Supplementary Information for *A machine learning model to classify dynamic processes in liquid water*

Jie Huang,[1] Gang Huang*,[2] and Shiben Li*[1]

[1)] *Jie Huang, Prof. Shiben Li*
Department of Physics, Wenzhou University, Wenzhou, Zhejiang 325035, China; E-mail: shibenli@wzu.edu.cn

[2)] *Gang Huang*
Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100190, China; E-mail: hg08@lzu.edu.cn

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

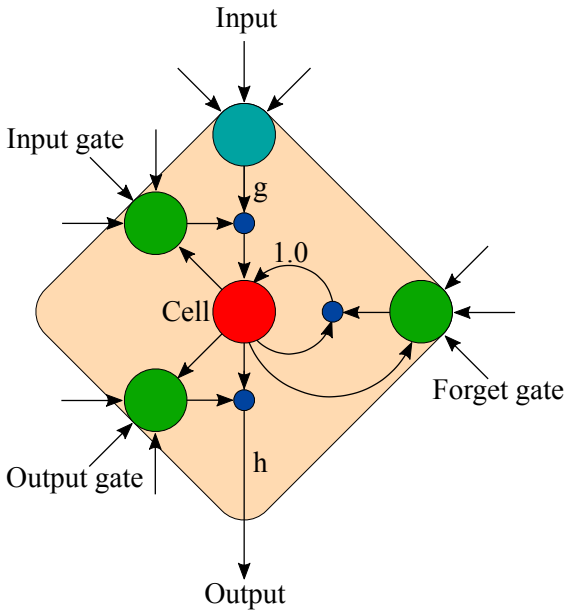

FIG. S1. LSTM unit. The outcome of the gates is to allow the cell to store and access information over long periods[3].

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 same as layers 2 and 1, respectively. The last layer is the time distributed layer,

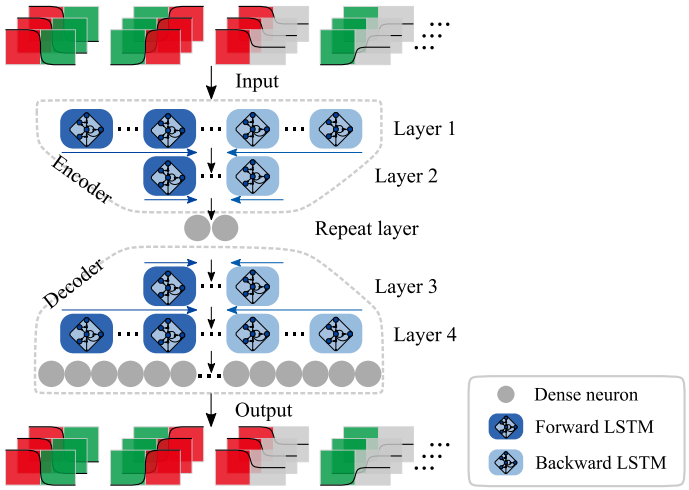

FIG. S2. The structure of BLSTM AE. This kind of design is used to identify positive and negative $\tilde{h}$ sequences. The training data for the BLSTM AE are the filtered positive $\tilde{h}$ sequences. Since $\tilde{h}$ sequences are time-varying sequences, we choose to use the LSTM unit as the building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

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 breakageerent thresholds. We notice that these three values increase first and then decrease in the interval $[0.01, 0.03]$. When $\mathscr{L} = \mathscr{L}_{\mathrm{T}} = 0.019$, its accuracy, balanced accuracy, and F1 score achieve maximum values. Therefore, $\mathscr{L}_{\mathrm{T}}$ is chosen as the threshold of the classifier.

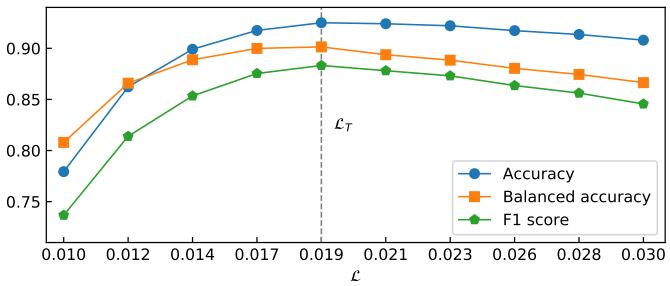

FIG. S3. The accuracy, balanced accuracy, and F1 score of the BLSTM autoencoder classifier in the testing data at different reconstruction error thresholds. The reconstruction error $\mathscr{L}_{\mathrm{T}} = 0.019$ is chosen as the threshold for the BLSTM autoencoder classifier corresponding to the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

3. Classification demonstrations

Figure S4 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 interchanges; (D), (E), and (F) are breakages; (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 interchange and breakage sequences almost coincide, which means the BLSTM AE can reconstruct interchange and breakage 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 interchange or breakage. Figure S5 shows the reconstruction errors and ranges corresponding to different $\tilde{h}$ sequences in Fig. S4. The background colors represent the predictions of the BLSTM AE classifier. Red, blue, and green denote the interchange, breakage, and negative process, respectively, indicating that the BLSTM AE classifier can correctly classify the H-bond configuration change processes.

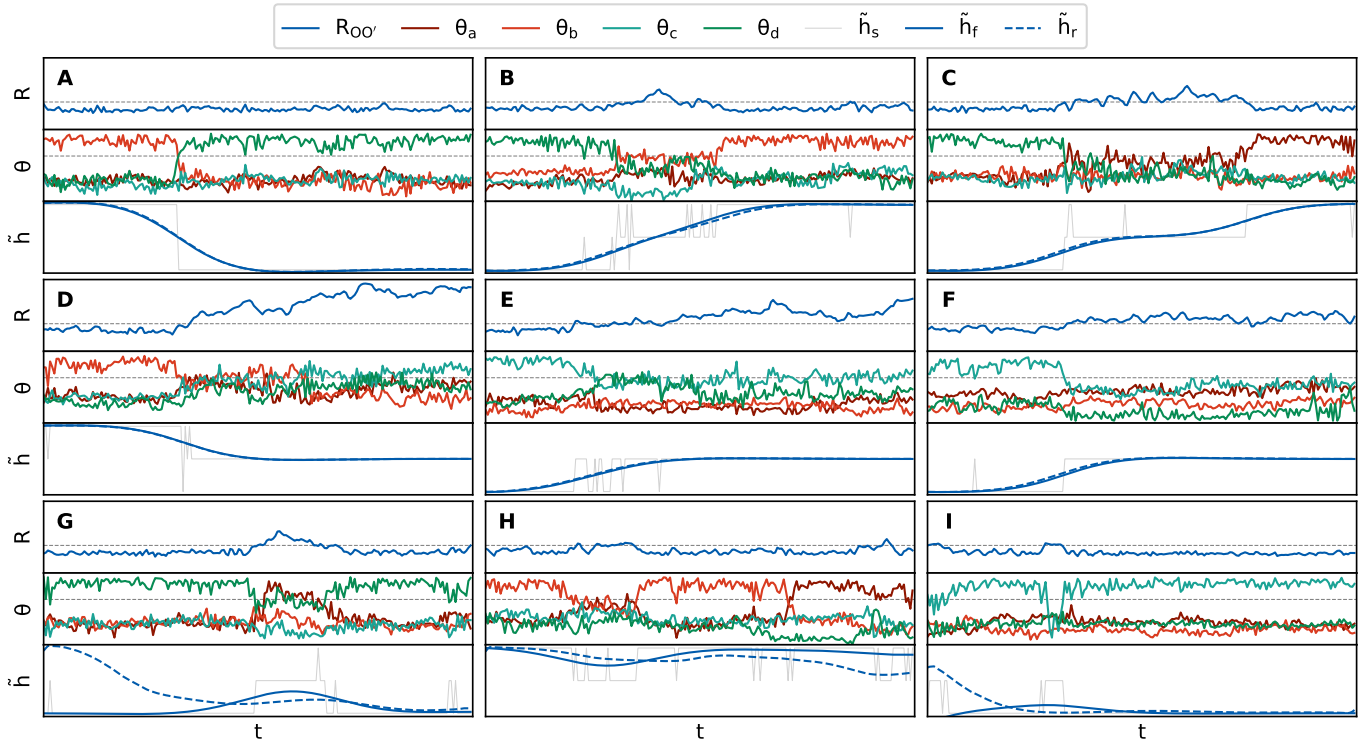

FIG. S4. Different types of H-bond configuration change processes. (A), (B), and (C) are interchange processes; (D), (E), and (F) are breakage processes; (G), (H), and (I) are negative processes. The corresponding simulation time for each sequence is 8 ps.

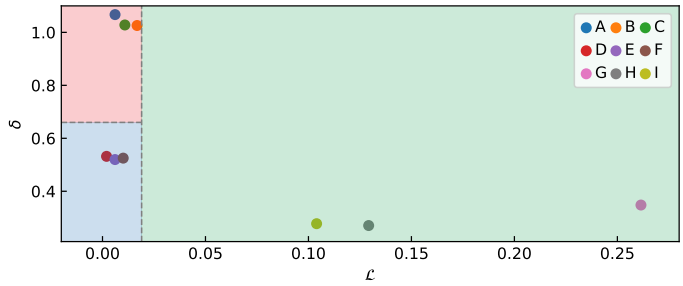

FIG. S5. Classification results for sequences in Fig. S4.

4. 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 interchange and breakage processes. As shown in Fig. S6, we find that this relative ratio is almost unaffected by the step size of the sliding window.

5. Velocity autocorrelation function and vibrational density of states

In this work, we have simulated bulk water in a large temperature range by using DFTMD with DFTD3 correction. To prove that dynamic properties are evaluated with accuracy, we calculate the velocity autocorrelation function (VACF) and its Fourier transform, the vibrational density of states (VDOS) of the water molecules in the simulated bulk water systems. For a system containing M atoms, the VACF $C(t)$ can be expressed as

$$C(t) = \frac{\left\langle \sum_{i=1}^{M} \mathbf{v}_i(t) \cdot \mathbf{v}_i(0) \right\rangle}{\left\langle \sum_{i=1}^{M} \mathbf{v}_i(0) \cdot \mathbf{v}_i(0) \right\rangle} \tag{S1}$$

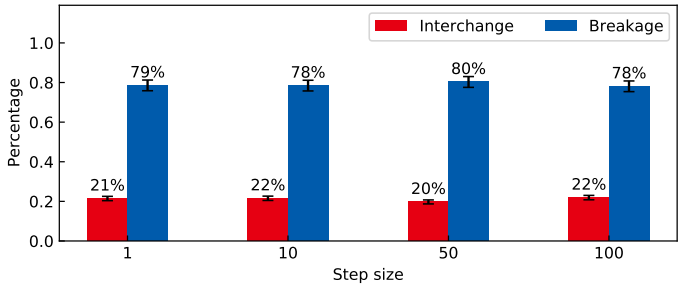

FIG. S6. The relative ratios of interchange and breakage processes under different step sizes for the simulation bulk water at 310 K.

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 S7 (A) and (B) show the VACF and VDOS of bulk water systems at different temperatures, respectively.

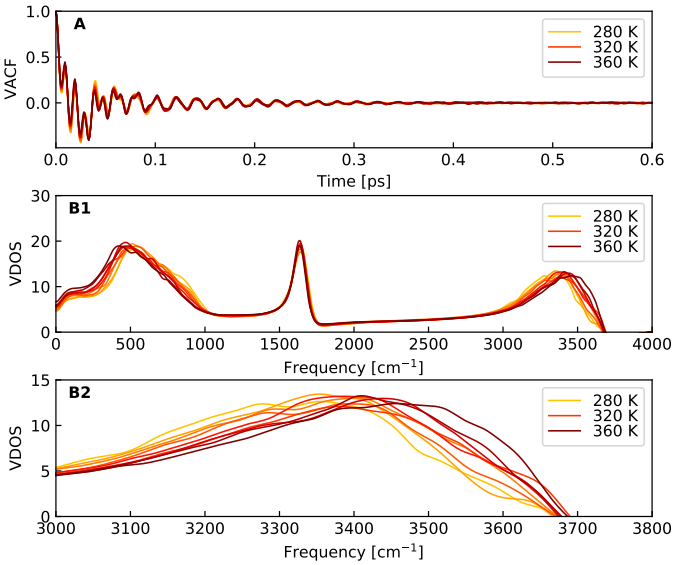

FIG. S7. Velocity autocorrelation function (A) and vibrational density (B) of states.

The VDOS we obtained is in good agreement with previous experiments and *ab initio* all-electron calculations[6,7], which laid a foundation for us to observe the motion of water molecules from the microscopic level.

To display the information of OH stretching more clearly, we made Fig. S7 (B2), just zooming in on the third band in Fig. S7 (B1). The position of the third peak represents the OH stretching vibration frequency of water molecules. From Fig. S7 (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 $\langle n_{HB} \rangle$ as temperature increases.

6. Videos

Video 1 (DA1_09_14.mp4): A typical interchange process in bulk water; Video 2 (DF2_23_35.mp4): A typical breakage process in bulk water)

[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**, 2319–2327 (2011).
[2]S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation **9**, 1735–1780 (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**, 855–868 (2009).
[4]M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," IEEE Transactions on Signal Processing **45**, 2673–2681 (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**, 937–946 (1999).
[6]T. D. Kühne, M. Krack, F. R. Mohamed, and M. Parrinello, "Efficient and accurate car-parrinello-like approach to born-oppenheimer molecular dynamics," Physical Review Letters **98** (2007), 10.1103/physrevlett.98.066401.
[7]M. Krack and M. Parrinello, "All-electron ab-initio molecular dynamics," Physical Chemistry Chemical Physics **2**, 2105–2112 (2000).
[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**, 14171–14174 (2005).
[9]S. Garrett-Roe and P. Hamm, "The oh stretch vibration of liquid water reveals hydrogen-bond clusters," Physical Chemistry Chemical Physics **12**, 11263–11266 (2010).