专题: 生物分子模拟中的机器学习 • 封面文章

使用中间层受监督的自编码器探索蛋白质的构象空间*

陈光临 张志勇†

(中国科学技术大学物理系, 合肥 230026)

(2023 年 6 月 28 日收到; 2023 年 7 月 29 日收到修改稿)

蛋白质的功能往往与其结构和动态变化密切相关. 分子动力学模拟是研究蛋白质结构变化的有效方法, 然而使用分子动力学模拟对蛋白质的构象空间进行采样需要花费很长的时间. 近年来的一些研究表明, 使用简单的机器学习模型——自编码器及其改进型, 可以在有限采样的情况下, 快速完成对蛋白质构象空间的探索. 该模型通过训练神经网络, 完成对隐变量的提取, 同时根据其产生构象, 但是由于提取出的隐变量没有直观的含义, 探索构象空间的方向会受到影响. 本工作通过引入反应坐标 (如质心距离等), 建立了一个中间层受监督的自编码器模型, 以解决上述问题. 该模型应用于噬菌体 T4 溶菌酶和腺苷酸激酶两个蛋白质分子, 结果表明, 仅使用短时间分子动力学模拟作为训练数据, 就可以探索到这两种蛋白分子的多种典型构象. 有监督 (合理的反应坐标或者实验数据等) 的自编码器模型有望成为探索蛋白质构象空间的有效工具.

关键词: 蛋白质构象空间, 分子动力学模拟, 机器学习, 自编码器

PACS: 87.15.ap, 87.15.hp **DOI:** 10.7498/aps.72.20231060

1 引 言

蛋白质的功能与其结构和动态构象变化密切相关[1]. 为了获得蛋白质分子的结构, 人们开发出了各种实验和预测技术. X 射线晶体衍射[2] 和冷冻电镜技术[3] 可以解析高分辨率的蛋白质分子结构, 而核磁共振[4] 可以提供分子中的原子距离等信息. 此外, 小角 X 射线散射[5]、化学交联[6] 和荧光共振能量转移[7] 等技术可以从不同的角度给出蛋白质分子的各种结构信息. 基于人工智能的结构预测方法, 如 AlphaFold2[8] 和 RoseTTAFold[9], 可以直接根据氨基酸序列预测蛋白质的结构. 这些方法在获取蛋白质静态结构时十分有效, 但是不易得到蛋白质的动态变化信息.

计算模拟方法, 例如分子动力学 (molecular dynamics, MD) 模拟, 是研究蛋白质分子动态变化的重要工具[10]. MD 方法用半经验的能量函数来描述原子间的相互作用, 在经典力学的框架下对蛋白质分子进行模拟. 从一个已知的分子结构出发, 通过迭代求解运动方程, 得到分子动态变化的过程. 为了确保结果的可靠性, 通常要求对整个构象空间充分采样. 但由于分子模拟的结果服从玻尔兹曼统计, 在生理条件下, 对高能构象的采样十分困难, 这一问题通常需要引入增强采样等方法来解决[11]. 模拟的另一个问题来自分子力场, 它是对分子间相互作用的一种近似描述, 因而必然存在一定的误差. 力场选择不合适可能会导致模拟结果表现出与实际情况不同的倾向[12], 即使经过大量计算后达到了充分采样的要求, 也无法正确描述生物大分子

* 国家重点研发计划 (批准号: 2021YFA1301504)、国家自然科学基金 (批准号: 91953101) 和中国科学院战略性先导科技专项 (B 类)(批准号: XDB37040202) 资助的课题.

† 通信作者. E-mail: zzyzhang@ustc.edu.cn

© 2023 中国物理学会 Chinese Physical Society

的动态变化. 这种情况下, 可以先尽可能多地产生不同的构象, 再验证其合理性.

近年来, 机器学习方法的快速发展为解决分子模拟中的采样和力场问题提供了新思路[13,14]. 自编码器是一种生成神经网络, 最初用于计算机图形领域[15], 目前也应用于探索蛋白质分子的构象空间[16]. 自编码器由编码器和解码器组成, 高维的蛋白质结构信息经过编码器压缩得到低维空间的隐变量, 再经过解码器重构出蛋白质结构, 同时要求重构的结构与输入的结构尽可能一致. 训练完成后, 只需要向解码器输入随机数据, 就可以构建出不同的蛋白质构象. 由于自编码器在训练过程中只要求数据成功重构, 中间层的隐变量没有明确的含义, 而构象生成是从中间层的数据开始的, 因此探索构象空间的方向也是不确定的, 有时可以找到各种不同的构象, 有时只能得到不感兴趣或不合理的构象. 为了解决上述问题, 一种常用的方案是对中间层的结果进行一些限制.

本研究设计了一个有监督的自编码器模型. 将一些反应坐标引入到自编码器中, 要求其在重构蛋白质结构的同时, 中间层的数据要与给定的反应坐标接近, 从而使构象空间的探索在给定的方向上进行. 将该模型运用到两个多结构域蛋白, 噬菌体 T4 溶菌酶和腺苷酸激酶, 探索得到的蛋白质构象空间覆盖了目前已知的实验结构. 通过引入合理的反应坐标和实验数据, 建立有监督的自编码器模型, 有望成为探索蛋白质构象空间的有效工具.

2 方 法

2.1 中间层受监督的自编码器模型

为了实现在给定方向的构象空间探索, 使用 Pytorch2.0 设计了一个中间层受监督的自编码器 (图 1). 该模型的整体结构与普通的自编码器相似, 由编码器和解码器组成. 其中编码器是一个多层的全连接神经网络, 在输入层之后每一层的维数分别是 2048, 512, 128, 32, 8, 2, 解码器也是多层全连接网络, 其结构与编码器对称, 每一层的维数依次是 2, 8, 32, 128, 512, 2048, 输出层的维数与编码器输入层相同. 除了最后一层外, 编码器和解码器的每一层都使用了 ReLU 作为激活函数, 而最后一层则使用 Sigmoid 激活函数, 以控制输出结果的范围. 这一模型的参数量很少, 对计算资源的要求较低.

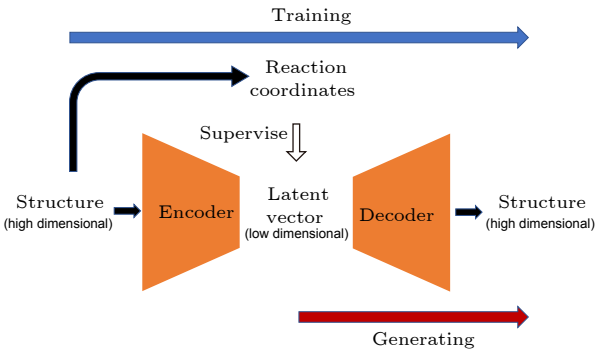

图 1 中间层受监督的自编码器示意图

Fig. 1. Schematic of supervised-AE.

不同于无监督的自编码器, 将监督学习引入自编码器的中间层, 训练时使用的损失函数形式如下:

$$L = \mathcal{L}_{\text{output}} + \omega \mathcal{L}_{\text{midden}}, \tag{1}$$

其中 $\mathcal{L}_{\text{output}}$ 为输出层的损失函数, 用来描述重构后的结构与输入结构之间的差距; $\mathcal{L}_{\text{midden}}$ 为中间层的损失函数, 描述中间层数据与输入结构对应的反应坐标之间的差距. 只使用反应坐标往往不能准确地描述和重构整个分子结构, 只能反映结构的某些特征, 因此模型需要在正确提取反应坐标和成功重构分子结构之间找到平衡. 引入了权重因子 ω 来调整两者对损失函数的贡献, ω 较大时, 中间层对损失函数的贡献更大, 模型会倾向得到给定的反应坐标, 而重构分子结构的效果会变差, 反之, ω 较小时, 模型可以完成分子结构的重构, 但中间层的数值不一定接近给定的反应坐标. 本文中, 该因子的值设定为 100.

2.2 数据获取

训练模型的数据来自 MD 模拟. 模拟的体系分别是噬菌体 T4 溶菌酶 (T4 lysozyme, T4L) 和大肠杆菌腺苷酸激酶 (adenylate kinase, AdK). T4L 及其突变体在 PDB 数据库中有大量晶体结构, 其结构变化主要体现在 N 端结构域和 C 端结构域之间口袋的打开和关闭 (图 2(a)). AdK 可以分为 CORE, LID 以及 AMPbd 三个结构域, 分别在 CORE 和 LID, 以及 CORE 和 AMPbd 之间形成两个口袋. 在酶的催化过程中, 口袋的打开和关闭十分重要 (图 2(b)). 这两个蛋白分子的动态构象变化已经研究得比较充分, 适合用来验证我们的模型.

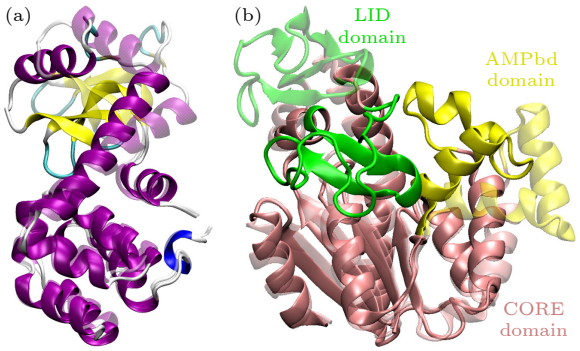

图 2 本研究中使用的两种蛋白质分子的不同结构 (a) T4L 的闭合 (不透明) 和打开 (透明) 结构, 紫色为 α 螺旋, 黄色为 β 折叠; (b) AdK 的闭合 (不透明) 和打开 (透明) 结构, 不同颜色表示不同的结构域

Fig. 2. Different structures of the two proteins in the work. (a) The close (opaque) and open (transparent) state of T4L. α-helix is colored in purple and β-sheet is colored in yellow. (b) The close (opaque) and open (transparent) state of AdK. Different domains are colored in different colors.

根据蛋白质分子的结构变化特征, 计算相应的反应坐标作为监督引入到自编码器的中间层. 从 T4L 及其突变体的晶体结构中选取能够反映其构象变化的 41 个结构, 消除它们之间的平动和转动后, 使用 C_α 原子的坐标进行主成分分析. 特征值最大的 2 个主成分分别对应 T4L 的开闭和扭转运动, 其占比分别为 86% 和 6%. 因此使用这 2 个主成分作为反应坐标, 可以较好地描述 T4L 分子的运动 [17]. AdK 的结构变化主要表现为结构域的相对运动, 因此可以选取 CORE-LID 和 CORE-AMPbd 结构域的质心距离作为反应坐标 [18].

分子动力学模拟使用 GROMACS-2023 版本进行 [19]. 从 PDB 数据库中分别选取 T4L 的打开 (PDB 编号 2LZM[20]) 和关闭 (PDB 编号 178L[21]) 结构, 以及 AdK 的打开 (PDB 编号 1AKE[22]) 和关闭 (PDB 编号 4AKE[23]) 结构作为模拟的初始构象. 为了验证模型是否受分子力场的影响, 每一组模拟都分别使用了 AMBER99SB 力场/OPC 水模型的组合 [24,25] 以及 CHARMM36m 力场/TIP3P 水模型的组合 [26]. 将分子放入正十二面体的周期性盒子中, 同一分子的不同体系使用同样大小的盒子, 以避免盒子尺寸对模拟结果的影响. 向体系中填充水分子, 并加入离子直到电荷平衡. 先后用 2000 步最速下降法和 1000 步共轭梯度法进行能量最小化, 然后在 NPT 系综下进行 100 ps 的位置约束 MD 模拟, 以平衡系统的温度和压强, 随后进行 NPT 模拟以获取训练模型的数据. AdK 在没有结合配体时无法维持关闭状态, 因此在模拟中额外加入了结构域距离的位置限制. 所有模拟的步长均为 2 fs, 使用 LINCS 算法约束氢原子参与的化学键, 分别用 V-rescale[27] 和 C-rescale 算法控制系统的温度和压强, 非键相互作用中静电相互作用通过 PME[28] 算法计算, 范德瓦耳斯力则做截断处理, 截断距离为 1 nm.

由于不要求充分采样, 每组用于产生训练数据的模拟仅进行 100 ns, 每 10 ps 保存一个结构, 共保存 10000 个. 消除不同结构之间的平动和转动变化后, 提取主链部分的原子, 即 N, C_α, C, O 的笛卡尔坐标作为模型的输入, 同时计算出每个结构的反应坐标作为标签. 在开始训练之前, 还需要对数据进行归一化处理, 数据的每一个维度都分别被放缩到 0.2 与 0.8 之间, 这一区间 Sigmoid 函数的斜率较大, 有利于模型训练更快达到收敛.

2.3 利用有监督的自编码器探索蛋白质构象空间

将模拟轨迹整理为数据集, 从中随机取出 80% 作为训练集, 剩余的 20% 作为测试集. 以平方误差作为损失函数, 用 Adam 优化器 [29] 进行训练, 遍历训练集 500 次, 初始学习率为 1×10^{-4}, 并随着遍历次数以 1×10^{-8} 的速率减小. 完成训练后, 在 $[0.05, 0.95]\times[0.05, 0.95]$ 的范围内均匀选取 10000 个点作为自编码器中间层隐空间的数据点, 将这些点输入解码器构建出对应的蛋白质分子主链结构. 模型训练和数据生成的相关运算在 RTX 3090Ti 上运行.

由于生成的结构并不总是合理的, 通过两种判据对其进行筛选. 其一是蛋白质的主链二面角取值需要满足一定的规律, 这一规律通常用 Ramachandran 图来描述, 将大量已知蛋白质结构的 Ramachandran 图的统计结果 [30] 作为参考, 与模型生成的蛋白质结构的 Ramachandran 图进行比较, 若 90% 以上处于合理区间, 则认为该结构的主链二面角分布是合理的. 其二是不同原子之间不能存在空间冲突, 使用分子模拟工具 OpenMM[31] 对分子结构进行一小段能量最小化, 如果最终原子间的力比较小, 就可以认为该分子不存在空间冲突. 考虑到这一步需要频繁进行, 与其他分子模拟工具相比, 使用直接运行在 Python 中的 OpenMM 可以节省大量用于初始化模拟引擎的时间. 由于模型仅

产生主链部分的原子坐标, 并非完整的分子, 用 ParmEd 工具[32] 将力场参数中非主链的部分删去, 同时将所有原子的电荷设置为 0, 在能量最小化时仅保留化学键和范德瓦耳斯力. 能量最小化不仅可以筛选掉明显不合理的结构, 还可以对结构中的一些键长键角的错误进行修正.

模拟得到的构象空间分布十分有限, 在此基础上进行构象空间探索也因此受到限制. 为了进一步扩大构象空间探索的范围, 将模型生成的结构与原有数据集的一半合并成新的数据集, 并重复进行模型训练和构象空间探索. 随着探索范围逐渐扩大, 模型生成的不合理结构逐渐增加, 构象空间的探索效率也随之下降, 因此只重复上述流程 3 次.

3 结果与讨论

3.1 T4L 构象空间探索结果

以 T4L 的模拟轨迹作为训练集, 进行训练以及构象空间探索, 整个流程耗时仅 20 min. 探索结果如图 3 所示, 由于使用不同力场得到的模拟轨迹不同, 构象空间探索的区域也有所不同, 整体上看使用 AMBER99SB 力场/OPC 水模型的探索范围更大. 不过使用两种力场得到的探索范围都可以覆盖包括所有参考晶体结构在内的训练集附近的区域, 例如可以找到与 PDB 编号为 173L 晶体结构[21] 十分相似的构象 (图 4(a)), RMSD 为 0.7 Å. 此外, 探索结果中还可以看到大幅度的构象变化, 例如闭合状态与打开状态的不同 (图 4(b)), 以及两个结构域的相对转动角度不同 (图 4(c)).

虽然模型生成的结构都通过了二面角分布的检验, 以及键长键角和空间冲突的修正, 但依然存在一些不合理的情况, 如生成的结构中二级结构含量显著低于晶体结构和模拟轨迹中二级结构的含量. 为了验证模型产生结构的合理性, 我们使用 kmeans 算法, 根据反应坐标将探索结果分为 50 组, 取每一组最靠近中心的构象作为代表, 用 tleap 补全侧链, 然后进行 100 ns 约束 C_α 原子的 MD 模拟, 从而在不改变反应坐标的情况下修复二级结构. 除少数情况由于侧链存在空间冲突而失败外, 大部分代表构象的二级结构得到修复 (图 5(a) 和图 5(b)), DSSP[33] 计算表明修复后二级结构含量基本可以接近模拟轨迹的水平 (图 5(c)). 还计算了每个代表构象与同组各构象的主链 RMSD, 所有 RMSD 数值都小于 2 Å (图 5(a) 和图 5(b)), 这说明二级结构的缺失只是由一些局部的偏差

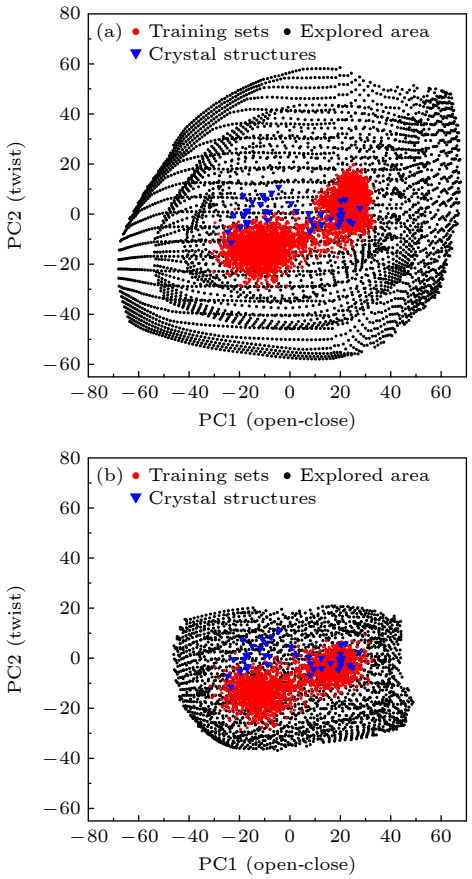

图 3 T4L 的构象空间探索结果 (a) 使用 AMBER99SB 力场/OPC 水模型; (b) 使用 CHARMM36m 力场/TIP3P 水模型

Fig. 3. Results of conformational space exploration of T4L: (a) With AMBER99SB/OPC; (b) with CHARMM36m/TIP3P.

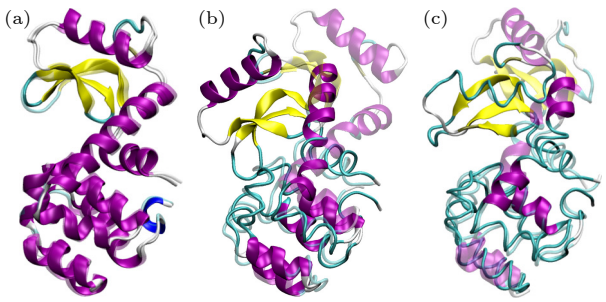

图 4 探索到的不同 T4L 构象 (a) PDB 编号 173L 的晶体结构 (不透明) 与探索到的相似结构 (透明); (b) 开合程度不同的两个构象; (c) 扭动情况不同的两个构象; 紫色为 α 螺旋, 黄色为 β 折叠

Fig. 4. Different T4L conformations explored: (a) PDB:173L (opaque) and a similar structure explored; (b) two conformations with different degrees of opening and closing; (c) two conformations with different degrees of twisting. α-helix is colored in purple and β-sheet is colored in yellow.

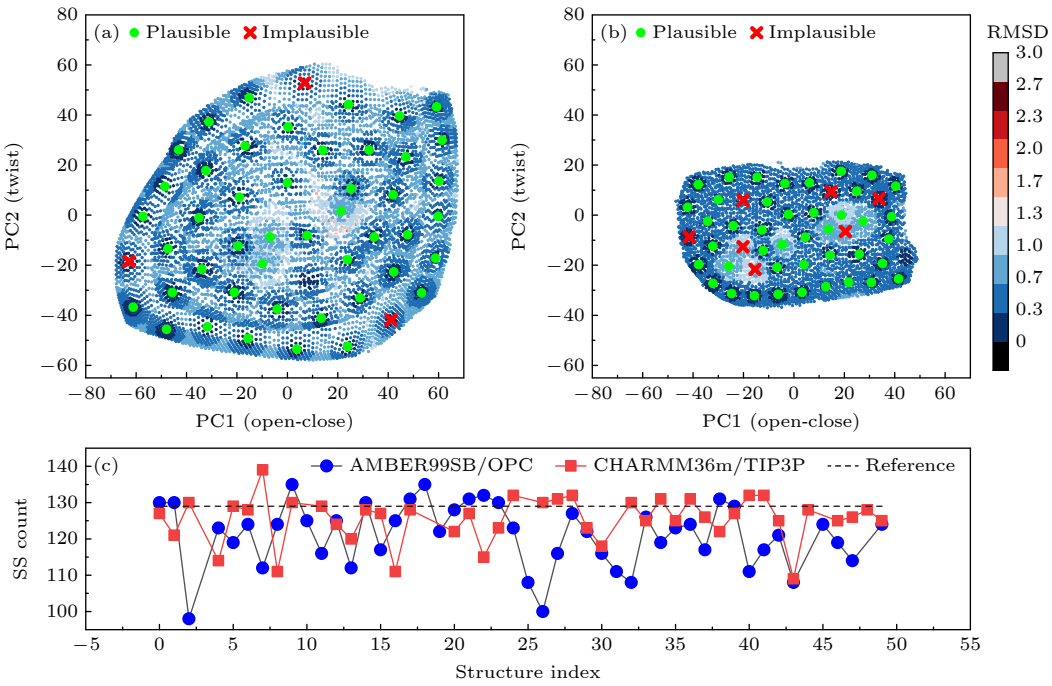

图 5 T4L 构象探索结果的合理性检验 (a) 使用 AMBER99SB 力场/OPC 水模型; (b) 使用 CHARMM36m 力场/TIP3P 水模型; (c) 修复后各代表构象的二级结构含量, 参考值为模拟轨迹的平均值

Fig. 5. Plausibility check of T4L conformational exploration results: (a) With AMBER99SB/OPC; (b) with CHARMM36m/TIP3P; (c) secondary structure counts of each representative conformation after fixing, the reference is the average value of the simulated trajectory.

导致的, 模型生成的大多数结构都可以通过简单修正得到合理的结果, 而侧链可能存在空间冲突的情况则需要进一步改进模型来解决.

在上述流程中, 闭合与打开两段模拟轨迹都被用于模型的训练. 还测试了仅使用打开状态的模拟轨迹训练的情况 (图 6), 虽然探索区域由于训练集减少而缩小, 但是仍然可以覆盖包括闭合状态在内的各个晶体结构.

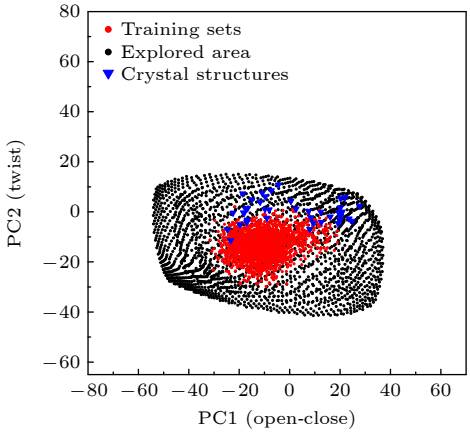

图 6 仅从打开状态出发的 T4L 构象探索结果

Fig. 6. Results of T4L conformational exploration from the open state only.

3.2 AdK 构象空间探索结果

以 AdK 的模拟轨迹作为训练集, 进行训练以及构象空间探索. 结果如图 7 所示, 除了训练集中包含的完全关闭和完全打开状态外, 还可以从中找到 LID 结构域单独打开 (图 8(a)) 和 AMPbd 结构域单独打开的结构 (图 8(b)).

对 AdK 构象探索结果的合理性进行了检验, 结果如图 9 所示. 在使用 CHARMM36m 力场/TIP3P 水模型时, 修复后二级结构含量与模拟轨迹相当, 而使用 AMBER99SB 力场/OPC 水模型时, 虽然也能修复到较高的水平, 但与前者相比显著偏低. 这表明与 CHARMM36m 相比, AMBER99SB 力场/OPC 水模型的组合使蛋白质结构更加容易发生变化, 探索构象空间的范围更大, 同时二级结构也会有一定的破坏, 更适用于柔性较强的蛋白质分子.

值得注意的是, 大部分构象与其所在组的中心构象之间的 RMSD 较小, 除少数不合理构象外, 大部分 RMSD 较大的构象都在模拟产生的训练集中. 这意味着模型产生的构象仅包含反应坐标相关的信息, 而在与反应坐标正交的自由度上没有表现

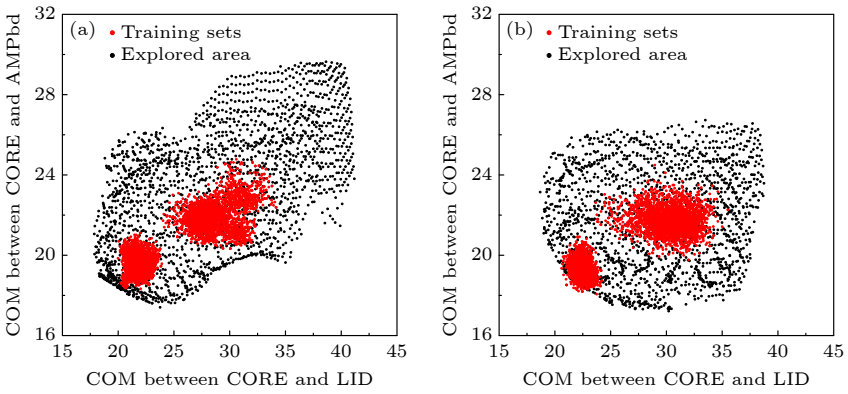

图 7 AdK 的构象空间探索结果 (a) 使用 AMBER99SB力场/OPC 水模型; (b) 使用 CHARMM36m 力场/TIP3P 水模型

Fig. 7. Results of conformational space exploration of AdK: (a) With AMBER99SB/OPC; (b) with CHARMM36m/TIP3P.

出差异. 这是由自编码器自身的性质决定的, 对于相同的输入总是会给出相同的输出, 而实际上如模拟轨迹反映的一样, 相同的反应坐标下, 构象仍应该有一定的变化空间, 这些空间是自编码器无法探索的. 因此, 反应坐标的选取对该模型的效果至关重要. 若要解决这一问题, 可以将自编码器换成变分自编码器, 学习构象系综而非单个分子的特征, 从而体现相同反应坐标下的差异.

以上结果是使用常规的自编码器难以获得的. 将引入反应坐标监督的自编码器换成无监督的自

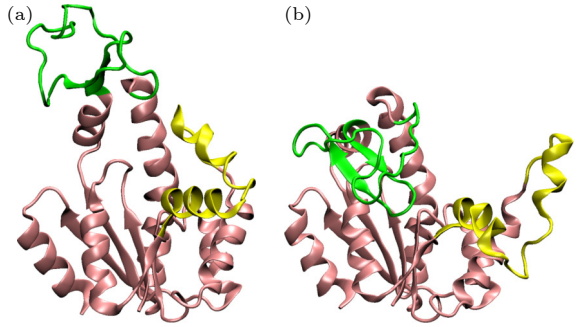

图 8 探索到的不同 AdK 构象

Fig. 8. Different AdK conformations explored.

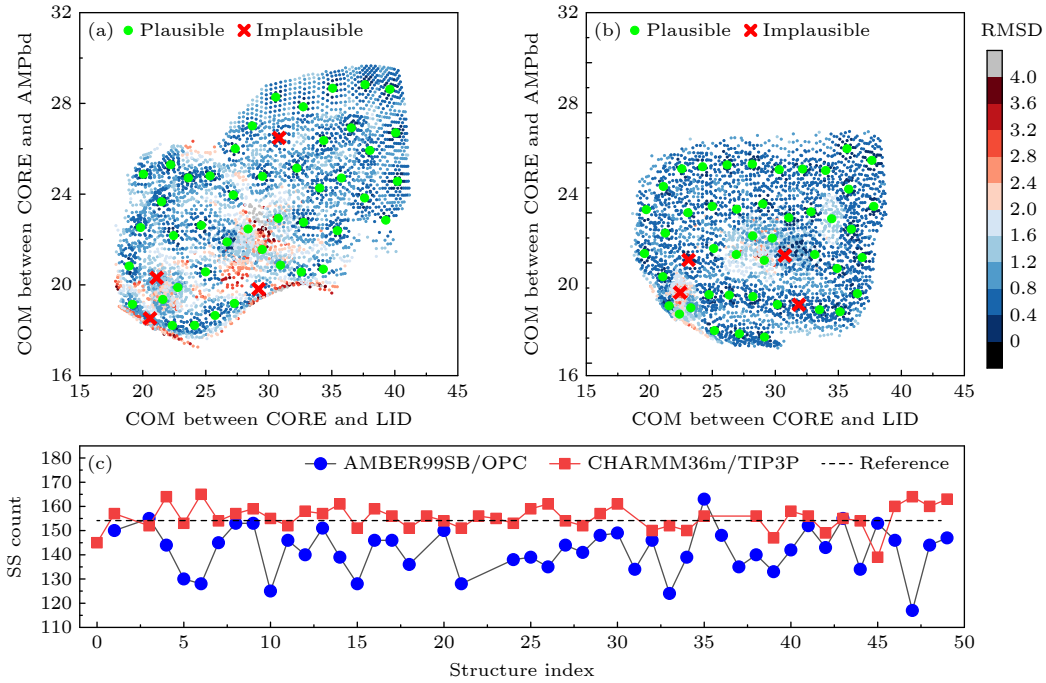

图 9 AdK 构象探索结果的合理性检验 (a) 使用 AMBER99SB 力场/OPC 水模型; (b) 使用 CHARMM36m 力场/TIP3P 水模型; (c) 修复后各代表构象的二级结构含量, 参考值为模拟轨迹的平均值

Fig. 9. Plausibility check of AdK conformational exploration results: (a) With AMBER99SB/OPC; (b) with CHARMM36m/TIP3P; (c) secondary structure counts of each representative conformation after fixing, the reference is the average value of the simulated trajectory.

编码器, 对 AdK 的构象空间进行探索, 结果如图 10 所示. 自编码器需要从训练集中学习反应坐标, 这在采样不足的情况下非常困难. 通常情况下, 自编码器只能提取两组轨迹的差异, 并完成对两种状态之间的构象空间探索, 但是无法探索其他区域, 例如图 8 所示的单个结构域打开的构象. 引入反应坐标作为监督的改进, 使得自编码器不再需要提取反应坐标, 从而可以在采样不足的情况下工作.

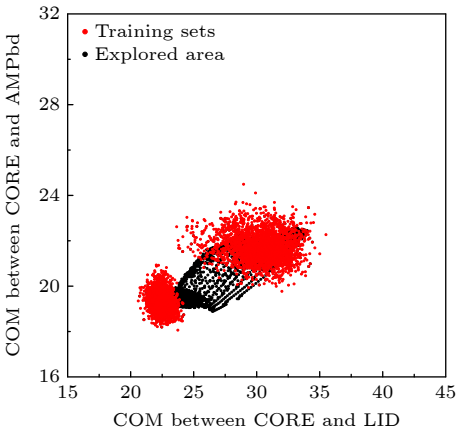

图 10 使用普通自编码器探索 AdK 的构象空间

Fig. 10. Exploring the conformational space of AdK with a common self-encoder.

4 结 论

本文对使用自编码器探索蛋白质构象空间的方法进行了改进, 将监督学习引入自编码器的中间层, 并使用改进后的方法对 T4L 和 AdK 的构象空间进行探索, 达到了预期的效果. 结果表明这一改进使该方法可以在有限采样的情况下, 仅使用很少的计算资源, 就可以大范围探索蛋白质的构象空间.

虽然模型只能生成构象, 并不能给出构象的生物学意义以及动力学过程, 但是如果对特定体系引入实验信息, 就可以筛选出具有生物学意义的构象, 以便进行下一步的研究. 对于实验信息较少的蛋白质分子, 可以直接通过模型生成有潜在研究价值的构象, 然后从这些构象出发进行 MD 模拟, 研究蛋白质分子的动态过程, 进而预测可能的生物学意义. 这种策略与仅依靠 MD 模拟的构象空间采样相比, 效率更高.

在测试模型时, 发现了进一步的改进空间. 通过对模型生成构象的筛选和修正, 可以确保构象的合理性, 但同时也降低了生成构象的效率. 考虑直接将对构象合理性的要求引入模型的损失函数中, 从而省去筛选和修正的过程. 由于模型中只有蛋白质的主链部分, 有可能出现侧链不合理情况, 需要对不同氨基酸残基做不同修正或在模型中使用完整的蛋白质分子. 对于模型生成的构象无法表现出反应坐标之外变化的问题, 可以尝试使用变分自编码器. 最后, 反应坐标决定了构象空间探索的方向, 结合实验数据选取合适的反应坐标对模型的效果十分重要. 基于这些思路, 将继续对该模型进行发展和完善.

感谢中国科学技术大学超算中心张运动提供的硬件和软件技术支持.

参考文献

[1] Chu X, Gan L, Wang E, Wang J 2013 *Proc. Natl. Acad. Sci. U.S.A.* **110** E2342
[2] Smyth M S, Martin J H 2000 *Mol. Pathol.* **53** 8
[3] Danev R, Yanagisawa H, Kikkawa M 2019 *Trends Biochem. Sci.* **44** 837
[4] Vincenzi M, Mercurio F A, Leone M 2021 *Curr. Med. Chem.* **28** 2729
[5] Kachala M, Valentini E, Svergun D I 2015 *Adv. Exp. Med. Biol.* **870** 261
[6] Chu F, Thornton D T, Nguyen H T 2018 *Methods* **144** 53
[7] Bhaumik S R 2021 *Emerg. Top Life Sci.* **5** 49
[8] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl S A A, Ballard A J, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior A W, Kavukcuoglu K, Kohli P, Hassabis D 2021 *Nature* **596** 583
[9] Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee G R, Wang J, Cong Q, Kinch L N, Schaeffer R D, Millán C, Park H, Adams C, Glassman C R, DeGiovanni A, Pereira J H, Rodrigues A V, van Dijk A A, Ebrecht A C, Opperman D J, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy M K, Dalwadi U, Yip C K, Burke J E, Garcia K C, Grishin N V, Adams P D, Read R J, Baker D 2021 *Science* **373** 871
[10] Karplus M, Kuriyan J 2005 *Proc. Natl. Acad. Sci.* **102** 6679
[11] Bernardi R C, Melo M C R, Schulten K 2015 *Biochim. Biophys. Acta* **1850** 872
[12] Mu J, Liu H, Zhang J, Luo R, Chen H F 2021 *J. Chem. Inf. Model.* **61** 1037
[13] Lemke T, Peter C 2019 *J. Chem. Theory Comput.* **15** 1209
[14] Zhu J, Wang J, Han W, Xu D 2022 *Nat. Commun.* **13** 1661
[15] Hinton G E, Salakhutdinov R R 2006 *Science* **313** 504
[16] Degiacomi M T 2019 *Structure* **27** 1034
[17] Wen B, Peng J, Zuo X, Gong Q, Zhang Z 2014 *Biophysical J.* **107** 956

[18] Giri Rao V V H, Gosavi S 2014 *PLOS Computational Biology* **10** e1003938
[19] Abraham M J, Murtola T, Schulz R, Páll S, Smith J C, Hess B, Lindahl E 2015 *SoftwareX* **1–2** 19
[20] Weaver L H, Matthews B W 1987 *J. Mol. Biol.* **193** 189
[21] Zhang X J, Wozniak J A, Matthews B W 1995 *J. Mol. Biol.* **250** 527
[22] Müller C W, Schulz G E 1992 *J. Mol. Biol.* **224** 159
[23] Müller C W, Schlauderer G J, Reinstein J, Schulz G E 1996 *Structure* **4** 147
[24] Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C 2006 *Proteins Struct. Funct. Bioinf.* **65** 712
[25] Izadi S, Anandakrishnan R, Onufriev A V 2014 *J. Phys. Chem. Lett.* **5** 3863
[26] Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot B L, Grubmüller H, MacKerell A D 2017 *Nat. Methods* **14** 71
[27] Bussi G, Donadio D, Parrinello M 2007 *J. Chem. Phys.* **126** 014101
[28] Essmann U, Perera L E, Berkowitz M L, Darden T A, Lee H C, Pedersen L G 1995 *J. Chem. Phys.* **103** 8577
[29] Kingma D P, Ba J 2014 arXiv:1412.6980 [cs.LG]
[30] Lovell S C, Davis I W, Arendall III W B, de Bakker P I W, Word J M, Prisant M G, Richardson J S, Richardson D C 2003 *Proteins Struct. Funct. Bioinf.* **50** 437
[31] Eastman P, Swails J, Chodera J D, McGibbon R T, Zhao Y, Beauchamp K A, Wang L P, Simmonett A C, Harrigan M P, Stern C D, Wiewiora R P, Brooks B R, Pande V S 2017 *PLoS Comput. Biol.* **13** e1005659
[32] Shirts M R, Klein C, Swails J M, Yin J, Gilson M K, Mobley D L, Case D A, Zhong E D 2017 *J. Comput. -Aided Mol. Des.* **31** 147
[33] Touw W G, Baakman C, Black J, te Beek T A, Krieger E, Joosten R P, Vriend G 2015 *Nucleic Acids Res.* **43** D364

SPECIAL TOPIC—Machine learning in biomolecular simulations • COVER ARTICLE

Exploring proten's conformational space by using encoding layer supervised auto-encoder*

Chen Guang-Lin Zhang Zhi-Yong†

(*Department of Physics, University of Science and Technology of China, Hefei* 230026, *China*)

(Received 28 June 2023; revised manuscript received 29 July 2023)

Abstract

Protein function is related to its structure and dynamic change. Molecular dynamics simulation is an important tool for studying protein dynamics by exploring its conformational space, however, conformational sampling is a nontrivial issue, because of the risk of missing key details during sampling. In recent years, deep learning methods, such as auto-encoder, can couple with MD to explore conformational space of protein. After being trained with the MD trajectories, auto-encoder can generate new conformations quickly by inputting random numbers in low dimension space. However, some problems still exist, such as requirements for the quality of the training set, the limitation of explorable area and the undefined sampling direction. In this work, we build a supervised auto-encoder, in which some reaction coordinates are used to guide conformational exploration along certain directions. We also try to expand the explorable area by training through the data generated by the model. Two multi-domain proteins, bacteriophage T4 lysozyme and adenylate kinase, are used to illustrate the method. In the case of the training set consisting of only under-sampled simulated trajectories, the supervised auto-encoder can still explore along the given reaction coordinates. The explored conformational space can cover all the experimental structures of the proteins and be extended to regions far from the training sets. Having been verified by molecular dynamics and secondary structure calculations, most of the conformations explored are found to be plausible. The supervised auto-encoder provides a way to efficiently expand the conformational space of a protein with limited computational resources, although some suitable reaction coordinates are required. By integrating appropriate reaction coordinates or experimental data, the supervised auto-encoder may serve as an efficient tool for exploring conformational space of proteins.

Keywords: protein conformational space, molecular dynamics simulation, machine learning, auto-encoder

PACS: 87.15.ap, 87.15.hp **DOI:** 10.7498/aps.72.20231060

* Project supported by the National Key Research and Development Program of China (Grant No. 2021YFA1301504), the National Natural Science Foundation of China (Grant No. 91953101), and the Strategic Priority Research Program (B) of the Chinese Academy of Sciences (Grant No. XDB37040202).

† Corresponding author. E-mail: zzyzhang@ustc.edu.cn

物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

使用中间层受监督的自编码器探索蛋白质的构象空间

陈光临　张志勇

Exploring proten’s conformational space by using encoding layer supervised auto-encoder

Chen Guang-Lin　Zhang Zhi-Yong

引用信息 Citation: Acta Physica Sinica, 72, 248705 (2023)　DOI: 10.7498/aps.72.20231060
在线阅读 View online: https://doi.org/10.7498/aps.72.20231060
当期内容 View table of contents: http://wulixb.iphy.ac.cn

您可能感兴趣的其他文章
Articles you may be interested in

基于波动与扩散物理系统的机器学习
Machine learning based on wave and diffusion physical systems
物理学报. 2021, 70(14): 144204　https://doi.org/10.7498/aps.70.20210879

结合机器学习的大气压介质阻挡放电数值模拟研究
Numerical study of discharge characteristics of atmospheric dielectric barrier discharges by integrating machine learning
物理学报. 2022, 71(24): 245201　https://doi.org/10.7498/aps.71.20221555

通过机器学习实现基于摩擦纳米发电机的自驱动智能传感及其应用
Self-powered sensing based on triboelectric nanogenerator through machine learning and its application
物理学报. 2022, 71(7): 078702　https://doi.org/10.7498/aps.71.20211632

基于机器学习J_1–J_2反铁磁海森伯自旋链相变点的识别方法
Identifying phase transition point of J_1–J_2 antiferromagnetic Heisenberg spin chain by machine learning
物理学报. 2021, 70(23): 230701　https://doi.org/10.7498/aps.70.20210711

基于机器学习和器件模拟对Cu(In,Ga)Se_2电池中Ga含量梯度的优化分析
Optimization of Ga content gradient in Cu(In,Ga)Se_2 solar cells through machine learning and device simulation
物理学报. 2021, 70(23): 238802　https://doi.org/10.7498/aps.70.20211234

机器学习辅助绝热量子算法设计
Machine learning assisted quantum adiabatic algorithm design
物理学报. 2021, 70(14): 140306　https://doi.org/10.7498/aps.70.20210831