分 类 号：O469　　UDC：530

学　　号：184511084146　　密级：公开

温州大学

硕 士 学 位 论 文

机器学习在体相水氢键动力学和高分子链结构因子中的应用

作 者 姓 名：黄杰

培 养 类 型：学术型

专 业 名 称：凝聚态物理

研 究 方 向：生物物理与高分子物理

指 导 教 师：李士本 教授

完 成 日 期：2021 年 3 月

温州大学学位委员会

温州大学学位论文独创性声明

本人郑重声明：所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得温州大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

论文作者签名：　　日期：　　年　　月　　日

温州大学学位论文使用授权声明

本人完全了解温州大学关于收集、保存、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权温州大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。本人在导师指导下完成的论文成果，知识产权归属温州大学。

保密论文在解密后遵守此规定。

论文作者签名：　　导师签名：

日期：　　年　　月　　日　日期：　　年　　月　　日

机器学习在体相水氢键动力学和高分子链结构因子中的应用

摘　要

水分子在液态水中的重定向和扩散在许多过程中都是必不可少的。我们使用基于密度泛函理论的分子动力学方法对液态体相水进行了模拟。我们发现液态水中一种具有明显特征的氢键构型改变过程: 氢键的供体-受体交换 (DA 交换)。为了找到 DA 交换在氢键构型改变过程中的占比，我们设计了一种基于循环神经网络的分类模型，对氢键构型改变过程进行分类。通过该模型，我们发现液态体相水在不同温度下的 DA 交换和扩散过程的相对比例约为 1:4。尽管 DA 交换和扩散过程的数量会随着温度的变化而改变，但是其相对比例基本上不变。该特征表明 DA 交换在液态体相水氢键网络动力学中的普遍性。

作为了解高分子链内部结构的重要物理量，结构因子无论是在理论上还是实验上都受到极大的关注。通过深度神经网络，我们获得了一个有效的模型来计算高分子链的结构因子，而无需考虑波数和链刚性的不同区域。此外，利用训练后的神经网络模型，结合散射实验数据，我们预测了一些聚合物链的链长和库恩长度，结果表明我们的模型可以得到相当合理的预测结果。这项工作提供了一种计算高分子链结构因子的方法，该方法的预测精度高，同时有很好计算效率。另外，我们也为实验研究人员提供了一种测量任意高分子链的链长和库恩长度的方法。

关键词： 氢键, 水的重定向, 密度泛函分子动力学，结构因子，高分子物理, 机器学习, 循环神经网络

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

ABSTRACT

The reorientation and diffusion of water molecules in liquid water are essential to a wide range of processes. Based on the Density Functional Theory-based Molecular Dynamics (DFTMD) simulation of bulk water, we designed a Recurrent Neural Network (RNN) based model to classify water molecule pairs' configuration changes as the hydrogen bond network evolves. Although we found that there are many types of configuration changes from the simulation trajectory, one has apparent characteristics: donor-acceptor (DA) exchange the configuration changes found in experiments or simulations related to water dimers. Besides, through our model, we determined the relative proportions of DA exchange and diffusion. Although the absolute number of the two processes will fluctuate with the increase of temperature, the relative proportion is basically invariable. This feature implies the universality of DA exchange processes of water molecules in the hydrogen bond network.

As a substantial physical quantity to understand polymer chains' internal structure, the structure factor is studied both in theory and experiment. In this work, by training a deep neural network (NN), we obtained an efficient model to calculate the structure factor of polymer chains without considering different wavenumber and chain rigidity regions. Furthermore, based on the trained neural network model, we predicted the contour and Kuhn length of some polymer chains using scattering experimental data. We found our model can get pretty reasonable predictions. This work provides a method to obtain structure factor for polymer chains, which is as

good as previous and more computationally efficient. Also, it provides a potential way for the experimental researchers to measure the contour and Kuhn length of polymer chains.

KEY WORDS: hydrogen bonds, water reorientation, DFTMD, structure factor, polymer physics, RNN, machine learning

目录

第一章　绪论

1　机器学习简介

1.1　机器学习

让我们来思考这样一个问题。对于图1-1中的手写数字[1]，我们想要得到一个算法: 给出一张手写数字的图片, 它能准确告诉我们这张图片上的数字。该如何实现这样的算法呢? 或许你看出数字 0 大概是一个圈，所以可以让算法建立这样的规则: 若图片中的黑色像素构成一个圈，那么这张图片里的数字就是 0。但你还发现，数字 9 和 6 可能也会构成一个圈，它们除了一个圈还有多余的部分。于是又可以让算法添加规则: 若圈的上方有黑色像素那么这个值就是 6，下方有黑色像素就是 9。接着还可以添加更多的规则。但这样的算法很天真！实际上，图片中数字的线条粗细不均匀，一些数字会倾斜，一些数字 0 不是完整的圈等等。这给算法识别其中的数字带来了极大的困难。为什么像这样简单的任务，算法却处理不好呢? 其中一个很重要的原因是，基于规则的算法很难考虑所有的情况。其实，我们注意到，虽然我们刚刚可以总结出那么多的规则，但当人类对一张图片中的数字进行识别的时候，我们仿佛没有思考，答案自动浮现。当人在做出判断之前，人对于数字的判断已经有了相当多的经验，这种经验来自我们看过很多数字的画面。在手写字符识别的例子中，最开始人们使用基于规则的算法，但是识别效果很难得到突破性的提升，算法的通用性也不强。因此人们放弃主动制定规则的思路。而尝试使用向学习系统演示正确示例的方法，让其自己掌握“认数字”这项技能。我们给一个判别系统演示大量图片及其对应数字的示例，并让其自己去提取需要的特征来做判断。这就引出了机器学习。

机器学习 (Machine Learning) 是通过数据的方式自动提升算法性能的研究方法。机器学习算法建立在数据之上。这些数据被叫做*训练数据*。训练数据的意义在于将人们从设计算法规则当中解脱出来，让模型通过数据来学习，进而自动得到算法规则。如今机器学习已经在各个领域有着非常广泛的应用，例如手写字符识别，人脸识别，垃圾邮件过滤，语音识别等。若使用传统基于规则的算法，那么这些应用几乎是无法完成的。

深度学习 (Deep Learning) 是建立在人工神经网络 (Artificial Neural Networks, ANNs) 上的机器学习的一部分。人工智能，机器学习和深度学习的关系如图1-2。如今深度学

图 1-1 一些手写数字的样本。
Figure 1-1 Some samples of handwritten digits.

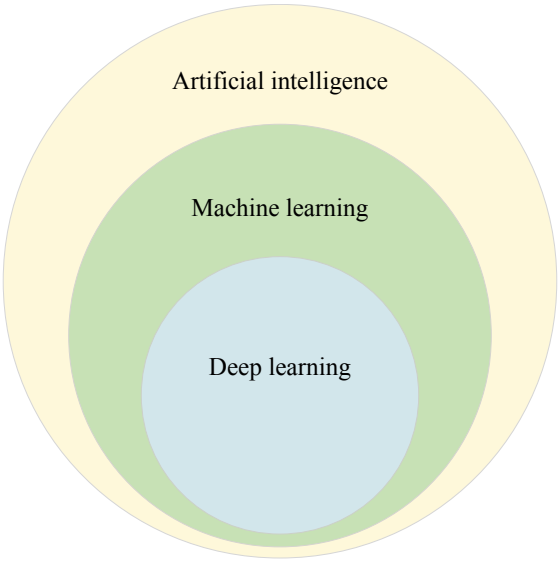

图 1-2 人工智能，机器学习和深度学习之间的关系。机器学习是人工智能的一个分支，深度学习是机器学习的一个分支[2]。不过也有人认为机器学习不全是人工智能，仅仅只有机器学习中智能的那一部分算为人工智能。

Figure 1-2 The relationship between artificial intelligence, machine learning and deep learning. Machine learning is a branch of artificial intelligence, and deep learning is a branch of machine learning.[2] However, some people think that machine learning is not all artificial intelligence, only the intelligent part of machine learning is counted as artificial intelligence.

习的结构有很多，诸如深度神经网络 (Deep Neural Network, DNN)，深度信念网络 (Deep Belief Network, DBN)，循环神经网络 (Recurrent Neural Network, RNN) 和卷积神经网络 (Convolutional Neural Network, CNN) 之类的深度学习架构已应用于机器视觉，语音识别，自然语言处理，音频识别，社交网络过滤，机器翻译，生物信息学，药物设计，医学图像分析和棋盘游戏等。它们产生的结果可与人类专家的表现相媲美，甚至在某些情况下超越人类专家[3–5]。

从学习模式上来说，机器学习可以是*有监督*的 (Supervised)，*半监督*的 (Semi-supervised) 或*无监督*的 (Unsupervised)。有监督的学习指的是训练计算模型的时候，我们将数据和数据对应的答案 (标签, Label), 同时输入到计算模型，让它学习数据和标签之间的映射关系。拿手写字符识别为例子，每一张图片都有一个对应的标签，范围从 0 到 9。一张图片的数据和它的标签构成的*数据-标签对*就构成了一个有监督的*训练数据* (Training data)。由训练数据构成的集合称为*训练集* (Training set)。在无监督学习中，数据是不带标签的，经过训练的模型能够对输入的数据进行分类或分群。常见的无监督学习算法有聚类分析 (Cluster analysis)[6,7]，降维 (Dimensionality reduce)[8,9] 等。半监督学习在训练过程中将少量标记的数据与大量未标记的数据结合在一起，它是介于无监督学习和监督学习之间的一种学习方法。

1.2 神经网络

人工神经网络 (Artificial Neural Network，ANN)，简称神经网络 (Neural Network，NN)，在机器学习和认知科学领域，是一种模仿生物神经网络结构和功能的计算模型，用于对函数进行估计或近似[10]。如今，神经网络的结构也是百花齐放。我们需要根据要解决的问题合理地选择最适合的神经网络模型。

最基本的神经网络结构是多层感知机 (MultiLayer Perceptron, MLP)。多层感知机是

一种前向结构的神经网络模型，它映射一组输入向量到一组输出向量。多层感知机可以被看作是一个有向图，由多个节点层组成，每一层上的每一个神经元都与下一层的每一个神经元相连接。除了输入节点，每个节点都是一个带有非线性激活函数的神经元 (或称处理单元)。一种被称为反向传播算法 (Backpropagation) 的监督学习方法常常被用来训练多层感知机。多层感知机遵循人类神经系统原理，学习并进行数据预测。它首先学习，然后使用权重存储数据，并使用类似反向传播的优化算法来调整权重并减少训练过程中实际值和预测值之间的误差。多层感知机的主要优势在于其快速解决复杂问题的能力。多层感知机的基本结构由输入层，隐藏层和输出层组成。多层感知机是感知机的推广，克服了感知机不能对线性不可分数据识别的弱点。

最擅长处理有序数据的神经网络结构是循环神经网络 (Recurrent Neural Network, RNN)。RNN 中节点之间的连接沿时间序列形成有向图。这使其表现出时间动态行为。RNN 来源于前馈神经网络，可以使用其内部状态来处理长度可变的输入序列[11]。这使得 RNN 适用于连续手写识别[12]，语音识别[13] 之类与时间序列相关的任务。长短期记忆[14](Long Short-Term Memory，LSTM) 是一种循环神经网络。由于独特的设计结构，LSTM 非常适合用来处理和预测时间序列中间隔和延迟非常长的重要事件, 它还解决了一般循环神经网络不可避免的梯度消失问题。

另外一种常见的神经网络结构是自编码 (Autoencoder, AE)，也称为自编码器。它是一种在无监督学习中用于有效编码的网络结构。自编码的目的是通过对一组数据忽略掉“噪声”的学习，得到这组数据的一种表示 (也称表征，编码)。自编码的一个应用是异常检测 (Anomaly detection)[15–18]。通过学习训练数据中最显著的特征，可以鼓励自编码学习精确地再现最频繁观察到的特征。当面对异常时，自编码应该恶化其重建性能。在大多数情况下，我们仅使用具有正常实例的数据来训练自编码。经过训练后的自编码将准确地重建“正常”数据，而对陌生的异常数据则无法重建[16]。原始数据与其重建数据之间的重建误差被用作检测的异常评分[16]。

近年来，人工神经网络在高分子学科中也得到了广泛的应用。例如，神经网络可以被用来对物质的相进行分类[19]，求解非线性偏微分方程 (Partial Differential Equations, PDE)[20]，预测大分子的结构[21]，对聚合物构型分类[22]，预测蛋白质结构[23] 等。

2 研究内容

本文介绍了机器学习在凝聚态物理中的两个具体应用。

一. 使用 *ab initio* 分子动力学 (Ab Initio Molecular Dynamics, AIMD) 模拟和深度学习的方法来对纯水中氢键构型变化进行分类。首先我们用 AIMD 模拟对液态体相水进行模拟，使用动态有向图的方法对模拟的体相水体系进行粗粒化建模。我们发现了液态水中的一种特殊的氢键构型改变: 供体-受体交换 (DA 交换)。为了找出氢键状态改变过程中的 DA 交换，我们设计了一个基于 LSTM 的自编码用来预测氢键构型状态改变的分类器。通过设计的分类器，我们找到了液态体相水中 DA 交换和扩散 (Diffusion) 的相对比例约为 1:4，这证明了 DA 交换过程在氢键动力学中普遍存在。最后我们探究了温度对氢键状态的影响。发现温度会导致水中的 DA 交换和扩散的绝对数目发生改变，但 DA 交换和扩散的相对比例基本保持不变。

二. 用神经网络建立高分子链的结构因子 (Structure Factor) 模型。我们首先利用全联接网络 (Fully Connected Neural Network, FCNN) 对在全刚性-波数 (L/a-ka) 空间下高分子

链的结构因子进行建模，得到了一个精确的结构因子模型。接着利用得到的结构因子模型, 我们设计了一种测量任意高分子链的链长 (Contour length) 和库恩长度 (Kuhn length) 的算法。我们对已有的一些高分子链的中子散射数据做预测，准确地预测出了高分子链的链长和库恩长度。基于神经网络的高分子链的结构因子模型为实验工作者提供了一种测量高分子链的有效工具。

第二章　液态体相水中的氢键构型改变

本章我将使用 *ab initio* 分子动力学模拟和深度学习的方法研究液态体相水 (Bulk water) 中的氢键动力学过程。使用基于密度泛函的分子动力学 (Density Functional Theory-based Molecular Dynamics, DFTMD) 模拟这种具体的 *ab initio* 分子动力学方法, 我们发现了体相水中一种特殊的氢键构型改变过程，即供体-受体交换过程 (DA 交换): 构成氢键的两个水分子迅速互换角色。这种构型改变过程在一些金属表面的二聚体水中已经被实验证实[24–26]，但尚未在体相水中被实验验证。为了确定这种 DA 交换过程在体相水中是否普遍存在，我们结合深度学习的方法，设计了一种基于循环神经网络 (RNN) 的模型来对氢键构型改变做分类。通过该方法，我们发现 DA 交换和水分子扩散 (Diffusion) 这两种氢键状态改变过程发生概率的相对比例约为 1:4，并且这个比例基本不随温度改变。这个结果表明 DA 交换过程在液态体相水中是大量存在的，具有普遍性。由于 DA 交换与水中氢键网络的生成和重建密切相关，因此它在诸如质子运输和大分子 (如蛋白质) 的水合等过程中可能起到重要作用。

1　引言

液态水是自然界中最常见的液体，它占据了大约 71% 的地球表面积。人们已经知道水的独特性质，例如温度下降密度降低，4 摄氏度的时候密度最大，很高的表面张力等都和水中的氢键网络的动态变化有着紧密的联系[27]，但很少有实验研究能够从原子的尺度去充分地理解水的性质[28]。然而理解水的结构，作为 21 世纪最重要的 125 个问题[29]之一，对于理解细胞，生物组织，以及生态系统都至关重要[30–33]。

水分子是如何运动的? 对这个问题，人们一直没有停止探索。Mitsui 等人使用扫描隧道显微镜 (Scanning tunneling microscopy, STM) 探究了 Pd(111) 表面的水分子的扩散和聚集行为，从原子层面观测到了二聚体水可以快速扩散[24]。随后，Ranea 等人利用 *ab initio* 分子动力学模拟的方法模拟了 Pd (111) 表面的二聚体水。他们发现水分子间的一种氢键供体与受体的互换过程 (*DA 交换*) 可以被用来解释二聚体水的快速扩散[25]。再后来，Kumagai 等人利用 STM 直接观测到了 Cu (111) 表面的 DA 交换过程，他们称之为氢键交换 (Hydrogen-Bond Exchange)，他们认为在 DA 交换过程中存在量子隧穿[26]。最近，Fang 等人使用 DFTMD 模拟探究了二聚体水在表面快速运动的起源。他们发现二聚体水快速扩散背后的原因涉及到多种不同的水分子运动过程，其中包括 DA 交换，距离扩散和旋转等[34]。这些研究从实验角度出发，研究了受限在金属表面的二聚体水的运动方式。人们发现 DA 交换对于研究水分子的运动，氢键的断裂和重组尤为重要。

然而，除了受限于金属表面的水之外还有很多不同环境下的水，例如其中最典型的液态体相水。仅仅研究二聚体水对于理解液态水的氢键网络的重组还是不够的。液态体相水中的水分子三维运动相比受限于金属表面的水分子的二维运动的自由度更高，尽管如今有时间分辨红外光谱 (IR)[35,36]，X 光散射，中子散射[37] 等实验手段，但要从原子尺度去观测液态体相水中的水分子的运动还是有很大的困难。因此人们把视线转到了计算机模拟。它允许人们从微观的角度去研究水分子的运动，而不受实验仪器的限制。计

算机模拟可以是实验的辅助手段，甚至在一定程度上其得到的结果会超前于实验。

Lagge 和 Hynes 正是用计算机模拟的方法研究了液态水中水分子的重定向问题。他们提出的大角度跳变模型[38] 对传统的 Debye 扩散理论做出挑战。Debye 扩散理论认为在水分子重定向的旋转扩散机制中，水分子以很小的角步长改变其取向[39–41]。然而，大角度跳变模型认为水分子的重定向涉及大角度跳变[38]。这个观点提出之后，Moilanen 等人使用超快 2D 红外振动回波化学交换光谱法 (CES) 对阴离子和水进行观测，直接观测到了水羟基氢键交换, 这对大角度跳变模型提供了有力的支撑[42]。另外，Ji 等人利用 2D 时间分辨光谱的方法在高氯酸盐水溶液中观察到氢键交换的大角度跳变机制[43], 再一次支持了大角度跳变模型。新的观念建立之后，人类对水的认识还在不断往前推进。Piskulich 等人利用分子动力学 (MD) 模拟的方法继续探讨了温度对于氢键交换动力学的影响，发现 EJM(Extended jump model, EJM) 模型可以很好地描述水分子重定向的原理细节[44]。

跳变模型涉及了多个水分子之间的重定向，人们也想要探究液态水中是否存在 DA 交换过程。尽管有一些模拟成果探究了水分子团簇中的 DA 交换过程[45–48]，但据我们了解，液态体相水中的 DA 交换过程还没有被研究过。我们好奇液态体相水中是否也存在这样的过程。如果答案是肯定的，那就表示 DA 交换过程和液态水中氢键网络的断裂和重建具有紧密的关系，那么 DA 交换过程就很有可能在质子运输和大分子 (例如蛋白质) 的水合作用等过程中起重要作用。接着一个自然的问题是: DA 交换过程在氢键的动态变化过程中的占比是多少?

为了回答这些问题，我们采用了 *ab initio* 分子动力学模拟[49] 的方法，对液态体相水进行了模拟。基于密度泛函的分子动力学 (DFTMD) 模拟是一种具体的 *ab initio* 分子动力学模拟方法，已经被用在了很多工作当中，并且很多模拟结果也在实验上得到了验证。相比其他基于势函数的动力学模拟来说，基于密度泛函的分子动力学模拟不对原子模型做任何假设，一切从头算起。因此模拟的精确度高，可以被用来模拟诸如液态水这样具有复杂多变的原子环境的系统。

通过分析由模拟生成的动力学轨迹，观测水分子之间的距离和相应的角度，我们发现体相水中除了存在水分子的扩散这样被人们熟知的动态过程外，确实还存在 DA 交换这类过程，即氢键的供体水和受体水的角色以一种耦合的方式发生互换。相比受限金属表面水分子的 DA 交换，体相水中的 DA 交换过程的自由度更高。为了找到液态水中的 DA 交换过程并以此验证 DA 交换普遍存在的猜测，我们结合深度学习设计了一种基于长短期记忆 (Long short-term memory, LSTM) 的方法。该方法能够对氢键状态改变过程进行分类，从而可以帮助我们找出液态体相水动力学过程中的 DA 交换。最终，我们得到了 DA 交换与水分子扩散在体相水的动力学过程中的相对比例，证明了 DA 交换过程在体相水系统中是普遍存在的。另外我们还探究了温度对于 DA 交换过程的影响。我们发现在 280K 到 360K 范围内，DA 交换和扩散过程的相对比例都基本保持为 1:4。

2 研究过程和结果

2.1 体相水系统的动态图模型

为了更方便地描述体相水系统中氢键网络的动态变化，我们使用图 (Graph) 来描述液态体相水系统。图2-1展示了包含 N 个水分子的模拟体相水系统中的三个不同时刻的

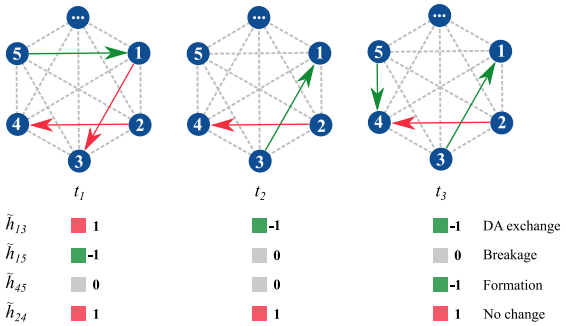

图 2-1 动态有向图用来表示各个时刻液态体相水中水分子之间氢键网络的变化。蓝色节点代表水分子。箭头表示形成氢键的方向，从供体指向受体。

Figure 2-1 A directed graph is used to demonstrate the H-bond network dynamics in the simulated bulk water system. The ordered blue nodes represent water molecules. An H-bond is indicated by an arrow, which is from the donor to the acceptor.

氢键网络连接状态图。第 i 个水分子和其他 $N-1$ 个水分子都有可能形成氢键。为了方便，我们称系统中任意两个水分子对 i, j 叫做一个准氢键 (Q-bond)，表示为 $\mathrm{QB}_{i,j}$，如图2-1虚线所示。于是液态水系统中氢键网络连接状态被表示成了一个有向图。由于氢键网络连接状态无时无刻都在改变，故有向图的连接状态随着时间不断发生改变。受到 Luzar 和 Chandler 提出的氢键布居[28] 的启发，我们定义一个类似的量 $\tilde{h}_{ij}(t)$，表示 t 时刻的准氢键 $\mathrm{QB}_{i,j}$ 的有向氢键布居：

$$\tilde{h}_{ij}(t) = \begin{cases} 1 & \text{存在氢键, } i \text{ 是供体, } j \text{ 是受体} \\ 0 & \text{不存在氢键} \\ -1 & \text{存在氢键, } i \text{ 是受体, } j \text{ 是供体} \end{cases} \tag{2-1}$$

有向氢键布居 $\tilde{h}$ 包含了氢键的方向信息，我们通过它不仅可以直接地看出氢键的供体和受体，还可以简化对氢键状态改变过程的描述。

2.2 氢键的几何定义

由于不是系统中所有的准氢键都能够形成氢键，因此我们主要关心的是那些更有可能形成氢键的准氢键。因为这些准氢键才和氢键网络的重建最相关。为了观测氢键构型的变化，并找到我们关心的准氢键，我们用下面的氢键几何标准[50–54] 来定义水分子间的氢键：

$$\begin{cases} R_{\mathrm{OO'}} < R_{\mathrm{cutoff}} = 3.5\ \text{Å} \\ \widehat{\mathrm{DHA}} > \theta_{\mathrm{cutoff}} = 120^\circ \end{cases} \tag{2-2}$$

这里的 $R_{\mathrm{OO'}}$ 是指供体氧原子和受体氧原子之间的距离，角度 $\widehat{\mathrm{DHA}}$ 是由供体氧原子，被贡献氢原子和受体氧原子组成的角度。如图2-2，为了研究液态水中的重定向和扩散等过程，我们观测准氢键 $\mathrm{QB}_{i,j}$ 的 5 个量，分别是两个水分子间的距离 $R_{\mathrm{OO'}}$，和四个角度 $\widehat{\mathrm{OH1O'}}$, $\widehat{\mathrm{OH2O'}}$, $\widehat{\mathrm{O'H'1O}}$ 和 $\widehat{\mathrm{O'H'2O}}$, 对应图中的 θ_{a}, θ_{b}, θ_{c} 和 θ_{d}。

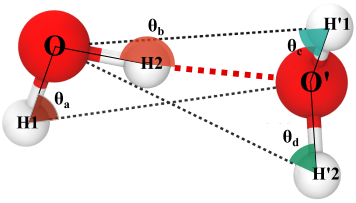

图 2-2 对于液态体相水中的任意一对水分子 (准氢键), 测量两个水分子之间的距离 $R_{OO'}$, 四个角度 $\theta_a, \theta_b, \theta_c, \theta_d$。我们使用距离阈值 $R_{cutoff} = 3.5$ Å 和角度阈值 $\theta_{cutoff} = 120°$ 来定义水分子之间的氢键。若 $R_{OO'} < R_{cutoff}$, 并且存在一个角 $\theta > \theta_{cutoff}$, $\theta \in \{\theta_a, \theta_b, \theta_c, \theta_d\}$, 那么认为这两个水分子之间存在氢键。图中展示了左侧水分子中的氧原子 O 贡献水分子的一个氢原子 H2 给右侧水分子的受体 O′。由于 $R_{OO'} < R_{cutoff}$ 并且 $\theta_b > \theta_{cutoff}$, 因此有向氢键布居 $\tilde{h}_{OO'}(t) = 1$, 即在这组水分子对中存在一个氢键，其方向是从氧原子 O 指向氧原子 O′。

Figure 2-2 For each selected water pair in simulated bulk water, a distance $R_{OO'}$, and four angles θ_a, θ_b, θ_c, θ_d are measured. To define a H-bond, a distance cutoff $R_{cutoff} = 3.5$ Å and an angle cutoff $\theta_{cutoff} = 120°$ are used. If $R_{OO'} < R_{cutoff}$, and any angle $\theta > \theta_{cutoff}$, $\theta \in \{\theta_a, \theta_b, \theta_c, \theta_d\}$, then a H-bond exists in this pair. This figure shows that the oxygen atom O of the water molecule on the left as a donor donates one of the hydrogen atoms H2 to the acceptor O′ on the right. Since $R_{OO'} < R_{cutoff}$ and $\theta_b > \theta_{cutoff}$, so $\tilde{h}_{OO'}(t) = 1$, i.e. an H-bond exists in this water pair at time t, and the direction is from O to O′.

2.3 氢键构型的改变

由 *ab initio* 分子动力学模拟产生的水分子的轨迹让我们可以观测液态体相水中水分子的运动细节。图 2-3展示了一个典型的准氢键的距离和角度的动力学过程。氢键判定标准中的距离阈值 R_{cutoff} 和角度阈值 θ_{cutoff} 也以水平虚线的方式展示出来。另外还同时展示了有向氢键布居 $\tilde{h}_{OO'}$ 的变化。在这个动态过程中, 我们注意到三种有趣的过程，分别对应图中的区间 I_1, I_2, 和 I_3。**DA 交换** (I_1): 我们注意到在区间前半部分有 $\theta_b > \theta_{cutoff}$, 在后半部分有 $\theta_d > \theta_{cutoff}$。另外, $\tilde{h}_{OO'}$ 从 1 变到 -1, 这就表示氢键的供体和受体发生了互换。**扩散** (I_2): 在区间的前半部分有 $\tilde{h}_{OO'} = -1$, 在区间后半部分的大部分时间 $\tilde{h}_{OO'} = 0$, 也就是没有氢键存在。距离 $R_{OO'}$ 明显增大。**HH 交换** (I_3): 由于一开始有 $\theta_c > \theta_{cutoff}$, 因此被捐献的形成氢键的氢原子是 H′1. 接着角度 θ_c 减小, 而角度 θ_d 则增大，直到 $\theta_d > \theta_{cutoff}$。也就是说, 在这一过程中供体的另一个氢原子 H′2 被捐献形成氢键，即供体的氢原子发生了交换。

从模拟的体相水轨迹中，我们看到 DA 交换和 HH 交换这两种过程。因此一个很自然的问题是：这些变换过程在液态水中普遍存在吗? 由于氢原子的全同性，人们很难区分 HH 交换前后的水分子构型的改变，因此我们认为 HH 交换对氢键网络的重建几乎没有影响。但是，对于参与 DA 交换的准氢键 $QB_{i,j}$，我们得到对应过程的有向氢键布居 $\tilde{h}_{ij}$，发现其值总是从 -1 变为 1，或从 1 变为 -1。该结果有力地证明了两个相邻的水分子以非常一致和快速的方式改变了它们的方向，并且同时改变了它们在与之相关的氢键中扮演的角色。另外，由于相邻水分子之间的 DA 交换，水分子的偶极矩方向会改变，因此水分子的微观构型也会改变。因此在这本文中，我们主要关注 DA 交换。接下来我将要讨论如何设计一个氢键状态变化过程的分类器，从而得到 DA 交换和扩散在液态体相水中的相对比例。

2.4 RNN 氢键构型改变分类器

我们构建分类器最主要动机是想要得到一个自动工具来对准氢键的氢键**构型改变**过程做出分类。由于从有向氢键布居 $\tilde{h}$ 中可以非常直观地看出 DA 交换和扩散这两种过程。

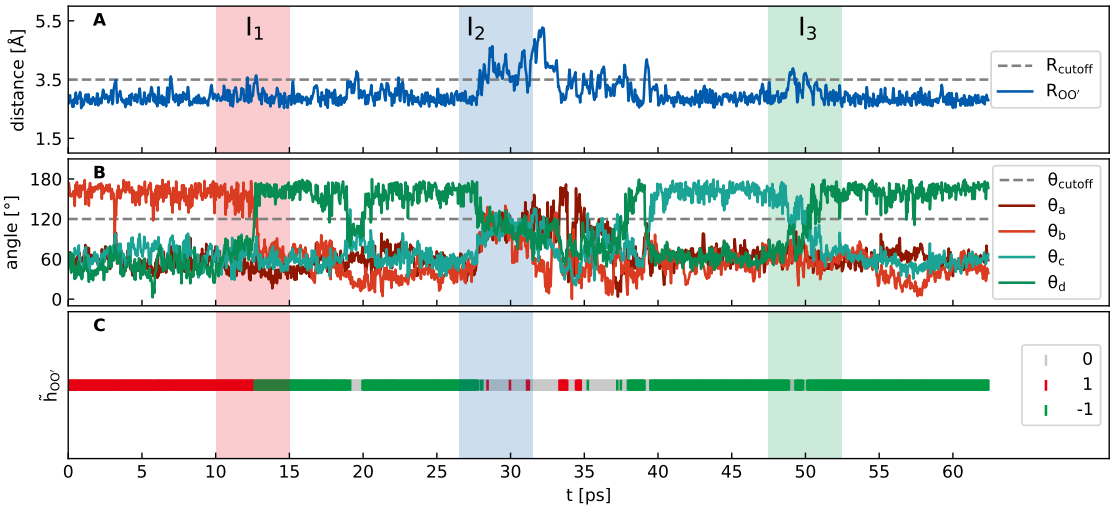

图 **2-3** 一个典型的准氢键的距离 $R_{OO'}$, 角度 θ, 以及有向氢键布居 $\tilde{h}$ 的动力学过程。图中展示了三种典型的变化过程: DA 交换 (I_1), 扩散 (I_2), 和 HH 交换 (I_3)。

Figure 2-3 Time dependence of distance $R_{OO'}$, angle θ, and directed H-bond population operator $\tilde{h}$ for one typical water pair. Three processes are spotted: DA exchange (I_1), translational diffusion (I_2), and HH exchange(I_3).

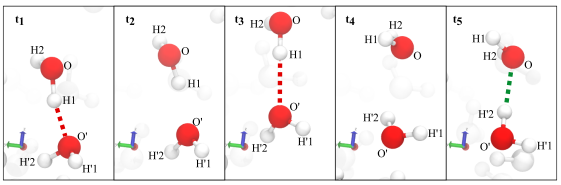

图 **2-4** 一个典型的 DA 交换过程。

Figure 2-4 A typical DA exchange demo.

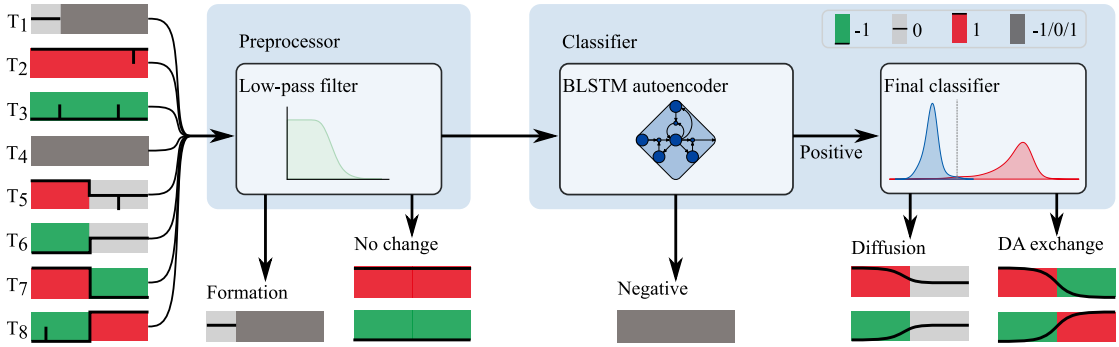

图 2-5 基于 RNN 的氢键状态分类器的处理流程. (i). 有向氢键布居的序列片段 $\tilde{h}$ 包含 8 种过程. T_1: 氢键形成 (Formation); T_2, T_3: 氢键状态不改变 (No change); T_5, T_6: 扩散 (Diffusion); T_7, T_8: DA 交换 (DA exhange); T_4: 其他。输入到预处理器的所有时间片段序列 $\tilde{h}_s[n]$，是氢键布居序列归一化后的结果, 每个时间片段对应的模拟时间是 8 皮秒. (ii). 预处理器的作用有两个。其一是过滤掉 $\tilde{h}_s$ 中的高频成分，其二是排除掉那些氢键形成 (T_1) 和氢键状态不发生 (T_2, T_3) 改变的序列。(iii) 分类器由两部分组成。第一部分是训练好的双向 LSTM 自编码器分类器。它的作用分辨出包含 DA 交换和扩散的阳性序列和其他的阴性序列。第二部分是区分 DA 交换和扩散过程的分类器。

Figure 2-5 The processing flow of the H-bond state change classifier based on RNN. (i). The time series fragment of $\tilde{h}$ contain several types of processes from T_1 to T_8. T_1: formation or not H-bonded; T_2, T_3: state unchanged; T_5, T_6: diffusion; T_7, T_8: DA exchange; T_4: other. The time corresponding to each sequence is 8 ps. $\tilde{h}$ fragments are normalized from [-1, 1] to [0, 1] before entering the preprocessor. (ii). The preprocessor has two functions. One is to filter out the high-frequency components of $\tilde{h}$. The second is to exclude those H-bond formations and no change processes. (iii). The RNN-based classifier consists of two parts. The first part is the trained bidirectional LSTM autoencoder. Its role is to separate the positive sequence containing diffusion or DA exchange and other negative sequences. The second part is the final classifier that distinguishes diffusion and DA exchange.

DA 交换反映在 $\tilde{h}$ 的变化是从 ±1 变到 ∓1；扩散过程反映在 $\tilde{h}$ 是从 ±1 变到 0。因此原则上来说，我们通过观测一小段时间 t_w 内有向氢键布居 $\tilde{h}$ 的序列，就可以判断出这段时间内氢键构型的变化情况，从而找到 DA 交换和扩散过程的相对比例。然而，纵使 DA 交换和扩散过程的变化有一定的规律可循，但由于变化序列本身有一定的涨落，因此对于分辨各种 $\tilde{h}$ 的序列所属的类型还是有很大的困难。为此我们设计了如图2-5所示的处理流程，来对氢键构型的改变过程做分类。

在预处理器中，首先我们使用一个低通滤波器对归一化后的有向氢键布居序列 $\tilde{h}_s[n]$ 进行滤波，从而过滤掉由于水分子的振动形成的高频涨落。得到滤波后的序列 $\tilde{h}_f[n]$。若 $\tilde{h}_f[0] - 0.5 < \alpha_T$, α_T 是一个较小的值, 表示序列一开始不存在氢键 (T_1) [1]；若 $\tilde{h}_f[n]$ 的标准差 σ 足够小即 $\sigma < \sigma_T$，我们认为氢键的状态没有发生改变 (T_2 和 T_3)。我们关注的重点是氢键状态的变化，因此我们在预处理阶段排除最开始不存在氢键的序列 (T_1) 和氢键状态不发生改变的序列 (T_2 和 T_3)。

接下来我们要解决的问题是一个典型的时间序列分类问题。但是我们面临一个非常大的困难: 除了 DA 交换和扩散这两类过程之外，还存在大量无规律的变化过程。为了方便说明，我们称 DA 交换和扩散这两类序列叫做*阳性序列* (Positive), 所有除这两类序列之外的序列叫做*阴性序列* (Negative)。由于需要对时间序列进行分类，我们很自然地想到了对有序数据建模[55–57] 的一类典型方法，循环神经网络 (RNN)[58,59]。

但是这里我们不能简单采集三类训练数据，DA 交换，扩散，和阴性序列，然后对循

[1] 虽然图中将 T_1 描述成氢键生成 (Formation), 但也包含后半部分没有氢键的情况, 即只要序列开始时没有氢键，这个序列就归为 T_1。

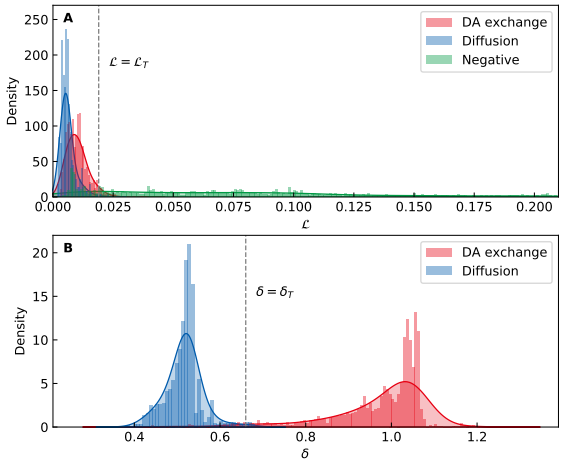

图 2-6 (A) DA 交换, 扩散, 和阴性测试数据经过双向 LSTM 自编码 (BLSTM AE) 的重建误差密度分布。(i) 训练了的自编码能够很好地还原出阳性的 DA 交换和扩散序列，因此重建误差小，大部分在图中的虚线左边。(ii) 由于阴性序列没有被用来训练自编码，因此很难被还原。故得到的重构误差大，大部分在图中虚线的右边。(iii) 找到一个合适的重建误差作为阈值，我们就可以得到一个阳性和阴性序列的分类器。(B) DA 交换和扩散过程序列的跨度观测值 δ 的分布密度。我们发现这两个分布区别非常明显。因此，选择 $\delta = \delta_{\mathrm{T}}$ 可以很好地区分 DA 交换和扩散过程。

Figure 2-6 (A) BLSTM Autoencoder reconstruction error ϵ densities for DA exchange, diffusion, and negative testing data. (i) The trained autoencoder can well restore the $\tilde{h}$ sequences corresponding to the positive samples. Hence, $\mathcal{L}$s for DA exchanges and diffusions are relatively small, mostly on the dashed line's left side. (ii) Since we don't use the negative samples to train the autoencoder, it is difficult for the autoencoder to restore the negative samples. Therefore, the $\mathcal{L}$s obtained are usually relatively large, mostly on the dotted line's right. (iii) Finding a suitable reconstruction error $\mathcal{L}_{\mathrm{T}}$ as the threshold, we can get a classifier that distinguishes between positive and negative. (B) The densities of δ for DA exchange and diffusion samples. The two distributions are seperated very well. Choose $\delta = \delta_{\mathrm{T}}$ as shown in dashed line, and the final classifier can distinguish DA exchange and diffusion samples.

环神经网络进行训练。原因是这里的阴性序列没有一定的模式，因此我们不能期待通过机器学习能够帮助我们分辨出阴性序列 (T_4)。但是，我们可以教会机器 DA 交换和扩散的正常序列 (Positive) 是什么样的。我们用仅包含 DA 交换和扩散的阳性序列去训练一个 RNN 自编码 (结构见图2-12)。训练完成后我们用自编码没有见过的异常序列 (Negative) 输入到自编码当中，它就很有可能不能够很好的还原输入，也就会导致自编码的重建误差相比正常序列 (Positive) 的大。通过这个重建误差我们可以判断输入序列是阳性 (重建误差小) 还是阴性 (重建误差大)。这样就得到了阳性序列和阴性序列的分类器。

如图2-6 (A) 所示，我们得到了 DA 交换序列, 扩散序列, 和阴性测试序列经过 BLSTM 自编码后的重建误差的密度分布。由于训练过的自编码能够很好地重建阳性的 DA 交换序列和扩散序列，从而得到的重建误差 $\mathcal{L}$ 较小，大部分在图中的虚线左边。然而阴性序列没有被用来训练自编码，因此自编码很难还原出阴性的序列。因此得到的重建误差 $\mathcal{L}$ 通常比较大，大部分在图中虚线的右边。我们只要找到一个合适的重建误差阈值 $\mathcal{L}_{\mathrm{T}}$，我们就可以得到一个区分阳性序列和阴性序列的分类器，即:

$$\begin{cases} h[n]\text{是阳性序列} & \text{若}\mathcal{L}(h_{\mathrm{f}}[n]) < \mathcal{L}_{\mathrm{T}} \\ h[n]\text{是阴性序列} & \text{若}\mathcal{L}(h_{\mathrm{f}}[n]) \geq \mathcal{L}_{\mathrm{T}} \end{cases} \tag{2-3}$$

重构误差阈值 $\mathcal{L}_{\mathrm{T}}$ 的确定见3.5。

接着我们使用图2-5中的最终分类器 (Final classifier) 区分阳性序列中的 DA 交换序列和扩散序列了。我们定义滤波后的序列 $\mathbf{x} = \tilde{h}_f[n]$ 的一个观测量 δ 来描述有向氢键布居序列 $\tilde{h}[n]$ 的跨度,

$$\delta(\mathbf{x}) = \max \mathbf{x} - \min \mathbf{x} \tag{2-4}$$

得益于最开始我们对有向氢键部居 $\tilde{h}$ 的定义, DA 交换过程的 $\tilde{h}$ 从 ± 1 变化到 ∓ 1, 而扩散过程的 $\tilde{h}$ 从 ± 1 变化到 0, 因此整体上 DA 交换过程的对应的 $\tilde{h}$ 跨度是扩散过程的两倍。如图2-6 (B) 所示，我们得到了 DA 交换和扩散过程对应的 δ 的密度分布。两类序列的 δ 的密度分布的均值差别非常明显。我们选择图示位置所示 $\delta_{\mathrm{T}} = 0.66$ 作为最后一步的 DA 交换和扩散的分类器的判断标准。我在附录3.6中，详细地展示了不同的氢键状态变化序列的类型判断过程和依据。

有了氢键状态变化的分类器，我们对系统中的准氢键的有向氢键布居 $\tilde{h}$ 的动力学轨迹作了分析。我们采用*滑动窗口*的方法对准氢键的动力学轨迹进行时间片段截取，得到 $\tilde{h}[n]$ 的序列；然后使用我们得到的氢键状态分类器对这些序列作预测；接着统计 DA 交换序列和扩散序列的个数。我们得到液态体相水中 DA 交换和扩散对应的相对比例约为 1:4。附录3.7中的图2-16说明了滑动窗口步长对该比例的影响不大。

至此我们得到了一个基于 RNN 自编码的氢键状态分类器。在预处理阶段，我们对归一化的有向氢键布居序列 $\tilde{h}_{\mathrm{s}}[n]$ 作低通滤波处理，排除掉氢键生成序列和氢键状态不改变的序列, 将通过预处理器的序列 $\tilde{h}_{\mathrm{f}}[n]$ 送到下一级的分类器中。我们仅仅使用 DA 交换和扩散过程的阳性序列训练出了一个能够很好还原阳性序列的 BLSTM 自编码。没有经过训练的阴性序列不能够很好地被 BLSTM 自编码器还原，因此这些阴性序列对应的还原误差通常会很大。由此我们区分出序列是阳性还是阴性。最后一步 DA 交换序列和扩散序列的分类器几乎不引入误差的原因是有向氢键布居 $\tilde{h}$ 的定义，这个定义导致了 DA 交换和扩散序列的“跨度”δ 截然不同，这为我们区分 DA 交换和扩散过程提供了最直接的帮助。接着我们对液态体相水中各种氢键变化状态做分类，得到了 DA 交换和扩散过程的相对比例约为 1:4。这个比例不小说明了 DA 交换在液态体相水中具有普遍性。接下来我们会探究温度对于氢键状态改变的影响。

2.5 温度对氢键构型改变的影响

为了探究温度对氢键构型改变的影响，我们模拟了 9 个包含 64 个水分子的体相水系统，温度范围是从 280K 到 360K, 温度的间隔为 10K。对不同温度下的动力学轨迹，利用得到的氢键状态改变分类器对有向氢键布居 $\tilde{h}$ 动力学轨迹的采样窗口序列做分类，统计出各个温度下 DA 交换和扩散过程序列的数目，得到了它们相对比例。如图2-7，我们看到，随着温度上升，DA 交换和扩散过程的个数呈现出“先上升，后波动”的趋势，即从 280K 到 330K 整体有上升的趋势，但随着温度继续升高，检测到的个数出现震荡的趋势。另一方面，尽管统计到的 DA 交换和扩散过程的数目在不同温度下存在较大的变化，但两者的相对比例却几乎不变。DA 交换和扩散过程的比例仍然约为 1:4。这个结果表示 DA 交换和扩散过程的相对比例对温度的依赖很低，且 DA 交换这种水分子间的运动过程在不同温度下都具有普遍性。

为了去理解图2-7 (A) 中 DA 交换和扩散过程数目随着温度的变化趋势，我们首先计

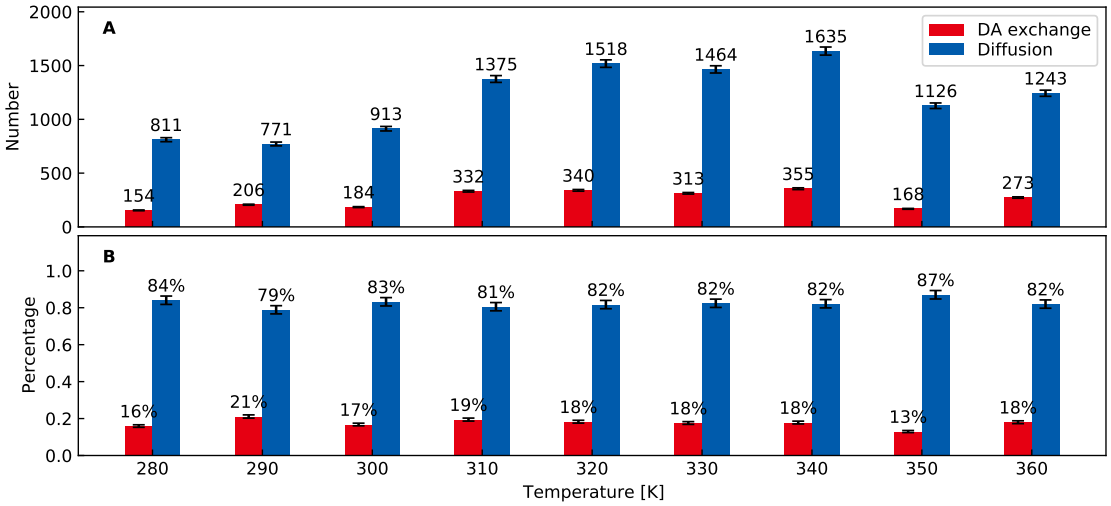

图 2-7 不同温度下氢键状态分类器检测出的 DA 交换和扩散过程的数目和它们的相对比例。
Figure 2-7 The detected number and proportion of DA exchange, and diffusion at different temperatures.

算了系统中水分子的平均氢键个数。在时刻 t, 水分子的平均氢键数目可以表示成

$$n_{\mathrm{HB}}(t) = \frac{2}{N}\sum_{i=1}^{N}\sum_{j>i}^{N}|\tilde{h}_{i,j}(t)| \tag{2-5}$$

对不同的时刻做统计，我们可以得到 n_{HB} 的分布。此外，为了表征氢键状态变化的快慢，我们使用一个 L 维的向量 $\tilde{\boldsymbol{h}}(t)$ 来表示纯水系统构成的有向图的氢键连接状态,

$$\tilde{\boldsymbol{h}}(t) = (\tilde{h}_{1,2}(t), \tilde{h}_{1,3}(t), \cdots, \tilde{h}_{i,j}(t), \cdots, \tilde{h}_{N-1,N}(t)) \tag{2-6}$$

其中向量的维度 $L = N \times (N-1)/2$, 它也是系统中的准氢键的数目。于是在一段时间内我们得到一个动态图状态向量 $\tilde{\boldsymbol{h}}(t)$ 的集合

$$H = \{\tilde{\boldsymbol{h}}(t) \mid t = t_{\mathrm{start}} + k \times \Delta t, k \in \{0, 1, \cdots, S\}\} \tag{2-7}$$

其中 t_{start} 表示的是观测窗口的起始时刻，Δt 是相邻两帧之间的时间间隔，S 是一个可变量, S 越大表示观测持续的时间越长。在长度为 $S \times \Delta t$ 的观测窗口中，我们统计不同的动态图状态向量 $\tilde{\boldsymbol{h}}(t)$ 的个数 Ω (即 H 集合的大小) 作为此观测时间内的状态数目。于是有

$$\Omega = |H| \tag{2-8}$$

Ω 表征了氢键状态改变的速率，越大表示氢键状态改变得越快。持续时间为 $S \times \Delta t$ 的观测窗口中状态数目的理论最大值为 $S+1$, 在这种情况下观测到的每一个 L 维向量 $\tilde{\boldsymbol{h}}(t)$ 都不同。通过改变窗口的的起始位置 t_{start}, 我们可以得到在不同起始时刻下 Ω 的分布。

图2-8分别展示了 (A) 平均氢键个数的分布和 (B) 在时间观测窗口宽度 $S \times \Delta t$ 为 1 皮秒, $\Delta t = 0.001$ 皮秒时的微观状态数目的对数 $\ln\Omega$ 的分布随温度的变化。温度较低时，图2-8 (A) 展示出氢的键数目较大，而图2-8 (B) 中的 $\ln\Omega$ 相对较小。这表示在温度较低时，水中氢键的数目多，但氢键的状态改变速率较低。这就解释了为什么图2-7 (A) 中在

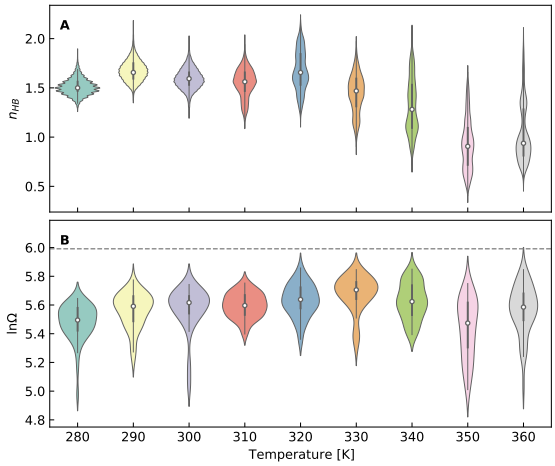

图 2-8 (A) 平均氢键数目 n_{HB} 在不同时刻下的分布随温度的变化趋势。(B) 不同温度下观测时间为 1 皮秒情况下的氢键网络图的个数的对数 $\ln\Omega$ 的分布。其中的虚线代表理论上的最大值。

Figure 2-8 (A) The number of H-bonds per water at different temperatures. (B) The coarse-grained entropy of the H-bond network in bulk water at different temperatures. Ω is the number of coarse-grained microscopic states within 1 ps.

温度较低时，DA 交换和扩散过程的数目较小。

此外，在温度较高时，n_{HB} 分布的中位数呈现下降的趋势，这也直接导致了图2-7 (A) 中高温部分 DA 交换和扩散过程的数目下降。因为分类器检测的是氢键状态的改变，即不考虑序列开始没有形成氢键的有向氢键布居序列 (这部分序列已经被预处理器排除)。因此，自然地，在高温下的 DA 交换和扩散过程数目就会减少。此外，接下来，我们还探究了不同氢键定义下的平均氢键个数和弛豫率常数随温度的变化。为了探究水分子在不同温度下的振动情况，我们计算了在不同温度下的速度自关联函数和振动态密度。

2.6 不同氢键定义下的平均氢键个数和弛豫率常数随温度的变化

基于两种不同的氢键的定义，我们计算了平均每个水拥有的氢键个数以及氢键布居 h[28] 的弛豫率常数。两种氢键定义的不同在于所依据的几何条件不同。

定义 1: 若 $R_{\mathrm{OO'}} < 3.5$ Å, $\widehat{\mathrm{OHO'}} > 120°$，则水分子间存在氢键;

定义 2: 若 $R_{\mathrm{OO'}} < 3.5$ Å, $\widehat{\mathrm{HOO'}} < 30°$，则水分子间存在氢键。[28]

图2-9(A) 给出了在不同温度下平均每个水分子的氢键的个数。可以看到两种氢键定义方式下得到的平均氢键个数 $\langle n^1_{\mathrm{HB}}\rangle$，$\langle n^2_{\mathrm{HB}}\rangle$ 随温度的变化趋势基本相同，即随着温度的增加呈现出下降的趋势。同时我们也看到 $\langle n^1_{\mathrm{HB}}\rangle$ 和 $\langle n^2_{\mathrm{HB}}\rangle$ 在数量上有一定的差别, $\langle n^1_{\mathrm{HB}}\rangle$ 的值小于 $\langle n^2_{\mathrm{HB}}\rangle$。弛豫率常数 Γ, 图2-9(B), 描述了氢键的生成断裂变化的剧烈程度，这个值越大代表氢键的“反应”越剧烈。从图中可以看出，随着温度的上升，氢键的运动更剧烈。两种氢键定义下，在温度小于 300K 的情况下，两种定义得到的弛豫率常数 Γ_1, Γ_2 非常接近；随着温度的升高两种定义下的 Γ_1, Γ_2 的差别逐渐变大，Γ_1 会稍高。图2-9向我们展示了氢键的不同定义尽管可能会导致的一些观测量的差异，但是两种定义下得到的 $\langle n_{\mathrm{HB}}\rangle$ 和 Γ 随温度的变化关系是一致的。

2.7 不同温度下的速度自关联函数和振动态密度

我们可以使用从 AIMD 轨迹当中得到速度, 并利用速度自关联函数 (Velocities autocorrelation functions, VACFs) 来得到体相水系统的振动性质。振动态密度 (Vibrational den-

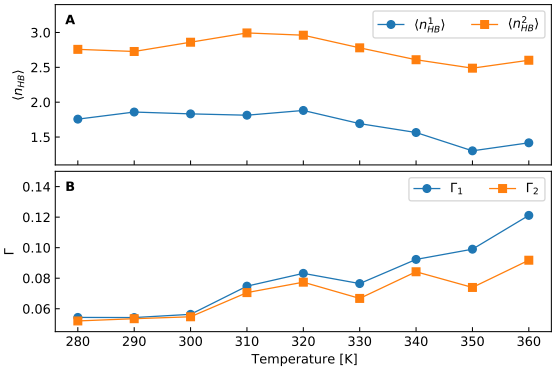

图 2-9 不同氢键定义方式，不同温度下的 (A) 平均氢键个数和 (B) 弛豫率常数。
Figure 2-9 (A) Mean number of H-bonds per water. (B) Rlaxation rate constant Γ.

sity of states, VDOS) 能够提供水分子的 OH 伸缩运动 (OH-stretching) 的信息。对于一个包含 M 个原子的系统, 对一个分子的速度自关联函数 (VACF) $C(t)$ 可以表示成

$$C(t) = \frac{\left\langle \sum_{i=1}^{M} \mathbf{v}_i(t) \cdot \mathbf{v}_i(0) \right\rangle}{\left\langle \sum_{i=1}^{M} \mathbf{v}_i(0) \cdot \mathbf{v}_i(0) \right\rangle} \tag{2-9}$$

其中 $\langle \cdots \rangle$ 表示对时间起点做平均，t 是时间间隔，$\mathbf{v}_i$ 表示的是第 i 个原子的速度。所选原子的振动态密度 $g(\nu)$ 可以表示成速度自关联函数 (VACF) 的傅立叶变换。因此是原子振动频率 ν 的函数。在平衡态有 $C(-t) = C(t)$, 故 $g(\nu)$ 是一个实函数。因此 $g(\nu)$ 可以写成

$$g(\nu) = \sqrt{\frac{2}{\pi}} \int_0^\infty \mathrm{d}t \cos(2\pi\nu t) C(t) \tag{2-10}$$

图2-10(A) 和 (B) 分别展示了不同温度下的速度自关联函数和振动态密度。从图 A 中我们可以看出在 0.6 皮秒以内，所有温度 [280,360]K 下的体相水系统的速度自关联都衰减为 0。B1 展示了频率范围为 [0, 4000] cm^{-1} 的振动态密度。为了将 OH 伸缩振动的信息展示得更清楚我做了图 B2, 仅仅是将 B1 中的第三个峰放大展示出来。从 B2 我们可以看出随着温度的上升，VDOS 的第三个峰值出现蓝移。第三个峰的位置代表的是水分子的 OH 伸缩振动频率。这就表示温度的升高会导致水分子的 OH 伸缩频率提高。这与弛豫率常数 Γ 随温度的升高而升高的变化趋势是一致的。

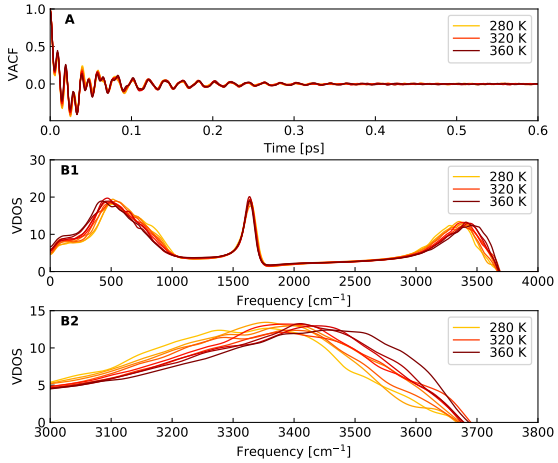

图 2-10 不同温度下液态体相水的 (A) 速度自关联函数；(B1, B2) 振动状态密度。
Figure 2-10 (A) Velocity auto-correlation function and (B1, B2) vibrational density of states.

3 研究方法

3.1 AIMD 模拟

我们通过 CP2K/QUICKSTEP (v7.1) 软件[60] 对体相水的正则系综 (NVT) 进行 DFTMD 模拟。模拟温度的范围是 [280, 360]K, 温度间隔是 10K, 模拟系统由边长为 12.4295 Å 的周期性边界盒子组成, 其中包含了 64 个水分子，其密度为 0.997 g/cm ^{3}。离散积分时间步长 Δt 设置为 0.5 飞秒, 模拟时间是 60 皮秒。使用了 BLYP 交换关联泛函，该泛函由 Becke 非本地交换泛函[61] 和 Lee-Yang-Parr 关联泛函[62] 组成。价电子与离子核之间的相互作用由 GTH 赝势[63,64] 描述。这种方法对波函数使用高斯基组，对密度使用辅助平面波基组。对所有原子使用 DZVP-GTH 基组，并且电荷密度[60] 的截止值为 280 Ry。为了让温度恒定，我们使用了 Nosé-Hoover 链温控器[65]。在模拟中我们使用了 Grimme 发展的 DFT-D3 校正方法[66] , 它可以很好地描述色散相互作用。

3.2 动力学轨迹分析

我们使用了 MDAnalysis (v1.0.0)[53,54] 分析处理液态体相水的模拟轨迹，包括加载坐标系，找氢键，计算准氢键的观测值距离和角度等。首先，我们去除前 10 皮秒非平衡的的动力学轨迹，对剩下的 50 皮秒的轨迹每隔 80 帧进行采样。原本两帧之间的时间间隔为 $\Delta t = 0.5$ 飞秒, 经过采样后动力学轨迹的时间间隔为 $80 \times \Delta t = 40$ 飞秒。接着，我们使用 MDAnalysis 中的 HydrogenBondAnalysis 模块找出动力学轨迹中每一帧的氢键，得到氢键供体氧原子，被贡献的氢原子，受体氧原子在模拟体系中对应的原子序号，这些信息帮助我们对模拟水系统进行动态图的建模。进而我们得到了准氢键的有向氢键布居 $\tilde{h}$ 动力学轨迹，同时也利用 MDAnalysis 提供的角度和距离的计算模块得到了距离 $R_{\mathrm{OO}'}$ 和角度 $\theta_{\mathrm{a}}, \theta_{\mathrm{b}}, \theta_{\mathrm{c}}, \theta_{\mathrm{d}}$ 的动力学轨迹。

3.3 $\tilde{h}$ 序列收集与预处理

通过观测准氢键 $\mathrm{QB}_{i,j}$ 中距离 $R_{\mathrm{OO}'}$, 角度 $\theta_a, \theta_b, \theta_c, \theta_a$ 和有向氢键布居 $\tilde{h}$ 的动力学过程，我们收集了 $\mathrm{QB}_{i,j}$ 动力学过程中的 DA 交换和扩散过程的时间序列, 图例见附录

图2-14中的 (A) 至 (F)。固定时间窗口的长度是 8 皮秒, 则对应的 $\tilde{h}$ 序列长度是 200。选择 8 皮秒的原因是这个时间长度足以来判断在这个时间区间内的扩散过程。另外对于 DA 交换过程，我们仅收集在 8 皮秒以内发生一次 DA 交换的时间序列。收集 DA 交换和扩散过程序列是为了训练 BLSTM 自编码，使其能够“认识”DA 交换和扩散过程这些阳性序列。另外，我们也收集了不属于 DA 交换和扩散过程的阴性 $\tilde{h}$ 序列，用来对 BLSTM 自编码分类器做评估。用作训练自编码器的阳性序列有 6786 个，其中 DA 交换和扩散过程各占一半，每个温度下的阳性数据 754 个。总共用来测试阴性序列有 18931 个。

根据向氢键布居的定义，可知 $\tilde{h}$ 序列的范围是 [-1, 1]。为了方便训练 BLSTM 自编码，我们将其值归一化到 [0, 1]，记为 $\tilde{h}_{\mathrm{s}}$, 归一化后 0 和 1 都代表存在氢键, 但方向相反，0.5 代表不存在氢键。由于 $\tilde{h}[n]$ 存在一些高频涨落，DA 交换和扩散过程对应的是 $\tilde{h}[n]$ 低频的变化，因此我们对窗口序列 $\tilde{h}[n]$ 做了低通滤波的处理得到滤波后的序列 $\tilde{h}_{\mathrm{f}}[n]$。低通滤波器采用的是二阶巴特沃斯 (Butterworth) 低通数字滤波器，使用 Scipy 中信号 (Signal) 模块实现。另外，我们采用了两个参数 $\alpha_{\mathrm{T}} = 0.15, \sigma_{\mathrm{T}} = 0.1$，用于排除氢键形成 (Formation) 的序列和氢键状态不变的序列，即图2-5中的 T_1, T_2, T_3。若 $\tilde{h}_{\mathrm{f}}[0] - 0.5 < \alpha_{\mathrm{T}}$, 即序列最开始是没有氢键的，不管后面是否有氢键生成，我们都不关注这一类变化，把这类序列统一归类到 T_1，这类序列不在进入后续判别器中。另外，我们计算 $\tilde{h}_{\mathrm{f}}[n]$ 的标准差 σ, 若 $\sigma < \sigma_{\mathrm{T}}$ 则认为这个氢键布居序列没有发生改变，即在这 8 皮秒时间内准氢键 (Q-bond) 中的氢键状态是没有改变的。这类序列 (T_2, T_3) 也不进入后续的判别器。

3.4 双向 LSTM 自编码

为了让机器还原出阳性的 DA 交换序列和扩散序列，我们使用阳性序列来训练一个自编码。由于有向氢键布居序列是天然的时间序列，因此我们使用了非常擅长处理时间序列 LSTM 单元来搭建自编码网络。考虑到在氢键状态改变的识别中，有向氢键布居序列从前向后和从后向前观测都有助于识别氢键构型的改变，因此我们采用了双向的 LSTM 网络结构。自编码的结构包含了两个部分，编码器和解码器。双向 LSTM 自编码利用 Tensorflow(2.2.0) 的 Keras 模块实现。编码器和解码器可以被表示成两个变换 ϕ 和 ψ

$$\phi : \mathcal{X} \to \mathcal{F}$$
$$\psi : \mathcal{F} \to \mathcal{X}$$

其中 ϕ, ψ 满足

$$\phi, \psi = \underset{\phi,\psi}{\arg\min} \|X - (\psi \circ \phi)X\|^2 \tag{2-11}$$

在我们的工作中, 特征空间 $\mathcal{F}$ 的维数小于输入空间 $\mathcal{X}$ 的维数, 因此特征向量 $\phi(\mathbf{x})$ 可以视为输入 $\mathbf{x}$ 的压缩表示。这种自编码叫做**不完整自编码** (Undercomplete autoencoders)。这类自编码“开始大中间小”的设计有效地避免了自编码仅仅只是简单地复制信号，并迫使自编码近似地重建输入，仅保留原始数据最相关部分。将 LSTM 和自编码结合我们可以得到 LSTM 自编码，采用双向的结构我们可以得到双向 LSTM 自编码 (Bidirectional LSTM Autoencoder, BLSTM AE)。

BLSTM AE 的输入 $\mathbf{x}$ 是归一化和滤波后的有向氢键布居序列 $\tilde{h}_f[n]$, 其中每个序列长度固定为 200, 对应的模拟时间是 8 皮秒。训练 BLSTM AE 的目的是让它能够在输出端

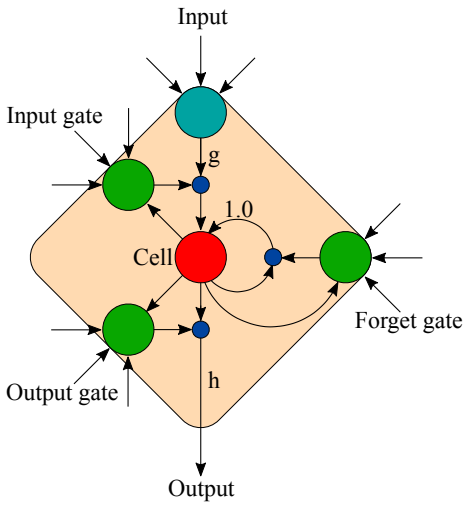

图 2-11 LSTM 基本单元。它的输出由三个乘法门控制：输入门，遗忘门和输出门。
Figure 2-11 The LSTM unit whose activation is controlled by three multiplicative gates: the input gate, forget gate and output gate.

尽可能地还原出输入。定义 BLSTM AE 对一个序列 $\mathbf{x} = \tilde{h}_f[n]$ 的重构误差为

$$\mathcal{L}_{\omega,\omega'}(\mathbf{x}) = \|\mathbf{x} - \psi_{\omega'}(\phi_{\omega}(\mathbf{x}))\|^2 \tag{2-12}$$

其中 ω, ω' 分别表示编码器和解码器的参数。训练的目的是得到最优的 ω, ω', 即

$$\omega^*, \omega'^* = \arg\min_{\omega,\omega'} \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}_{\omega,\omega'}(\mathbf{x}^i) \tag{2-13}$$

这里的 $\mathbf{x}^i$ 表示的是第 i 个序列。从式2-12和式2-13可以看出，训练 BLSTM 自编码仅仅需要输入序列 $\mathbf{x}$, 也就是说 BLSTM 自编码是无监督学习。

标准 RNN 可以访问的上下文信息的范围非常有限。它的主要问题在于网络的输入对隐藏层和网络输出的影响。网络的输出会随着网络的循环连接而呈指数级衰减或爆炸。这个缺点通常被称为消失梯度问题，它使得标准的 RNN 很难处理超过 10 个时间步长的序列。长短时记忆 (Long Short-Term Memory, LSTM) 正是专门为了解决梯度消失问题的一类循环神经网络。LSTM 隐藏层由循环连接的子网组成，称为存储块。每个存储块包含一组内部单元或单元，它们的激活由三个乘法门控制：输入门 (Input gate)，遗忘门 (Forget gate) 和输出门 (Output gate)。图2-11详细地展示了具有单个单元的 LSTM 存储模块[12]。门的作用是允许单元长时间存储和访问信息。例如，只要输入门保持关闭, 即具有接近于 0 的激活，则单元的激活将不会被到达网络中的新输入所覆盖。同样，只有在输出门打开时，单元激活才可用于网络的其余部分，并且通过遗忘门可以打开和关闭该单元的循环连接。

对许多任务，得到将来和过去的信息非常有用。例如，在氢键状态改变的识别中，有向氢键布居序列从前向后，和从后向前观测都有助于识别氢键构型的改变。双向循环神经网络 (Bidirectional recurrent neural networks, BRNNs)[96,97], 能够沿着输入序列在两个方向上访问上下文信息。BRNN 包含两个独立的隐藏层，其中一个隐藏层处理正向输入序列，而另一个隐藏层处理反向序列。两个隐藏层都连接到同一输出层，从而使它可以访问序列中每个点的过去和将来的信息。在一些序列学习任务中，BRNN 的性能优于标准

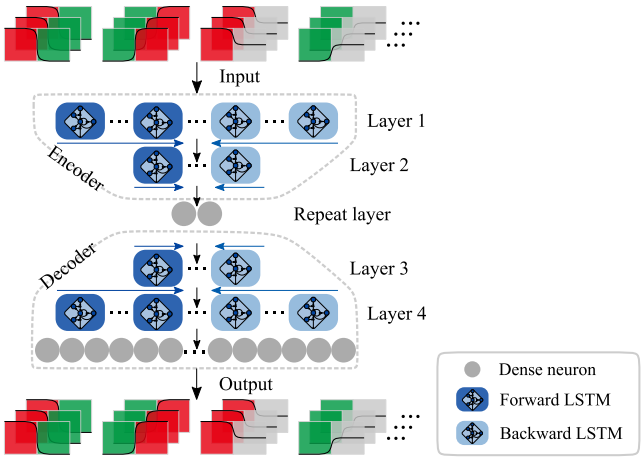

图 2-12 双向 LSTM 自编码的结构。(i) 我们使用双向 LSTM 自编码来区分阳性序列和阴性序列。(ii) 训练双向 LSTM 自编码用到的数据 $\tilde{h}_f[n]$ 仅包含 DA 交换和扩散序列。(iii) 由于 $\tilde{h}_f[n]$ 是随时间变化的序列, 因此我们选择广泛采用的的 LSTM 单元作为自编码的基本单元。此外, 由于 $\tilde{h}_f[n]$ 的开始和结尾对于分辨这个序列所属的类型同样重要，因此我采用来双向的网络结构。

Figure 2-12 Bidirectional LSTM autoencoder. (i) We use the BLSTM autoencoder to identify positive and negative $\tilde{h}$ fragments. The positives refer to the DA exchange and diffusion processes. (ii) The input data used to train the BLSTM autoencoder the filtered positives $\tilde{h}$ fragments only. (iii) Since $\tilde{h}$ is a time-varying sequence, we choose to use the LSTM unit as the primary network unit. Besides, the $\tilde{h}$ segment's start and end are equally crucial for classifying the $\tilde{h}$ segment type, so we select a bidirectional network structure.

RNN，尤其是蛋白质结构预测[98] 和语音处理[96,99]。将 BRNN 和 LSTM 结合起来就得到了双向 LSTM(Bidirectional LSTM, BLSM)。

双向 LSTM 自编码 (Bidirectional LSTM Autoencoder, BLSTM AE) 的构构如图2-12所示。第一层 (BLSTM layer 1) 包含正向，反向两个 LSTM 层，每层有 64 个 LSTM 单元; 编码器第二层 (BLSTM layer2) 同样有正向，反向的两个 LSTM 层，每层有 32 个 LSTM 单元；重复向量 (Repeat vector) 的次数是 2；接下来是解码器，其网络结构和编码器关于重复向量层对称。即解码器的第一层 (BLSTM layer 3) 的网络结构和编码器第二层 (BLSTM layer 2) 相同, 解码器的第二层 (BLSTM layer 4) 的网络结构和编码器第一层 (BLSTM layer 1) 相同；最后一层是时间分布 (Time distributed) 层，包含 200 个全联接神经元。训练过程采用的优化器是 Adam, 损失函数采用的是 MAE, 批大小 (Batch size) 为 32, Droput 率为 0.1，Epochs 为 500。

3.5 自编码分类器

选择一个合理的重建误差作为判断阳性和阴性的标准，我们就可以得到一个自编码分类器 (AE classifier)。为了找到合适的重建误差作为分类标准, 如图2-13 所示, 我们测量了在不同阈值下的自编码分类器 (AE classifier) 的准确率 (Accuracy), 平衡准确率 (Balanced accuracy) 和 F1 得分。可以看到这三个值在重建误差区间 $[0.01, 0.03]$ 呈现出先增加后减小的趋势，在图示虚线位置的重建误差阈值 $\mathcal{L}_\mathrm{T}$ 对应的准确率，平衡准确率，F1 得分都取得极大值。因此我们选择此时的阈值作为自编码分类器的阈值。通过计算 DA 交换序列，扩散序列对应的重建误差大于 $\mathcal{L}_\mathrm{T}$ 的密度面积和阴性序列对应的重建误差小于 $\mathcal{L}_\mathrm{T}$ 的密度面积我们可以得到自编码分类器的误差。

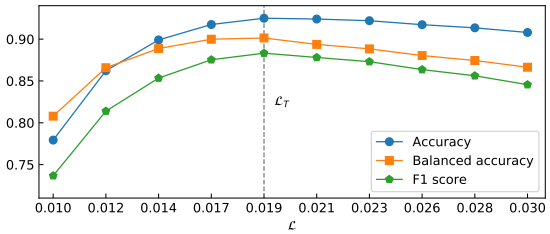

图 2-13 BLSTM 自编码分类器在测试数据集上的准确率 (Accuracy), 平衡准确率 (Balanced accuracy) 和 F1 得分随着重构误差阈值的变化。为了使得 BLSTM 自编码器具有较好的分类的效果，我们选择如图所示的重构误差 $\mathcal{L}_{\mathrm{T}}$ 作为 BLSTM 自编码器区分阴性和阳性的阈值，在图示位置的重构误差 $\mathcal{L}_{\mathrm{T}}$ 使得 BLSTM 自编码器分类器在这三个标准下都具有最大值。

Figure 2-13 The accuracy, balanced accuracy, and F1 score of the BLSTM autoencoder classifier in the testing data at diffenrent reconstruction error thresholds. To find a suitable $\mathcal{L}_{\mathrm{T}}$ for the BLSTM autoencoder classifier, we choose the reconstruction error shown in the figure, which corresponding to the maximums of the three scores.

3.6 判别示例

图2-14展示了不同类型的氢键状态变化过程。利用距离 R 和角度 θ 我们可以清楚知道这个过程的类型。其中 (A), (B), (C) 是 DA 交换过程，(D), (E), (F) 是扩散过程，(G), (H), (I) 是其他过程，我们将其归为阴性过程。$\tilde{h}_s$ 是有向氢键布居 $\tilde{h}$ 归一化后的结果，即将值域范围从 [-1, 1] 线性地映射到了 [0, 1]。$\tilde{h}_f$ 是 $\tilde{h}_s$ 经过低通滤波器的输出，也是 BLSTM 自编码的输入。$\tilde{h}_r$ 是经过自编码重建的序列。从图中我们可以清楚地看到，DA 交换序列和扩散序列对应的自编码还原序列 $\tilde{h}_r$ 和它们的自编码输入序列 $\tilde{h}_f$ 几乎重合。也就是说训练后的自编码器能够很好地还原 DA 交换序列和扩散序列。另一方面, (G)(H)(I) 的 $\tilde{h}_r$ 和 $\tilde{h}_f$ 却相去甚远。这是由于我们在训练自编码器的时候没有使用这类数据去训练。通过重建误差的大小我们可以判断出序列是阳性序列还是阴性序列。然后我们把阳性序列输入到后一级的最终判别器中。最终判别器通过 $\tilde{h}_f$ 的跨度参量 δ 判断出序列的类型是 DA 交换还是扩散过程。图2-15展示了不同氢键状态变化过程对应的重建误差 $\mathcal{L}$ 和跨度参量 σ。背景的颜色代表了 BLSTM 自编码分类器的预测结果。红色代表 DA 交换过程，蓝色代表扩散过程，绿色代表阴性过程。这表明我们设计的 BLSTM 自编码分类器能够很好对氢键状态变化过程做分类。

3.7 滑动窗口步长的影响

如图2-16所示，我们可以看到在 310K 的纯水中，DA 交换和扩散两个过程的相对比例约为 1:4, 这表示 DA 交换这种过程不是偶然观测到的现象，而是在纯水系统中大量存在。由于对有向氢键布居 $\tilde{h}$ 动力学轨迹的采样使用了滑动窗口的方法，因此我们以滑动步长作为参数观察了在不同步长下的 DA 交换和扩散的相对比例。我们发现这个相对比例几乎不受滑动窗口步长的影响。

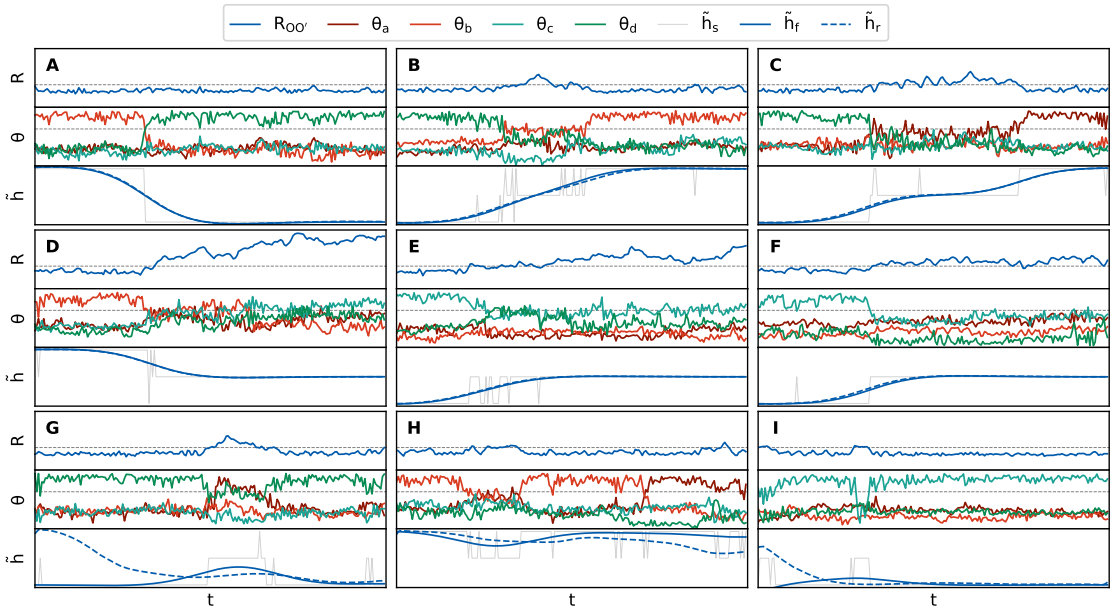

图 **2-14** 不同类型的氢键状态变化过程。其中 A, B, C 是 DA 交换; D, E, F 是扩散过程; G, H, I 是阴性样本。

Figure 2-14 Different types of H-bond change process. (A, B, C) DA exchange. (D, E, F) Diffusion; (G, H, I) Negative

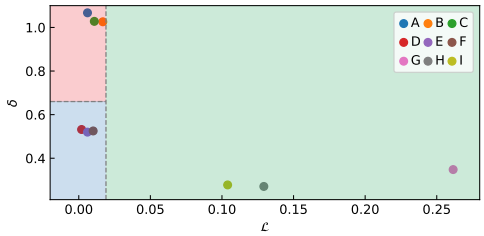

图 **2-15** 图2-14中各个示例对应的重建误差 $\mathcal{L}$ 和有向氢键布居跨度参量 δ。背景的颜色代表了 BLSTM 自编码分类器的预测结果。红色代表 DA 交换过程，蓝色代表扩散过程，绿色代表阴性过程。

Figure 2-15 The reconstruction error $\mathcal{L}$ and the directed hydrogen bond population span parameter δ corresponding to each example in the figure 2-14. The color of the background represents the prediction result of the BLSTM AE classifier. Red represents the DA exchange process, blue represents the diffusion process, and green represents the negative processes.

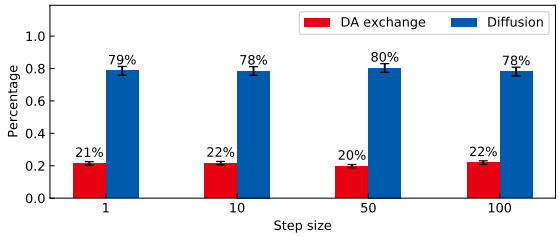

图 **2-16** 310K 液态体相水中，在不同滑动窗口步长下的 DA 交换和扩散过程的相对比例。

Figure 2-16 The relative proportion of DA exchange and diffusion processes in bulk water under different sliding window steps at 310K.

4 结论和意义

从物理上来说，我们可以得到如下结论。1. 通过 DFTMD 模拟，我们发现了在液态水中除了扩散过程以外的一种独特的水分子重定向机制: DA 交换。2. 为了找到液态体相水中 DA 交换过程，并确定该过程和水分子扩散过程的相对比例，通过训练一个基于循环神经网络的模型，我们得到了一个氢键状态变化过程的分类器。我们统计了 DA 交换和扩散两种变化过程的数目，并算出了其比例约为 1:4。该比例并不小这意味着体相水中存在大量的 DA 交换过程。3. 我们探究了温度对于氢键构型的影响。我们发现在不同温度下 DA 交换和扩散过程的相对比例几乎不变, 也基本都维持在 1:4。这说明这两种氢键状态改变过程的相对比例几乎不依赖于温度，且 DA 交换在不同温度下的液态体相水中是普遍存在的。

从方法上来说，我们总结如下。1. 我们使用如今精度很高的 DFTMD 方法模拟了不同温度下的液态体相水，并研究了液态水中氢键构型的改变。2. 我们定义的有向氢键布居 $\tilde{h}$，将氢键的方向信息包含在其中，让我们仅仅通过观测 $\tilde{h}$ 变化就可以直接观测氢键构型的改变。3. 利用准氢键和有向图的方式对模拟水体系进行描述，在数据处理和分析上可以给我们带来便利。用图的方法描述模拟的水系统，让我们从另一个角度去看待水中的氢键网络。4. 利用深度学习，我们设计了一个双向 LSTM 自编码分类器。它能够很好地对氢键构型的改变过程进行分类，从而帮助我们找出液态水中的 DA 交换过程，且在不同温度下都有很好的分类效果。深度学习的方法也给我们研究液态水的氢键动力学提供了一种新的思路。

第三章　高分子链的结构因子

本章我将展示如何构造高分子链结构因子的神经网络模型。结合中子散射的实验强度数据，利用神经网络模型，预测高分子链的链长和库恩长度。

1　引言

1.1　结构因子

高分子系统的结构因子 (Structure factor)

$$S(\mathbf{k}) = \frac{1}{\rho}\int_V \langle\rho(\mathbf{r})\rho(0)\rangle \exp(i\mathbf{k}\cdot\mathbf{r})\mathrm{d}\mathbf{r} \tag{3-1}$$

是一个可以测量的物理量，它表征了系统的密度-密度 (Density–density) 关联函数[67]。从理论上讲，结构因子可用于场论计算。在高斯涨落理论[68,69] 中, 相互作用系统的结构因子是使用理想链 (Ideal chain) 的结构因子来计算的。此外，在动态平均场理论中，扩散过程[70–72] 中的链间相关特性也由理想链的结构因子来描述。实验上，可以通过将散射数据与结构因子进行拟合来分析聚合物的基本特征，例如聚合度，刚性和手性。至于计算，可以从微观链模型预测结构因子。高斯链模型众所周知的表达式具有 Debye 函数的形式[73]，可以被用来从实验上求出中小波数范围 $ka \leq 1$ 的 θ 点稀释聚合物溶液的结构因子，这里的 a 是库恩长度。

此外，还有一大类半刚性聚合物链，其中有限刚性的影响不可忽视，这些聚合物链是不能用高斯链模型来描述的。蠕虫状链模型是最好的半柔链模型之一。在此模型中，聚合物是不可拉伸的线，并受到了线性弹性弯曲能的影响[74]。总长度为 L 的蠕虫状链的构型由一条光滑的空间曲线描述，其坐标由 $\mathbf{R}(s)$ 指定，其中 s 是一个弧形变量，从一端（$s=0$）到另一端（$s=1$）连续变化[73,75,76]。这种结构的玻尔兹曼权重由下式给出：

$$\mathscr{W}[\mathbf{R}(s)] = \exp\left[-\beta H_0\right] \tag{3-2}$$

式中，

$$\beta H_0 = \frac{a}{4L}\int_0^1 \mathrm{d}s\left|\frac{\mathrm{d}\mathbf{u}(s)}{\mathrm{d}s}\right|^2 + \frac{L}{a}\int_0^1 \mathrm{d}s w[\mathbf{R}(s),\mathbf{u}(s)] \tag{3-3}$$

切向量 $\mathbf{u}(s) \equiv (1/L)\mathrm{d}\mathbf{R}(s)/\mathrm{d}s$ 指定了聚合物链在位置 s 的局部方向。$u(s)$ 是单位向量，由于局部不可扩展的约束使得 $|u(s)| = 1$。第一项描述弯曲曲线的能量损失。最初，弯曲能模量 $\beta\varepsilon$ 被写为系数[75]; 一旦得到自由空间的均方回旋半径和高斯链的均方根回旋半径后，在较大的刚性 L/a 的范围中, 我们可以证明该前置因子可以用这样的形式编写，其中对于三维系统有

$$a = 2\beta\varepsilon \tag{3-4}$$

库恩长度 a 在这里直接用于与根据高斯链模型计算的结果进行比较。蠕虫状链模型涉及

两个特征长度：链长度 L 和有效库恩长度 a。

1.2 研究背景

计算结构因子的关键是等式3-3中格林函数（$w = 0$）的计算。事实证明，蠕虫状链模型是没有格林函数的解析表达式的[69,77–80]。Kholodenko 在刚性和柔性极限情况下探究了半柔性聚合物模型的格林函数与狄拉克费米子的传播子之间的相似性[81,82]。高斯链 (Gaussian- chain) 和棒 (Rod) 表达式的极限可以从公式中得出。与之前的 Yoshizaki 和 Yamakawa[83] 以及后来的 Pedersen 和 Schurtenberger 提出的近似值相比，它是最简单的[84] 。Pedersen 和 Schurtenberger 对这种半柔性链进行了一系列的蒙特卡洛模拟，模拟包括了带和不带单体之间排除体积的相互作用。然后可以从模拟中数值获得结构因子。[84] 他们提供了一个经验公式来表示其模拟数据。最近，Hsu 及其同事使用蒙特卡洛模拟方法在简单的立方晶格上计算了半柔性链模型的结构因子[85,86]。Spakowitz 和 Wang 提出了另一种方法。通过计算受约束的一维随机行走 (Random walk) 问题，他们正式获得了蠕虫状链的格林函数[79]。他们将力矩扩展可以表示为无限的连续分数。连续分数问题的计算等效于求逆与 Stepanow 工作中使用的矩阵具有相同格式的矩阵。然而，要找到结构因子，必须回到数值处理上。尤其需要逆数值拉普拉斯变换[79,87]。在我们先前的工作中[69]，获得了一种基于标准蠕虫链模型的均质蠕虫状聚合物溶液结构因子的数值计算方法。我们使用修正后的扩散方程（MDE)[76,88] 的形式解来计算依赖于 s 的格林函数。该方法在数值上比最近提到的其他方法更直接。该解决方案能够在 L/a 和 ka 的整个参数空间中捕获结构因子的正确物理行为。

1.3 神经网络在高分子学科中的应用

近年来，作为机器学习 (Machine learning, ML) 的重要分支的神经网络（Neural networks , NNs）在高分子物理学中得到了广泛的应用。例如，神经网络可以被用来对物质的相进行分类[19]，求解非线性偏微分方程（Partial differential equations, PDE)[20]，预测大分子的结构[21] 和对聚合物构象分类[22]。由于已经证明了神经网络几乎可以近似任何函数[89]，我们不需要从猜测解析公式的角度来查找结构因子，而只需使用 NN 来代替它。为了确保神经网络模型的精度度，需要足够的数据集来训练该 NN。幸运的是，使用参考文献[69] 中的方法，我们可以获得具有不同 ka 和 L/a 的大量精确结构因子数据。通过这些数据对神经网络进行训练，我们可以获得高精度的 NN 模型，以轻松计算结构因子。

1.4 研究内容

这项工作的动机有两个。首先，我们尝试在 L/a 和 ka 的整个参数空间中为蠕虫状链找到一种更有效，更直接的结构因子表述，它既不需要像 Monte Carlo 模拟那样进行大量计算，也不需要求解偏微分方程。其次，我们想为聚合物链的散射实验建立一种可能的测量工具。如果给出散射强度数据，则可以容易地获得高分子的链长 L 和库恩长度 a。

这一章的概述如下。我们首先在第2部分中展示如何将 NN 应用于结构因子的拟合，包括 NN 的基本介绍，训练目标，训练过程以及 NN 的体系结构对拟合结果的影响。此后，我们建立了一种方法，通过使用训练后的 NN 来预测3部分聚合物链的轮廓长度 L 和库恩长度 a。

2 结构因子的拟合

2.1 深度神经网络

首先，我们介绍神经网络的一些基本概念。神经网络 (NN)，也称为人工神经网络（Artificial neural networks, ANNs），是一种计算系统，可以通过一般地考虑示例来学习执行不同的任务，而无需使用特定于任务的规则进行编程。NN 是基于称为神经元 (Neurons) 连接构成的集合。如在图3-1中，从神经元 i 到神经元 j 的连接用于将激活输出 a_i 从 i 传

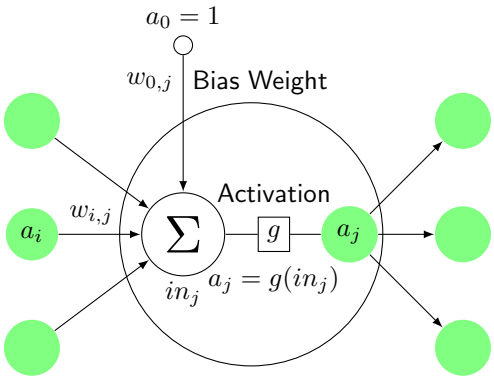

图 3-1 单个神经元的数学模型。神经元的输出为 $a_j = g(\sum_{i=0}^{n} w_{i,j}a_i)$，这里的 x_i 是第 i 个神经元的输出，$w_{i,j}$ 是从神经元 i 到这个神经元的连接的权重。

Figure 3-1 The mathematical model for a neuron. The unit's output activation is $a_j = g(\sum_{i=0}^{n} w_{i,j}a_i)$, where x_i is the output of unit i and $w_{i,j}$ is the weight on the link from unit i to this unit.

播到 j。每条连线还具有与之关联的权重 $w_{i,j}$，它表示连接的强度和正负。每个神经元都有一个虚拟输入 $a_0 = 1$，且具有与之关联的权重 $w_{0,j}$。每个神经元 j 首先计算其输入的加权和：

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

然后将这个加权和输入到一个激活函数 g，得到输出[90]：

$$a_j = g(in_j) = g\left(\sum_{i=0}^{n} w_{i,j}a_i\right) \tag{3-5}$$

非线性激活函数是 NN 具有强大功能的关键，它可以使 NN 几乎可以逼近任何函数。在这项工作中，我使用了 Sigmoid 函数：

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

确定了单个神经元的数学模型后，可以通过将它们连接起来获得全连接的 NN。如图3-2所示，全连接 NN 分层排列，它可以分为输入层，许多个隐藏层和输出层。其中只有输入层不参与方程式3-5中的计算。每层上可以有多个神经元。隐藏层和输出层中的任意一个神经元都连接到上一层中的所有神经元。NN 可以看作是从输入 $\mathbf{x}$ 到输出 $h_{\mathbf{w}}(\mathbf{x})$ 的映射 h，其中 $\mathbf{w}$ 是此 NN 中所有权重的集合。变化 $\mathbf{w}$，则可以获得不同的 h。调整 $\mathbf{w}$ 以逼近另一个函数 f 的过程称为**训练**神经网络。我们通过向 NN 反复展示大量输入/输出对 (Input–output pairs) 来训练 NN，以便通过调整 $\mathbf{w}$ 逐渐**学习**从输入到输出的映射。输入输

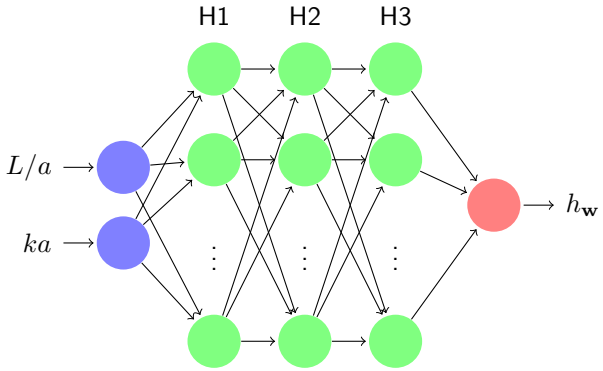

图 3-2 一个拥有三层隐藏层的全联接网络
Figure 3-2 A fully connected NN with 3 hidden layers

出对构成**训练集**，这种带答案的学习称为**监督**学习。

更具体地说，训练任务可以描述如下。给定 N 个示例输入输出对 $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_j, y_j), \ldots, (\mathbf{x}_N, y_N)$, 其中 $\mathbf{x}_j = ((L/a)_j, (ka)_j)^T$，而 y_j 是通过参考文献[69] 中的方法生成的。

$$y = f(\mathbf{x}) = (L/a)(ka)^2 S(L/a, ka), \tag{3-6}$$

找到一个函数 h 用来近似函数 f。训练神经网络的方法如下。首先我们定义个损失函数:

$$(\mathbf{w}) = \frac{1}{N} \sum_x \| f(x) - h_{\mathbf{w}}(x) \|^2 \tag{3-7}$$

损失函数表明 h 与目标函数 f 相距多远。通过训练神经网络，我们希望找到权重 $\mathbf{w}^*$，以便可以将示例中的损失函数最小化，即

$$\mathbf{w}^* = \arg\min_{\mathbf{w}}(\mathbf{w}) \tag{3-8}$$

找 $\mathbf{w}^*$ 的算法 (Optimizer) 很多，但是基本思想可以表示如下。权重 w_i 通过以下方式更新

$$w_i \leftarrow w_i - \alpha \frac{\partial(\mathbf{w})}{\partial w_i} \tag{3-9}$$

其中 α 是学习率。在我们的工作中，我们使用了一个名为 Adam[91] 的自适应学习速率优化器，该优化器专门用于训练神经网络。为了提高训练效率，训练集通常分为许多小批。每次使用小批量数据更新 $\mathbf{w}$。此外，训练集多次用于更新 $\mathbf{w}$。使用掉整个训练集更新 $\mathbf{w}$ 的过程也称为一个 *episode*。

这项工作的结构因子训练集来自求解扩散方程[69] 中的数值方法，该方法在 Spakowitz 和 Wang[79] 用无限连续分数法计算的结构因子之间取得了很好的一致性。与 Dirac 传播器方法[81] 或 Monte Carlo 模拟[84] 等其他方法相比，此方法即使是在很小和很大的刚性 L/a 范围下也可精确确定整个 L/a-ka 空间中的结构因子（$L/a, ka \in [10^{-2}, 10^3]$），从而为可靠的训练集奠定了基础。

在 Zhang 的工作[69] 中，多项式 $(L/a)(ka)^2 S$ 是一个以 L/a 和 ka 作为自变量的函数。我们将 $\mathbf{x} = (L/a, ka)^T$ 和相应的 $(L/a)(ka)^2 S$ 作为训练样本。对每个 L/a，在 $\ln(ka)$ 上均匀采集 $[10^{-2}, 10^3]$ 中的 100 个 ka 样本点。类似地，在 $\ln(L/a)$ 上均匀采集 $[10^{-2}, 10^3]$

中 100 个 L/a 样本点。我们获得了 10,000 个训练样本，它涵盖了蠕虫链模型描述的 ka 和 L/a 的整个区域。

2.2 拟合结果

通过使用 Tensorflow，我们选择了 Adam 优化器，其固定学习率为 10^{-5}，4 层隐藏层和每层 25 神经元。经过 6×10^5 的训练后，损失值 L 降低为 6.9×10^{-7}。

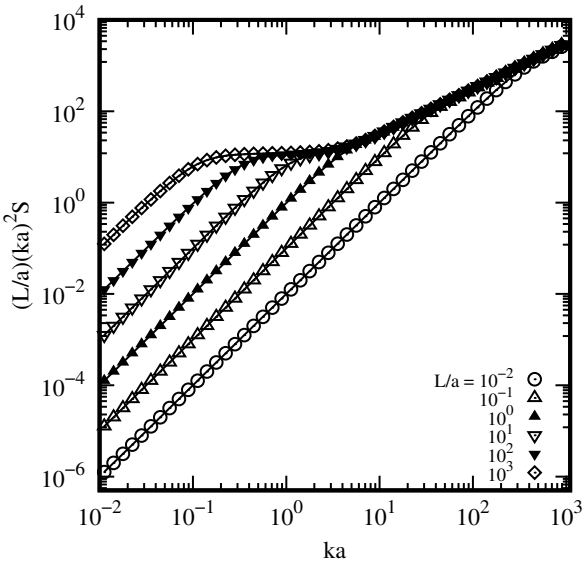

图 3-3 目标值（圆，三角形等）与训练后的神经网络预测值（直线）之间的结构因子比较。神经网络为 4 隐藏层，而每个隐藏层上有 25 个神经元，训练最后的损失函数 $= 6.92 \times 10^{-7}$ 。

Figure 3-3 Structure factor comparison between the target values (circles, triangles, etc) and trained neural network predictions (lines) in logarithmic coordinates with 4 hidden layers and 25 nodes on each hidden layer with $= 6.92 \times 10^{-7}$.

为了证明神经网络形式的结构因子的有效性，在图3-3中，我们绘制了 5 种不同刚性的神经网络模型预测值，其中 $L/a = 10^{-1}, 10^0, 10^1, 10^2$ 和 10^3。为了进行比较，我们还对求解扩散方程[69] 给出的结构因子进行了求解，并用圆圈，三角形等表示。神经网络可以很好地代表整个波数 k 范围内不同刚性的结构因子，并且与求解扩散方程[69] 提出的方法获得的精确结果一致。

为了进一步验证训练后的神经网络的拟合结果，我们在线性坐标的不同比例下进行了比较。如图3-4中所示，a，b，c 和 d 分别对应于不同的 ka 范围 $[10^{-2}, 10^{-0}]$, $[10^0, 10^1]$, $[10^1, 10^2]$, $[10^2, 10^3]$，其中实线是神经网络模型给定的值，圆形，三角形等是求解扩散方程[69] 得出的解。可以得出结论，神经网络模型可以在**整个** $L/a - ka$ 空间中给出高度准确的结构因子值。因此，我们的模型对刚性和半刚性聚合物链也有很好的描述，如波动理论和散射实验等具有实际意义。

更重要的是，我们的训练后的神经网络模型可以通过有限的离散训练样本在整个 L/a-ka 空间中提供结构因子的连续函数。如图 3-5所示，在对数坐标中获得连续的 $(L/a)(ka)^2S$ 平面。该结果表明，给定**任意** L/a-ka 对，可以直接预测结构因子。同样，我们可以轻松地同时计算多个结构因子的值，因此神经网络模型大大提高了计算效率。

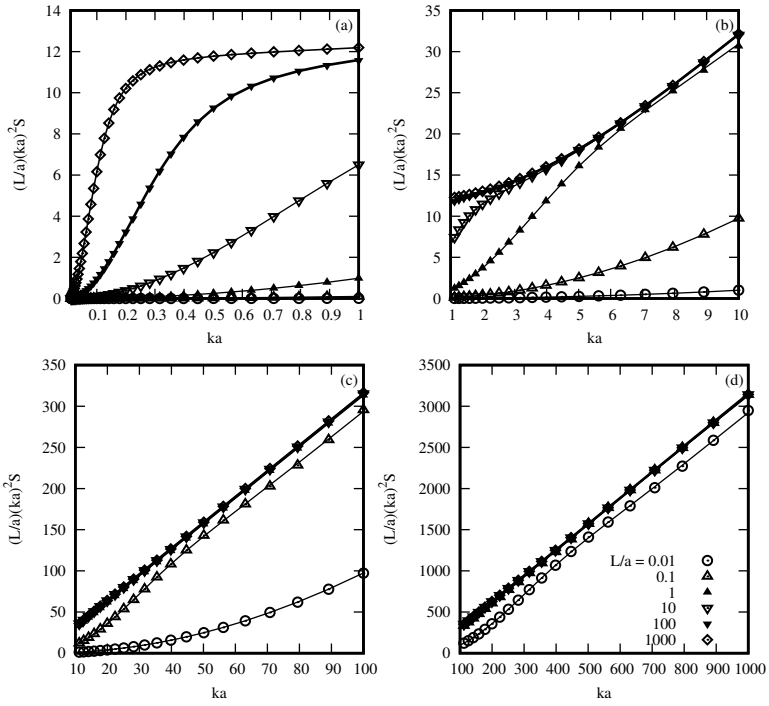

图 3-4 目标值（圆形，三角形等）与线性坐标中经过训练的神经网络预测（线）之间的结构因子比较。(a) $ka < 1$; (b) $ka \in [1, 10]$; (c) $ka \in [10, 100]$; (d) $ka \in [100, 1000]$.

Figure 3-4 Structure factor comparison between the target values (circles, triangles,etc.) and trained 神经网络 predictions (lines) in linear coordinates. (a) $ka < 1$; (b) $ka \in [1, 10]$; (c) $ka \in [10, 100]$; (d) $ka \in [100, 1000]$.

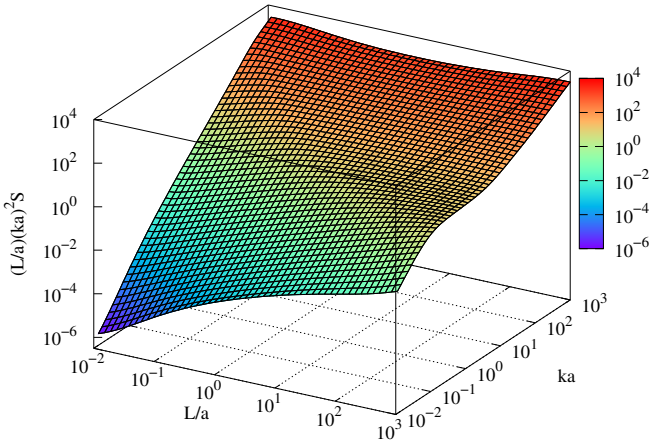

图 3-5 由训练后的神经网络模型生成的 $(L/a)(ka)^2S$ 的连续曲面。

Figure 3-5 The continuous surface of $(L/a)(ka)^2S$ generated by the trained neural network model.

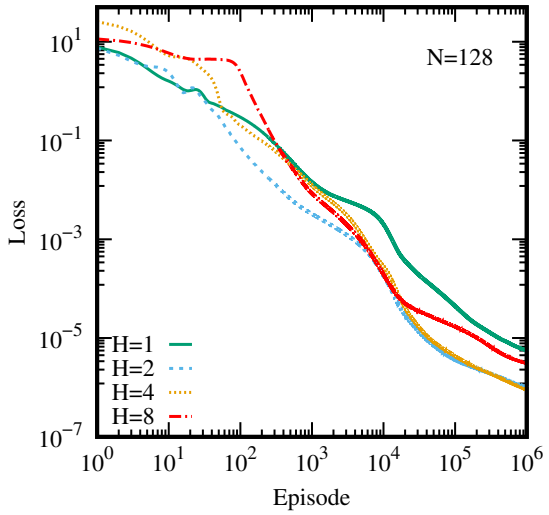

图 3-6 损失函数对不同数量的隐藏层 H（$N = 128$）随时间的变化。
Figure 3-6 Time dependence of the loss function for different number H of hidden layers ($N = 128$).

2.3 网络结构的选择

超参数，例如神经元的数量，隐藏层的数量，优化器和学习率，在训练中也很重要。在这项工作中，我们集中于隐藏层数 H 和每个隐藏层中神经元数 N 的影响。

为了简化参数调整过程，我们使每个隐藏层上的节点数相同。为了研究 H 对拟合结果的影响，我们固定了 N，然后我们使用 H 的四个值（$1, 2, 4, 8$）获得了四个网络结构。最后，我们分别训练了神经网络。

图3-6显示了当隐藏层节点的总数 N 固定为 128 时，四个独立训练中 $Loss$ 的变化。在训练结束时，Episod ~ 900000，$H = 2, 4$ 的损失函数小于 $H = 1, 8$ 的损失函数。该结果表明，当固定 N 时，太多或太少的隐藏层将降低训练后的神经网络模型的性能。因此，选择适当的 H 可以帮助使模型收敛得更快。

	$N = 32$	$N = 64$	$N = 128$	$N = 256$
$H = 1$	1.531×10^{-5}	8.059×10^{-6}	5.607×10^{-6}	7.129×10^{-6}
$H = 2$	1.423×10^{-6}	6.197×10^{-7}	1.024×10^{-6}	1.644×10^{-6}
$H = 4$	3.847×10^{-6}	1.680×10^{-6}	9.017×10^{-7}	5.411×10^{-7}
$H = 8$	2.597×10^{-4}	1.918×10^{-5}	3.109×10^{-6}	2.005×10^{-6}

表 3-1 在 9×10^5 次训练后，不同 N 和 H 的损失函数的比较。
Table 3-1 The comparison of $Loss$ for different N and H after 9×10^5 episodes of training.

为了进一步测试 N 的效果，我们分别用 $N = 32, 64, 128$ 和 256 训练了神经网络。在附录 2中，显示了不同 N 的损失函数随训练时间的变化。在表 3-1中，我们列出了不同 N 和 H 情况下，当 Episode $= 900000$ 时的损失函数值。从表中我们注意到，对于所有 N，$H = 2, 4$ 的损失函数小于 $H = 1, 8$ 的损失函数。因此，在 $[H_{\text{min}}, H_{\text{max}}]$ 之间必然有一个最优的 H 值，在我们的例子中为 $[1, 8]$。另外，随着 N 的增加，损失值减少得更多，这意味着训练过程可以收敛得更快，并且可以获得更准确的预测结果。

3 预测高分子链的链长和库恩长度

小角中子散射（SANS）是研究聚合物结构的一种广泛使用的技术。在 SANS 中，散射强度 I 被测量为散射矢量 q 的函数。这个工作中得到的神经网络中蠕虫状链的结构因子是一个精确的解。先前用于分析散射实验的任何近似构造都可以直接由神经网络代替。在这一部分中，我们以来自 SANS 实验的一些公共的聚合物散射强度数据为例，来说明训练后的神经网络模型的用法，然后预测了聚合物链的两个重要参数，即轮廓长度 L 和库恩长度 a。

3.1 方法

聚合物链的散射强度可以使用以下公式拟合[84,92]

$$I_p(q) = cP(q)S(q) \tag{3-10}$$

其中 c 是比例因子，$S(q)$ 是由神经网络模型获得的结构因子，并且

$$P(q) = \left[\frac{2J_1(Rq)}{Rq}\right]^2 \tag{3-11}$$

是形状因子，其中 $J_1(x)$ 是一阶贝塞尔函数，而 R 是横截面半径。我们将聚合物链的有限横截面近似为一个半径为 R，长度为 a 的圆柱体[84]。结构因子 S 由参数 L 和 a 确定，而形状因子 P 由 R 确定。因此，这里有四个拟合参数：轮廓长度 L，库恩长度 a，半径 R 和缩放系数 c。我们可以使用等式3-10通过更改参数 L, a, R, c 来拟合 SANS 数据。定义

$$\epsilon(a, L, R, c) = \frac{1}{N}\sum_{i=1}^{N}\left(I_p^i(a, L, R, c) - I^i\right)^2 \tag{3-12}$$

作为优化目标，其中 I 是由 SANS 获得的聚合物链的散射强度，而 I_p 是由我们的神经网络模型预测的散射强度。需要调整参数 a, L, R, c，以使预测的散射强度 $I_p(q)$ 和实验测得的 $I(q)$ 尽可能接近。当 ϵ 足够小时，可以获得最优参数 L^*, a^*, R^*, c^*。因此，我们预测了聚合物链的轮廓长度（L^*）和库恩长度（a^*）。

公式3-12中有多个拟合参数，这将给拟合实验数据带来困难，尤其是在具有波动的散射数据中。通过修改理想的链结构因子，可以近似得出溶液中链的近似结构因子 Eq.3-12。为了描述链厚度以及溶剂-单体和单体-单体相互作用等的影响，我们不得不在公式中引入一些其他参数。在很多文献[84,93,94] 中也有许多近似的公式用于散射实验。这些公式的前提是理想的蠕虫状链模型的结构因子。而本工作的主要目的是得到一种理想的蠕虫状链模型的精确结构因子公式。

3.2 讨论

使用3.1中的方法，给定聚合物链的 SANS 强度数据，我们可以确定轮廓长度和库恩长度。下面给出两个例子。

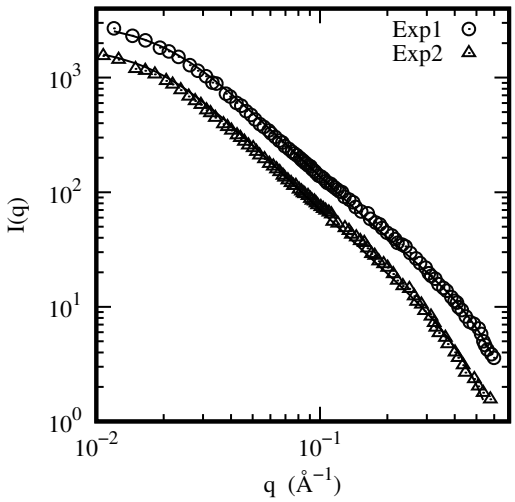

图 3-7 在 CS_2 中，分子量 M_w 为 50000 的 PS 的 SANS 数据和经过训练的神经网络模型预测之间的散射强度比较。

Figure 3-7 Scattering intensities comparison between the SANS data and the trained 神经网络 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.

3.2.1 Polystyrene

Rawiso，Duplessix 和 Picot[95] 已使用 SANS 确定了具有不同选择性氘化聚合物的无规聚苯乙烯（PS）在二硫化碳（CS2）中的散射强度数据。如图3-7所示，我们对氘化（Exp1：圆点）和完全氘化（Exp2：三角点）PS 的苯环这两组散射强度数据做连预测。在表3-2中，确定的库恩长度 a 分别为 22.38 和 22.17 Å，这与先前 SANS 确定（22 ∼ 27 Å）很吻合。对于 Exp1 和 Exp2，轮廓长度 L 分别为 1300 和 1574 Å。这些结果也与 Pedersen 得到的[84] 的 1360 和 1810 Å 值高度吻合。

	Upper	Lower
Reasonable Target Value L [1]	1360	1810
神经网络 L	1300.03	1573.76
Reasonable target value a	22 ∼ 27	22 ∼ 27
神经网络 a	22.38	22.17
ϵ [2]	0.00068	0.0013

表 3-2 图3-7 对 PS 的 L 和 a 的预测值。(单位 Å)

Table 3-2 Fig. 3-7 The predictions of L and a (unit: Å)for PS.

3.2.2 Poly(3-(2'-ethyl)hexylthiophene)

另一组散射强度数据来自 McCulloch 等人的对聚合物链 P3EHT4 做的实验[94]。他们提出了一个非常新颖的 SANS 强度模型，即经多分散性校正的蠕虫状链模型:

$$I(q) = K \int_{n=0}^{n=\infty} w_i g\left(u_{ni}\right) n_i \mathrm{d}n_i + I_{\text{inc}} \tag{3-13}$$

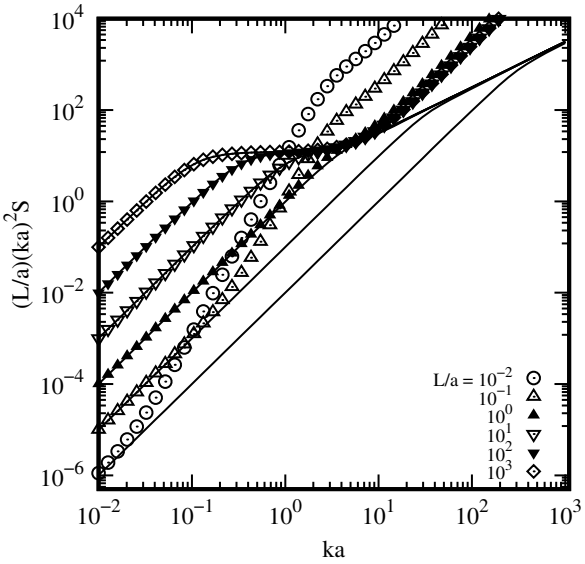

图 3-8 训练后的神经网络模型（线）和蠕虫状链模型 g（圆形，三角形等）之间的结构因子比较[94]。

Figure 3-8 The structure factor comparison between trained neural network model (lines) and the worm-like chain model g (circles, triangles, and etc.)[94].

其中 McCulloch 得到的[94] 结构因子由 g 表示

$$g(u) = \frac{2}{u^2}\left(u - 1 + \mathrm{e}^{-u}\right) + \frac{2}{5q^2L^2}\left[4u - 11u\mathrm{e}^{-u} + 7\left(1 - \mathrm{e}^{-u}\right)\right] \tag{3-14}$$

w_i 是特定分子量下的重量分数，并且

$$u = q^2R_{\mathrm{g}}^2 = q^2\left[\frac{Ll_{\mathrm{p}}}{3} - l_{\mathrm{p}} + \frac{2l_{\mathrm{p}}^3}{L}\left(1 - \frac{l_{\mathrm{p}}}{L} + \frac{l_{\mathrm{p}}}{L}\mathrm{e}^{-L/l_{\mathrm{p}}}\right)\right]$$

这里 l_p 是持久长度。

原则上，我们可以直接使用我们训练过的神经网络模型获得的结构因子 S 来代替等式3-13中的 g，然后拟合 P3EHT4 的轮廓长度和库恩长度。但是，由于缺少 P3ETH4 的原始绝对分子量分布，因此我们没有使用等式3-13来拟合 SANS 强度数据。为此，我们制作了 $a = 10$ 时，$(L/a)(ka)^2S$ 和 $(L/a)(ka)^2g$ 比较图来对比神经网络模型的 S 和等式3-14中的 g。

我们发现，当 $L/a > 1$，$ka < 10$，神经网络模型的 S 和等式3-14中的 g 匹配得很好，如图所示3-8。具体来说，因为 g 是根据柔性链模型的公式推导的，并且在大 k 的情况下使用近似形式。因此，对于柔性聚合物链（$L/a = 10^1, 10^2, 10^3$），当 $ka < 10^0$ 时，g 很好地描述了结构因子。但是对于较大的 ka ($ka > 10^1$), g 不能很好地描述高分子链。另外，对于半刚性聚合物链（$L/a = 10^0$），当 k 小时，g 能够较好地反映结构因子，但是当 $k > 10^0$ 时，它是不够的。实际上，在许多情况下，半刚性链却是关键。此外，g 不能描述刚性链的结构因子（$L/a < 10^{-1}$），因为刚性链的近似值不够好。这里的结构因子 g 的形式与 Pedersen[84] 和 Kholodenko[93] 提出的结构因子非常相似。但是后者却可以在刚性极限和较大的 k 极限方面提供更好的描述。

我们的神经网络模型可以在整个 L/a-ka 空间中给出 S 的精确预测。因此，如果将

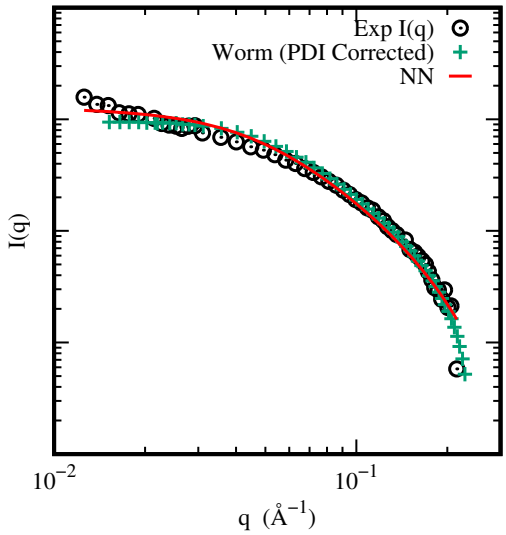

图 3-9 SANS 数据与不同模型的预测之间的散射强度比较，包括 Debye，蠕虫链（PDI 校正）[94] 和神经网络模型对于 P3EHT-4 的预测。

Figure 3-9 Scattering intensities comparison between the SANS data and the prediction of different model including Debye, Wormlike chain(PDI Corrected) [94] and the neural network model for P3EHT-4.

等式3-13中的 g 替换为 S，我们期望的拟合结果与 McCulloch 得到的结果[94] 相同。另外，当 ϵ 最小时, 由我们在等式3-10中使用的强度模型确定的 P3EHT4 的库恩长度 a 为 5.013 Å。如图3-9所示，根据我们的神经网络模型计算出的强度 $I_p(q)$ 非常适合 SANS 强度 $I(q)$。

4 结论和意义

我们通过训练全连接神经网络，为蠕虫状链状聚合物的结构因子开发了一种有效的模型。我们的神经网络模型具有以下特征：(a) 可以轻松获得整个 L/a-ka 空间中的高精度连续数值解；(b) 与先前的数值和分析方法[69] 中的计算高度一致。此外，(c) 我们还提出了该模型的一种应用。结合 SANS 强度数据，我们可以确定聚合物链的轮廓长度和库恩长度。因此，我们的神经网络模型为实验研究人员提供了一种探索聚合物链特性的潜在工具。

第四章　总结与展望

本文包含机器学习在凝聚态物理中的两个具体应用。在第一个应用中，我们利用AIMD 的方法模拟了液态体相水，从微观的角度对水分子的运动方式进行观测。通过机器学习的方法我们发现了液态体相水中存在大量的 DA 交换过程。另外我们还探究了温度对于 DA 交换的影响，发现在 280K 到 360K 范围内 DA 交换和扩散过程的比例都约为1:4。在方法处理上使用 AIMD 模拟的结果比经典基于力场的动力学模拟更精确。有向氢键布居的定义帮助我们观测包含氢键方向信息的构型改变，从而将观测角度和距离转变成观测有向氢键布居随时间的变化。准氢键和有向图的方法很好地帮助我们分析模拟的水系统。基于双向 LSTM 自编码器分类器能够对氢键的状态改变过程进行预测。当然这里使用的机器方法是一次尝试，以后可能可以使用卷积神经网络去构建分类器。另外在这个工作中我们仅仅研究了液态体相水，我们同样也可以用本文中提到的方法去研究气液界面，固液界面的水。水作为生命之源是人类不可或缺的一部分。水结构的探索作为21 世纪最重要的 125 个科学问题之一仍然是人类面临的重大问题。

在第二个应用中，我们建立了高分子链的结构因子神经网络模型，训练好的神经网络总结了以往的求解扩散方程得到结构因子的经验，可以得到任意刚性和波数下结构因子高精度的预测值。在此基础上，我们尝试对已有的中子散射实验数据进行库恩长度和链长的预测。结果表面，结合中子散射实验数据，利用神经网络模型，我们可以很好地预测出高分子链的链长和库恩长度。为实验上测量高分子链提供了一种有效的方法。机器学习作为一种通用的工具相信它在未来会有更广泛的应用，我们也期待它能够帮助人类更好地探索未知世界。

附录

1 通过插值算法得到的结构因子

由于 $(L/a)(ka)^2S$ 曲面具有单调性，我们还可以通过使用合适的插值算法来获得结构因子。除了使用神经网络外，我们还使用了两种插值算法来加速结构因子的计算。在这两种插值算法中，我们使用了与2.1中所述相同的数据点。我们发现，插值算法在整个 ka，L/a 空间中近似给出了结构因子的数值解。如图所示: (a). 使用最近邻方法 (Nearest interpolation)，(b) 使用三次样条方法 (Cubic interpolation)。通过图0-1中的最近邻方法获得的 $(L/a)(ka)^2S$ 表面不如通过图0-2中的三次样条方法获得的表面光滑。通过三次样条法获得的结构因子面更接近 MDE 和 NN 模型的解。插值算法在大 k 和小 k 范围内均能很好地拟合，在那里可以很好地确定蠕虫状链的分形维数。对于中等范围的 k 条件和半柔性条件，S 变化很快。插值算法则需要更多的数据。

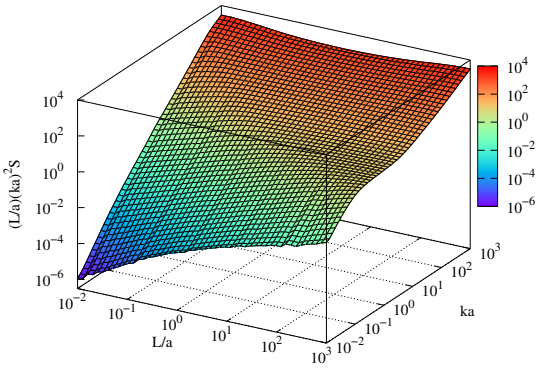

图 0-1 通过最近邻法获得的 $(L/a)(ka)^2S$ 的表面。
Figure 0-1 The surface of $(L/a)(ka)^2S$ obtained by nearest-neighbor interpolation method.

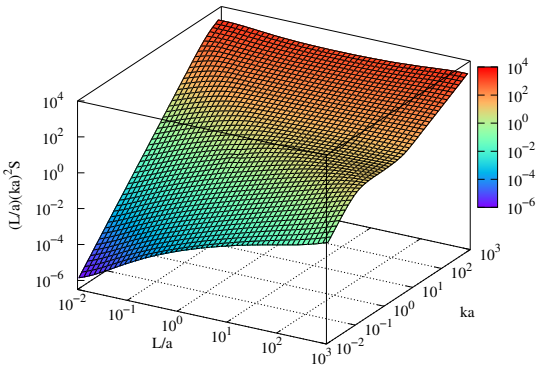

图 0-2 通过三次样条法获得的 $(L/a)(ka)^2S$ 的表面。
Figure 0-2 The surface of $(L/a)(ka)^2S$ obtained by cubic-spline interpolation method.

2 不同网络结构的损失函数

图0-3 显示了损失函数对于不同的 N 的变化。

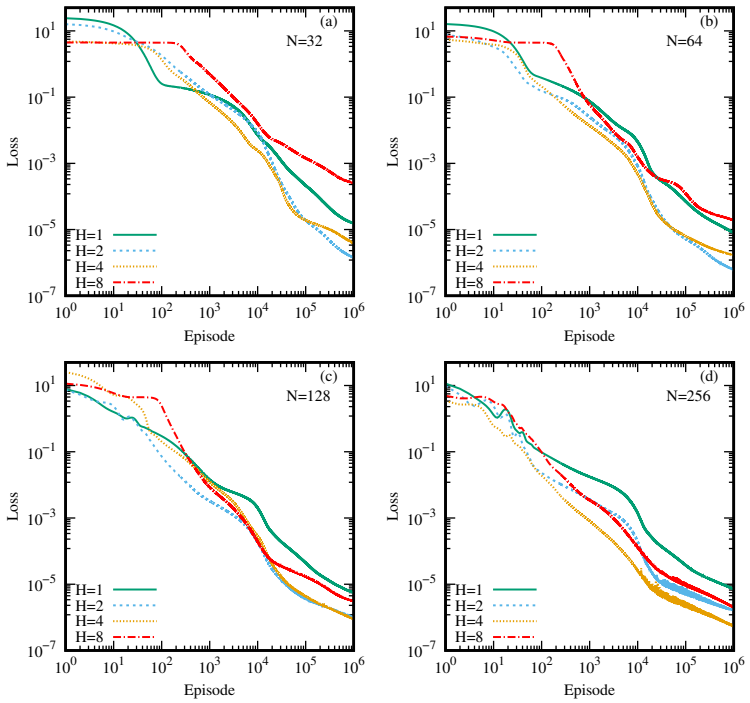

图 0-3 不同 H 和 N 参数下损失函数随训练时间的变化。(a) $N = 32$; (b) $N = 64$; (c) $N = 128$ 和 (d) $N = 256$。

Figure 0-3 Time dependence of the loss function for different H and N of the NN. (a) $N = 32$; (b) $N = 64$; (c) $N = 128$ and (d) $N = 256$.

3 其他结构因子模型

在本附录中，我们列出了不同方法中结构因子的一些解析表达式。

3.1 Kholodenko

在 Kholodenko 的工作[81] 中，获得了结构因子 S，该结构因子正确地再现了刚性杆和随机线圈的极限，并由下式给出

$$S(k) = \frac{2}{x}\left[I_{(1)}(x) - \frac{1}{x}I_{(2)}(x)\right] \tag{A.1}$$

这里 $I_{(\mathrm{n})}(x) = \int_0^x f(z)z^{n-1}\mathrm{d}z$, $n = 1, 2$, $x = 3L/a$,

$$f(z) = \begin{cases} \frac{1}{E}\frac{\sinh(Ez)}{\sinh z} & (k \le 3/2\mathrm{a}) \\ \frac{1}{\hat{E}}\frac{\sin(\hat{E}z)}{\sinh z} & (k > 3/2\mathrm{a}) \end{cases}$$

且

$$E = \left[1 - \left(\frac{2}{3}ak\right)^2\right]^{1/2}, \hat{E} = \left[\left(\frac{2}{3}ak\right)^2 - 1\right]^{1/2}.$$

3.2 Pederson 和 Schurtenberger

在 Pedersen 和 Schurtenberger 的工作[84] 中, 半柔性链的结构因子由下式给出

$$S = S_{\text{SB}}P + S_{\text{loc}}(1 - P) \tag{A.2}$$

这里

$$S_{\text{loc}} = \frac{c_1}{Laq^2} + \frac{\pi}{Lq} \tag{A.3}$$

是 Burchard 和 Kajiwara[100] 建议的在高 q 处的近似散射函数，

$$S_{\text{SB}} = S_{\text{Debye}} + \frac{c_2 a}{L}\left[\frac{4}{15} + \frac{7}{15x} - \left(\frac{11}{15} + \frac{7}{15x}\right)\exp(-x)\right] \tag{A.4}$$

是 Sharp 和 Bloomfield[101] 为 Daniels 近似计算的散射函数，并且

$$P = \exp\left[-\left(\frac{qa}{q_1}\right)^{p_1}\right]$$

这里 q_1 和 p_1 和经验函数。在等式A.4中,

$$S_{\text{Debye}}(x) = \frac{2}{x^2}[\exp(-x) + x - 1]$$

是由 Debye 函数[102] 给出的散射函数, $x \equiv R_g^2 q^2$,

$$R_g^2 = \frac{La}{6}\left\{1 - \frac{3a}{2L} + \frac{3a^2}{2L^2} - \frac{3a^3}{4L^3}\left[1 - \exp\left(-\frac{2L}{a}\right)\right]\right\}$$

这里的参数依赖于 L/a. 对 $L/a > 2, c_1 = 1, c_2 = 1, p_1 = 5.33, q_1 = 5.53, R_g^2 = La/6$. 对 $L/a \leq 2, c_1 = 0.0625, c_2 = 0, p1 = 3.95, q_1 = 11.7a/L$.

参考文献

[1] LECUN Y, CORTES C. MNIST handwritten digit database[Z]. [S.l.: s.n.], 2010.

[2] V S. An Empirical Science Research on Bioinformatics in Machine Learning[J/OL]. Journal of Mechanics of Continua and Mathematical Sciences, 2020, spl7(1). https://doi.org/10.26782/jmcms.spl.7/2020.02.00006.

[3] HU J, NIU H, CARRASCO J, et al. Voronoi-based multi-robot autonomous exploration in unknown environments via deep reinforcement learning[J/OL]. IEEE Transactions on Vehicular Technology, 2020, 69(12): 14413–14423. https://doi.org/10.1109/tvt.2020.3034800.

[4] CIREGAN D, MEIER U, SCHMIDHUBER J. Multi-column deep neural networks for image classification[J]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012(February): 3642–3649.

[5] ZHANG Y, GAO J, ZHOU H. Breeds classification with deep convolutional neural network [C/OL]//Proceedings of the 2020 12th International Conference on Machine Learning and Computing. ACM, 2020. https://doi.org/10.1145/3383972.3383975.

[6] LLOYD S. Least squares quantization in PCM[J/OL]. IEEE Transactions on Information Theory, 1982, 28(2): 129–137. https://doi.org/10.1109/tit.1982.1056489.

[7] SIBSON R. SLINK: An optimally efficient algorithm for the single-link cluster method[J/OL]. The Computer Journal, 1973, 16(1): 30–34. https://doi.org/10.1093/comjnl/16.1.30.

[8] PEARSON K. LIII. on lines and planes of closest fit to systems of points in space[J/OL]. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1901, 2(11): 559–572. https://doi.org/10.1080/14786440109462720.

[9] JAMIESON A R, GIGER M L, DRUKKER K, et al. Exploring nonlinear feature space dimension reduction and data representation in breast CADx with laplacian eigenmaps and t-SNE[J/OL]. Medical Physics, 2009, 37(1): 339–351. https://doi.org/10.1118/1.3267037.

[10] CHEN Y Y, LIN Y H, KUNG C C, et al. Design and implementation of cloud analytics-assisted smart power meters considering advanced artificial intelligence as edge analytics in demand-side management for smart homes[J/OL]. Sensors (Basel, Switzerland), 2019, 19(9): 2047 (2019/05/02). https://pubmed.ncbi.nlm.nih.gov/31052502. DOI: 10.3390/s19092047.

[11] TEALAB A. Time series forecasting using artificial neural networks methodologies: A systematic review[J/OL]. Future Computing and Informatics Journal, 2018, 3(2): 334–340. https://doi.org/10.1016/j.fcij.2018.10.003.

[12] GRAVES A, LIWICKI M, FERNANDEZ S, et al. A novel connectionist system for unconstrained handwriting recognition[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(5): 855–868. https://doi.org/10.1109/tpami.2008.137.

[13] LI X, WU X. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition[J/OL]. CoRR, 2014, abs/1410.4281. http://arxiv.org/abs/1410.4281.

[14] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J/OL]. Neural Computation, 1997, 9(8): 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

[15] SAKURADA M, YAIRI T. Anomaly detection using autoencoders with nonlinear dimensionality reduction[C/OL]//Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis - MLSDA'14. ACM Press, 2014. https://doi.org/10.1145/2689746.2689747.

[16] AN J, CHO S. Variational autoencoder based anomaly detection using reconstruction probability [C]//[S.l.: s.n.], 2015.

[17] ZHOU C, PAFFENROTH R C. Anomaly detection with robust deep autoencoders[C/OL]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017. https://doi.org/10.1145/3097983.3098052.

[18] RIBEIRO M, LAZZARETTI A E, LOPES H S. A study of deep convolutional auto-encoders for anomaly detection in videos[J/OL]. Pattern Recognition Letters, 2018, 105: 13–22. https://doi.org/10.1016/j.patrec.2017.07.016.

[19] VAN NIEUWENBURG E P L, LIU Y H, HUBER S D. Learning phase transitions by confusion [J/OL]. Nature Physics, 2017, 13(5): 435–439. https://doi.org/10.1038/nphys4037.

[20] RAISSI M, PERDIKARIS P, KARNIADAKIS G E. Physics Informed Deep Learning (Part I): Data-driven Solutions of Nonlinear Partial Differential Equations[R/OL]. https://arxiv.org/pdf/1711.10561.pdf.

[21] LI J, ZHANG H, CHEN J Z. Structural prediction and inverse design by a strongly correlated neural network[J/OL]. Physical Review Letters, 2019, 123(10): 108002. https://doi.org/10.1103/PhysRevLett.123.108002.

[22] VANDANS O, YANG K, WU Z, et al. Identifying knot types of polymer conformations by machine learning[J]. Physical Review E, 2020, 101(2): 1–10. DOI: 10.1103/PhysRevE.101.022502.

[23] ALQURAISHI M. AlphaFold at CASP13[J]. Bioinformatics, 2019, 35(22): 4862–4865.

[24] MITSUI T. Water diffusion and clustering on pd(111)[J/OL]. Science, 2002, 297(5588): 1850–1852. https://doi.org/10.1126/science.1075095.

[25] RANEA V A, MICHAELIDES A, RAMÍREZ R, et al. Water dimer diffusion on Pd{111} assisted by an H-bond donor-acceptor tunneling exchange[J]. Physical Review Letters, 2004, 92(13): 1–4. DOI: 10.1103/PhysRevLett.92.136104.

[26] KUMAGAI T, KAIZU M, HATTA S, et al. Direct observation of hydrogen-bond exchange within a single water dimer[J]. Physical Review Letters, 2008, 100(16): 1–4. DOI: 10.1103/PhysRevLett.100.166101.

[27] NILSSON A, PETTERSSON L G M. The structural origin of anomalous properties of liquid water [J/OL]. Nature Communications, 2015, 6(1): 8998. https://doi.org/10.1038/ncomms9998.

[28] LUZAR A, CHANDLER D. Effect of environment on hydrogen bond dynamics in liquid water [J]. Physical Review Letters, 1996, 76(6): 928–931. DOI: 10.1103/PhysRevLett.76.928.

[29] KENNEDY D. What don't we know?[J/OL]. Science, 2005, 309(5731): 75–75. https://doi.org/10.1126/science.309.5731.75.

[30] PAL S K, ZEWAIL A H. Dynamics of water in biological recognition[J]. Chemical Reviews, 2004, 104(4): 2099–2123. DOI: 10.1021/cr020689l.

[31] FOGARTY A C, DUBOUÉ-DIJON E, STERPONE F, et al. Biomolecular hydration dynamics: A jump model perspective[J]. Chemical Society Reviews, 2013, 42(13): 5672–5683. DOI: 10.1039/c3cs60091b.

[32] CHAPLIN M. Do we underestimate the importance of water in cell biology?[J/OL]. Nature Reviews Molecular Cell Biology, 2006, 7(11): 861–866. https://doi.org/10.1038/nrm2021.

[33] BALL P. Water is an active matrix of life for cell and molecular biology[J/OL]. Proceedings of the National Academy of Sciences, 2017, 114(51): 13327–13335. https://www.pnas.org/content/114/51/13327. DOI: 10.1073/pnas.1703781114.

[34] FANG W, CHEN J, PEDEVILLA P, et al. Origins of fast diffusion of water dimers on surfaces[J/OL]. Nature Communications, 2020, 11(1): 1–9. http://dx.doi.org/10.1038/s41467-020-15377-8.

[35] FECKO C J, EAVES J D, LOPARO J J, et al. Ultrafast hydrogen-bond dynamics in the infrared spectroscopy of water[J/OL]. Science, 2003, 301(5640): 1698–1702. https://science.sciencemag.org/content/301/5640/1698. DOI: 10.1126/science.1087251.

[36] BAKKER H J, NIENHUYS H K. Delocalization of protons in liquid water[J/OL]. Science, 2002, 297(5581): 587–590. https://science.sciencemag.org/content/297/5581/587. DOI: 10.1126/science.1073298.

[37] HEAD-GORDON T, HURA G. Water structure from scattering experiments and simulation[J/OL]. Chemical Reviews, 2002, 102(8): 2651–2670(2002/08/01). https://doi.org/10.1021/cr0006831.

[38] LAAGE D. A molecular jump mechanism of water reorientation[J/OL]. Science, 2006, 311(5762): 832–835. https://doi.org/10.1126/science.1122154.

[39] DEBYE P. Dover books on chemistry and physical chemistry: Polar molecules[M/OL]. Dover Publ., 1970. https://books.google.co.jp/books?id=f70ingEACAAJ.

[40] LAAGE D, STIRNEMANN G, STERPONE F, et al. Reorientation and allied dynamics in water and aqueous solutions[J/OL]. Annual Review of Physical Chemistry, 2011, 62(1): 395–416. https://doi.org/10.1146/annurev.physchem.012809.103503.

[41] MADHAVI W A, WEERASINGHE S, MOMOT K I. Effects of Hydrogen Bonding on the Rotational Dynamics of Water-Like Molecules in Liquids: Insights from Molecular Dynamics Simulations[J]. Australian Journal of Chemistry, 2020, 73(8): 734–742. DOI: 10.1071/CH19537.

[42] MOILANEN D E, WONG D, ROSENFELD D E, et al. Ion–water hydrogen-bond switching observed with 2d ir vibrational echo chemical exchange spectroscopy[J/OL]. Proceedings of the National Academy of Sciences, 2009, 106(2): 375–380. https://www.pnas.org/content/106/2/375. DOI: 10.1073/pnas.0811489106.

[43] JI M, ODELIUS M, GAFFNEY K J. Large angular jump mechanism observed for hydrogen bond exchange in aqueous perchlorate solution[J]. Science, 2010, 328(5981): 1003–1005. DOI: 10.1126/science.1187707.

[44] PISKULICH Z A, LAAGE D, THOMPSON W H. Activation energies and the extended jump model: How temperature affects reorientation and hydrogen-bond exchange dynamics in water [J/OL]. Journal of Chemical Physics, 2020, 153(7). https://doi.org/10.1063/5.0020015.

[45] FELLERS R S, LEFORESTIER C, BRALY L B, et al. Spectroscopic determination of the water pair potential[J/OL]. Science, 1999, 284(5416): 945–948. https://science.sciencemag.org/content/284/5416/945. DOI: 10.1126/science.284.5416.945.

[46] KEUTSCH F N, SAYKALLY R J. Water clusters: Untangling the mysteries of the liquid, one molecule at a time[J]. Proceedings of the National Academy of Sciences of the United States of America, 2001, 98(19): 10533–10540. DOI: 10.1073/pnas.191266498.

[47] SCHULZ R, Von Hansen Y, DALDROP J O, et al. Collective hydrogen-bond rearrangement dynamics in liquid water[J]. Journal of Chemical Physics, 2018, 149(24). DOI: 10.1063/1.5054267.

[48] MÉNDEZ E, LARIA D. Nuclear quantum effects on the hydrogen bond donor–acceptor exchange in water–water and water–methanol dimers[J/OL]. The Journal of Chemical Physics, 2020, 153 (5): 054302. https://doi.org/10.1063/5.0016122.

[49] KÜHNE T D, IANNUZZI M, Del Ben M, et al. CP2K: An electronic structure and molecular dynamics software package -Quickstep: Efficient and accurate electronic structure calculations [J/OL]. Journal of Chemical Physics, 2020, 152(19). https://doi.org/10.1063/5.0007045.

[50] SCIORTINO F, FORNILI S L. Hydrogen bond cooperativity in simulated water: Time dependence analysis of pair interactions[J]. The Journal of Chemical Physics, 1989, 90(5): 2786–2792. DOI: 10.1063/1.455927.

[51] GREGORET L M, RADER S D, FLETTERICK R J, et al. Hydrogen bonds involving sulfur atoms in proteins[J]. Proteins: Structure, Function, and Bioinformatics, 1991, 9(2): 99–107. DOI: 10.1002/prot.340090204.

[52] BALASUBRAMANIAN S, PAL S, BAGCHI B. Hydrogen-Bond Dynamics near a Micellar Surface: Origin of the Universal Slow Relaxation at Complex Aqueous Interfaces[J]. Physical Review Letters, 2002, 89(11): 9–12. DOI: 10.1103/PhysRevLett.89.115505.

[53] RICHARD J. GOWERS, MAX LINKE, JONATHAN BARNOUD, et al. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations[C]//SEBASTIAN BENTHALL, SCOTT ROSTRUP. Proceedings of the 15th Python in Science Conference. 2016: 98 – 105. DOI: 10.25080/Majora-629e541a-00e.

[54] MICHAUD-AGRAWAL N, J. Denning E, B. Woolf T, et al. MDAnalysis: A Toolkit for the

Analysis of Molecular Dynamics Simulations[J]. Journal of computational chemistry, 2011, 32 (10): 2319–2327. DOI: 10.1002/jcc.

[55] HUGHES T W, WILLIAMSON I A D, MINKOV M, et al. Wave physics as an analog recurrent neural network[J/OL]. Science Advances, 2019, 5(12). https://advances.sciencemag.org/content/5/12/eaay6946. DOI: 10.1126/sciadv.aay6946.

[56] RANK N, PFAHRINGER B, KEMPFERT J, et al. Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance[J/OL]. npj Digital Medicine, 2020, 3(1): 139. https://doi.org/10.1038/s41746-020-00346-8.

[57] TSAI S T, KUO E J, TIWARY P. Learning molecular dynamics with simple language model built upon long short-term memory neural network[J/OL]. Nature Communications, 2020, 11(1): 5115. https://doi.org/10.1038/s41467-020-18959-8.

[58] HOPFIELD J J. Neural networks and physical systems with emergent collective computational abilities[J/OL]. Proceedings of the National Academy of Sciences, 1982, 79(8): 2554–2558. https://www.pnas.org/content/79/8/2554. DOI: 10.1073/pnas.79.8.2554.

[59] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[60] VANDEVONDELE J, KRACK M, MOHAMED F, et al. Quickstep: Fast and accurate density functional calculations using a mixed gaussian and plane waves approach[J/OL]. Computer Physics Communications, 2005, 167(2): 103–128. https://doi.org/10.1016/j.cpc.2004.12.014.

[61] BECKE A D. Density-functional exchange-energy approximation with correct asymptotic behavior[J/OL]. Physical Review A, 1988, 38(6): 3098–3100. https://doi.org/10.1103/physreva.38.3098.

[62] LEE C, YANG W, PARR R G. Development of the colle-salvetti correlation-energy formula into a functional of the electron density[J/OL]. Physical Review B, 1988, 37(2): 785–789. https://doi.org/10.1103/physrevb.37.785.

[63] HARTWIGSEN C, GOEDECKER S, HUTTER J. Relativistic separable dual-space gaussian pseudopotentials from H to Rn[J/OL]. Phys. Rev. B, 1998, 58: 3641–3662. http://link.aps.org/doi/10.1103/PhysRevB.58.3641.

[64] LIPPERT G, HUTTER J, PARRINELLO M. The gaussian and augmented-plane-wave density functional method for ab initio molecular dynamics simulations[J/OL]. Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta), 1999, 103(2): 124–140. https://doi.org/10.1007/s002140050523.

[65] MARTYNA G J, KLEIN M L, TUCKERMAN M. Nosé–hoover chains: The canonical ensemble via continuous dynamics[J/OL]. The Journal of Chemical Physics, 1992, 97(4): 2635–2643. https://doi.org/10.1063/1.463940.

[66] GRIMME S, ANTONY J, EHRLICH S, et al. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu[J/OL]. J. Chem. Phys., 2010, 132: 154104. https://aip.scitation.org/doi/abs/10.1063/1.3382344.

[67] SVERGUN D I. Structure Analysis by Small-Angle X-Ray and Neutron Scattering[M]. [S.l.: s.n.], 1987.

[68] BU X, ZHANG X. Scattering and gaussian fluctuation theory for semiflexible polymers[J]. Polymers, 2016, 8(9). DOI: 10.3390/polym8090301.

[69] ZHANG X, JIANG Y, MIAO B, et al. The structure factor of a wormlike chain and the random-phase-approximation solution for the spinodal line of a diblock copolymer melt[J]. Soft Matter, 2014, 10(29): 5405–5416. DOI: 10.1039/c4sm00374h.

[70] CHEN X, QI S, ZHANG X, et al. Influence of Small-Scale Correlation on the Interface Evolution of Semiflexible Homopolymer Blends[J]. ACS Omega, 2020, 5(13): 7593–7600. DOI: 10.1021/acsomega.0c00421.

[71] QI S, ZHANG X, YAN D. External potential dynamic studies on the formation of interface in polydisperse polymer blends[J]. Journal of Chemical Physics, 2010, 132(6). DOI: 10.1063/1.3314730.
[72] ZHANG X, QI S, YAN D. Spinodal assisted growing dynamics of critical nucleus in polymer blends[J]. Journal of Chemical Physics, 2012, 137(18). DOI: 10.1063/1.4765371.
[73] M. DOI; S. F. EDWARDS. The theory of polymer dynamics, Oxford University Press: volume 27 [M/OL]. 1986: 391. http://doi.wiley.com/10.1002/pol.1989.140270706.
[74] LANDAU L D, LIFSHITZ E M, SYKES J B, et al. Theory of Elasticity[M/OL]. New York: Pergamon, 1986. http://www.amazon.com/dp/075062633X. DOI: 10.1063/1.3057037.
[75] SAITÔ N, TAKAHASHI K, YUNOKI Y. The Statistical Mechanical Theory of Stiff Chains[J]. Journal of the Physical Society of Japan, 1967, 22(1): 219–226. DOI: 10.1143/JPSJ.22.219.
[76] FREED K F. Advances in chemical physics: volume 22 Advances in Chemical Physics [EB/OL]. Hoboken, NJ, USA: John Wiley & Sons, Inc., 1972. http://doi.wiley.com/10.1002/9780470143728.
[77] STEPANOW S. Statistical mechanics of semiflexible polymers[J]. European Physical Journal B, 2004, 39(4): 499–512. DOI: 10.1140/epjb/e2004-00223-9.
[78] STEPANOW S. On the behaviour of the short Kratky–Porod chain[J]. J. Phys.: Condens. Matter, 2005, 17: 1799–1807. DOI: 10.1088/0953-8984/17/20/009.
[79] SPAKOWITZ A J, WANG Z G. Exact Results for a Semiflexible Polymer Chain in an Aligning Field[J/OL]. Macromolecules, 2004, 37(15): 5814–5823. https://pubs.acs.org/doi/10.1021/ma049958v.
[80] YANG Y, BU X Y, ZHANG X. Structure factor based on the wormlike-chain model of single semiflexible polymer[J]. Acta Polymerica Sinica, 2016: 1002–1010. DOI: 10.11777/j.issn1000-3304.2016.16066.
[81] KHOLODENKO A L. Analytical calculation of the scattering function for polymers of arbitrary flexibility using the Dirac propagator[J/OL]. Macromolecules, 1993, 26(16): 4179–4183. https://pubs.acs.org/doi/abs/10.1021/ma00068a017.
[82] KHOLODENKO A L. Fermi-bose transmutation: From semiflexible polymers to superstrings[J]. Annals of Physics, 1990, 202(1): 186–225. DOI: 10.1016/0003-4916(90)90344-N.
[83] YOSHIZAKI T, YAMAKAWA H. Scattering Functions of Wormlike and Helical Wormlike Chains [R/OL]//Macromolecules: volume 13. 1980: 1518–1525. https://pubs.acs.org/sharingguidelines.
[84] PEDERSEN J S, SCHURTENBERGER P. Scattering functions of semiflexible polymers with and without excluded volume effects[J]. Macromolecules, 1996, 29(23): 7602–7612. DOI: 10.1021/ma9607630.
[85] HSU H P, PAUL W, BINDER K. Scattering function of semiflexible polymer chains under good solvent conditions[J]. Journal of Chemical Physics, 2012, 137(17). DOI: 10.1063/1.4764300.
[86] HSU H P, PAUL W, BINDER K. Estimation of persistence lengths of semiflexible polymers: Insight from simulations[J]. Polymer Science - Series C, 2013, 55(1): 39–59. DOI: 10.1134/S1811238213060027.
[87] MEHRAEEN S, SUDHANSHU B, KOSLOVER E F, et al. End-to-end distribution for a wormlike chain in arbitrary dimensions[J]. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics, 2008, 77(6). DOI: 10.1103/PhysRevE.77.061803.
[88] LIANG Q, LI J, ZHANG P, et al. Modified diffusion equation for the wormlike-chain statistics in curvilinear coordinates[J]. Journal of Chemical Physics, 2013, 138(24). DOI: 10.1063/1.4811515.
[89] CYBENKO G. Approximation by superpositions of a sigmoidal function[J]. Mathematics of Control, Signals, and Systems, 1989, 2(4): 303–314. DOI: 10.1007/BF02551274.
[90] NORVIG P, RUSSELL S J. Artificial Intelligence A modern Approach[M]. Third ed. US: Pearson Education, 2010.
[91] KINGMA D P, BA J. Adam: A method for stochastic optimization[Z]. [S.l.: s.n.], 2017.

[92] HANSEN S. Approximation of the structure factor for nonspherical hard bodies using polydisperse spheres[J]. Journal of Applied Crystallography, 2013, 46(4): 1008–1016. DOI: 10.1107/S0021889813015392.

[93] KHOLODENKO A L. Persistence length and related conformational properties of semiflexible polymers from Dirac propagator[J]. The Journal of Chemical Physics, 1992, 96(1): 700–713. DOI: 10.1063/1.462455.

[94] MCCULLOCH B, HO V, HOARFROST M, et al. Polymer chain shape of poly(3-alkylthiophenes) in solution using small-angle neutron scattering[J]. Macromolecules, 2013, 46(5): 1899–1907. DOI: 10.1021/ma302463d.

[95] RAWISO M, DUPLESSIX R, PICOT C. Scattering Function of Polystyrene[J]. Macromolecules, 1987, 20(3): 630–648. DOI: 10.1021/ma00169a028.

[96] SCHUSTER M, PALIWAL K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673–2681. DOI: 10.1109/78.650093.

[97] BALDI P, BRUNAK S, FRASCONI P, et al. Exploiting the past and the future in protein secondary structure prediction[J/OL]. Bioinformatics, 1999, 15(11): 937–946. https://doi.org/10.1093/bioinformatics/15.11.937.

[98] BALDI P, BRUNAK S, FRASCONI P, et al. Bidirectional dynamics for protein secondary structure prediction[J/OL]. Lecture Notes in Computer Science, 2001, 1828: 80–104. citeseer.ist.psu.edu/baldi99bidirectional.html.

[99] FUKADA T, SCHUSTER M, SAGISAKA Y. Phoneme boundary estimation using bidirectional recurrent neural networks and its applications[J/OL]. Systems and Computers in Japan, 1999, 30 (4): 20–30. https://doi.org/10.1002/(sici)1520-684x(199904)30:4<20::aid-scj3>3.0.co;2-e.

[100] BURCHARD W, KAJIWARA K. The statistics of stiff chain molecules I. The particle scattering factor[J]. Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences, 1970, 316(1525): 185–199. DOI: 10.1098/rspa.1970.0074.

[101] SHARP P, BLOOMFIELD V A. Light scattering from wormlike chains with excluded volume effects[J/OL]. Biopolymers, 1968, 6(8): 1201–1211. https://doi.org/10.1002/bip.1968.360060814.

[102] DEBYE P. Molecular-weight determination by light scattering[J]. Journal of Physical and Colloid Chemistry, 1947, 51(1): 18–32. DOI: 10.1021/j150451a002.

致　谢

我要感谢我的导师李士本教授，感谢他一直一来对我的鼓励; 感谢蒋滢研究员，陈征宇教授，感谢他们给我去北京航空航天大学软物质中心学习的机会; 感谢中国科学院理论物理研究所的黄刚博士，感谢他在论文设计，问题讨论等等方面给我的帮助；感谢北京交通大学的张兴华教授，感谢他在高分子链结构因子课题中对我的指导；感谢兰州大学黄韬博士，感谢他在北航给我的帮助；感谢温州大学何林李老师，温州职业技术学院王向红老师, 感谢她们一直以来的关怀；感谢温州大学季永运老师，感谢他给我提供计算机集群的支持; 感谢温州大学同组的同学戴晓勇，杨艾，李丰庆，感谢他们的陪伴; 感谢温州大学陈炎英，龚书楠师妹，感谢她们带来的欢乐; 感谢温州大学我的室友吴欣学，陈昌足，俞蔡阳，感谢他们在生活上对我的关照; 感谢我的家人，感谢他们一直以来的支持；感谢帮助过我的人，感谢你们!

攻读硕士期间发表的论文

[1] Jie Huang, Gang Huang†, and Shiben Li†, A machine learning model to classify dynamic processes in liquid water, https://arxiv.org/abs/2104.07965.

[2] Jie Huang, Shiben Li†, Xinghua Zhang†, and Gang Huang, Neural Network Model for Structure Factor of Polymer Systems, The Journal of Chemical Physics, 2020 153 (12) 124902.