comput* **1997**, *9*, 1735–1780.
[49] J. VandeVondele, M. Krack, F. Mohamed, M. Parrinello, T. Chassaing, J. Hutter, *Comput. Phys. Commun.* **2005**, *167*, 103–128.
[50] A. D. Becke, *Phys. Rev. A* **1988**, *38*, 3098.
[51] C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **1988**, *37*, 785.
[52] C. Hartwigsen, S. Goedecker, J. Hutter, *Phys. Rev. B* **1998**, *58*, 3641–3662.
[53] J. H. G. Lippert, M. Parrinello, *Theor. Chem. Acc.* **1999**, *103*, 124.
[54] J. VandeVondele, J. Hutter, *J. Chem. Phys.* **2007**, *127*, 114105.
[55] G. J. Martyna, M. L. Klein, M. Tuckerman, *J. Chem. Phys.* **1992**, *97*, 2635.
[56] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, *J. Chem. Phys.* **2010**, *132*, 154104.

[57] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, *Nat. Methods* **2020**, *17*, 261–272.
[58] M. Matsumoto, A. Baba, I. Ohmine, *J. Chem. Phys.* **2007**, *127*, 134504.
[59] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, *J. Stat. Mech.: Theory Exp.* **2008**, *2008*, P10008.
[60] J. A. Bilbrey, J. P. Heindel, M. Schram, P. Bandyopadhyay, S. S. Xantheas, S. Choudhury, *J. Chem. Phys.* **2020**, *153*, 024302.
[61] A. C. Belch, S. A. Rice, *J. Chem. Phys.* **1987**, *86*, 5676–5682.
[62] Y. Ding, A. A. Hassanali, M. Parrinello, *Proc. Nat. Acad. Sci.* **2014**, *111*, 3310–3315.

Manuscript received: August 13, 2021
Revised manuscript received: October 16, 2021
Accepted manuscript online: October 18, 2021
Version of record online: November 11, 2021