本科毕业论文（设计）

题　　目：利用机器学习研究抗冻蛋白的抑制结晶机制

学生姓名：晏浩洋

学　　号：81913758

入学年份：2017

所在学院：物质科学与技术学院

攻读专业：物理学

指导教师：孙兆茹

上海科技大学

2021 年 5 月

THESIS

Subject: 利用机器学习研究抗冻蛋白的抑制结晶机制

Student Name: Haoyang Yan

Student ID: 81913758

Year of Entrance: 2017

School: School of Physical Science and Technology

Major: Physics

Advisor: Zhaoru Sun

ShanghaiTech University

Date: 2021 / 5

利用机器学习研究抗冻蛋白的抑制结晶机制

摘要

本文首先介绍了抗冻蛋白的发现历史，其热滞活性与冰重结晶抑制活性，与其分子结构。然后我们概述了对抗冻蛋白抑制结晶机制的历史研究，分析了抗冻蛋白与冰晶表面结合机制的各种假说，近年来分子动力学对抗冻蛋白抑制结晶过程的模拟，以及机器学习，尤其是神经网络，在此领域中的应用，并着重介绍了一篇利用神经网络预测不同抗冻蛋白的热滞活性的文献。在下半篇中，本文介绍了近期非常热门的分子动力学方法，深度势能分子动力学，探讨了其与从头算分子动力学和经验力场的优势所在，详细解释了其计算原理，包括转换描述符，构建和训练神经网络，并给出了深度势能分子动力学套件的使用方法。最后，我们首次尝试将深度势能分子动力学引入到模拟抗冻蛋白抑制结晶的过程中，指出了这一方法目前遇到的难点和可能的前景。

关键词：抗冻蛋白，深度势能分子动力学

USING MACHINE LEARNING TO STUDY THE ANTIFREEZE PROTEINS' MECHANISM OF INHIBITING CRYSTALLIZATION

ABSTRACT

In this article, we firstly introduce the discovering history of antifreeze proteins, their thermal hysteresis activity and ice recrystallization inhibition activity, and their molecular structure. Then, we conclude the historical researches about the antifreeze proteins' mechanism of inhibiting crystallization. We analyze the hypothesis of how antifreeze proteins combined with the ice surface, the molecular dynamics simulation of the process of antifreeze proteins inhibiting crystallization, and the application of machine learning, especially neural network, in this field. We share a paper about predicting the thermal hysteresis activity of different types of antifreeze proteins by neural network in detail. In the second half of this article, we introduce a molecular dynamics method, Deep Potential Molecular Dynamics. We indicate its advantages compared with Ab Initio Molecular Dynamics and Empirical Force Fields Molecular Dynamics. We represent its theory, including the transition to descriptor, the construction and training of its neural network. We also provide the application steps of its package, DeePMD-kit. Last, we initiate trying to introduce Deep Potential Molecular Dynamics into the simulation of the process of antifreeze proteins inhibiting crystallization. We point out some difficulties the method facing and some possible prospect.

Key words: antifreeze proteins, Deep Potential Molecular Dynamics

目录

第一章 绪论----------1
1.1 研究背景----------1
1.2 本文结构----------1
第二章 抗冻蛋白----------2
2.1 抗冻蛋白的简介----------2
2.1.1 抗冻蛋白的发现历史----------2
2.1.2 抗冻蛋白的基本性质----------2
2.1.3 抗冻蛋白的结构----------2
2.2 抗冻蛋白抑制结晶机制的历史研究----------3
2.2.1 抗冻蛋白的“吸附-抑制”模型----------3
2.2.2 抗冻蛋白与冰晶表面的结合机制----------4
2.2.3 分子动力学模拟与机器学习对抗冻蛋白的研究----------5
第三章 深度势能分子动力学（DPMD）----------7
3.1 分子动力学----------7
3.1.1 从头算分子动力学（AIMD）----------7
3.1.2 经验力场（EFF）----------7
3.2 深度势能分子动力学（DPMD）的引入----------7
3.3 深度势能分子动力学（DPMD）的原理----------8
3.3.1 将原子位置坐标转换为描述符----------8
3.3.2 通过神经网络计算势函数----------9
3.3.3 训练神经网络----------10
3.4 深度势能分子动力学套件（DeePMD-kit）的使用----------11
3.5 利用深度势能分子动力学模拟抗冻蛋白抑制结晶过程----------12
参考文献----------14
致谢----------16

第一章 绪论

1.1 研究背景

抗冻蛋白自 1969 年在南极鱼类中被首次发现之后，就成为了热门研究领域。生物体中的抗冻蛋白能够在低温下防止水的结晶，从而保护细胞不受低温伤害。将提取的抗冻蛋白作为冷冻保护剂，抑制冰的生长会在食品工业，冷冻保存等领域具有巨大的应用前景。因此，探究并理解抗冻蛋白是如何抑制结晶的在理论和应用上都具有重大意义。另一方面，随着计算物理学的发展，越来越复杂和精确的分子动力学模拟方法被应用在抗冻蛋白的研究上，而这些分子动力学模拟产生的海量数据，可以由机器学习的方法进行处理，优化和预测。在 2017-18 年被提出并发展的深度势能分子动力学（DPMD）即是近年来最热门的分子动力学模拟方法之一，它通过构建训练深度神经网络，能够同时兼顾从头算分子动力学的精确和经验力场方法的快速。如果能将深度势能分子动力学应用于抗冻蛋白的研究，可能能对抗冻蛋白抑制结晶的机制做出更完善的解释。

1.2 本文结构

本文接下来的结构如下：

在第二章中，2.1 节是对抗冻蛋白的基本介绍，先是介绍其发现历史，然后是抗冻蛋白的特殊性质，包含热滞（TH）活性与冰重结晶抑制（IRI）活性等，最后对部分抗冻蛋白的分子结构做了简单介绍。2.2 节中我们介绍了历史上对抗冻蛋白抑制结晶机制的研究，从最早最经典的“吸附-抑制”模型开始，然后是对于抗冻蛋白与冰晶表面结合机制的疑点，包括结合是否可逆等，给出了更多的实验现象和前人的假说，最后介绍了近年来对抗冻蛋白的分子动力学模拟，其结果是否能与之前结合机制的假说相印证，以及机器学习方法是如何影响分子动力学发展的，并介绍了一个采用机器学习预测抗冻蛋白热滞活性的研究。

在第三章中，3.1 节介绍了传统的分子动力学，包括基于密度泛函理论的从头算分子动力学（AIMD）和经验力场（EFF）两大类，以及它们各自的优点和局限性。3.2 节介绍了深度势能分子动力学（DPMD）的发展历史。3.3 节介绍了深度势能分子动力学的计算原理，包含了将原子坐标转换为描述符，通过神经网络计算势函数，训练神经网络三大步骤。3.4 节介绍了深度势能分子动力学套件（DeePMD-kit）的使用方法，其所需的输入文件和一些有关的网络资源。3.5 节介绍了将深度势能分子动力学应用到抗冻蛋白的模拟上时遇到的困难与可能的前景。

第二章 抗冻蛋白

2.1 抗冻蛋白的简介

冰结合蛋白，英文名 Ice - Binding Proteins，简称 IBPs，是生物体中具有热滞活性、能够与冰结合、吸附到冰上并干扰其生长、改变冰晶生长特性、抑制冰晶重结晶（ice recrystallization，IR）的蛋白质的统称。冰结合蛋白在食品工业，冷冻保存等技术中具有广泛应用。

2.1.1 抗冻蛋白的发现历史

根据发现历史，冰结合蛋白（IBPs）又可分为抗冻蛋白（Antifreeze Proteins，AFPs），热滞蛋白（Thermal Hysteresis Proteins，THPs），冰结构蛋白（Ice Structuring Proteins，ISPs），冰重结晶抑制蛋白（Ice Recrystallization Inhibition Proteins， IRIPs）。抗冻蛋白最早发现于鱼类[1]，热滞蛋白最早发现于昆虫中[2]，它们具有独立的起源，此外，在植物，真菌，细菌中也有冰结合蛋白的出现。传统上将所有冰结合蛋白统称为抗冻蛋白，但随着抗冻剂在食品添加剂中的应用，抗冻蛋白容易令人联想到负面含义，因此有文献建议使用冰结合蛋白或是冰结构蛋白来代替抗冻蛋白的称呼[3]。本文方便起见，还是采用抗冻蛋白来代指所有的冰结合蛋白。

2.1.2 抗冻蛋白的基本性质

抗冻蛋白具有热滞活性（Thermal Hysteresis Activity，THA），热滞指的是溶液的冰点和熔点出现差值，热滞包含凝固滞后（Freezing Hysteresis，FH）和熔化滞后（Melting Hysteresis，MH）。凝固滞后指的是凝固点降低，是次级成核作用被阻止所导致的，熔化滞后指的是熔点升高，是由表面吸附作用导致的。一般来说凝固滞后的作用是熔化滞后的 10 倍以上，所以热滞活性通常仅指凝固滞后。

抗冻蛋白可保护生物免受低温的侵害。当植物内的组织液达到冰点时，由于异质成核剂的存在，许多小冰晶会在细胞外形成。 随着时间的流逝，大冰晶倾向于生长，而小冰晶会产生 Ostwald 熟化效应而熔化，具体到这里可称为冰重结晶（Ice Recrystallization，IR）。由于会导致膜的机械损坏或脱水，大冰晶的生长是有害的，即使它发生在细胞外。抗冻蛋白是出色的冰重结晶抑制剂，这种活性被认为是抗冻植物和某些生活在海冰中的微生物所含有的抗冻蛋白的主要作用。所有已知的抗冻蛋白都具有冰重结晶抑制（Ice Recrystallization Inhibition，IRI）活性，这与其吸附抑制机制一致。 但是，冰重结晶抑制活性和热滞活性的效力之间没有明显的相关性[4]。

2.1.3 抗冻蛋白的结构

一般来说，抗冻蛋白分子不是太大，这反映了它们需要在毫摩尔浓度下起防冻剂的作用。从结构上来说，尽管抗冻蛋白都具有与冰结合的能力，但它们具有高度的结构多样性。迄今为止，通过 X 射线晶体学已解决的结构包括一个 α 螺旋，一个 β 电磁体，四个螺旋束，聚脯氨酸 II 型螺旋束和小球状蛋白[5]。多样的结构也使得我们对其抗冻机制的探索变得更加困难。

抗冻糖蛋白（Antifreeze Glycoproteins，AFGPs）是抗冻蛋白中被研究的较为充分的一类。抗冻糖蛋白按照分子大小可以从大到小分为 AFGP1 到 AFGP8 共八类，它们全部由丙

氨酸-丙氨酸-苏氨酸的三肽重复序列（Ala - Ala - Thr）组成，其中苏氨酸的羟基被β-D-半乳糖基-（1,3）-α-N-乙酰基-D-半乳糖胺糖基化，下图所示即 AFGP8 结构，其中肽链分别由丙氨酸，丙氨酸，被乙酰基化的苏氨酸构成。在较小的抗冻糖蛋白中，第一个丙氨酸和糖基化的苏氨酸有时分别被脯氨酸和精氨酸取代[6][7]。

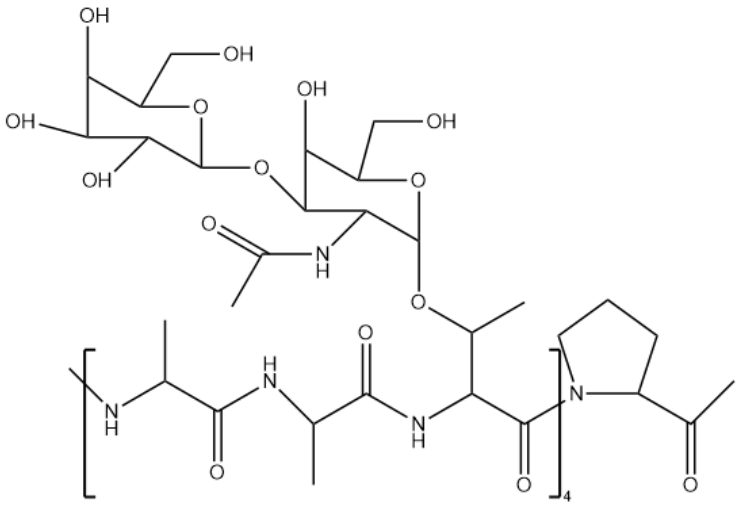

图 2-1 AFGP8 的结构示意图

2.2 抗冻蛋白抑制结晶机制的历史研究

2.2.1 抗冻蛋白的“吸附-抑制”模型

为了探究抗冻蛋白具有热滞活性从而抑制冰结晶生长的原因，1977 年，Raymond JA 和 DeVries AL 提出了“吸附-抑制”假说（Adsorption-Inhibition Hypothesis）[8]。“吸附-抑制”模型基于吉布斯-汤姆森效应（Gibbs-Thomson Effect）解释了抗冻蛋白如何抑制冰生长。吉布斯-汤姆森效应指的是整个曲面或界面上的蒸气压或化学势的变化，平衡相变参量随界面曲率变化而变化。对于冰晶表面来说，其具体方程式为：

$$\Delta T_m(r) = T_{mB} - T_m(r) = T_{mB}\frac{\alpha_p \sigma_{sl} cos\theta}{H_f \rho_s r}$$

$$= T_{mB}\frac{\alpha_p \gamma_{sl} v cos\theta}{H_m r}$$

$$= \frac{k_{GT}}{r} \qquad (2-1)$$

其中ΔT_m为熔点降低量，T_{mB}为体熔点，即界面完全平坦时的熔点，α_p为几何常数，圆柱形冰帽时α_p取 2，球形时取 4，σ_{sl}为单位面积的界面能，θ为冰帽的接触角，H_f为单位质量熔化焓，ρ_s为密度，r为界面的曲率半径，γ_{sl}为冰晶-液体的表面张力，v为摩尔体积，H_m为摩尔熔化焓。

根据实验结果，代入数值，上式也可写为[9]：

$$T_m(r) = T_{mB} - \frac{50\ nm * ℃}{r} \qquad (2-2)$$

上式表明，半径较大的冰晶具有较高的熔点，因此可以解释，在一定温度下，大于一定尺寸的冰晶会继续生长，而较小的冰晶会熔化的现象。从微观结构的角度说，抗冻蛋白分子结合正在生长的冰晶的表面，使得水分子仅能在被结合的抗冻蛋白分子之间形成冰

晶。从无限曲率半径的平坦表面开始，冰继续生长几纳米，随着冰的生长，该表面变成圆形表面，最终达到临界半径，在该临界半径上，能量的增长变得不利。在此阶段，生长停止，并且只有在温度下降时才重新开始生长。概括地说，抗冻蛋白通过结合至冰晶面，阻碍冰晶对应位点的生长，使得曲率半径减小，导致生长面冰点下降。

2.2.2 抗冻蛋白与冰晶表面的结合机制

在“吸附-抑制”模型的基础上，假设抗冻蛋白与冰晶的结合是不可逆的，更多更详细和精确的模型被提出以模拟抗冻蛋白抑制结晶的机制。例如，“床垫纽扣”模型（Mattress Button Model）认为，抗冻蛋白分子阻碍了冰晶在与表面垂直方向上的生长[10]。但是依然还有一些问题存在，比如是什么样的作用力导致抗冻蛋白能够结合到冰晶表面，对应的抗冻蛋白基团是什么，结合的位点在哪里，最关键的是，这个结合是否的确是不可逆的。

抗冻蛋白与冰晶表面的不可逆结合意味着，一旦抗冻蛋白附着在冰的表面，它将被束缚直至冰熔化。在光学显微镜下对抗冻蛋白的观察表明，低于原凝固点晶体在一段时间内没有明显的生长，这支持了不可逆结合的模式。如果这种结合是可逆的，那么围绕冰晶的水将不可避免地导致滞后间隙中冰的增长[8]。但另一方面，有实验结果表明，抗冻蛋白的热滞活性与其浓度是相关的，大致是与其浓度平方成正比关系。

$$\Delta T = \alpha C^{\frac{1}{2}} \qquad (2-3)$$

其中α为常数，C 为抗冻蛋白浓度。而且不同的抗冻蛋白在相同的浓度下，产生的热滞活性也具有差异[11][12]，这说明在冰晶表面和在溶液中的抗冻蛋白可能是存在一个动态平衡的，这给抗冻蛋白与冰晶表面的结合是可逆的提供了证据。此外，实验结果还表明，热滞效应的大小随时间变化程度并不大[13]，也就是说，附着在冰晶表面的抗冻蛋白分子数量并没有随着时间进行而不断增加，这也对抗冻蛋白与冰晶表面的不可逆结合产生了合理怀疑。

抗冻蛋白和冰晶表面之间的结合力是理解抗冻蛋白抑制结晶机制的关键。冰冻蚀刻（Freezing Etching）等技术可以帮助我们研究在抗冻蛋白的作用下冰晶形态的改变，但却无法帮助我们观察抗冻蛋白在冰晶表面结合的细节。其它研究大分子间相互作用的方法，例如酶-底物和蛋白质-配体分析等方法，在这里都无法实现。现有的对于抗冻蛋白和冰晶表面之间的结合力的探究，大都是理论上的假设推导，或是在计算机上的分子动力学模拟。

早期的假说认为氢键是抗冻蛋白与冰晶表面结合的作用力。DeVries AL 和 Lin Y 在 1977 年提出，水性残基苏氨酸和天冬氨酸在螺旋上有规律的间隔，其中苏氨酸的羟基和天冬氨酸的羧基之间的距离为 4.5 埃，这个距离与冰晶平面上氧原子的间距正好符合形成氢键的条件，从而使抗冻蛋白结合到冰晶表面[14]。但是，氢键的强度并不完全足以支撑抗冻蛋白与冰晶表面的结合是不可逆的假设。

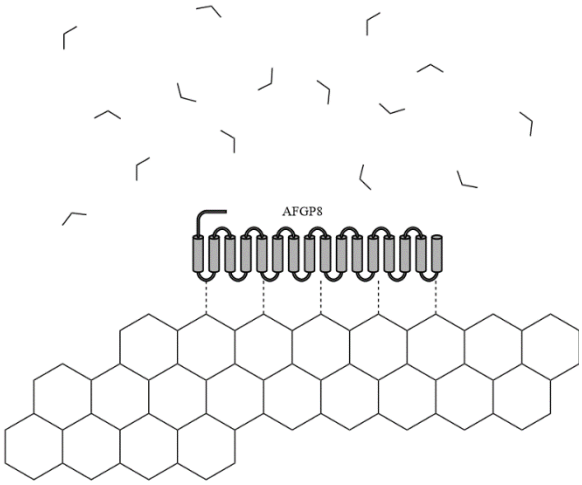

图 2-2 抗冻蛋白分子结合在冰晶表面抑制结晶过程的示意图

Kristiansen E 和 Zachariassen KE 在 2005 年提出的假说认为，接近于熔点处的温度的表面-溶液平衡决定了抗冻蛋白在冰晶表面的聚集，当温度降低时，聚集停止，抗冻蛋白分子“冻结”在冰上[15]。Knight CA 和 DeVries AL 在 2009 年提出的假说认为，热滞效应是由抗冻蛋白阻止基平面上新形成的冰生长的能力决定的，而基平面没有被蛋白质覆盖，这种新形成的冰层可以被看作是被抗冻蛋白所抑制的二次成核事件，这个步骤除了需要附着在冰面上的抗冻蛋白分子外，还需要溶液中有游离的抗冻蛋白分子[16]。这两个假说解释了抗冻蛋白与冰晶表面的不可逆结合与热滞活性与抗冻蛋白浓度成正相关以及热滞效应不随时间而改变之间的矛盾。

Garnham CP，Campbell RL 和 Davies PL 在 2011 年提出的包合物锚定模型（Anchored Clathrate Model）认为，抗冻蛋白分子通过疏水作用使水分子按晶格方式排列，并通过氢键锚定水分子晶格。锚定的水分子晶格又通过匹配特定的冰晶表面使抗冻蛋白分子与冰晶结合[17]。包合物锚定模型解释了抗冻蛋白如何与特定冰平面的特异性结合，而无法直接接触其配体，因为液体层充当了蛋白质表面和冰晶格之间的屏障。将水放置在两层中，其灵活性有助于它们的合并。X 射线晶体衍射学的研究也为包合物锚定假说提供了更多证据[18]。

2.2.3 分子动力学模拟与机器学习对抗冻蛋白的研究

随着计算机科学的发展，计算能力的进步，更多的研究开始采用分子动力学模拟的方法来探究抗冻蛋白抑制结晶的机制，比较常用的方法是，通过模拟计算与冰面结合的羟基或者说氢键的数量随时间变化规律，判断结合是否可逆。比较广泛被承认并应用的分子动力学模拟软件包括 CHARMM（Chemistry at HARvard Macromolecular Mechanics），LAMMPS（Large-scale Atomic/Molecular Massively Parallel Simulator，大规模原子分子并行模拟器）等等，一般是基于密度泛函理论或是经验力场去模拟，更详细的分子动力学模拟原理请见下一章。

2015 年，Kuiper MJ 等人使用 NAMD2.9 和 CHARMM22 力场的无偏分子动力学模拟了云杉芽虫抗冻蛋白的结构-功能机制，包括立体特异性结合以及相应的融化和冻结抑制作用。抗冻蛋白通过结构上不同于冰的有序水分子的线性阵列间接结合到棱柱形冰面上，抗冻蛋白与冰晶结合的表面的突变破坏了水的有序性并废除了活性[19]。该模拟说明了抗冻蛋白与冰晶表面的结合是不可逆的，并且冰的生长抑制与吉布斯-汤姆森定律是一致的。

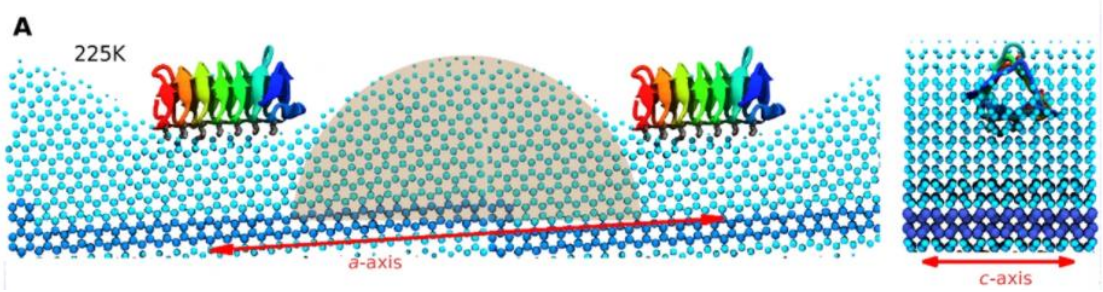

图 2-3 被吸附的抗冻蛋白抑制的稳态平衡冰面的侧视图，引自[19]

但是，也有模拟认为抗冻蛋白和冰晶表面的结合是可逆的。2018 年，Kenji Mochizuki 等人使用 CHARMM27 的碳水化合物力场，用 GROMACS 对 AFGP8 抑制结晶过程进行了分子动力学模拟，尝试解释 AFGP8 与冰结合的机理和结合力，并在模拟溶液中的 AFGP8 时采用了 PPII 螺旋二级结构。模拟表明，结合是通过肽和二糖的甲基吸附到冰上而发生的，这是由于疏水基团嵌套在冰表面的空腔中而受到的。模拟估计了 AFGP8 和更长的 AFGPs4-6 结合的自由能，发现该能量符合可逆结合的假设。模拟显示 AFGP8 通过无数种构象与冰结合，它用于在冰面中扩散并找到冰阶，并强烈吸附在冰阶上。研究认为抗冻蛋白的冰重结晶抑制活性的关键在于抗冻蛋白与冰晶间存在多个弱结合位点[20]。

近几年，随着机器学习，尤其是深度神经网络（Deep Neural Network，DNN）在各研究领域的大量应用，分子动力学模拟和抗冻蛋白机制的研究也被注入了新的活力。例如，Jörg Behler 等人在 2007 年介绍的密度泛函势能面的新型神经网络表示形式，它提供了任意大小的系统中所有原子位置的函数的能量和力，并且比密度泛函模拟快了几个数量级[21]。再例如，Matthias Rupp 等人在 2012 年引入了一种机器学习模型，仅基于核电荷和原子位置来预测各种有机分子的原子化能。将求解分子薛定谔方程的问题映射到降低复杂性的非线性统计回归问题上[22]。

2018 年，Daniel J. Kozuch 等人提出了一种利用神经网络从分子模拟中预测抗冻蛋白的抗冻活性的方法[23]。他们将氢键寿命定义为氢键自相关函数衰减到 0.1 所需平均时间，采用 GROMACS，对已知的具有随浓度变化的热滞曲线（抗冻活性谱）的 17 种抗冻蛋白进行了分子动力学模拟，并将 5 种结构相似的蛋白质作为对照组。通过分析蛋白质表面附近水的动态行为和蛋白质的几何结构，引入了一种自动检测抗冻蛋白与冰晶表面结合的方法。根据这些数据构建了一个具有 1 个规范化层，4 个 6 节点隐藏层的神经网络，采用 Adam 算法进行训练并使用五重交叉验证。他们认为与预测抗冻活性相关的抗冻蛋白的主要特征是两个表面，即平面的，冰结合表面（面向冰晶），与非平面的，非冰结合表面（面向液态水防止更多结晶）。并将这两个表面的三个参数作为神经网络的输入参数：第一，预期的冰结合表面的面积A，该面积越大，抗冻蛋白越有可能吸附；第二，冰结合表面的氢键寿命L_B，对冰结合表面识别并附着更好的抗冻蛋白一般也拥有更大的L_B；第三，非冰结合表面的氢键寿命L_N，如果L_N较大，则说明该抗冻蛋白更可能造成冰晶在非冰结合表面上的过度生长，抗冻活性较差。同时，将热滞活性Δ_T作为被预测的输出值。该网络能够从这三组物理量定量预测实验观察到的热滞。他们的结果支持了抗冻蛋白能通过一个被结合表面上的氢键的寿命所定义的水层去识别并结合冰晶表面。

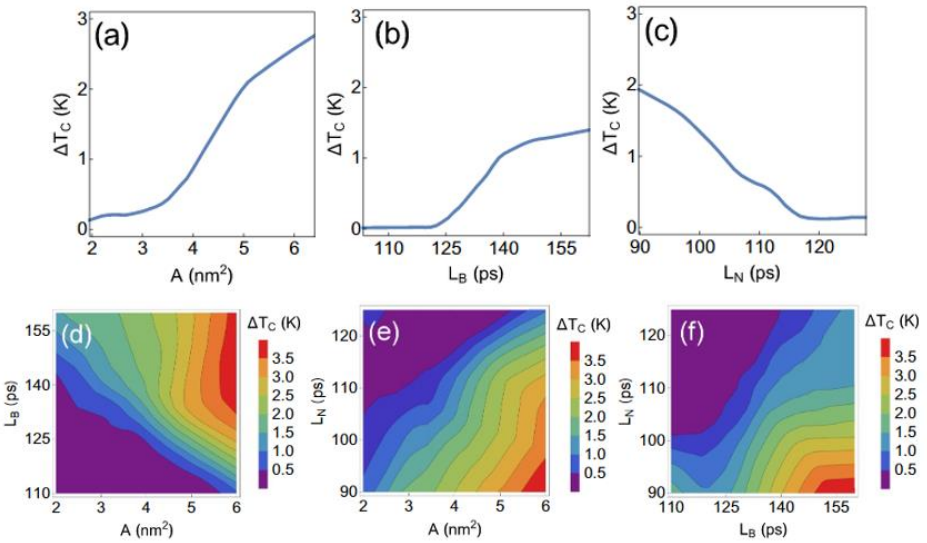

图 2-4 冰结合表面面积、氢键寿命与非冰结合表面氢键寿命预测热滞活性，引自[23]

总的来说，分子动力学模拟与机器学习改变了传统的对于抗冻蛋白抑制结晶机制的研究方式。更多更复杂更精确的分子动力学力场被构建，计算了更庞大更复杂的模型，同时，各种机器学习方法能够对这些复杂模型的数据进行预测和解释，并改善运行速度。可以预见的是，未来会有更多的研究专注于此方向，并利用计算结果对实验现象做出更完备的解释，让我们对抗冻蛋白抑制结晶的机制有更深入的了解。

第三章 深度势能分子动力学（DPMD）

3.1 分子动力学

分子动力学（Molecular Dynamics, MD）在材料，能源，物理，化学，生物等多门学科中均有重要应用。传统的分子动力学的三要素是分子构型，积分算法，力场（势函数）。其中势函数规定了原子处在一定空间的能量，将其约束在势能面上。当体系中原子数量增加，计算其势能面也变得更为复杂。

3.1.1 从头算分子动力学（AIMD）

传统上通过第一性原理（First Principle）和密度泛函理论（Density Function Theory, DFT）的从头算分子动力学（Ab Initio Molecular Dynamics, AIMD）可以算出原子在空间中任何位置的能量和力，准确性较高。但是，其计算复杂度与原子数量三次方成正比，随着体系原子数和复杂度增加，其计算成本急剧增大。因此，在应用中，一般从头算分子动力学所模拟的模型一般被限制在数百个原子内，时间尺度在 100ps 内。因此，从头算分子动力学对于大分子蛋白的模拟是难以实现的。

3.1.2 经验力场（EFF）

经验力场（Empirical Force Fields）在一定程度上能解决大体系分子动力学模拟的问题。通过密度泛函理论模拟数据的拟合，加以一定的物理直觉，从而得到一个相对简洁的势函数。

常见的经验力场，例如勒让德-琼斯势（Lennard - Jones Potential），描述了两个电中性原子间的相互作用势能，在模拟惰性气体时较为准确。

$$V(r) = 4\epsilon[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}]$$

$$= \epsilon\left[\left(\frac{r_{min}}{r}\right)^{12} - 2\left(\frac{r_{min}}{r}\right)^{6}\right] \quad (3-1)$$

例如莫尔斯势（Morse Potential），能够较好地模拟分子振动的微细结构。

$$V(r) = -D_e + D_e\left(1 - e^{-a(r-r_e)}\right)^2 \quad (3-2)$$

再例如嵌入原子势（Embedded Atom Method, EAM），将能量分解为嵌入能和斥能，在金属体系及表面的模拟上表现较好。

$$E = \sum_{i=1}^{N} G_i(\rho_{h,i}) + \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i}\varphi_{ij}(r_{ij}) \quad (3-3)$$

总的来说，经验力场方法能够用简洁的势函数模拟大规模体系，在特定的情况下表现优异，但在大多数情况下，简洁的势函数造成了体系中部分信息的丢失，带来模拟结果的误差。

3.2 深度势能分子动力学（DPMD）的引入

为了更好地兼顾分子动力学模拟中的精确度和计算成本，Han Wang，Linfeng Zhang，

Jiequn Han 等人与 2017-2018 年开发并发布了深度势能分子动力学（Deep Potential Molecular Dynamics，简称 DeePMD 或 DPMD）的深度学习套件[24][25]。

深度势能分子动力学是一种基于神经网络的分子动力学模拟方法，它克服了对称函数，库伦矩阵等辅助量的局限性。在深度势能分子动力学中，每个原子分配了一个局部参考系和一个局部环境，每个环境都包含有限数量的原子，这些原子的局部坐标是按照深度势能方法的规定以对称性保持的方式排列的，该方法被设计为仅用势能训练神经网络。通过灵活的损失函数族，深度势能分子动力学所构造的 NN 势能在扩展和有限系统中能够较为准确地复现从头算分子动力学模拟的结果。总的来说，深度势能分子动力学模拟的计算成本与系统大小成线性相关，因此相较从头算分子动力学，其计算成本低了几个数量级。

3.3 深度势能分子动力学（DPMD）的原理

3.3.1 将原子位置坐标转换为描述符

在一个拥有N个原子的体系中，设原子位置为$\{\boldsymbol{R_1},\boldsymbol{R_2},\cdots,\boldsymbol{R_N}\}$，其对应能量为$\{E_1,E_2,\cdots,E_N\}$，体系总能量为：

$$E=\sum_i E_i \tag{3-4}$$

对于一个原子来说，只需要考虑其周边的与其距离小于截断半径R_c的原子的相互作用能，因此：

$$E_i = E_{s(i)}\left(\boldsymbol{R_i},\{\boldsymbol{R_j}\mid j\in N_{R_c}(i)\}\right) \tag{3-5}$$

以上两点，保证了体系在平移，旋转，置换时具有对称性，使得计算出来的势函数具有可扩展性，即原子中心框架（Atom - Centered Frame）。

具体一点，原子位置坐标$\{\boldsymbol{R_1},\boldsymbol{R_2},\cdots,\boldsymbol{R_N}\}$是不具有对称性的，即当体系平移，旋转，置换时，由原子位置坐标计算的能量会改变，因此原子位置坐标无法直接代入体系计算，需要将其转化为具有对称性的描述符。

将位置坐标转化为描述符的方法有很多，传统上有通过 Zernike 矩等方法：

$$\rho_i(r)=\sum_{j\neq i\ 且\ R_{ij}<R_c}\eta_j\delta(\boldsymbol{r}-\boldsymbol{R_{ij}})f_c(\|\boldsymbol{R_{ij}}\|) \tag{3-6}$$

具体到深度势能分子动力学模拟中，其采用了如下方法来得到描述符：
对于任意两相对距离小于截断半径R_c的原子，其相对坐标$\boldsymbol{R_{ij}}$，相对距离R_{ij}。

$$R_{ij}=|\boldsymbol{R_{ij}}|=|\boldsymbol{R_i}-\boldsymbol{R_j}|<R_c \tag{3-7}$$

在传统三维直角坐标系下：

$$\boldsymbol{R_{ij}}=x_{ij}\boldsymbol{e_x}+y_{ij}\boldsymbol{e_y}+z_{ij}\boldsymbol{e_z} \tag{3-8}$$

将传统三维直角坐标$\{x_{ij},y_{ij},z_{ij}\}$，转化为：

$$\{s(R_{ij}),\hat{x}_{ij},\hat{y}_{ij},\hat{z}_{ij}\} \tag{3-9}$$

其中：

$$s(R_{ij})=\begin{cases}\dfrac{1}{R_{ij}},R_{ij}<R_{cs}\\ \dfrac{1}{R_{ij}}\left[\dfrac{1}{2}\cos\left(\pi\dfrac{R_{ij}-R_{cs}}{R_c-R_{cs}}\right)+\dfrac{1}{2}\right],R_{cs}<R_{ij}<R_c\\ 0,r_{ij}>r_c\end{cases} \tag{3-10}$$

$$\hat{x}_{ij} = s(R_{ij})x_{ij} \quad (3-11)$$

$$\hat{y}_{ij} = s(R_{ij})y_{ij} \quad (3-12)$$

$$\hat{z}_{ij} = s(R_{ij})z_{ij} \quad (3-13)$$

为了方便，我们这里考虑$R_{ij} < R_{cs}$的情形，将其表示为：

$$\{D_{ij}^{\alpha}\} = \left\{\frac{1}{R_{ij}}, \frac{x_{ij}}{R_{ij}}, \frac{y_{ij}}{R_{ij}}, \frac{z_{ij}}{R_{ij}}\right\} \quad (3-14)$$

当$\alpha = 0,1,2,3$时，我们同时具有半径和角度的完整信息，当$\alpha = 0$时，我们仅使用半径信息，不使用角度信息。共价键相互作用，如键的拉伸和弯曲，以及二面角力，是由输入数据中前两个相邻壳层的完整坐标信息描述的。距离较远的邻居的半径信息已足够准确地捕获了范德华力等更长距离的相互作用。因此，为了提高效率，通常先考虑完整信息找到最接近的邻居和固定的邻居球壳，然后再在截断半径内仅通过半径信息去计算其它邻居。而对于一个纯净物的体系来说，最接近的邻居一般是由材料的分子构型可以完全确定的。

3.3.2 通过神经网络计算势函数

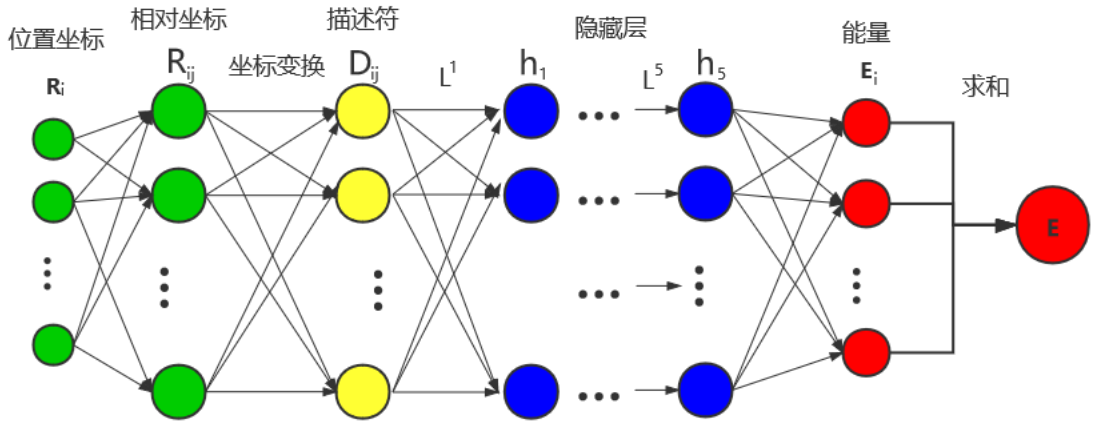

图 3-1 深度势能分子动力学神经网络示意图

在 3.3.1 节中，对于每一个中心原子$\boldsymbol{R_i}$，我们将其周边距其小于截断半径的原子坐标转化为它们距中心原子的距离向量$\boldsymbol{R_{ij}}$，即上图绿色部分，再将其转化为具有对称性的描述符D_{ij}，即上图黄色部分。之后就是将描述符$\boldsymbol{D_{ij}}$作为神经网络的输入层，经过隐藏层一系列线性和非线性的变换，得到输出层的各原子局部能量E_i，再对各原子局部能量E_i求和即可得到体系总能量E。

通过给定的神经网络初始参数求得体系总能量表达式，再对其求梯度得到受力，求出其能量和力与从头算分子动力学模拟的能量，力和位力（Virial）的差值作为神经网络的损失函数，即可不断训练此神经网络，优化其参数。实践表明，对于该损失函数中能量和力的差值，在训练刚开始时，力的差值的权重系数较大，能量差值和位力差值的权重系数较小，随着训练进行，增大能量差值和位力差值的权重系数，减小力的差值的权重系数，可以训练出更好的结果。

该神经网络为前馈型神经网络，数学表述形式为：

$$E_{s(i)} = N_{s(i)}(D_i) = L_{s(i)}^{out} \circ L_{s(i)}^{N_h} \circ L_{s(i)}^{N_{h-1}} \circ \cdots \circ L_{s(i)}^{1}(D_i) \quad (3-15)$$

其中，h为神经网络层数，$\circ$表示复合函数，$L_{s(i)}^{p}$表示神经网络中第$p-1$层到第p层的激活函数，具体来说：

$$\boldsymbol{d}_i^p = L_{s(i)}^p \boldsymbol{d}_i^{p-1} = \varphi\left(\boldsymbol{W}_{s(i)}^p \boldsymbol{d}_i^{p-1} + \boldsymbol{b}_{s(i)}^p\right) \quad (3-16)$$

其中，$\boldsymbol{W}$和$\boldsymbol{b}$为神经网络中线性变换的参数，φ为神经元之间的非线性变换函数，在深度势能分子动力学中采用了双曲正切函数 tanh。此外，在深度势能分子动力学中，给出的隐藏层数为 5，节点数分别为 240, 120, 60, 30, 10，最后一个隐藏层到输出层，只采用线性变换，不采用非线性变换，即：

$$E_{s(i)} = L_{s(i)}^{out} \boldsymbol{d}_i^{N_h} = \boldsymbol{W}_{s(i)}^{out} \boldsymbol{d}_i^{N_h} + \boldsymbol{b}_{s(i)}^{out} \quad (3-17)$$

$$E = \sum_i N_{\alpha_i}\left(D_{\alpha_i}\left(r_i, \{r_j\}_{j\in N(i)}\right)\right) \quad (3-18)$$

通过对能量求梯度可求得原子受力F，进而求得位力Ξ：

$$\begin{aligned}\boldsymbol{F_i} &= -\nabla_{R_i} E \\ &= -\sum_{j\in N(i),\alpha} \frac{\partial N_{s(i)}}{\partial D_{ij}^\alpha}\frac{\partial D_{ij}^\alpha}{\partial \boldsymbol{R}_i} - \sum_{j\neq i}\sum_{k\in N(j),\alpha} \delta_{i,a(j)} \frac{\partial N_{s(j)}}{\partial D_{jk}^\alpha}\frac{\partial D_{jk}^\alpha}{\partial \boldsymbol{R}_i} \\ &\quad -\sum_{j\neq i}\sum_{k\in N(j),\alpha} \delta_{i,b(j)} \frac{\partial N_{s(j)}}{\partial D_{jk}^\alpha}\frac{\partial D_{jk}^\alpha}{\partial \boldsymbol{R}_i} - \sum_{j\neq i}\sum_{k\in N(j)-\{a(j),b(j)\},\alpha} \delta_{i,k} \frac{\partial N_{s(j)}}{\partial D_{jk}^\alpha}\frac{\partial D_{jk}^\alpha}{\partial \boldsymbol{R}_i}\end{aligned} \quad (3-19)$$

$$\Xi = -\sum_{i\neq j} \boldsymbol{R}_{ij} \sum_\alpha \frac{\partial N_{s(i)}}{\partial D_{ij}^\alpha}\frac{\partial D_{ij}^\alpha}{\partial \boldsymbol{R}_{ij}} - \sum_{i\neq j} \delta_{j,a(i)} \boldsymbol{R}_{ij} \sum_{q,\alpha} \frac{\partial N_{s(i)}}{\partial D_{iq}^\alpha}\frac{\partial D_{iq}^\alpha}{\partial \boldsymbol{R}_{ij}} - \sum_{i\neq j} \delta_{j,b(i)} \boldsymbol{R}_{ij} \sum_{q,\alpha} \frac{\partial N_{s(i)}}{\partial D_{iq}^\alpha}\frac{\partial D_{iq}^\alpha}{\partial \boldsymbol{R}_{ij}} \quad (3-20)$$

将神经网络求得的能量，力，位力与从头算分子动力学模拟得到的结果求方均根误差ΔE，ΔF，$\Delta \Xi$ 乘上权重即可作为损失函数：

$$L(p_\varepsilon, p_f, p_\xi) = \frac{p_\varepsilon}{N}\Delta E^2 + \frac{p_f}{3N}\sum_i |\Delta \boldsymbol{F}_i|^2 + \frac{p_\xi}{9N}\|\Delta \Xi\|^2 \quad (3-21)$$

其中p_ε，p_f，p_ξ为可调的权重参数（prefactor），当位力信息缺失时，取$p_\xi = 0$。权重参数的更新机制为：

$$p(t) = p^{limit}\left[1 - \frac{r_l(t)}{r_l^0}\right] + p^{start}\left[\frac{r_l(t)}{r_l^0}\right] \quad (3-22)$$

其中r_l为学习率，学习率越大，训练速度越快，和旧模型相关性越小。随着训练进行，权重参数逐渐从开始时的p^{start}变成结束时的p^{limit}。

3.3.3 训练神经网络

我们采取 Adam（适应性矩估计，Adaptive moment estimation）算法训练神经网络参数$\boldsymbol{W}$和$\boldsymbol{b}$从而最小化损失函数L。Adam 是一种可以替代传统随机梯度下降过程的一阶优化算法，它能基于训练数据迭代地更新神经网络权重[26]。具体来说，Adam 算法在梯度下降中的优化步骤是：

1. 计算原目标函数对参数的梯度。
2. 计算梯度的一阶矩m_t，即历史梯度与当前梯度的平均。
3. 计算梯度的二阶矩v_t，即历史梯度的平方与当前梯度的平方的平均。
4. 对一阶矩m_t进行校正，减少其初始值向 0 偏置的影响，$m_t := \frac{m_t}{1-\beta_1}$，其中$\beta_1$为一阶矩估计的指数衰减率。

5. 对二阶矩v_t进行校正，减少其初始值向 0 偏置的影响，$v_t := \frac{v_t}{1-\beta_2^2}$，其中$\beta_2$为二阶矩估计的指数衰减率。

6. 更新参数，$\theta_t := \theta_{t-1} - \alpha\frac{m_t}{\sqrt{v_t}+\varepsilon}$，其中$\varepsilon$为一个极小常数。

Adam 算法同时具有适应性梯度算法（AdaGrad）和均方根传播（RMSProp）的优势，既基于一阶矩均值计算适应性参数学习率，同时还充分利用了梯度的二阶矩均值。具体来说，一方面，Adam 算法记录了梯度的一阶矩，即过往所有梯度与当前梯度的平均，使得每一次更新时，上一次更新的梯度与当前更新的梯度不会相差太大，即梯度平滑、稳定的过渡，可以适应不稳定的目标函数；另一方面，Adam 算法记录了梯度的二阶矩，即过往梯度平方与当前梯度平方的平均，这体现了环境感知能力，为不同参数产生自适应的学习速率；同时，Adam 算法的超参数$\alpha,\beta_1,\beta_2,\varepsilon$具有很好的可解释性，并且一般无需调整。

3.4 深度势能分子动力学套件（DeePMD-kit）的使用

DeepMD-kit（深度势能分子动力学套件）是用于多体势能表示和分子动力学的深度学习套件。主要由 Python 和 C++语言写成，目的是通过神经网络尽可能减少计算势能，力场和分子动力学模拟的计算量，利用深度学习解决传统多尺度建模中的高维灾难问题。DeePMD-kit 与 TensorFlow 有接口，使得训练过程高度自动化和高效，另一方面，DeePMD-kit 与高性能分子动力学和量子分子动力学的软件包，例如 LAMMPS 和 i-PI 等，也有接口，这使得 DeePMD-kit 可以学习并模拟各种系统和模型[25]。

DeePMD-kit 另一个巨大优势是它是开源的。它的源代码和大量数据集可以在官网 http://www.deepmd.org/ 上找到。除此之外，关于 DeePMD-kit 的 github 社区 DeepModeling，地址 https://github.com/deepmodeling/ ，汇聚了大量分子动力学模拟方向的研究者，不断探索开发深度势能分子动力学的新应用。在 DeePMD-kit 之后，还有的较为成功的项目包括 dpdata（将其它分子动力学软件，例如 VASP 或者 LAMMPS，的计算结果，例如 OUTCAR 文件或 lmp 文件，转换为 DeepMD-kit 能识别的数据格式），dpgen（基于同步学习的深度势能生成器，能够生成初始数据集并全自动化地完成训练），deepks-kit（用于开发基于机器学习的精确的能量和密度泛函模型）等等，这里不做过多介绍。

DeePMD-kit 的安装方法有三种，分别是：通过 conda 联网安装，有 CPU 和 GPU 两个版本，在 github 获取离线安装包，下载源代码编译安装。安装完成后系统获得 dp 命令。dp 命令下有三个子命令，分别是：dp train，dp freeze， dp test。其作用分别是：训练势函数，冻结势函数，测试训练获得势函数与文件之间的差值。

DeePMD-kit 最重要的输入参数文件格式为“json”，其中的参数包含了原子种类，截断半径，神经网络节点数，嵌入矩阵大小，权重参数的起始值和终止值，学习率，等等。这里以水的 json 文件作为示例，如图 3-2 所示。

除了参数文件 json 外，DeePMD-kit 的训练还需要 raw 文件和 npy 文件。raw 文件包含了系统（system）和框架（frame）的所有信息，系统指原子类型，坐标，盒子大小及温度的所有信息，框架指所有系统信息在不同时间点的状态。raw 文件可以由从头算分子动力学或者经验力场模拟软件的结果，例如 OUTCAR，或者 lammps，经过 dpdata 处理转化而来。npy 文件则是一个包含力，能量等数据的二进制文件，便于 TensorFlow 读取。

water.json

```
{
    "with_distrib": false,
    "_comment": " model parameters",
    "model":{
    "type_map":     ["O", "H"],
    "descriptor": {
        "type":      "loc_frame",
        "sel_a":        [16, 32],
        "sel_r":        [30, 60],
        "rcut":     6.00,
        "axis_rule":    [0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0],
        "_comment": " default rule: []",
        "_comment": " user defined rule: for each type provides two axes, ",
        "_comment": "                    for each axis: (a_or_r, type, idx)",
        "_comment": "                    if type < 0, exclude type -(type+1)",
        "_comment": "                    for water (O:0, H:1) it can be",
        "_comment": "                    [0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0]",
        "_comment": " that's all"
    },
    "fitting_net": {
        "neuron":       [240, 120, 60, 30, 10],
        "resnet_dt":    true,
        "seed":     1,
        "_comment":     " that's all"
    },
    "_comment": " that's all"
    },

    "learning_rate" :{
    "type":     "exp",
    "start_lr": 0.001,
    "decay_steps":  5000,
    "decay_rate":   0.95,
    "_comment": "that's all"
    },

    "loss" : {
```

water - DD.json

```
    "start_pref_e": 0.02,
    "limit_pref_e": 8,
    "start_pref_f": 1000,
    "limit_pref_f": 1,
    "start_pref_v": 0,
    "limit_pref_v": 0,
    "_comment": "that's all"
    },

    "_comment": " traing controls",
    "training": {
    "systems":  ["../data/"],
    "set_prefix":   "set",
    "stop_batch":   1000000,
    "batch_size":   [4],

    "seed":     1,

    "_comment": " display and restart",
    "_comment": " frequencies counted in batch",
    "disp_file":    "lcurve.out",
    "disp_freq":    100,
    "numb_test":    10,
    "save_freq":    1000,
    "save_ckpt":    "model.ckpt",
    "load_ckpt":    "model.ckpt",
    "disp_training":true,
    "time_training":true,
    "_comment": "that's all"
    },

    "_comment":     "that's all"
}
```

图 3-2 水的参数文件 “water.json”

3.5 利用深度势能分子动力学模拟抗冻蛋白抑制结晶过程

尽管深度势能分子动力学具有广泛的应用范围，近两年来已经被应用在一阶相变[27]，拉曼光谱[28]，硅的燃烧模拟[29]，固体电解质[30]等等研究领域中，但将其应用在抗冻蛋白抑制结晶的模拟中还是存在一些需要克服的难点，具体如下：

第一，抗冻蛋白是一个至少由四种元素和上百个原子构成的生物大分子，要模拟它抑制结晶的过程还需要在体系中加入数以百计的水分子，对于这个复杂度混合物，它的参数文件 json 需要作何改变才能使训练达到较好的效果？

第二，更为困难的是，这个由抗冻蛋白和水组成的体系，其 raw 文件和 npy 文件该如何书写或生成？假设使用 dpdata 从传统分子动力学模拟的结果转化，那么是用从头算分子动力学模拟的结果还是经验力场模拟的结果？

假设我们像引用文献[19][20]里面一样采用 CHARMM 内含的经验力场计算，或是即使采用具有与 DeePMD-kit 有接口的 LAMMPS 中的经验力场计算，如果直接将经验力场算出的结果代入 DPMD 的神经网络，在经验力场给出的训练集本身就不完全可靠的情况下，其外生性带来的误差是难以通过神经网络训练而消除的，即泛化后的结果难以比训练数据本身更好，这样通过深度势能分子动力学模拟得到的结果并不会比 CHARMM 或是 LAMMPS 的更好。

假设我们使用从头算分子动力学模拟的数据来训练，这将足够精确，但用从头算分子动力学模拟如此复杂的体系本身就是一个难以完成的任务。首先，其输入文件，类似 POSCAR，INCAR 等该如何得到？更关键的是，这个体系的从头算分子动力学计算量无比庞大，如何通过深度势能分子动力学的同步训练减少其计算量？在一个超大规模的 AIMD 精度的 DPMD 模拟实验中，作者提到：尽管从原理上讲通过机器学习训练的分子动力学模拟使得以经验力场效率实现从头算分子动力学精度成为可能，但这在实践中还没有实现[31]。

一个潜在的可能有效的思路是，通过 AIMD 先模拟一个丙氨酸或是苏氨酸在水中的情形，如果计算资源足够可以计算一个丙氨酸-丙氨酸-苏氨酸链，这也是抗冻蛋白最基本的组成部分。再用这个计算结果训练神经网络。对于石墨烯或者硅等周期性较强的固体材料，是可以通过对周期性边界条件的控制实现从较小的 AIMD 结果训练较大的 DPMD 体系。但由于单个的氨基酸不具有热滞活性和重结晶抑制性，而抗冻蛋白具有的复杂的空间结构和可能的不同

结合位点，这些都是无法通过训练一个氨基酸的数据得到的。可能的解决这一问题的办法是，利用强化学习（Reinforcement Learning）的思路，将训练的前馈型神经网络改进为深度 Q 学习网络(Deep Q-learning Network,DQN)或是循环型神经网络(Recurrent Neural Network，RNN)，并将与抗冻活性成正相关的物理量，比如冰结合表面的氢键数量或寿命等，作为网络的奖赏因子，使得通过一个小基团训练出来的数据越来越向可能具有抗冻活性的方向发展。据说，DPMD 的开发者 Han Wang 等人近期正在致力于将强化学习应用到 DPMD 中，期待他们的成果能够在抗冻蛋白上具有更好的应用。

总的来说，利用深度势能分子动力学模拟抗冻蛋白抑制结晶过程目前还是存在一些难点，很遗憾也很惭愧在这次的工作中这些问题并没有得到解决。希望其他有志于此领域的研究者们能够尽快克服这些难点，早日在有关大分子蛋白质的实践中同时实现从头算分子动力学的精度与经验力场的效率，进而更精确地模拟抗冻蛋白吸附并抑制结晶的过程，更深入地解释其抑制结晶的机制。

参考文献

[1] DeVries AL. 1971. Glycoproteins as biological antifreeze agents in Antarctic fishes. Science. 172:1152－1155.
[2] Duman JG, Patterson JL. 1978. Role of thermal-hysteresis-proteins in low-temperature tolerance of insects and spiders. Cryobiology 15:683－84
[3] Clarke CJ, Buckley SL, Lindner N. 2002. Ice structuring proteins—a new name for antifreeze proteins. Cryo Letters 23:89－92
[4] Olijve LL, Meiser K, DeVries AL, Duman J, Guo S, et al. 2016. Blocking rapid ice crystal growth through nonbasal plane adsorption of antifreeze proteins. PNAS113:3740－45
[5] Davies PL. 2014. Ice-binding proteins: a remarkable diversity of structures for stopping and starting ice growth. Trends Biochem. Sci.39:548－5
[6] Harding MM, Anderberg PI, Haymet AD. 2003. J.Eur. J.Biochem. 270(7), 1381-1392.
[7] Lin Y., Duman JG, DeVries AL, Biochem. 1972. Biophys. Res.Commun. 46(1), 87-92
[8] Raymond JA, DeVries AL. 1977. Adsorption inhibition as a mechanism of freezing resistance in polar fishes. PNAS 74:2589－93
[9] Liu ZH, Muldrew K, Wan RG, Elliott JAW. 2003. Measurement of freezing point depression of water in glass capillaries and the associated ice front shape. Phys.Rev.E67:061602
[10] Knight CA, DeVries AL. 1989. Melting inhibition and superheating of ice by an antifreeze glycopeptide. Science 245:505－7
[11] Wen DY, Laursen RA. 1992. Structure-function-relationships in an antifreeze polypeptide: the role of neutral, polar amino-acids. J. Biol. Chem.267:14102－8
[12] Scotter AJ, Marshall CB, Graham LA, Gilbert JA, Garnham CP, Davies PL. 2006. The basis for hyper-activity of antifreeze proteins. Cryobiology 53:229－3
[13] Takamichi M, Nishimiya Y, Miura A, Tsuda S. 2007. Effect of annealing time of an ice crystal on the activity of type III antifreeze protein. FEBS J.274:6469－76
[14] Devries AL, Lin Y. 1977. Structure of a peptide antifreeze and mechanism of adsorption to ice. Biochim Biophys Acta, 495, 2, p. 388-392 5 p.
[15] Kristiansen E, Zachariassen KE. 2005. The mechanism by which fish antifreeze proteins cause thermal hysteresis. Cryobiology 51:262－80
[16] Knight CA, DeVries AL. 2009. Ice growth in supercooled solutions of a biological “antifreeze,” AFGP1－5: an explanation in terms of adsorption rate for the concentration dependence of the freezing point. Phys. Chem. Chem. Phys.11:5749－61
[17] Garnham CP, Campbell RL, Davies PL. 2011. Anchored clathrate waters bind antifreeze proteins to ice. PNAS108:7363－67
[18] Sun TJ, Lin FH, Campbell RL, Allingham JS, Davies PL. 2014. An antifreeze

protein folds with an interior network of more than 400 semi-clathrate waters. Science 343:795－98
[19] Kuiper MJ, Morton CJ, Abraham SE, Gray-WealeA. 2015. The biological function of an insect antifreeze protein simulated by molecular dynamics. eLife4: e05142
[20] Mochizuki K, Molinero V. 2018. Antifreeze Glycoproteins Bind Reversibly to Ice via Hydrophobic Groups. J Am Chem Soc. 2018 Apr 11;140(14):4803-4811. doi: 10.1021/jacs.7b13630. Epub 2018 Feb 14. PMID: 29392937.
[21] J.Behler, M.Parrinello. 2007. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. Phys. Rev. Lett.98(14)(2007)146401.
[22] M.Rupp, A.Tkatchenko, K.-R.Müller, O.A.VonLilienfeld. 2012. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. Phys. Rev. Lett. 108(5)(2012)058301.
[23] Kozuch Daniel, Stillinger Frank, Debenedetti Pablo. 2018. Combined molecular dynamics and neural network method for predicting protein antifreeze activity. Proceedings of the National Academy of Sciences. 115. 201814945. 10.1073/pnas.1814945115.
[24] Zhang L, Han J, Wang H, Car R, E W. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. Phys Rev Lett. 2018 Apr 6;120(14):143001. doi: 10.1103/PhysRevLett.120.143001. PMID: 29694129.
[25] Han Wang, Linfeng Zhang, Jiequn Han, Weinan E, DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics, Computer Physics Communications, Volume 228, 2018, Pages 178-184, ISSN 0010-4655.
[26] Kingma Diederik, Ba Jimmy. 2014. Adam: A Method for Stochastic Optimization. International Conference on Learning Representations.
[27] Luigi Bonati, Michele Parrinello. 2018. Silicon liquid structure and crystal nucleation from ab initio deep metadynamics. Physical review letters,121(26):265701.
[28] Grace M. Sommers, Marcos F. Calegari Andrade, Linfeng Zhang, HanWang, Roberto Car. 2020. Raman spectrum and polarizability of liquid water from deep neural networks. Phys. Chem. Chem. Phys., 22:10592－10602.
[29] Jinzhe Zeng, Liqun Cao, Mingyuan Xu, Tong Zhu, John ZH Zhang. 2019. Neural network based in silico simulation of combustion reactions. arXiv preprint arXiv:1911.12252.
[30] Aris Marcolongo, Tobias Binninger, Federico Zipoli, Teodoro Laino. 2019. Simulating diffusion properties of solid-state electrolytes via a neural network potential: Performance and training scheme. ChemSystemsChem.
[31] Denghui Lu, Han Wang, Mohan Chen, Lin Lin, Roberto Car, Weinan E, Weile Jia, Linfeng Zhang. 2021. 86 PFLOPS Deep Potential Molecular Dynamics simulation of 100 million atoms with ab initio accuracy. Computer Physics. Communications. Volume 259, 107624, ISSN 0010-4655.

致谢

感谢物质学院孙兆茹老师及其课题组的学长学姐们，指导我完成毕业设计并提高我的狼人杀技巧。感谢信息学院王浩老师及其课题组的学长学姐们，在我什么都不懂时带我感受科研并让我的减重计划从来没有成功过。感谢上海科技大学提供的高质量多元化的课程与全年免费的健身房。感谢我的魔方与扫雷游戏带给我无数次刺激与愤怒。感谢我的父母养育我长大并每月准时给我转钱。最后要感谢我的女朋友，她在我 21 年的生命中从未出现过，使我避免了将有限的时间和精力浪费在无意义的谈情说爱上。

本科四年的物理系生涯就这么告一段落了，应该说，之后也不会再从事物理相关的科研工作了，不出意外的话之后应该会读一个 data science 的 master，再之后呢？说不准。我花了四年时间，学了一堆杂七杂八也许有用也许没用的东西，排除了一个我学不懂的方向，希望我还有更多可能，学更多更有意思的东西。