Supplementary Information for Hydrogen bond donor-acceptor exchange in water

Jie Huang and Shiben Li[*]
Department of Physics, Wenzhou University, Wenzhou, Zhejiang 325035, China

Gang Huang[†]
Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100190, China

(Dated: April 16, 2021)

1. TRAJECTORY ANALYSIS

MDAnalysis (v1.0.0) [1] is used to analyze the simulation trajectories. The first 10 ps non-equilibrium trajectory is removed, and the remaining 50 ps trajectory is sampled every 80 frames. So the time interval after sampling is $80\Delta t = 40$ fs. Next, we use the HydrogenBondAnalysis module to find the atom IDs of the H-bond donor, acceptor, and the contributed hydrogen in each frame used to model the dynamic graph.

2. BLSTM AE CLASSIFIER

LSTM is a type of RNN architecture specifically designed to solve the vanishing gradient problem of standard RNNs. It can learn to model time intervals over 1000 steps even in noisy input sequences without losing short time lag capabilities [2]. The LSTM hidden layer is composed of recurrently connected memory blocks. Each block contains a group of internal units whose activation is controlled by three multiplication gates: input gate, forget gate, and output gate [3]. Figure S1 shows a LSTM memory block with a single unit in detail. For the classification of H-bond configuration change processes, it is useful to access future and past contexts. Bidirectional RNNs [4, 5] can access contextual information in two directions along the input sequence. BRNNs contain two independent hidden layers; one hidden layer processes the forward input sequence, and the other hidden layer processes the reverse sequence. Both hidden layers are connected to the same output layer to access the past and future information of each point in the sequence. Combining BRNNs, LSTM, and autoencoder gives bidirectional LSTM autoencoder (BLSTM AE) as shown in Fig. S2. The BLSTM AE is implemented using the Keras module of Tensorflow (2.2.0). Layer 1 contains two LSTM layers, forward and reverse, each with 64 LSTM units; Layer 2 also has two LSTM layers, each with 32 LSTM units; The parameter of the repeat vector is 2; The network structures of the encoder and decoder are symmetrical about the repeat vector layer. Hence, layers 3 and 4 are the same as layers 2 and 1, respectively. The last layer is the

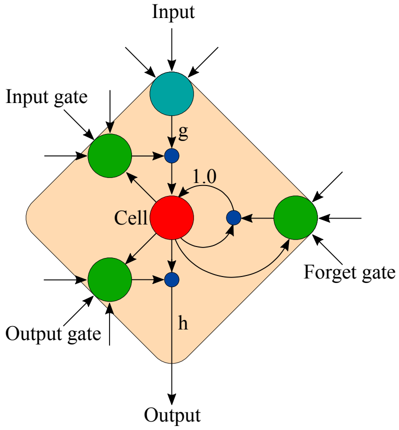

FIG. S1. LSTM unit. The outcome of the gates is to allow the cell to store and access information over long periods [3].

[*] shibenli@wzu.edu.cn
[†] hg08@lzu.edu.cn

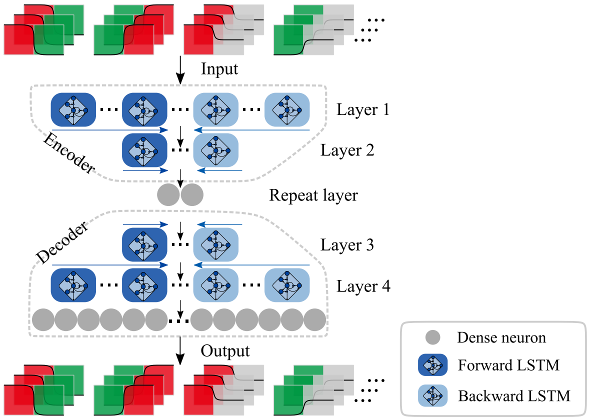

FIG. S2. The structure of BLSTM AE. This kind of design is used to identify positive and negative $\tilde{h}$ sequences. The training data for the BLSTM AE are the filtered positive $\tilde{h}$ sequences. Since $\tilde{h}$ sequences are time-varying sequences, we choose to use the LSTM unit as the building block. The $\tilde{h}$ sequence's start and end are equally crucial for the classification, so we use a bidirectional network structure.

time distributed layer, which contains 200 neurons. The optimizer for training is Adam; the loss function is MAE; the batch size is 32; the dropout rate is 0.1, and the epoch number is 500.

A BSLTM AE classifier is obtained by choosing a reasonable reconstruction error as the threshold for classifying positive and negative sequences. As shown in Fig. S3, we measure the classifier's accuracy, balanced accuracy, and F1 score under different thresholds. We notice that these three values increase first and then decrease in the interval $[0.01, 0.03]$. When $\mathcal{L} = \mathcal{L}_{\mathrm{T}} = 0.019$, its accuracy, balanced accuracy, and F1 score achieve maximum values. Therefore, $\mathcal{L}_{\mathrm{T}}$ is chosen as the threshold of the classifier.

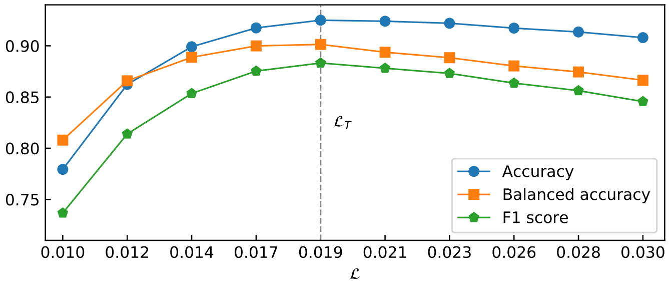

FIG. S3. The accuracy, balanced accuracy, and F1 score of the BLSTM autoencoder classifier in the testing data at different reconstruction error thresholds. The reconstruction error $\mathcal{L}_{\mathrm{T}} = 0.019$ is chosen as the threshold for the BLSTM autoencoder classifier corresponding to the highest values of the accuracy, balanced accuracy, and F1 score.

3. STEP SIZE EFFECT OF THE SLIDING WINDOW

Since we use the sliding window method for sampling the dynamic trajectory of $\tilde{h}$ to obtain the 8 ps sequences, we take the sliding step as a parameter to observe the relative ratios of DA exchange and diffusion processes. As shown in Fig. S4, we find that this relative ratio is almost unaffected by the step size of the sliding window.

4. CLASSIFICATION DEMONSTRATIONS

Figure S5 shows different types of dynamic processes of H-bond configuration. We can classify those processes by looking at the O-O distance and angles. (A), (B), and (C) are DA exchanges; (D), (E), and (F) are diffusions; (G), (H), and (I) are negative processes. $\tilde{h}_s$ is the normalized result of $\tilde{h}$; $\tilde{h}_f$ is the filtered $\tilde{h}_s$; $\tilde{h}_r$ is the sequence reconstructed by the BLSTM AE. We see that the $\tilde{h}_r$ and $\tilde{h}_f$ of DA exchange and diffusion sequences almost coincide,

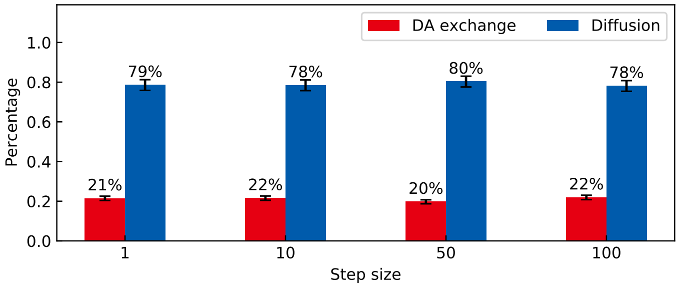

FIG. S4. The relative ratios of DA exchange and diffusion processes under different step sizes for the simulation bulk water at 310 K.

which means the BLSTM AE can reconstruct DA exchange and diffusion sequences very well. However, the $\tilde{h}_f$ of negative sequences can not be reconstructed well. Finally, the positive sequences are inputted into the final classifier; and the range δ of $\tilde{h}_f$ is used to determine whether the sequence is DA exchange or diffusion. Figure S6 shows the reconstruction errors and ranges corresponding to different $\tilde{h}$ sequences in Fig. S5. The background colors represent the predictions of the BLSTM AE classifier. Red, blue, and green represent the DA exchange, diffusion, and negative process, respectively, indicating that the BLSTM AE classifier can correctly classify the H-bond configuration change processes.

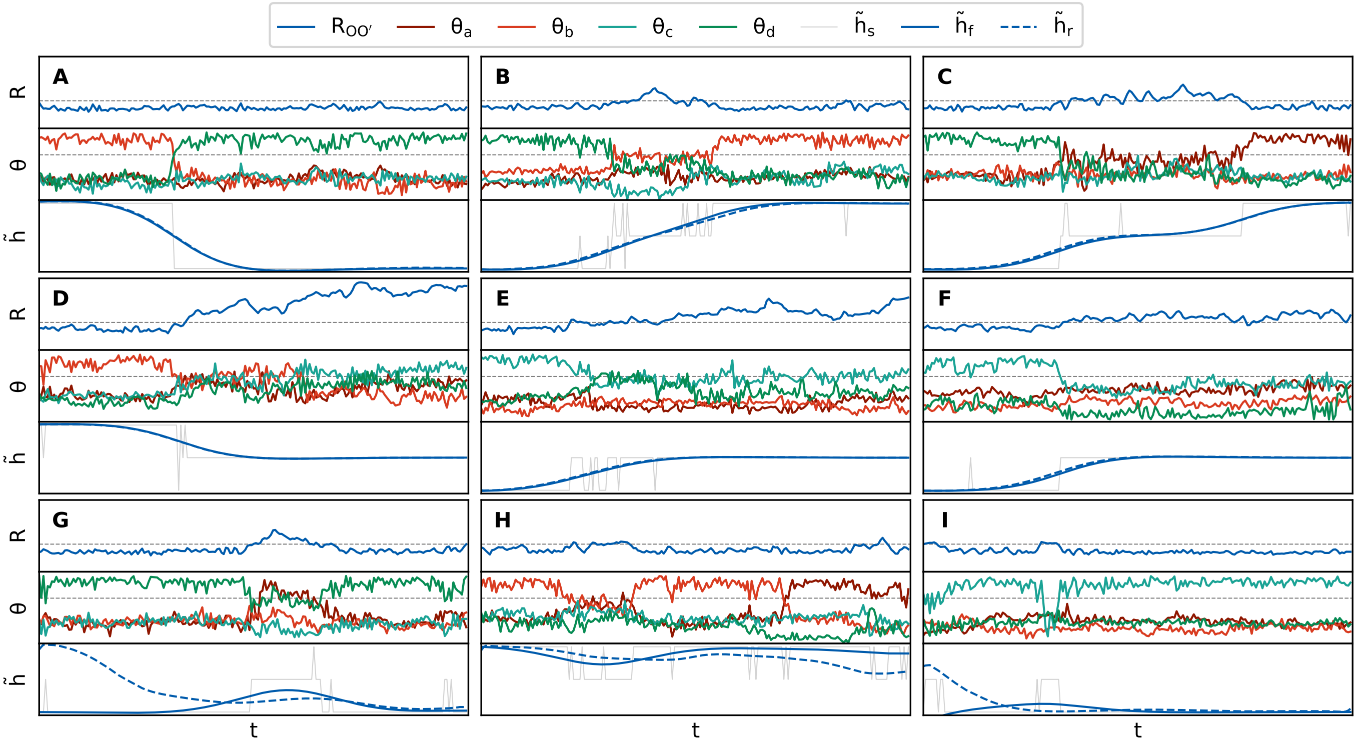

FIG. S5. Different types of H-bond configuration change processes. (A), (B), and (C) are DA exchange processes; (D), (E), and (F) are diffusion processes; (G), (H), and (I) are negative processes. The corresponding simulation time for each sequence is 8 ps.

5. THE NUMBER OF H-BONDS PER MOLECULE AND H-BOND RELAXATION TIME.

Based on two different H-bond definitions, we calculate the average number $\langle n_{\mathrm{HB}} \rangle$ of H-bonds per molecule and H-bond relaxation time τ_{R}. The average number $\langle n_{\mathrm{HB}} \rangle$ of H-bonds in bulk water of N water molecules is $\frac{1}{2}N(N-1)\langle h \rangle$,

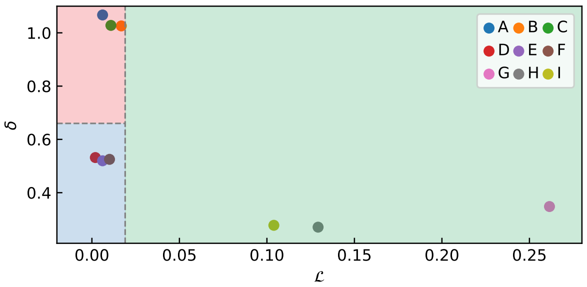

FIG. S6. Classification results for sequences in Fig. S5.

where $\langle h \rangle$ denotes the average of Luzar-Chandler's H-bond population operator h [6]. The τ_{R} is computed by [7]

$$\tau_{\mathrm{R}} = \frac{\int t C_{\mathrm{HB}}(t)\mathrm{d}t}{\int C_{\mathrm{HB}}(t)\mathrm{d}t} \tag{S1}$$

where $C_{\mathrm{HB}}(t) = \langle h(0)h(t)\rangle / \langle h \rangle$ is the autocorrelation of h. The difference between the two H-bond definitions lies in the different geometric conditions, which are as follows. Definition 1: $R_{\mathrm{OO'}} < 3.5$ Å, $\widehat{\mathrm{OHO'}} > 120°$; Definition 2: $R_{\mathrm{OO'}} < 3.5$ Å, $\widehat{\mathrm{HOO'}} < 30°$. Figure S7 (A) shows the temperature dependence of $\langle n_{\mathrm{HB}} \rangle$ under two different geometric H-bond definitions. For both definitions, the $\langle n_{\mathrm{HB}} \rangle$ shows a downward trend as the temperature increases. In general, the $\langle \tau_{\mathrm{R}} \rangle$ also decreases as the temperature increases for both geometric definitions. Figure S7 shows that the different definitions of H-bond may cause some differences in observations. However, the relationship of $\langle n_{\mathrm{HB}} \rangle$ and τ_{R} with temperature changes is consistent. We can understand this similarity as follows. Shorter H-bond relaxation time at higher temperatures means that each water molecule has fewer H-bonds on average.

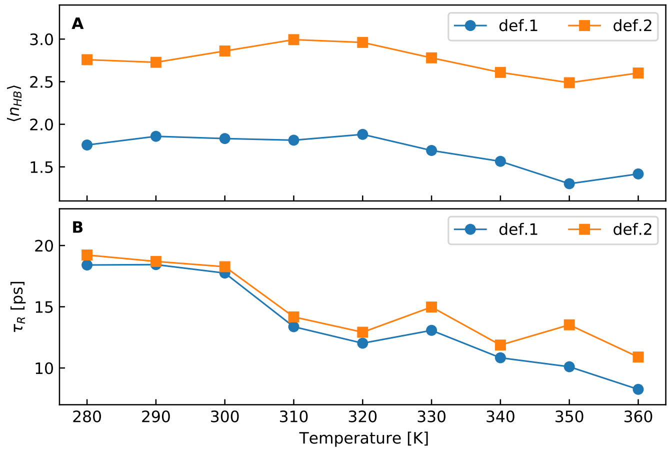

FIG. S7. (A) Mean number $\langle n_{\mathrm{HB}} \rangle$ of H-bonds per water. (B) Relaxation time τ_{R} of H-bonds. Both $\langle n_{\mathrm{HB}} \rangle$ and τ_{R} are calculated for two geometric definitions.

6. VELOCITY AUTOCORRELATION FUNCTION AND VIBRATIONAL DENSITY OF STATES

We use the velocity autocorrelation function (VACF) to obtain the bulk water system's vibration properties. For a system containing M atoms, the VACF $C(t)$ can be expressed as

$$C(t) = \frac{\left\langle \sum_{i=1}^{M} \mathbf{v}_i(t) \cdot \mathbf{v}_i(0) \right\rangle}{\left\langle \sum_{i=1}^{M} \mathbf{v}_i(0) \cdot \mathbf{v}_i(0) \right\rangle} \tag{S2}$$

where $\langle \cdots \rangle$ represents the averaging over all the time starting points, t is the time interval, and $\mathbf{v}_i$ represents the velocity of the i-th atom. Figures S8 (A) and (B) show the VACF and its Fourier transform, i.e., the vibrational

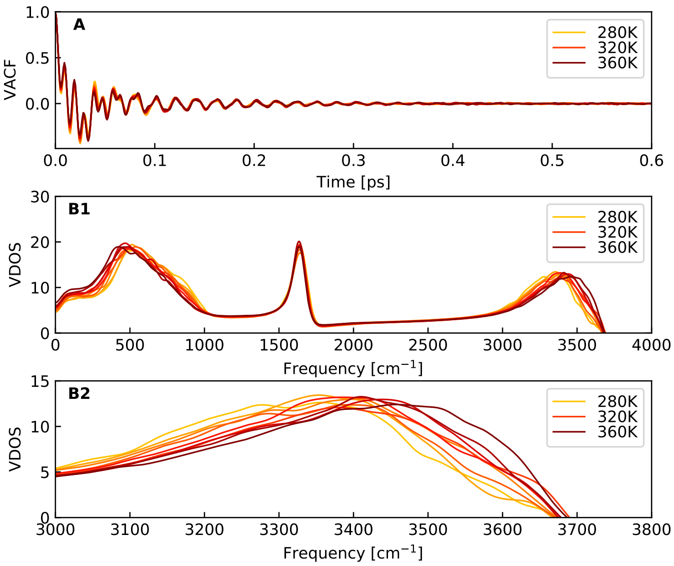

FIG. S8. Velocity autocorrelation function (A) and vibrational density (B) of states.

density of state (VDOS) of bulk water systems at different temperatures, respectively. To display the information of OH stretching more clearly, we made Fig. S8 (B2), just zooming in on the third band in Fig. S8 (B1). The position of the third peak represents the OH stretching vibration frequency of water molecules. From Fig. S8 (B2), we can see that with the increase of temperature, the peaks of OH stretching bands are blue-shifted. This result means that the increasing temperature causes a higher frequency of OH stretching. As OH stretch frequency is correlated to the strength of H-bonds in which the OH bonds are involved [8, 9], the blue-shifted OH stretch band has been assigned to weakly H-bonded water. Therefore, both the shorter relaxation time τ_{R} and blue-shifted OH stretch frequency are consistent with the smaller $\langle n_{\mathrm{HB}} \rangle$ as temperature increases.

[1] N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein, MDAnalysis: A toolkit for the analysis of molecular dynamics simulations, Journal of Computational Chemistry **32**, 2319 (2011).

[2] S. Hochreiter and J. Schmidhuber, Long short-term memory, Neural Computation **9**, 1735 (1997).

[3] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber, A novel connectionist system for unconstrained handwriting recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence **31**, 855 (2009).

[4] M. Schuster and K. Paliwal, Bidirectional recurrent neural networks, IEEE Transactions on Signal Processing **45**, 2673 (1997).

[5] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri, Exploiting the past and the future in protein secondary structure prediction, Bioinformatics **15**, 937 (1999).

[6] A. Luzar and D. Chandler, Hydrogen-bond kinetics in liquid water, Nature **379**, 55 (1996).

[7] R. Z. Khaliullin and T. D. Kühne, Microscopic properties of liquid water from combined ab initio molecular dynamics and energy decomposition studies, Phys. Chem. Chem. Phys. **15**, 15746 (2013).

[8] J. D. Smith, C. D. Cappa, K. R. Wilson, R. C. Cohen, P. L. Geissler, and R. J. Saykally, Unified description of temperature-dependent hydrogen-bond rearrangements in liquid water, Proceedings of the National Academy of Sciences **102**, 14171 (2005).

[9] S. Garrett-Roe and P. Hamm, The oh stretch vibration of liquid water reveals hydrogen-bond clusters, Physical Chemistry Chemical Physics **12**, 11263 (2010).