# Applications of machine learning in hydrogen-bond dynamics in water and polymer chains structure factor

Jie HUANG (黃杰)

August 31, 2022

# Schedule

# Hydrogen-bond dynamics in water

# The strangest liquid in the world

**Is superfluidity possible in a solid? If so, how?**
Despite hints in solid helium, nobody is sure whether a crystalline material can flow without resistance. If new types of experiments show that such outlandish behavior is possible, theorists would have to explain how.

**What is the structure of water?**
Researchers continue to tussle over how many bonds each $H_2O$ molecule makes with its nearest neighbors.

JUPITER IMAGES

**What is the nature of the glassy state?**
Molecules in a glass are arranged much like those in liquids but are more tightly packed. Where and why does liquid end and glass begin?

Figure: What is the structure of water? A big question over the next quarter-century[1].

[1] D. Kennedy, Science, 309, 75-75 (2005)

# Dynamic graph representation of H-bond networks

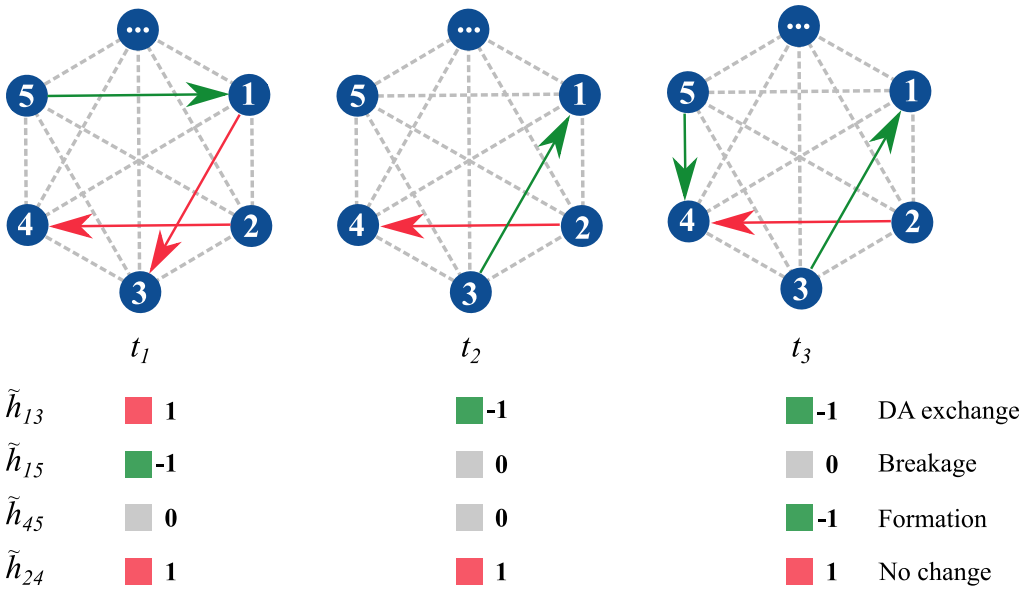


Figure: **Dynamic graph representation** of the H-bond network in simulated bulk water. Nodes represent **water molecules**; dashed grey lines represent **Q-bonds**; and solid red or green arrows represent **H-bonds**. The colors **red**, **grey**, and **green** indicate $\tilde{h}_{ij}$=1, $\tilde{h}_{ij}$=0,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

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

# Geometric criteria of H-bond

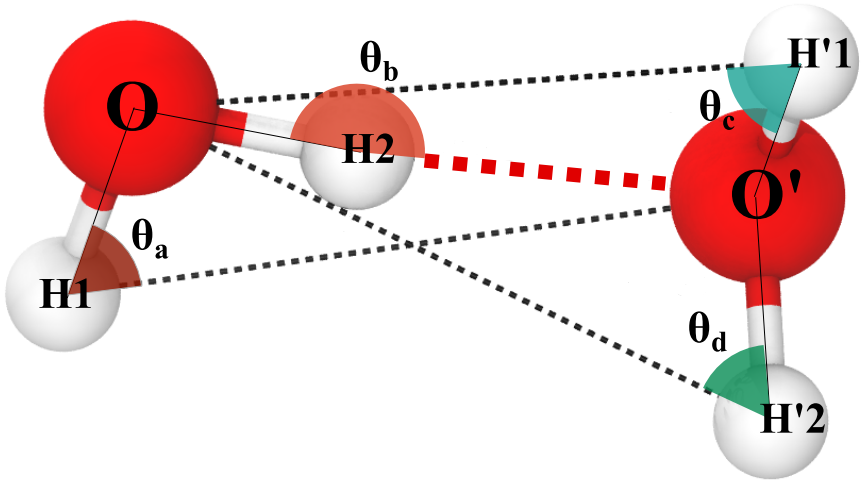


Figure: Scheme of the geometric coordinates. $R_{OO'}$ is the O-O **distance**. Four **angles** $\widehat{\mathrm{OH1O'}}$, $\widehat{\mathrm{OH2O'}}$, $\widehat{\mathrm{O'H'1O}}$, and $\widehat{\mathrm{O'H'2O}}$ are represented as $\theta_a$, $\theta_b$, $\theta_c$, and $\theta_d$, respectively. If $R_{OO'} < 3.5$ 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. 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**. 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 $QB_{OO'}$ at this time $t$ by $\tilde{h}_{OO'}(t) = 1$.

# H-bond configuration change processes

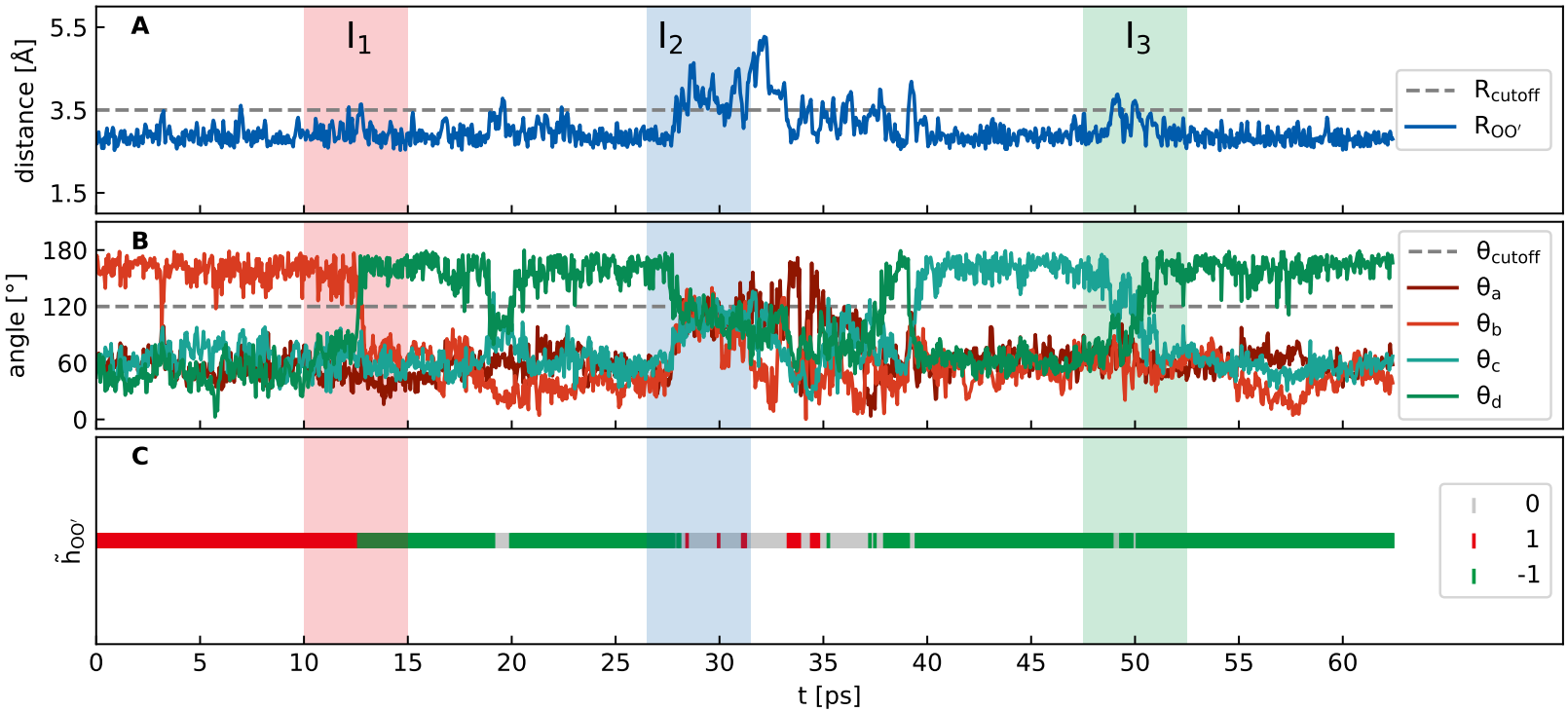


Figure: **Interchange** ($I_1$), **diffusion** ($I_2$), and **HH exchange** ($I_3$) process for one typical Q-bond in bulk water. Three typical processes are interchange, where the water molecule pairs exchange their roles as H-bond donor and acceptor; diffusion, where the H-bond is breaking as the distance increase of this water molecule pair; and HH exchange, 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. (Videos: **interchange** and **diffusion**)

# Typical interchange process

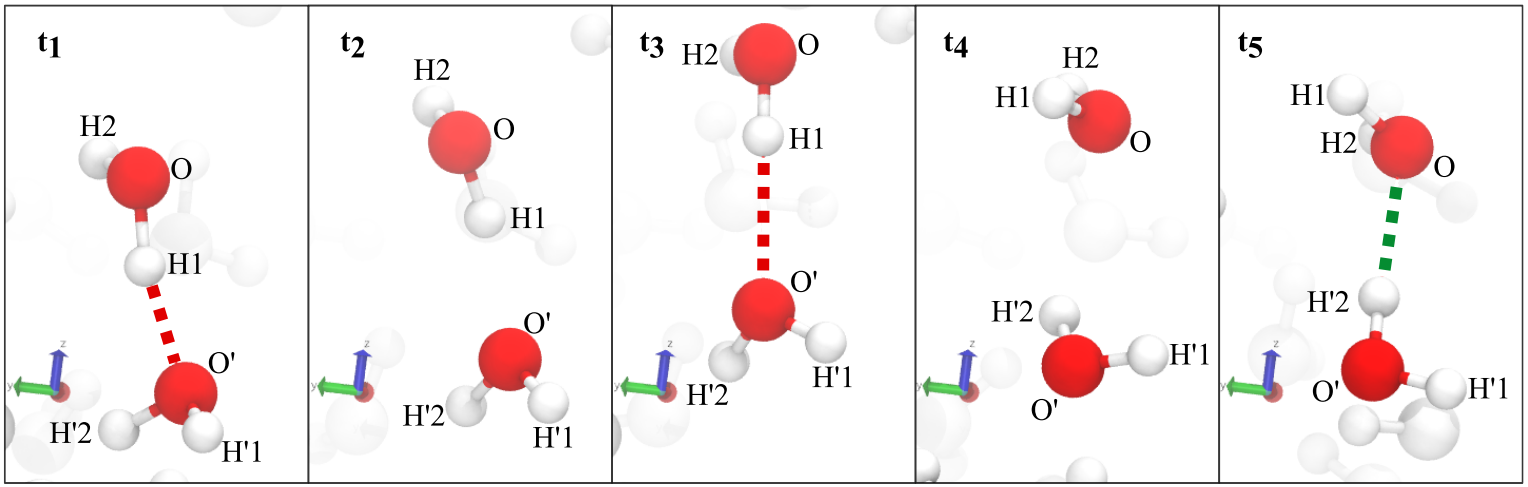


Figure: A typical interchange process, where two water molecules exchange their roles as H-bond donor and acceptor via water molecules' reorientation in an concerted manner. The donor oxygen atom has changed from the original O to O′ (**color of dashed line changed from red to green**). Besides, we have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

# RNN-based classifier for H-bond configuration change process

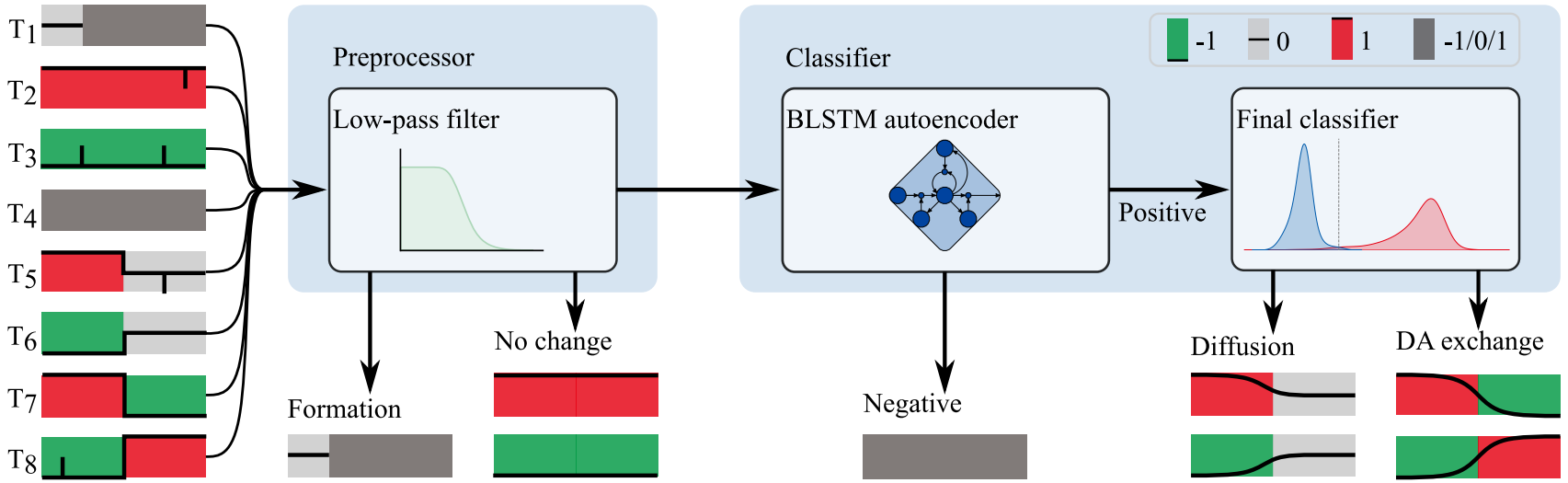


Figure: **The processing flow** of the H-bond configuration change classifier based on **RNN**. (i). **Different types** of $\tilde{h}$ sequences: $T_1$: Formation or no H-bond; $T_2$, $T_3$: No change; $T_4$: Negative sequence; $T_5$, $T_6$: Diffusion; $T_7$, $T_8$: interchange. We refer to the sequences of diffusion 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 diffusion and interchange sequences.

# Densities for different dynamic processes

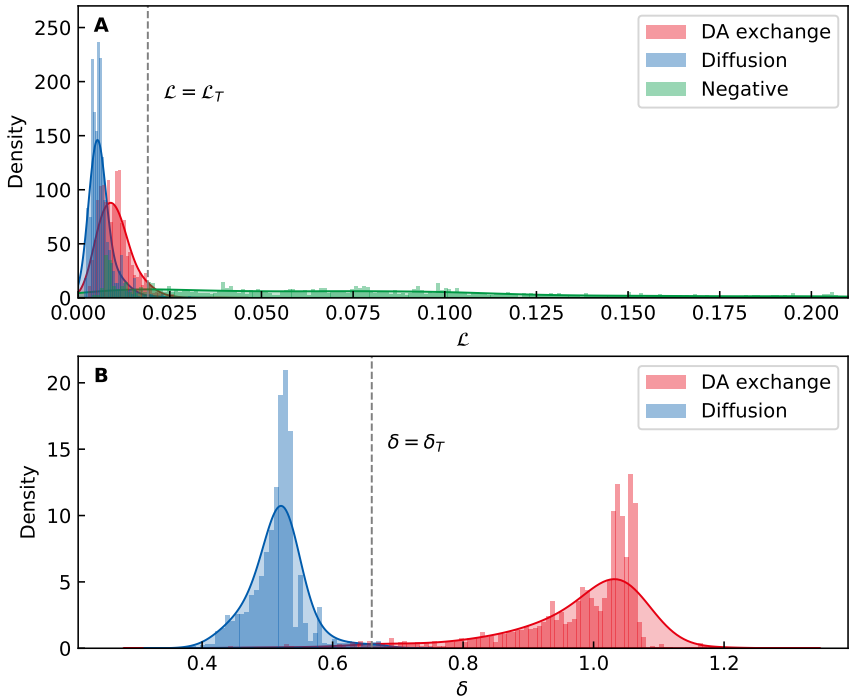


Figure: (A) Densities of reconstruction error $\mathcal{L}$ for interchange, diffusion, and negative sequences. (i) BLSTM AE can reconstruct positive sequences well. Hence, the reconstruction errors for interchange and diffusion sequences are relatively small, mainly less than $\mathcal{L}_T$. (ii) Since 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 $\delta$ for interchange and diffusion sequences. The two densities are significantly different from each other.

# Proportions of interchange

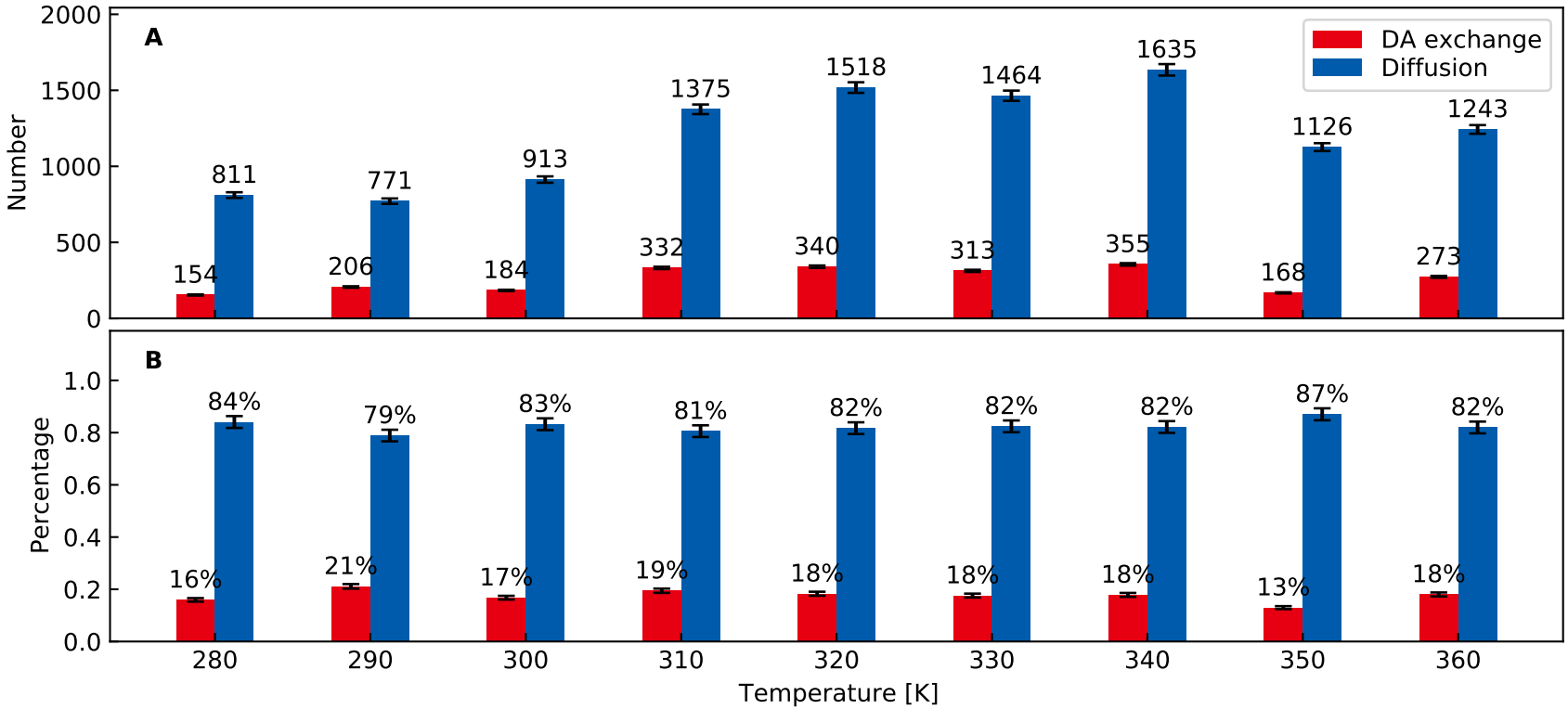


Figure: The **number** (A) and **proportion** (B) of interchange and diffusion processes determined by the RNN-based classifier at different temperatures. (i) With the temperature increasing, the number of interchange and diffusion processes **increases first and then decreases** on the whole. (ii) The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

# The trend of interchange and diffusion process number

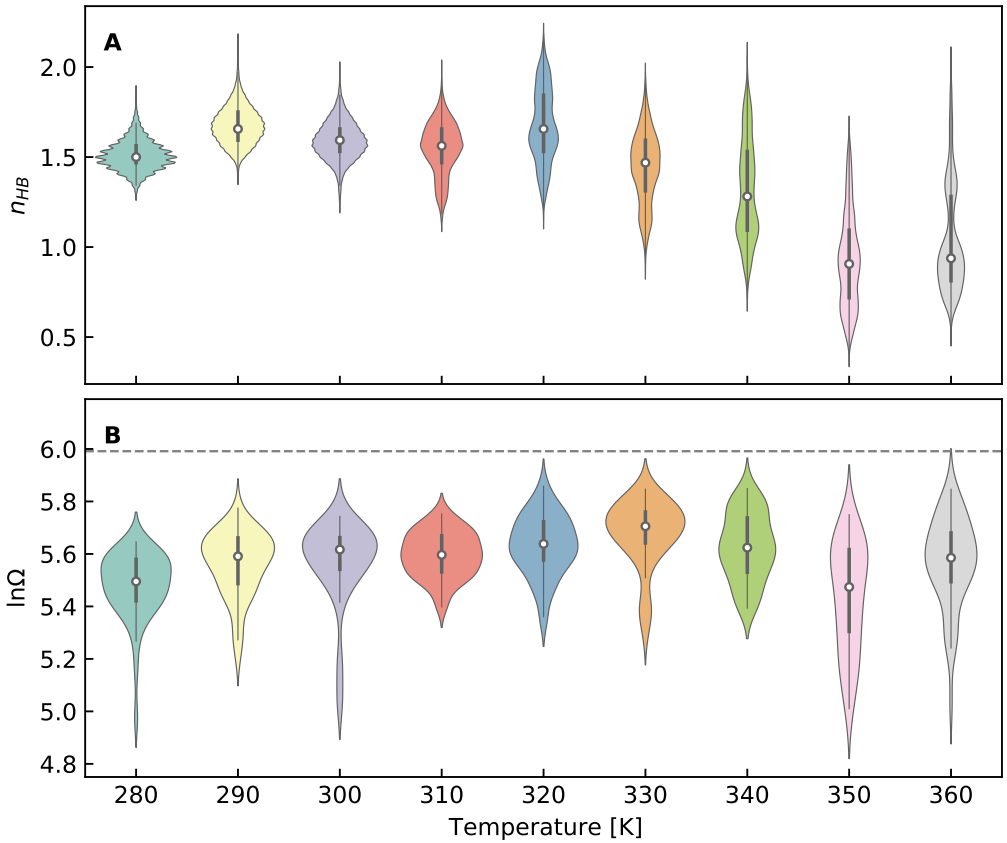


Figure: The temperature dependence of (A) The distributions of the number $n_{\mathrm{HB}}$ of H-bonds per molecule. (B) The distributions of $\ln \Omega$ characterizing the rate of H-bond breakage and reforming. The dashed line denotes the maximum value of  in the unit time of 1 ps.

# Conclusions

1. In this work, we **observed the interchange process in bulk water** by keeping our eyes on water molecule pairs. The **relative ratio of interchange and diffusion processes is approximately 1:4**. This ratio hardly depends on temperature, indicating the universality of the interchange process in water.
2. We used the dynamic graph and newly defined directed H-bond population to model the water system. This reasonable coarse-grained description of the H-bond network **simplifies the analysis of H-bond dynamics** dramatically.
3. Besides, the RNN-based method is used to successfully classify different types of H-bond population sequences, implying the **great potential to use deep learning to understand more complex dynamic processes** in water.

# Structure factor model of polymer chains

# Motivation

The structure factor of a polymer system defined as

$$S(\mathbf{k}) = \frac{1}{\rho}\int_V \langle \rho(\mathbf{r})\rho(0)\rangle \exp(i\mathbf{k}\cdot\mathbf{r})\mathrm{d}\mathbf{r}$$

is a measurable physical property, which characterizes the density-density correlation of the system.

1. First, we attempt to find a **more efficient formulation of structure factor for wormlike chains in the entire parameter space, but the more direct way** where we don't need to do any heavy calculation like Monte Carlo simulations or solve partial differential equations.
2. Second, we want to build **a possible measure tool for the scattering experiments of polymer chains**. If scattering intensity data is given, the contour length $L$ and Kuhn length $a$ can be easily obtained.

# Deep neural network-based structure factor model

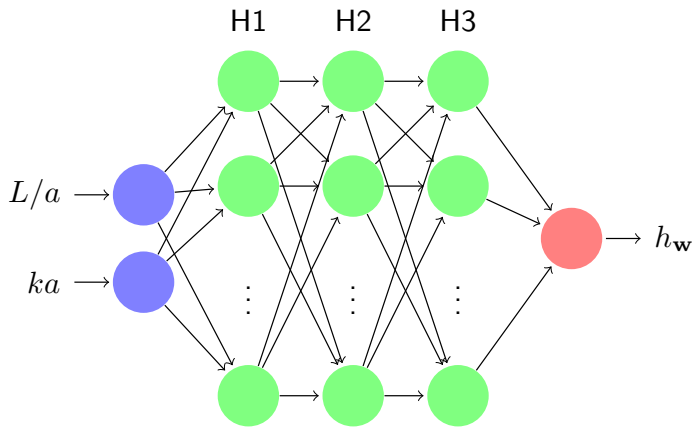


Figure: A fully connected NN with 3 hidden layers.

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{N} \sum_x \|f(x) - h_{\mathbf{w}}(x)\|^2$$

$f(\mathbf{x}) = (L/a)(ka)^2 S(L/a, ka)$ is the label, also the solution to the solution to the modified diffuse equation (MDE) [2], corresponds to the input $\mathbf{x} = (L/a, ka)$

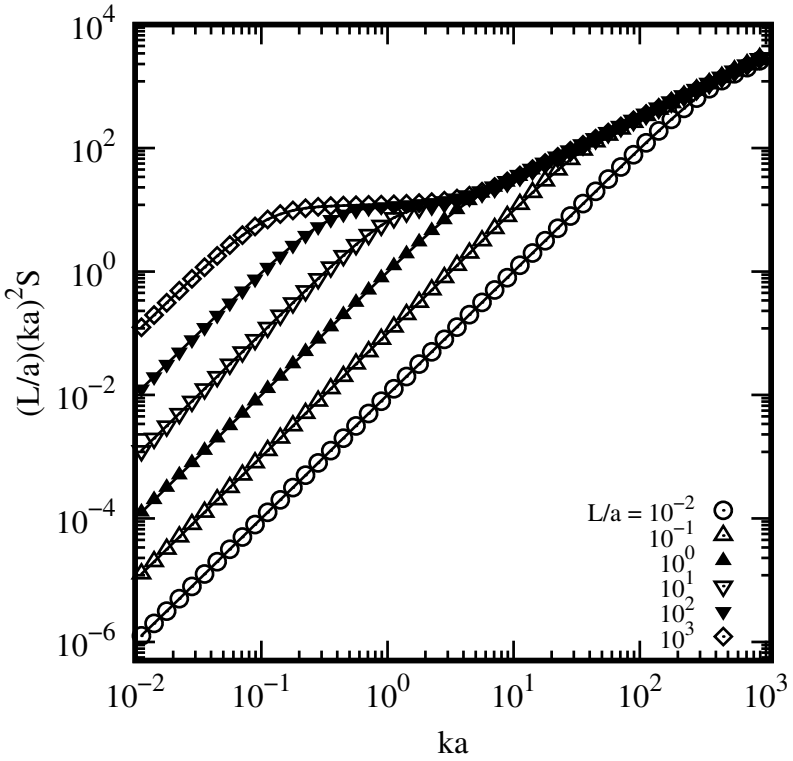


Figure: Structure factor comparison between the target values (circles, triangles, etc) [2] and trained NN predictions (lines) in logarithmic coordinates with 4 hidden layers and 25 nodes on each hidden layer with $Loss = 6.92 \times 10^{-7}$.

[2] Zhang et al, Soft Matter 10, 5405 (2014).

# Predict the contour length and Kuhn length of polymer chains

$$I_p(q) = cP(q)S(q)$$

$$P(q) = \left[\frac{2J_1(Rq)}{Rq}\right]^2$$

$$\epsilon(a, L, R, c) = \frac{1}{N}\sum_{i=1}^{N}\left(I_p^i(a, L, R, c) - I^i\right)^2$$

| | Upper | Lower |
|---|---|---|
| Target $L$ | 1360 | 1810 |
| NN $L$ | 1300.03 | 1573.76 |
| Target $a$ | $22 \sim 27$ | $22 \sim 27$ |
| NN $a$ | 22.38 | 22.17 |
| $\epsilon$ | 0.00068 | 0.0013 |

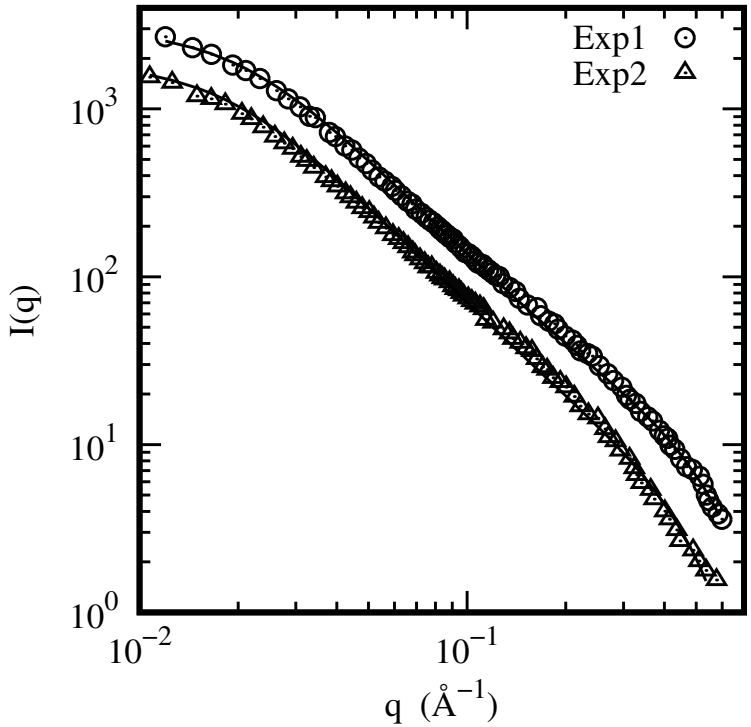


Figure: Scattering intensities comparison between the SANS data [3] and the trained NN model prediction for PS with molecular wight $M_w$ of 50000 in $CS_2$. The Exp1 is for the phenylring deuterated PS, and the Exp2 is for fully deuterated PS.

[3] M. Rawiso et al, Macromolecules, 20, 3 (1987)

# Conclusions

1. Our NN model is of the following characters: (a) High-precision, **continuous numerical solutions in the entire $L/a$–$ka$ space** can be obtained easily; (b) It is highly **consistent with the calculations in previous numerical and analytical method**.
2. Besides, we also proposed one application of the model. Combining SANS intensity data we can **determine the contour length and Kuhn length** of polymer chains.

# Publications

- **Jie Huang**, Gang Huang*, and Shiben Li*, "A machine learning model to classify dynamic processes in liquid water", ChemPhysChem 23, e202100599 (2022)
- **Jie Huang**, Shiben Li*, Xinghua Zhang*, and Gang Huang , "Neural network model for structure factor of polymer systems" , The Journal of Chemical Physics 153, 124902 (2020)

# Acknowledgment

- I would like to express my gratitude to **Prof. Shiben Li**, my supervisor, for his patient guidance, useful suggestions, and enthusiastic encouragement.
- I wish to express my appreciation for **Prof. Jeff Z. Y. Chen**, **Dr. Ying Jiang**, and **Prof. Xinghua Zhang** for their valuable suggestions and helpful guidance.
- I would also like to thank **Dr. Gang Huang**, for his advice and assistance.

# *Ab initio* molecular dynamics simulations

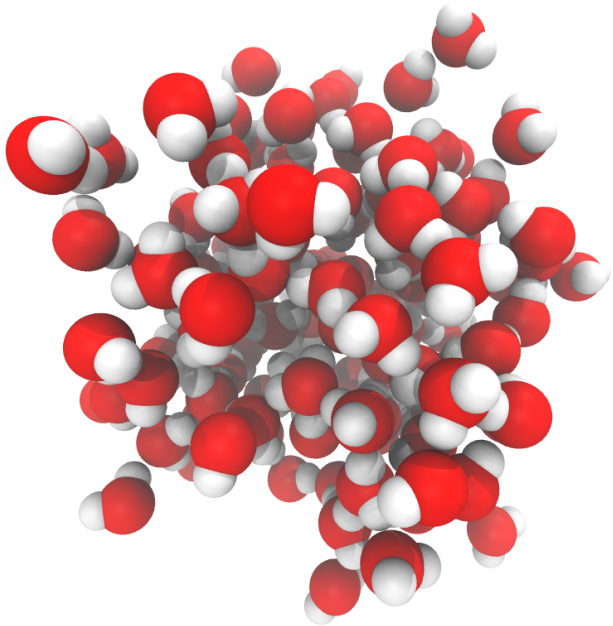

Figure: **AIMD simulations** were carried out for bulk water of **64 water molecules** within the canonical NVT ensemble using **CP2K/QUICKSTEP** (v7.1). The number $N$ of water molecules was 64 for all bulk water systems at different temperatures from **280 to 360 K**. The length of the periodic cubic box was **12.4295** Å. The discretized integration time step $\Delta t$ was set to 0.5 fs. The simulation time was **60 ps**.

# Trajectory analysis


Figure: MDAnalysis (v1.0.0) 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.

# BLSTM AE classifier

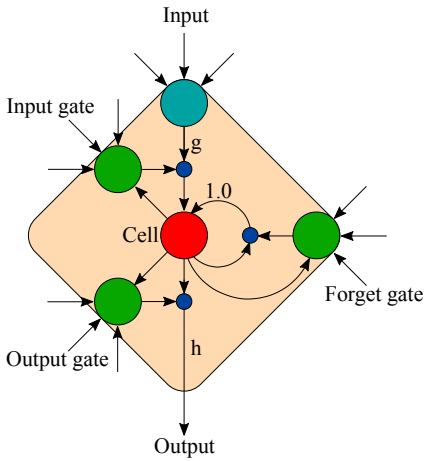


Figure: **LSTM unit**. The outcome of the gates is to allow the cell to store and access information over long periods[4].

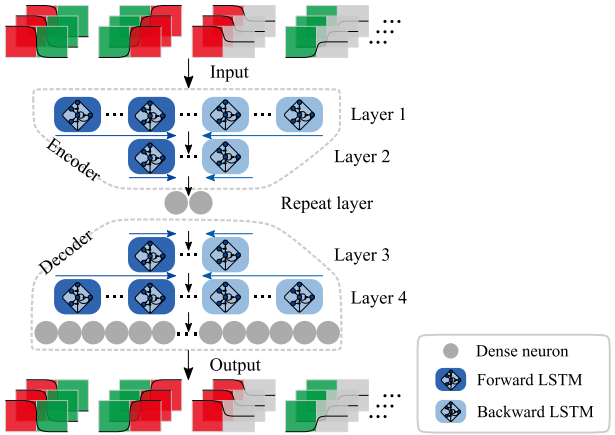


Figure: The structure of **BLSTM AE**. This kind of design is used to **identify positive and negative** $\tilde{h}$ sequences. The **training data** for the BLSTM AE are the filtered positive $\tilde{h}$ sequences. Since $\tilde{h}$ sequences are time-varying sequences, we choose to use the LSTM unit 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.

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

[3] A. Graves et al., vol. 31, no. 5, pp. 855-868

# Demonstration 1

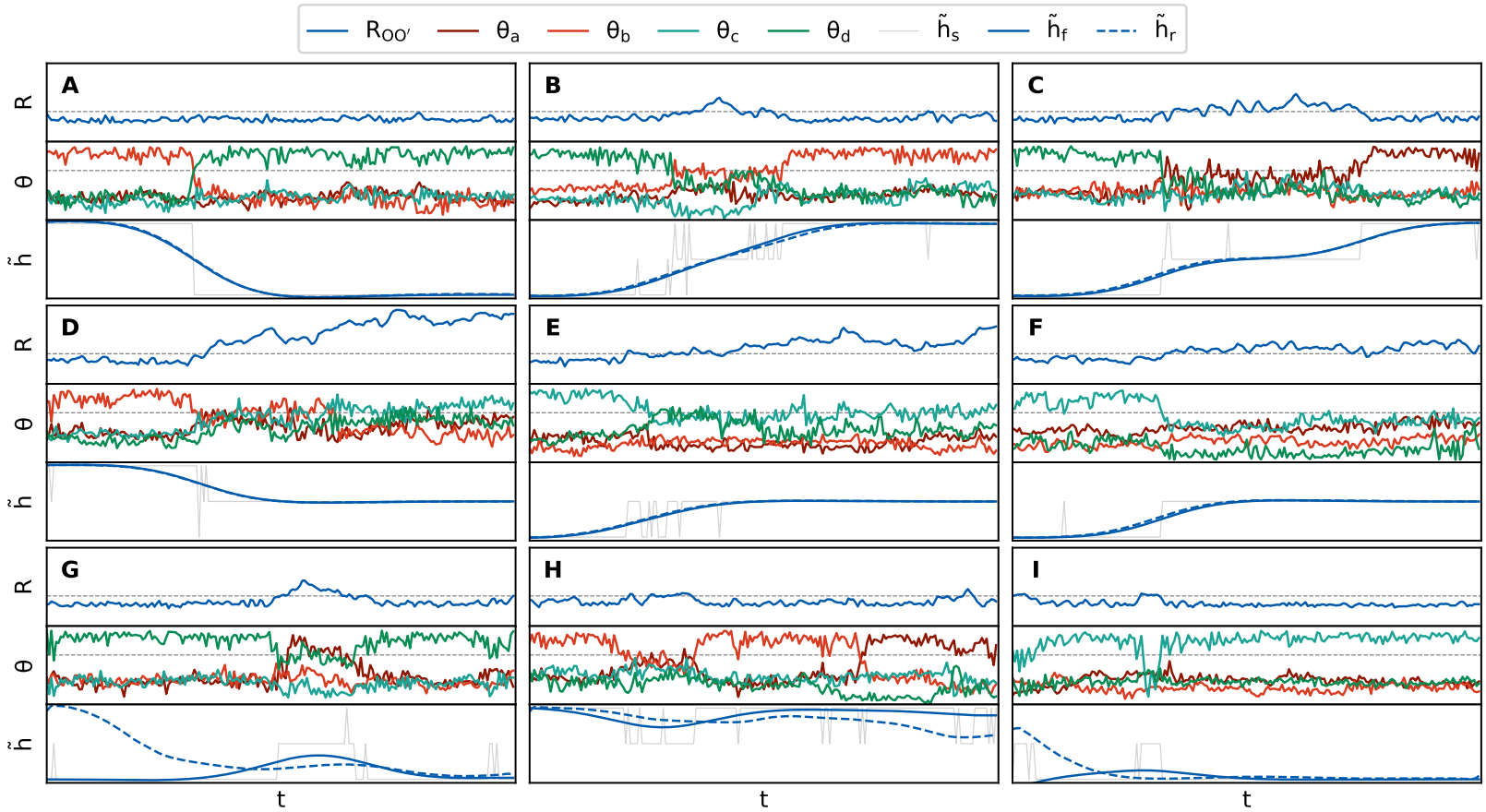


Figure: Different types of H-bond configuration change processes. (A), (B), and (C) are **interchange** processes; (D), (E), and (F) are **diffusion** processes; (G), (H), and (I) are **negative** processes.

# Demonstration 2

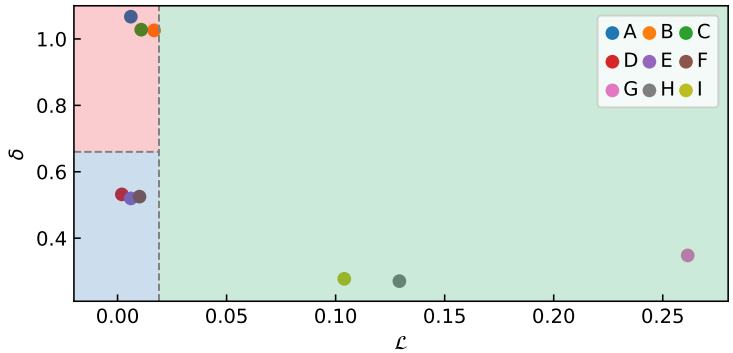


Figure: Classification results for sequences. The background colors represent the predictions of the BLSTM AE classifier. **Red, blue, and green denote the interchange, diffusion, and negative process**, respectively, indicating that the BLSTM AE classifier can correctly classify the H-bond configuration change processes.

# Step size effect of the sliding window

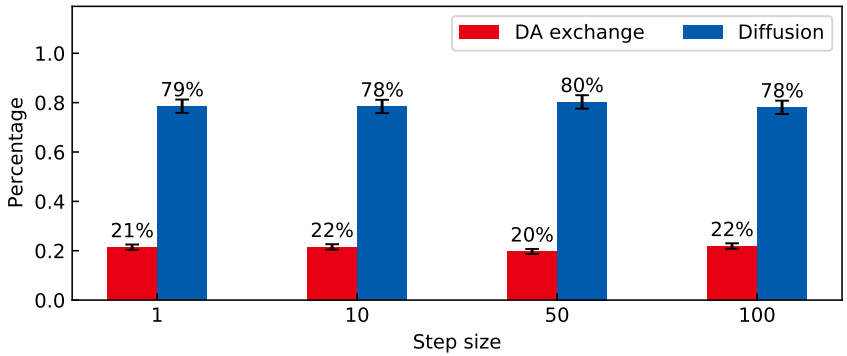


Figure: The relative ratios of interchange and diffusion processes under different step sizes for the simulation bulk water at 310 K. Since we use the sliding window method for sampling the dynamic trajectory of $\tilde{h}$ to obtain the 8 ps sequences, we take the sliding step as a parameter to observe the relative ratios of interchange and diffusion processes. We find that this relative ratio is almost unaffected by the step size of the sliding window.

# Number of H-bonds per molecule and H-bond relaxation time

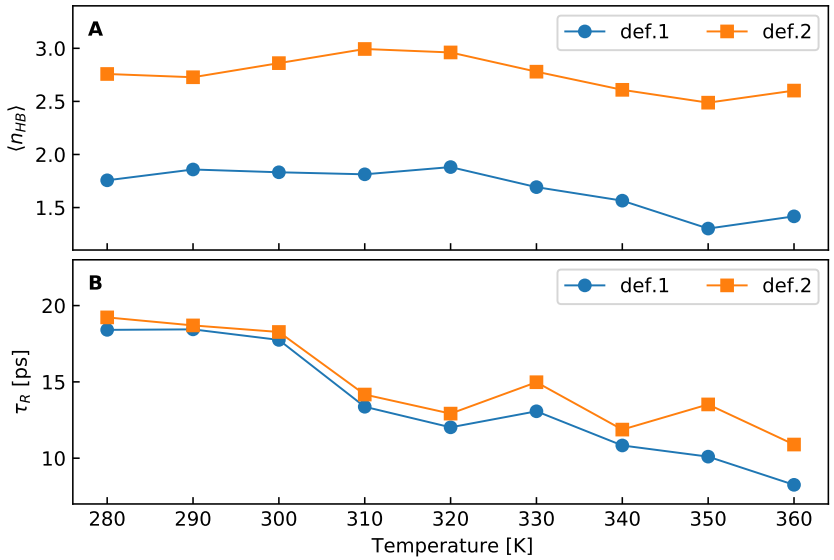


Figure: (A) Mean number $\langle n_{\mathrm{HB}} \rangle$ of H-bonds per water. (B) Relaxationtime $\tau_{\mathrm{R}}$ of H-bonds. Both $\langle n_{\mathrm{HB}} \rangle$ and $\tau_{\mathrm{R}}$ are calculated for two geometric definitions. Figure 21 shows that the different definitions of H-bond may cause some differences in observations. However, the relationship of $\langle n_{\mathrm{HB}} \rangle$ and $\tau_{\mathrm{R}}$ with temperature changes is consistent. Definition 1: $R_{\mathrm{OO}'} < 3.5$ Å, $\widehat{\mathrm{OHO}'} > 120^{\circ}$; Definition 2: $R_{\mathrm{OO}'} < 3.5$ Å, $\widehat{\mathrm{HOO}'} < 30^{\circ}$.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.

# Velocity autocorrelation function and vibrational density of states

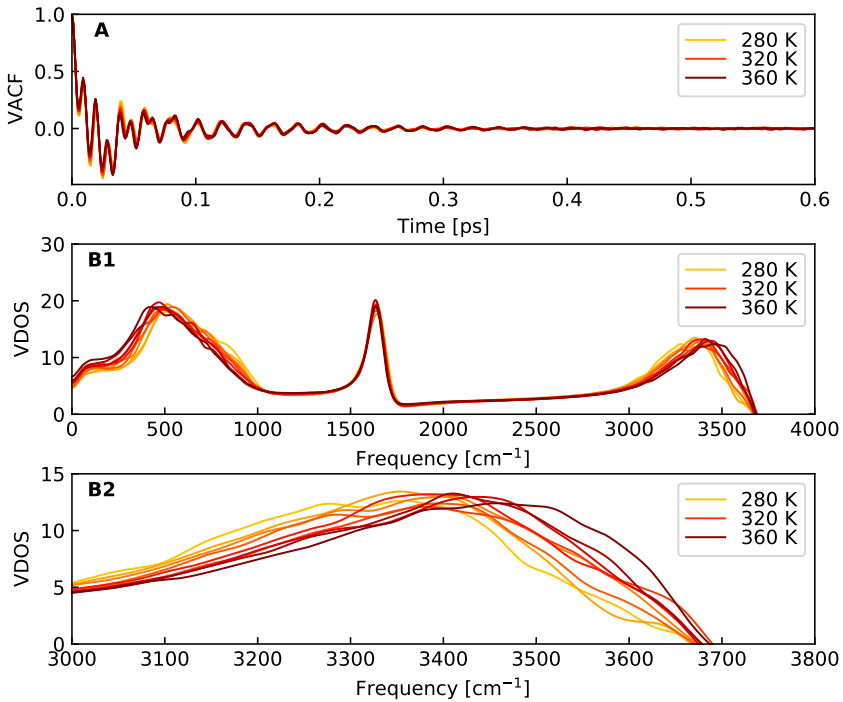


Figure: Velocity autocorrelation function (A) and vibrational density (B) of states. We can see that with the increase of temperature, the peaks of OH stretching bands are blue-shifted. This result means that the increasing temperature causes a higher frequency of OH stretching. As OH stretch frequency is correlated to the strength of H-bonds in which the OH bonds are involved, the blue-shifted OH stretch band has been assigned to weakly H-bonded water.

# The mathematical model for a neuron

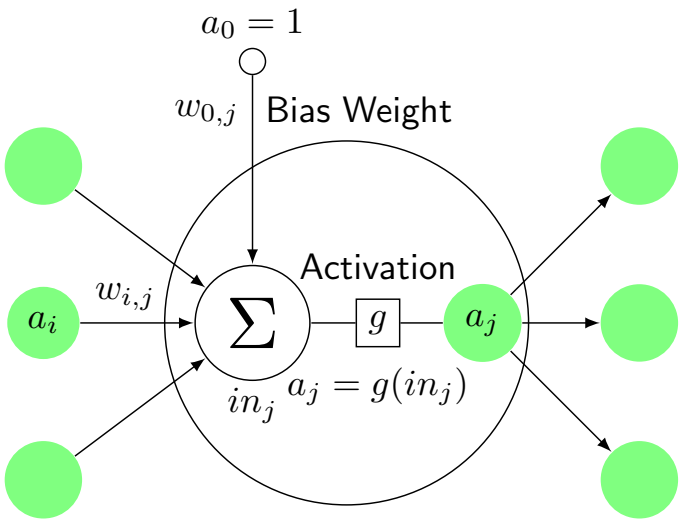


Figure: The mathematical model for a neuron.

$$in_j = \sum_{i=0}^{n} w_{i,j} a_i.$$

$$a_j = g(in_j) = g\left(\sum_{i=0}^{n} w_{i,j} a_i\right)$$

$$g(z) = \frac{1}{1 + e^{-z}}.$$