

专题: 生物分子模拟中的机器学习

分子体系自由能地貌图的变分分析及 AI 算法实现*

杜泊船¹⁾ 田圃^{1)2)†}

1) (吉林大学生命科学学院, 长春 130012)

2) (吉林大学人工智能学院, 长春 130012)

(2023 年 11 月 14 日收到; 2024 年 1 月 18 日收到修改稿)

精确描述复杂分子体系的自由能地貌图是理解和操控其行为, 并进一步实现分子设计制造工业化的重要基础. 刻画高维空间自由能地貌图的主要挑战是其往往在不同时空尺度上具有多个层次, 每个层次都可能不止一个亚稳态被相应的自由能垒分开, 且跨越路径有可能不止一条. 另外很多体系涉及非线性行为, 这使得理论解析和直接使用分子模拟都有很大困难. 针对这些挑战, 多年来研究者们发展了多种多样的增强采样方法, 但往往需要很多经验选择和操作, 从而一方面使得研究进程较为缓慢, 另一方面也让误差控制成为困难. 变分虽然在物理、统计和工程中已经被广泛应用并取得巨大成功, 但在复杂分子体系中的应用却随着神经网络的发展刚刚开始. 本文将对这些探索性工作的主要方向、进展和局限进行简要总结, 也对将来的可能发展给出展望, 希望能够激发更多对基于变分的分子体系自由能地貌图人工智能算法的关注和努力, 促进大分子药物、分子生物机器等实践应用的发展.

关键词: 变分, 神经网络, 复杂分子体系, 自由能地貌图**PACS:** 87.80.-y, 87.15.A-**DOI:** 10.7498/aps.73.20231800

1 引言

大多数复杂分子, 尤其是生物大分子体系, 都是通过构象变化或者在一定尺度上的相变实现其功能的^[1-6]. 和诸多分子的实验合成与表征测试过程相比较, 一方面分子模拟的代价往往更低廉; 另一方面很多生物大分子复合体的大量合成非常困难甚至不可能, 或者在能够获取的前提下动态表征很难实现. 因此分子模拟被广泛用于研究复杂分子体系^[7-9]. 决定分子体系各种行为的基础是对应的自由能地貌图, 因此对其准确刻画成为必要. 实现这一目标的主要挑战是复杂分子体系一般不止一个亚稳态并且相互之间有较高的自由能垒. 所以对典型的复杂分子体系 (如核糖体), 想要从全原子分子模拟中完成所有亚稳态的充分采样, 观察对应的

构象变化过程往往需要生成毫秒级甚至更长时间的模拟轨迹^[10]. 这对百万或更多原子的分子体系一方面算力需求很难满足, 另一方面在高维空间中理解所生成的轨迹也很不容易. 因此人们发展了各种各样的增强采样方法^[11-26]和轨迹降维分析方法^[27]. 增强采样方法大致可以分为两大类, 一类是保持分子体系的玻尔兹曼分布不变, 通过改变温度加速分子体系跨越能垒的方法^[12,13]. 另外一类则是通过加持偏置力/势 (bias force/potential) (如元动力学方法^[8]、自适应偏置力方法^[28]), 这类方法的主要依据是虽然一般分子体系的总自由度数目成千上万甚至更多, 但在跨越能垒的时间尺度上很多局部的原子运动都由于时间尺度的分离而成为近似白噪声, 使得体系在对应时空尺度的运动可以用较少的反应坐标 (reaction coordinates, RC) 或者集合自由度 (collective variable, CV) 成功描述, 下

* 吉林大学“学科交叉融合创新”项目 (批准号: JLUXKJC2021ZZ05) 资助的课题.

† 通信作者. E-mail: tianpu@jlu.edu.cn

文中统称集合自由度 (CV). 这类采样算法的主要困难是集合自由度的构建设没有系统的方法和步骤, 研究者往往依靠物理直觉选择部分体系自由度进行组合尝试. 由于我们生活中感受到的都是三维空间中的物理存在, 所以在体系维度升高后直觉判断的准确性会大打折扣. 如何准确地构建有效的 CV 是目前复杂分子体系模拟中尚未解决的重大挑战之一. 集合自由度空间中主要有 3 类互相关联的问题, 其一是准确描述体系的集合自由度的构建; 其二是绘制出该空间内主要亚稳态所在的构象空间位置和统计权重, 并计算不同亚稳态之间的转化速率; 其三是构建不同亚稳态之间的过渡路径. 这几类问题的传统应对策略已经被多个优秀综述覆盖 [14,29–41], 本文主要简述变分及其神经网络实现在这些领域的应用, 限于作者所熟悉研究工作的范围, 会遗漏一些优秀的研究进展, 在此表示歉意.

本文的内容组织如下, 首先将对 CV、变分和神经网络及自动微分进行简要说明, 其次对目前已有的针对复杂分子体系自由能地貌图的主要变分构造方法加以讨论, 再次对这些基于变分的和其他 CV 相关的神经网络方法进行比较分析, 最后展望将来的发展.

2 集合变量、相关神经网络架构、自动微分和变分简介

对一个在给定温度 T 和势能 $U(\mathbf{R})$ 下的分子体系, 用 $\mathbf{R}$ 表示其 $3N - 3$ 维坐标, 则平衡态玻尔兹曼分布为 $\mu(\mathbf{R}) = e^{-\beta U(\mathbf{R})}/Z$, 其中 $\beta = (k_B T)^{-1}$ 为逆温度, $Z = \int d\mathbf{R} e^{-\beta U(\mathbf{R})}$ 为配分函数, k_B 为玻尔兹曼常数. 在较长时间尺度上, 这个分子体系的动力学一般可以使用比 $3N$ 维度低很多也平滑很多的 d ($d \ll N$) 维自由能面描述, 对应一组由原来坐标 $\mathbf{R}$ 的函数构建的新变量 $\mathbf{s}(\mathbf{R}) = (s_1(\mathbf{R}), s_2(\mathbf{R}), \dots, s_d(\mathbf{R}))$, 分子体系自由能在这个低维空间也可表示为

$$F(\mathbf{s}) = -(1/\beta) \log \int d\mathbf{R} \delta(\mathbf{s} - \mathbf{s}(\mathbf{R})) e^{-\beta U(\mathbf{R})}, \quad (1)$$

人们通常称这组新变量 $\mathbf{s}(\mathbf{R})$ 为集合变量, $\delta(\cdot)$ 表示 δ 函数.

神经网络是目前人工智能技术浪潮的核心理论方法, 简而言之是由多个神经元组成的复合函数网

络. 每个神经元可以接受不同维度的输入, 经过线性组合和非线性激活函数作用后输出. 虽然原则上神经元之间的连接可以是任意的, 但受视神经分层分布的启发和随之带来的并行计算方便, 常用的各种神经网络架构都是层状结构. 神经网络最有力的特点是只需要一个隐藏层, 足够多神经元组成的网络就可以无限逼近任意函数映射, 这就是著名的万能逼近理论 (universal approximation theorem) [42–44]. 但这个理论并没有指出如何在有限的神经元数目的情况下有效拟合各种映射, 所以其发现虽然在很大程度上增强了人们使用神经网络拟合各种函数映射的信心, 却并没有迅速推动其在诸多实际问题中的应用. 后来多种神经网络架构 (卷积 [45]、循环 [45]、残差 [46]、注意力机制 transformer [47] 和扩散模型 [48]) 的发展推动了神经网络在多个学科领域应用的爆发. 当然另外一个不可或缺的基础是自动微分的发现 [49] 和在神经网络中的成功应用 [50], 这使得理论上基于任意阶导数的优化方法都能够被有效用来训练神经网络参数, 当然实际应用中由于算力和内存限制, 人们往往限于使用基于一阶和二阶导数的优化方法, 诸多具体实例和相关文献可以参考 PyTorch 中的 Optim 模块. 如下所述, 在众多神经网络架构中, 分子体系自由能地貌图刻画中应用最为广泛的是自编码器 (auto-encoder) 架构 [51] (如图 1 所示), 该架构把高维输入映射到一个低维

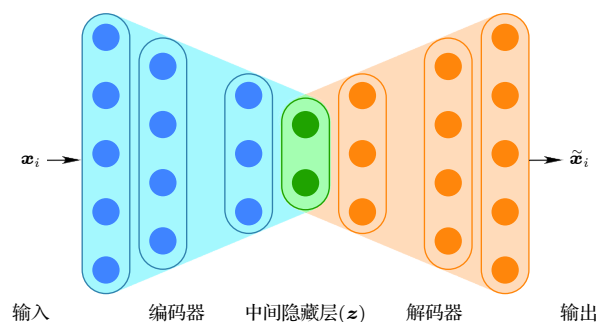


图 1 自编码器神经网络架构示意图, 蓝色部分表示编码器 (encoder) 函数 $f(\cdot)$, 橙色部分表示解码器 (decoder) 函数 $g(\cdot)$, 维度最低绿色表示中间隐藏层 (z). 对自编码器, 损失函数是输出 ($\tilde{x}_i$) 与输入 x_i 的差别的函数 (也可以加正则化项, 如参考文献 [58] (5) 式所示), 每一个输入数据点对应隐藏层空间的一个点

Fig. 1. Schematic representation of an auto-encoder neural network. The blue part on the left represents the encoder, the orange part on the right represents the decoder, and the middle green layer is the hidden layer (z). The loss is always a function of the difference between the input and the output vectors (x_i and $\tilde{x}_i$), one may add some form of regularization when necessary (e.g. Eq. (5) in Ref. [58]).

空间的降维部分被称为编码器 (encoder), 而随后从低维逆向映射到高维 (一般与输入同维度以方便训练) 空间的部分则被称为解码器 (decoder). 这显然与人们试图在更低维度空间理解复杂分子体系的目标在形式上较为吻合. 虽然在架构形式上非常相似, 但变分自编码器 (variational auto-encoder, VAE)^[52] 的目标和训练过程却与自编码器显著不同, 其中的隐变量 (z) 是个概率分布而非特定构型. 如果分别用 ϕ 和 ψ 表示编码器和解码器网络中的参数, $q_\phi(z|x)$ 和 $p_\theta(x)$ 表示隐变量 (z) 和 (x) 的分布, 则似然函数可表述如下:

$$\begin{aligned} \log p_\theta(x) &= \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x)] \\ &= \mathbb{E}_{q_\phi(z|x)} \left[\log \left[\frac{p_\theta(x, z)}{p_\theta(z|x)} \right] \right] \\ &= \mathbb{E}_{q_\phi(z|x)} \left[\log \left[\frac{p_\theta(x, z) q_\phi(z|x)}{q_\phi(z|x) p_\theta(z|x)} \right] \right] \\ &= \underbrace{\mathbb{E}_{q_\phi(z|x)} \left[\log \left[\frac{p_\theta(x, z)}{q_\phi(z|x)} \right] \right]}_{=\mathcal{L}_{\theta, \phi}(x) \text{ (ELBO)}} \\ &\quad + \underbrace{\mathbb{E}_{q_\phi(z|x)} \left[\log \left[\frac{q_\phi(z|x)}{p_\theta(z|x)} \right] \right]}_{=D_{\text{KL}}(q_\phi(z|x)||p_\theta(z|x))}, \end{aligned} \quad (2)$$

其中, $D_{\text{KL}}(q_\phi(z|x)||p_\theta(z|x)) \geq 0$, 所以 $\mathcal{L}_{\theta, \phi}(x) = \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x, z) - \log q_\phi(z|x)]$ 就是似然函数的下界, 也称为证据下界 (evidence lower bound, ELBO) 或变分下界, 是变分优化的目标, 而非自编码器中解码器输出构型与数据中实际构型差别的函数. 为了对随机隐变量 (z) 对自动微分, Waterfall 等^[53] 发展了二次参数化技巧 (reparameterization trick).

变分的历史非常悠久, 也是诸多理工科研究生的必修课程内容. 变分在物理、统计和工程领域都已经取得了非常广泛和成功的应用^[49,54], 如量子力学中的 Releigh-Ritz 方法^[55] 也正是本文中要讨论的分子体系变分计算的基础. 另外统计学中的大量应用展示了变分推断方法同采样计算相比高效、收敛性较好和更容易扩展的特点^[56,57]. 在神经网络广泛应用之前, 由于各种未知统计分布的解析和 (或) 参数化构造较为困难, 因此基于平均场的变分成为统计变分分析中最为常用的近似^[56]. 但在分子模拟及其增强采样中的应用却在最近十多年才陆续发生. 原因主要有两点, 其一是和很多统计模型与工程应用不同, 分子体系中的集合变量很难找

到直接的方程或模型解析描述, 其二是传统数值拟合方法 (如最小二乘法^[58]) 中导数计算昂贵且精度不高, 各种优化方法实现困难, 而且在变量较多 (大于 10 个) 时会收敛困难^[53]. 不过最近十多年以来基于自动微分^[49] 的多个人工智能框架 Pytorch^[59], Tensorflow^[60], PaddlePaddle^[61] 迅速发展成熟, 与之伴随的神经网络架构^[62] 也得到了迅猛发展. 这使得在拥有较为充足数据的前提下, 任意函数的稳健拟合成为可能, 因此增强采样和轨迹分析的变分应用也随之发展. 传统上人们探索复杂分子体系自由能地貌图的主要手段是 (加速) 采样, 变分的突出优点是用优化取代采样过程, 从而显著提高效率. 现代神经网络架构的强大拟合能力和基于自动微分的各种优化方法的结合为变分在复杂分子体系中的应用提供了巨大的潜力空间. 这也正是本文想要讨论的话题.

3 分子体系集合变量空间的变分方法

同物理学、工程和统计应用比较, 变分在复杂分子体系自由能地貌图应用相对较少, 主要是近十多年的工作, 不过目前正在迅速增长中. 目前的发展大致可以分为利用转移矩阵算子特征值和特征向量频谱分解分析 (spectral decomposition analysis) 的变分构建^[63-68]; 基于自由能垒跨越概率时间关联函数的变分^[69-71]; 利用偏置势 (bias potential) 的变分构建^[72]; 不受线性假设局限的可汇集性 (lumpability) 与可分解性 (decomposability) 泛函变分构建^[73]; 基于过去-将来信息瓶颈的变分构建^[74,75]; 同时考虑粗粒化、集合变量和增强采样的自适应^[76]; 以及直接利用变分自编码器的分析^[77], 这些方法的简要总结比较见表 1. 具体如下所述.

3.1 频谱分解分析

在严格马尔可夫过程和细致平衡假设下, 针对给定的子态构象空间划分, Perez-Hernandez 等^[63] 发展了利用演化算子 P (propagator) 特征函数自相关构建的变分实现了对最慢动力学过程集合变量 (CV) 的逼近, 分子体系动力学可以被下式表述为演化算子特征函数 ϕ_i ($i = 1, 2, \dots, \infty$) 的叠加:

$$\rho_{t+\tau}(y) = P(\tau)\rho_t(x) = \sum_{i=1}^{\infty} e^{-\frac{\tau}{\tau_i}} \langle \psi_i, \rho_t \rangle \phi_i, \quad (3)$$

表 1 复杂分子体系低维隐空间的变分方法简要总结, 表中所述集合空间问题类别是指引言中提到的三类问题
Table 1. A brief summary of variational methods for low-dimensional hidden spaces in complex molecular systems. The category of collective space problems mentioned in the table refers to the three types of problems defined in the introduction.

变分方法	主要目标	关注的集合空间问题类别	特点或主要局限
频谱分解分析	基组线性组合	第1类、第2类	马尔可夫假设与线性基组局限, 需要人工划分构象空间子状态
	神经网络实现	第2类	马尔可夫假设, 没有解析表示的特征函数, 需要人工调整架构测试不同聚类数量
自由能垒跨越概率时间关联函数	基组线性组合	第3类	基组线性组合局限, 需要定义始末态
	神经网络实现	第3类	需要定义始末态
基于偏置势变分	基组线性组合	第2类	泛函受基组选择限制
	神经网络实现	第2类	泛函导数求解的采样需求导致偏置势(和对应自由能)的精度紧密相关, 收敛受KL散度非对称性限制
Lumpability 和 Decomposability	优化集合变量	第1类	有明确误差控制, 方差取决于隐空间维度, 两种定义的一致性要求可逆过程
信息瓶颈模型	求解信息瓶颈对应集合空间CV表示, 并利用偏置势加速自由能面采样	第2类	线性编码过程假设局限
变分自适应	结合粗粒化信息加速采样求解自由能面	第2类	总体架构较为复杂
变分自编码器	通过集合变量空间加速采样求解自由能面和聚类转化路径	第2类、第3类	特别关注隐空间

其中 t_i 是和第 i 个特征值 $\lambda_i(\tau) = e^{-\tau/t_i}$ 对应的时间尺度, 尖括号代表标量积, $\langle \psi_i, \rho_t \rangle$ 的结果表征概率密度 ρ_t 和 ψ_i 的重叠程度, 体现了第 i 个特征函数对总体动力学的贡献. 因为 $\psi_i = \mu^{-1}(x) \phi_i$, 也可以认为概率密度函数 ρ 是基于特征函数 ϕ_i 展开的. 显然随着 $\tau \rightarrow \infty$, 概率密度会趋于平衡态, (3) 式中只有第 1 项有贡献, 对应于 $\lambda_1 = 1$. 如果人们感兴趣的时间尺度 $\tau \gg t_{d+1}$, 则分子体系的动力学主要取决于对应于 $(\lambda_1, \lambda_2, \dots, \lambda_d)$ 的 d 个特征函数, 也对应于前面 (见方程 (1) 中的 $s(\mathbf{R})$ 定义) 所说的 d 个集合变量. (3) 式可近似为

$$\rho_{t+\tau} = P(\tau) \rho_t \approx \sum_{i=1}^d e^{-\frac{\tau}{t_i}} \langle \psi_i, \rho_t \rangle \phi_i. \quad (4)$$

对于分子体系坐标的任意函数 $f(x)$, 其自相关函数可以表述为

$$\langle f(\mathbf{x}_t) f(\mathbf{x}_{t+\tau}) \rangle_t = \sum_{i=1}^{\infty} e^{-\frac{\tau}{t_i}} \langle \phi_i, f \rangle^2. \quad (5)$$

显然如果取 $f = \psi_i(x)$, 则

$$\begin{aligned} \left[\lambda_i^\dagger(\tau) = \langle \psi_i(\mathbf{x}_t) \psi_i(\mathbf{x}_{t+\tau}) \rangle_t = e^{-\frac{\tau}{t_i}} \right], \\ \left[t_i^\dagger = -\tau / \ln |\lambda_i^\dagger(\tau)| = t_i \right]. \end{aligned} \quad (6)$$

对于近似的特征函数 $\psi_2^\dagger$:

$$\langle \psi_2^\dagger(\mathbf{x}_t) \psi_2^\dagger(\mathbf{x}_{t+\tau}) \rangle_t \leq e^{-\frac{\tau}{t_2}}, \quad t_2^\dagger \leq t_2. \quad (7)$$

因此近似时间尺度 $t_2^\dagger$ 可以作为变分目标. 文中时间尺度最大化结果的实现依靠选定基函数 (具体来说使用了分子体系构型在特定构象子空间的示性函数 (indicator function)) 的线性组合来近似特征函数. 该方法虽然在给定构象聚类的前提下能够给出较好的动力学常数和对应的特征函数 (集合变量) 估计, 但实现构象聚类的过程依然依靠简单的降维方法. 类似地 Tiwary 和 Bern^[78] 通过最大化频谱间距 (spectral gap) 过渡路径熵也展开了给定集合变量空间的线性组合优化, 不过没有使用变分方法. 后来为了拓展该理论的应用范围, Wu 和 Noé^[66] 把该方法推广到不需要细致平衡的一般马尔可夫过程. 尽管多数复杂分子体系的动力学过程一般都是非线性的, 但依据 Koopman 理论^[66,76] 可以构建新的隐空间 $\chi_0(\mathbf{x}) = (\chi_{01}(\mathbf{x}), \chi_{02}(\mathbf{x}), \dots, \chi_{0d}(\mathbf{x}))^T$ 和 $\chi_1(\mathbf{x}) = (\chi_{11}(\mathbf{x}), \chi_{12}(\mathbf{x}), \dots, \chi_{1d}(\mathbf{x}))^T$ 使得原来坐标系中的非线性变换被转化为如下式所示的线性变化:

$$\mathbb{E}[\chi_1(\mathbf{x}_{t+\tau})] \approx \mathbb{K}^T \mathbb{E}[\chi_0(\mathbf{x}_t)]. \quad (8)$$

显然 (8) 式中转化过程 χ_0 和 χ_1 一般是非线性并且未知的. 神经网络的可训练万能逼近能力为其构建

未知转换的提供了可能性. VAMPNets^[67]正是在这种思考下构建的. 对于给定的 χ_0 和 χ_1 变换可以构建 3 个方差矩阵:

$$\begin{aligned} C_{00} &= \mathbb{E}_t[\chi_0(\mathbf{x}_t)\chi_0(\mathbf{x}_t)^T], \\ C_{01} &= \mathbb{E}_t[\chi_0(\mathbf{x}_t)\chi_1(\mathbf{x}_{t+\tau})^T], \\ C_{11} &= \mathbb{E}_{t+\tau}[\chi_1(\mathbf{x}_{t+\tau})\chi_1(\mathbf{x}_{t+\tau})^T]. \end{aligned} \quad (9)$$

这些方差矩阵被用来构建了一个 VAMP-2 打分^[66]:

$$\hat{R}_2[\chi_0, \chi_1] = \left\| C_{00}^{-1/2} C_{01} C_{11}^{-1/2} \right\|_F^2. \quad (10)$$

该分值最大化时对应转化后的低 (d) 维空间分子构象分布被准确复现. 以这个分值作为损失函数的神经网络通过训练就有可能实现从体系原始高维坐标向低维空间较为准确的映射, 实际上起到了低维空间模糊分类器的功能, 消除了前述变分理论^[65]中对人工聚类及以前各个步骤的需求. 具体实现架构如图 2(a) 所示^[67]. 对丙氨酸二肽体系, Mardt 等^[67]利用图 2(b) 的特定架构, 设定低维空间类别数目为 6(也尝试了从 2—8 的其他类别数目), 首先从 250 ns 的分子动力学模拟轨迹中每皮秒提取一帧得到 250000 个构型, 并通过和第一帧对齐除去分子的整体平移和旋转. 使用十个重原子的三维空间 (即长度为 30 的向量) 坐标作为神经网络输入, 取延迟时间 $\tau=40$ ps (也尝试了从 4—32 ps 的其他延迟时间), 通过最大化 VAMP 打分, 成功实现了在二维二面角空间 (φ, ψ) 的构象聚类. 他们同时使用 MSM 流程聚类, 当构象类别数目小于 20 时得到 VAMP 打分都低于 VAMPnets 的结果. 此外 Mardt 等^[67]还尝试分析了简单双势阱和 NTL9 蛋白折叠轨迹, 均展示了和原来人工复杂流程可比拟的准确性, 也说明这个思路有望在将来通过逐步发展真正实现自动分析分子模拟轨迹得到动力学特性的可能性. 不过目前该方法还不够成熟, 尚不能用于多系综组合数据^[79–84], 也不能有效集成模拟轨迹与相关实验数据, 另外还缺乏严格清楚定义的误差估算指标^[85–87]. 但该研究结合变分理论和非线性的神经网络拟合, 取代了原来 MSM 方法管线中一系列复杂步骤, 并在简单体系中实现了首次成功应用, 是人工智能用于分析复杂分子体系轨迹的重要进展. VAMPNets 的神经网络架构较为简单, 鉴于图神经网络^[88–91]和注意力机制^[92]在网络型数据中的优异表现, 考虑到复杂分子体系可以被视为由相互作用的单元构成的网络图, Brooks 等^[93]

构建了包含这两种架构要素的 GraphVAMPnet, 该模型实现了更高精度的构象嵌入表示, 也能够通过注意力机制给出蛋白质中对结构聚类起决定性作用的重要氨基酸. 在 20-氨基酸的 Trp-cage 蛋白, 35-氨基酸的 Villin 蛋白和 NTL9 蛋白轨迹上的成功应用展示了这些神经网络构架改变的好处.

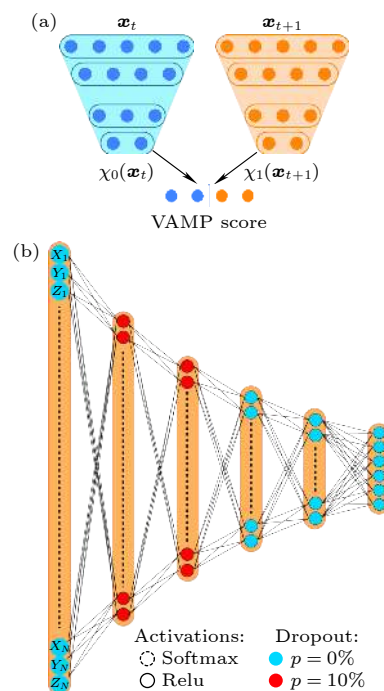


图 2 (a) VAMPnets 构建 VAMP 打分 ((10) 式) 的神经网络总体架构示意图; (b) 丙氨酸二肽轨迹分析实例中的典型神经网络架构, 各层神经元数目为 32-22-16-9-6, 前两层使用 10% 的 dropout, 除最后的 softmax 层外, 其余各层激活函数均使用 Relu^[67]

Fig. 2. (a) Schematic illustration of VAMP score construction from VAMPnets (see Eq. (10)). (b) A typical neural network architecture for alanine dipeptide analysis, with the number of neurons being 32-22-16-9-6 for five layers. The first two layers utilized a 10% dropout. Relu was selected as the activation function for all layers except the last softmax layer^[67].

随着人们使用电子显微镜解析生物大分子复合体的能力越来越强, 如何解释这些复合体的动力学过程变成了亟待解决的问题. 为了增进处理较大分子的能力并在将来能够有可能延伸到大复合体, Noé等^[94]结合独立马尔可夫分解方法 (independent Markov decomposition, IMD) 构建了由多个独立的 VAMPNets 构成的 iVAMPNets. 其中不同独立模块的划分由一个可训练的 MASK 实现, 通过竞争训练使每个不同的子网络仅处理不与其他子网络相互重叠的部分. 虽然该方法在 Syna-

ptotagmin-C2A 蛋白质分子中成功应用, 但显然这种处理仅适用于不同子模块间耦合程度较弱的情况, 距离准确描述不同组成分子之间有较强关联的复杂复合体仍然有较大距离. 利用 VAMPNets 输出的子构象空间 (状态) 概率, Kleiman 和 Shukla^[68] 尝试了结合 3 种不同后续处理, 包括最小计数 (least count, LC), 多目标强化学习 (multiagent reinforcement learning-based, MA REAP) 和最大熵 (MaxEnt), 显著促进了构象空间搜索能力. 这 3 种方法的宗旨基本一致, 就是利用前期生成的轨迹对 VAMPNets 进行初步训练后, 在后续的采样中按照上述不同标准聚焦前期采样最少访问的构象空间, 从而实现更进一步的增强采样. 其中最大熵和 VAMPNets 的结合在促进采样的同时消除了聚类步骤.

3.2 自由能垒跨越概率时间关联函数的变分

弦方法^[95-99]和过渡路径理论 (transition path theory, TPT)^[40] 致力于寻找不同亚稳态之间过渡路径及其过渡态的关键细节. 不过这些方法在得到最低自由能过渡路径的同时, 却不能直接给出人们非常感兴趣的路径上任意一点的自由能垒跨越概率. 针对此问题, 文献^[71,100] 基于两个亚稳态之间的净向前反应通量构造了自由能垒跨越概率时间关联函数, 发展了通过变分最小化该函数获得最佳过渡路径并同时给出自由能垒跨越概率的方法. 对两个亚稳态 A 和 B, 集合变量空间从 A 到 B 在时间步长 τ 基于算子 $P_\tau(s'|s)$ 的向前演化可表示为

$$\rho(s'; t + \tau) = \int ds P_\tau(s'|s) \rho(s, t), \quad (11)$$

其中 $\rho(s, t)$ 和 $\rho(s'; t + \tau)$ 分别对应于时刻 $t(t + \tau)$ 在路径位置 $s(s')$ 处的概率密度. 则自由能垒跨越概率 $q(s)$, 即从 s 开始最终到达亚稳态 B 并且在此前从未到达亚稳态 A 的所有过渡路径概率之和, 可定义如下:

$$q(s) = \int ds' q(s') P_\tau(z'|z). \quad (12)$$

则净向前 (从 A 到 B) 反应流为

$$J_{AB} = \frac{1}{2\tau} \int ds \int ds' (q(s') - q(s))^2 P_\tau(s'|s) \rho_{eq}(s). \quad (13)$$

也可以表达为自由能垒跨越概率的自相关函数:

$$J_{AB} = \frac{1}{2\tau} \langle (q(\tau) - q(0))^2 \rangle \\ = \frac{1}{\tau} (\langle q(0)q(0) \rangle - \langle q(\tau)q(0) \rangle), \quad (14)$$

其中二次形式可以作为任意给定始末态时尝试自由能垒跨越概率 $q(s')$ 的变分优化目标. 该方法使用基组展开, 通过优化系数来达到变分优化的目标, 在模型双势阱问题中展示了简化子空间 (CV 空间) 中理想一维反应坐标走向沿着自由能垒跨越概率梯度, 与高维空间中的 Kramers-Langer 理论^[101,102] 一致. 文献^[100] 是针对过渡路径变分构建的首次尝试, 并在双势阱问题和丙氨酸二肽中展示了应用. 由于变分函数限于选定基组函数的线性组合空间, 其结果显然会受到基组选择和线性组合的制约. Chipot 等^[69] 将自由能垒跨越概率时间关联函数的变分方法延伸到了神经网络 (variational committor-based neural networks, VCN), 从而可以拟合任意非线性映射. 同基于特征值变分优化的 VAMPNets 相比较, 在双势阱体系和 N-acetyl-N'-methylalanylamine 异构化过程中均得到一致结果. 不过显著不同的是 VCN 需要已知始末态, 针对的目标是一对始末态之间的过渡路径, 而 VAMPNets 则是从轨迹数据开始的无监督学习. 另外一点是有时候人们最感兴趣的慢过程可能不是分子体系中最慢的过程, 这种情况下显然 VCN 更为适合. 这两类方法可以协同使用从而结合其各自优势, 当然也有可能在未来集成到更复杂的神经网络架构中.

3.3 基于偏置势的变分

在给定 CV 的前提下, Valsson 和 Parrinello^[103] 构建了一个基于 CV 空间偏置势 $V(s)$ 的泛函:

$$\Omega[V] = \frac{1}{\beta} \log \frac{\int ds e^{-\beta[F(s) + V(s)]}}{\int ds e^{-\beta F(s)}} + \int ds p(s) V(s), \quad (15)$$

其中 $p(s)$ 是一个自由选择的目标分布, 这赋予人们使用该泛函的灵活性 (当然也伴随着选择的挑战). 该泛函是一个凸函数并且不随偏置势任意给定的有限常数的改变而变化. 用 $F(s)$ 表示体系自由能, 则当 $V(s) = -F(s) - (1/\beta) \log p(s)$ 时, 泛函 $\Omega[V]$ 取极小值, 因此在选定 $p(s)$ 的前提下通过参数化的 $V(s)$, 以 $\Omega[V]$ 极小值为目标的变分优化即

可求解自由能地貌图. 该方法使用线形基组组合在丙氨酸三肽分子中成功应用. 另外, 该泛函同 Kullback-Leibler (KL) 散度 (D_{KL}) 的关系如下所示 [46,104,105]:

$$\beta\Omega[V] = D_{\text{KL}}(p||P_V) - D_{\text{KL}}(p||P_0), \quad (16)$$

其中 P_V 和 P_0 分别是偏势为 V 和 0 时体系的概率密度分布. 由于凸函数特性, 使得偏势与自由能面有确定关系的驻点也是其极值点. 因此通过参数化偏势, 就可以对参数实施变分优化从而求解自由能面. 这在原理上比元动力学采样方法要高效很多, 不过, 其表现受限于所选 CV 在较长时间尺度上描述自由能面的能力. 为了克服对该泛函线性展开可能出现的一些麻烦 (比如自由能变化剧烈的区域需要很多项才能实现较好拟合, 集合变量增大时需要变分优化的参数空间指数增长), Bonati 等 [72] 用神经网络表示偏势泛函, 在给定的集合变量定义下通过优化神经网络参数实现, 如下所示:

$$\frac{\partial\Omega(\alpha)}{\partial\alpha_i} = -\left\langle \frac{\partial V(s; \alpha)}{\partial\alpha_i} \right\rangle_{V(\alpha)} + \left\langle \frac{\partial V(s; \alpha)}{\partial\alpha_i} \right\rangle_p. \quad (17)$$

泛函数值微分需要统计平均 ((17) 式中的尖括号表示系综平均), 因此需要采样获取. 直接高精度确定最低点较为困难, 因此 Bonati 等 [72] 在实现过程选择获得达到一定近似程度的偏势, 评判的标准选用了 $p_V(s)$ 和 $p(s)$ 在迭代次数 n 时的 KL 散度距离:

$$D_{\text{KL}}^{(n)}(p_V || p) = \sum_s p_V^{(n)}(s) \log \frac{p_V^{(n)}(s)}{p^{(t)}(s)}. \quad (18)$$

显然, 此过程在数值实现中需要选定两个参数, 一个是选定每次迭代计算 KL 散度之间的模拟更新次数, 另一个是每次更新时学习率调整的幅度. 为了集成 CV 构建和偏势优化, Bonati 等 [106] 利用 VAMPNets 的 VAMP 打分作为损失函数, 利用深度神经网络和 TICA (time-structure based independent component analysis) 结合生成 CV, 随后在更新的 CV 空间采用 OPES [100] 增强采样思路, 实现了 CV 优化和自由能地貌图收敛的迭代. 他们在丙氨酸二肽、chignolin 蛋白折叠和材料结晶过程的成功展示了该方法的应用 [106].

3.4 基于可汇集性 (lumpability) 和可分解性 (decomposability) 的非线性变分描述

由于马尔可夫假设和特征函数构建中的线性

假设, 基于频谱分解分析的变分优化无法正确处理非马尔科夫过程 [40] 和线性无关特征函数之间的非线性关联, 这些根本上的局限无法在后期变分优化中被消除. 针对这个问题, Bitttracher 等 [73] 通过延伸过渡流形理论 (transition manifold theory) 发展了不包含任何线性假设, 只关注于长时间尺度分子体系行为, 显式包含误差量且在可逆体系中互相等价的条件, lumpability 和 decomposability (详见文献 [73] 的 definition 3.2, 3.4), 这两个条件都可以作为损失函数变分. 此外该变分在近似损失函数时只要求在集合变量空间的稀疏采样, 而且损失函数的蒙特卡罗积分误差取决于集合变量空间而非原高维空间的方差, 这会带来巨大的算力节省. 该理论和过渡路径理论的连接仍然有待阐明. 另外这些理论上的优势在百万级甚至更大的复杂分子体系如何得以实现也有待于进一步探索.

3.5 过去-将来信息瓶颈模型

Wang 等 [74] 将分子体系中的集合变量空间视为其演化过程中的过去-将来信息瓶颈 (past-future information bottleneck, PIB [107,108]), 对给定分子体系任意时刻坐标 $\mathbf{X}$ 和下一时刻坐标 $\mathbf{X}_{\Delta t}$, 通过瓶颈变量 χ (与集合变量类似的分子体系低维空间描述) 分别和编码器 $P(\chi|\mathbf{X})$ 与解码器 $P(\mathbf{X}_{\Delta t}|\chi)$ 联系 (注意文献 [74] 中结果部分第 1 段把坐标 $\mathbf{X}$ 误解释为 N 个粒子体系中的 d 维 ($1 \ll d \ll N$) 表示, 容易引起混乱). PIB 的目标是瓶颈变量 χ 相对于过去应该尽量简单但对于将来则应该有尽可能好的预测力, Wang 等 [74] 据此构建了如下优化目标:

$$\mathcal{L} \equiv I(\chi, \mathbf{X}_{\Delta t}) - \gamma I(\mathbf{X}, \chi), \quad (19)$$

其中 $I(\chi, \mathbf{X}_{\Delta t})$ 和 $I(\mathbf{X}, \chi)$ 分别表示瓶颈变量与 $\mathbf{X}_{\Delta t}$ 和 $\mathbf{X}$ 的互信息, 常数 $\gamma \in [0, \infty)$ 用来平衡瓶颈变量 χ 的复杂程度和预测力. 进一步通过选择确定性的线性编码器, 则第 2 项可以忽略. 他们然后利用 Gibbs 不等式构建了可变分优化的 PIB 下限近似:

$$\begin{aligned} I(\chi, \mathbf{X}_{\Delta t}) &= H(P_\theta(\mathbf{X}_{\Delta t})) - H(P_\theta(\mathbf{X}_{\Delta t}|\chi)) \\ &\geq H(P_\theta(\mathbf{X}_{\Delta t})) - C(P_\theta(\mathbf{X}_{\Delta t}|\chi)|Q_\phi(\mathbf{X}_{\Delta t}|\chi)), \end{aligned} \quad (20)$$

其中 H 和 C 分别表示香农和交叉熵, Q_ϕ 为随机深度神经网络构建的解码器. 由于选择 P_θ 为确定性

线性编码器, 香农熵项退出优化目标, 可得更新变分下界:

$$\mathcal{L} \geq \mathcal{L}' = -C(P_\theta(\mathbf{X}_{\Delta t}|\chi)|Q_\phi(\mathbf{X}_{\Delta t}|\chi)), \quad (21)$$

其中 ϕ 为随机神经网络中的变分优化参数. 对平衡态轨迹 $\{\mathbf{X}^1, \dots, \mathbf{X}^{M+k}\}$ ($\mathbf{X}^n$ 和 $\mathbf{X}^{n+k}$ 之间的时间间隔为 Δt), 方程 (20) 可被离散为

$$\mathcal{L}' = \frac{1}{M} \sum_{n=1}^M \log Q(\mathbf{X}^{n+k}|\mathbf{X}^n), \quad (22)$$

其中 χ^n 从 $P(\chi^n|\mathbf{X}^n)$ 中采样得到. 对于有对应偏置势 $\{V^1, V^2, \dots, V^{M+k}\}$ 下模拟的轨迹则可在假设偏置势不改变解码器的情况下近似表述为

$$\mathcal{L}' = \left\{ \sum_{n=1}^M e^{\beta V^n} \right\}^{-1} \sum_{n=1}^M e^{\beta V^n} \log Q(\mathbf{X}^{n+k}|\mathbf{X}^n). \quad (23)$$

实际计算中 Wang 等 [74] 选择用坐标的线性基组合得到 CV, 首先对平衡态轨迹通过逐步增加 Δt 观察基组各项的权重变化, 并取其趋于稳定后最小的 Δt . 随后则按照方程 (24) 和 (25) 计算偏置势并重新估算机组系数, 反复迭代:

$$V_{\text{bias}}(\chi) = k_B T \log P^u(\chi), \quad (24)$$

$$P^u(\chi) \propto \frac{\langle w \delta(\chi - \chi(t)) \rangle_b}{\langle w \rangle_b} \equiv e^{-\beta F(x)}, \quad (25)$$

其中 $w = e^{\beta V_{\text{bias}}}$, $P^u(\chi)$ 是没有偏置势的情况下 χ 的平衡态分布. 简单的确定性线性编码器在带来方便的同时也在一定程度上限制了该方法的灵活性, 但 PIB 的优点之一是原则上没有其他线形假设, 不过在 PIB 思路下 (见 (19) 式) 使用非线性编码器后的变分优化方法仍有待发展. 该方法在苯-溶菌酶复合体模拟中获得了成功, 在几百纳秒的加速模拟中观察到了几百毫秒常规模拟所观察到的解离过程. Beyerle 等 [75] 后来使用该方法成功描述了双势阱模型和苯甲酸在双分子层膜中扩散这两个分别由能量和熵主导的过渡路径, 进一步展示了该方法的稳健性.

3.6 变分自适应

与前述变分方法主要关注集合变量和偏置势不同, 对有明确集合变量的情况, Zhang 等 [76] 结合生成式深度学习和基于能量模型 [109] (energy based models, EBM) 发展了变分对抗密度估计变分, 直接计算自由能地貌图中的概率密度分布. 将平衡态真实

自由能对应的概率分布记为 p , 在集合变量空间的参数化自由能地貌图和对应的分布分别记为 $F_\theta(s)$ 和 $p_\theta(s)$, 则 KL 散度 $D_{\text{KL}}(p||p_\theta)$ 对 θ 的导数可表示为

$$\nabla_\theta D_{\text{KL}}(p||p_\theta) = \langle \beta \nabla_\theta F_\theta(s) \rangle_{p(s)} - \langle \beta \nabla_\theta F_\theta(s) \rangle_{p_\theta(s)}, \quad (26)$$

其中 $\langle f(x) \rangle_{p(x)}$ 表示函数 $f(x)$ 在分布 $p(x)$ 下的期望值. (26) 式和对抗神经网络 [76,110] 高度相似, 因此在原文中被称为变分对抗密度估计 (variational adversarial density estimation, VADE). 在实际操作中可以用粗粒化实验数据 $P_{\text{FG}}(s)$ 取代真是分布 p . 再通过粗粒化模拟计算 $\langle \beta \nabla_\theta F_\theta(s) \rangle_{p_\theta(s)}$. 对于集合变量维度较高的情况, 由于直接采样计算代价过于昂贵, Zhang 等 [76] 通过加入可训练生成神经网络模块作为神经采样器 (neural sampler) q_ψ , 采用下式实现变分训练:

$$D_{\text{KL}}(q_\psi||p_\theta) = \langle \log q_\psi(s) + \beta F_\theta(s) \rangle_{q_\psi(s)} + \log Z_\theta. \quad (27)$$

加速分子体系自由能地貌图统计概率分布参数的训练. 对于没有集合变量函数的更一般情况, 通过加入了强化学习模式, 较好地解决了固定偏置势在动态采样中尴尬的同时, 实现了粗细两个不同粒度的有效采样补充. 这些方法都被集成在 SPONGE [111] 平台上.

3.7 变分自编码器的直接应用

上述构建显式变分优化目标函数的做法能够给出更有效的物理图像, 神经网络主要用于拟合其中未知非线性映射过程. 不过即使没有直接显式变分目标函数的构建, 变分的思想依然可以被利用. 最简单的做法就是直接使用变分自编码器 VAE 架构 [77] 对自己感兴趣的数值分布进行优化, 同时在生成的隐空间 (对应于分子体系集合变量空间) 展开一系列增强采样的操作, 必要时再引入迭代机制.

Ribeiro 等 [112] 发展了重配权变分贝叶斯增强采样 (reweighted auto encoded variational Bayes for enhanced sampling, RAVE) 方法, 通过隐空间分布和模拟的 KL 散度优化自编码器, 更新偏置势模拟后迭代优化直至收敛, 实现了独立于传统方法的隐空间增强采样. 针对在 MSM 模型中使用过渡路径理论方法时会得到大量量子状态之间的路径, 从

而使结果难以理解的困境, Qiu 等^[113,114] 利用 VAE 的数值分布变分优化, 在隐空间实现了类似过渡路径的合并. 该方法被成功应用在两个不同的简单体系, 分别是一对疏水粒子在水溶液中的聚集和 Fip35WW 结构域折叠路径的分析中. 利用 VAE 能够有效预测编码空间、隐空间和解码空间概率密度的特性, Monroe 和 Shen^[115] 发展了基于隐空间的蒙特卡罗移动建议方法, 再通过编码和解码, 从而在真实高维空间有效且高接受率的移动. 该方法的突出优点是直接满足细致平衡要求, 不需要一般偏置势加速采样生成轨迹后的权重调整, 从而避免了与之伴随的所有潜在问题和困难. 这个思路和通过粗粒化模拟促进 (细粒度) 全原子模拟^[76], 以及把低维子空间视为信息瓶颈^[74,75] 的具体方式虽然差别较大, 但总体基本思路一致. 不同粒度之间更加高效准确的构型映射和信息传递还有很大的方法学发展空间, 这方面的新发展也大概率会显著促进复杂分子体系高精度多尺度模型的构建.

4 其他神经网络方法在自由能地貌图相关研究中的应用

神经网络网络的万能逼近能力使得其在自由能面探索中从多个角度被加以应用. 其中很多工作都致力于获得更好的集合变量以改善复杂体系的增强采样. 早在 2005 年, Ma 和 Dinner^[116] 就开始使用神经网络用来寻找复杂体系的反应坐标. 针对各种传统降维方法不能直接把结果中的低维空间 (集合) 变量表达为原空间坐标的问题, Chen 和 Ferguson^[117] 利用自编码器可以实现从高维输入空间到低维隐空间之间的可训练映射, 把通过已有轨迹数据训练生成的隐空间自由度作为集合变量, 从而实现了集合变量偏置势通过自编码器对高维空间坐标的直接微分计算偏向受力, 集成了集合变量的神经网络构建和在加速采样中的直接应用, 该方法在丙氨酸二肽和 TrpCage 蛋白体系中被成功使用. 与此类似, Chen 等^[118] 也采用自编码器进行降维训练获得 CV, 然后通过自动微分把施加于 CV 上的偏置势传递到分子体系中去实现模拟采样和自由能计算.

与使用变分优化特征函数不同的另外一种思路是回归方法. Wehmeyer 和 Noé^[119] 尝试了选择对 N 个连续时间坐标序列 ($\mathbf{X}_t, \mathbf{X}_{t+\tau}, t = 1, 2, \dots, N$) 最小化回归误差^[120-122]:

$$\min_{D,E} \sum_t |\mathbf{X}_{t+\tau} - D(E(\mathbf{X}_t))|^2, \quad (28)$$

其中 D 和 E 分别为编码器和解码器. 在对已有轨迹数据的时间序列坐标构型按照下式进行均值归零 ((28) 式和 (29) 式) 和白化 ((30) 式和 (31) 式):

$$\mathbf{x}_t^{mf} = \mathbf{X}_t - \frac{1}{T-\tau} \sum_{s=1}^{T-\tau} \mathbf{X}_s, \quad (29)$$

$$\mathbf{y}_t^{mf} = \mathbf{x}_{t+\tau} - \frac{1}{T-\tau} \sum_{s=1}^{T-\tau} \mathbf{x}_{s+\tau}, \quad (30)$$

$$\tilde{\mathbf{x}}_t = C_{00}^{-\frac{1}{2}} \mathbf{x}_t^{mf}, \quad C_{00} = \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} \mathbf{x}_t^{mf} \mathbf{x}_t^{mf\top}, \quad (31)$$

$$\tilde{\mathbf{y}}_t = C_{\tau\tau}^{-\frac{1}{2}} \mathbf{y}_t^{mf}, \quad C_{\tau\tau} = \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} \mathbf{y}_t^{mf} \mathbf{y}_t^{mf\top}. \quad (32)$$

然后对处理后的坐标优化训练, 实现编码器降维和解码器对原空间的映射:

$$\min_{E,D} \sum_{t=1}^{T-\tau} \|\tilde{\mathbf{y}}_t - D(E(\tilde{\mathbf{x}}_t))\|_2^2. \quad (33)$$

通过训练过程中在输出端使用相对输入端 t 时刻的延后 $t + \tau$ 时刻坐标, 也实现了演化的预测. 对于在构象空间中线性可分的不同亚稳态, 该方法被证明同 Koopman 模型^[49,68] 等价. 但对非线性可分的体系, 与 PCA 和 TICA 及人工构造特征空间相比, 文献^[119] 的丙氨酸二肽体系显示通过编码器和解码器的深度学习拟合则可以更好地处理.

Zhang 和 Chen^[123] 针对不恰当的 CV 会在其正交空间出现亚稳态简并 (degeneracy) 从而导致对应方向不能加速采样的问题, 发展了利用随机动力学嵌入 (stochastic kinetic embedding, StKE) 的半监督学习方法增加对当前信息最匮乏区域 (current least informative regions, CLIRs) 的主动学习采样 (active enhanced sampling, AES), 这与 Kleiman 和 Shukla^[68] 在 VAMPNets 输出构象类型采样最少的部分增加后续采样的思路类似. 该方法成功在丙氨酸二肽和五肽 met-enkephalin 体系中从随意给定的无效 CV 开始, 以较短时间实现了对自由能地貌图的可靠采样. Rydzewski 和 Valsoson^[124] 提出的多尺度重配权重随机嵌入 (multi-scale reweighted stochastic embedding, MRSE) 则在此基础上更进一步, 通过高斯混合模型描述高维特征空间和重配权重, 实现对平衡态和偏置势采

样数据在训练中的有效使用. 该方法被 Rydzewski 和 Valsson 应用到 Müller-Brown Potential 以及丙氨酸二肽和四肽体系, 也已被整合到开源的 PLUMMD 软件包 (<https://www.plumed-nest.org/eggs/21/023/>). 类似地, Belkacemi 等^[125] 发展了利用自编码器的自由能偏置势迭代学习 (free energy biasing and iterative learning with auto encoders, FEBILAE), 该方法可以对在平衡态或者偏置势下采样的轨迹重配权重后作为自编码器的输入 (既可以是原来构象空间的, 也可以是某种转换之后的构型). 其中自编码器的瓶颈层确定了 CV 的维度, 但显然需要自行选择, 他们也给出了探索的建议. 可能的问题是迭代收敛的 CV 并不能保证自由能地貌图的全局充分采样. 和大多数类似研究一样, 这类编码过程不具备直接可解释性, 人们无从知道输入构型中不同的参数对 CV 的贡献. 虽然原则上可以间接从计算过程中的自动微分步骤获取一定信息, 但目前所有的方法中没有提供这种分析. 针对这个问题, Kikutsuji 等^[126] 利用模型无关的局部解释 (local interpretable model agnostic explanation, LIME) 和沙普利加和解释 (shapley additive explanations, SHAP) 框架, 给出各个输入量对 RC 的贡献, 能够在一定程度上增进我们对体系的直观物理认知.

Sun 等^[127] 发展了由一个降维编码器, 构象分类器和势能预测器组成的多任务 CV 学习构架, 在几个简单测试系统 (包括 5D Müller Brown model、丙氨酸二肽和金 (110) 晶面重建单元反应体系) 与单目标训练优化相比较展示了一定优势. 与很多应用中系统演化过程在原有高维空间进行不同, 隐空间模拟器^[128] (latent space simulator, LSS) 在训练产生编码器和解码器后, 在 CV 空间快速展开系统演化, 然后通过解码器生成原有高维空间的细粒度轨迹. 这些在隐空间或者集合变量空间进行操作的思路是很多工作中利用自编码器的重要方式. 比大多数方法在诸如丙氨酸二肽或类似模型体系中展示更进一步的是该方法在两个较大体系 (264 残基的 PROTAC 蛋白和 DNA 序列 5'-GCCGTTTCCGC-3' 对应的双螺旋结构) 获得了较为成功的应用.

Jung 等^[129] 以水溶液中离子的聚集和聚合物折叠为例, 集成了深度学习和过渡路径理论实现了复杂分子体系自组织模型的构建、验证和更新, 并

在此基础上通过符号回归总结出更容易理解的可观测量连接, 是分子复杂体系的深度学习和可解释性方面有意义的尝试和进步. 比变分求解自由能上界更进一步, Zhao 和 Wang^[130] 用流匹配 (flow matching) 同时求解上下界, 从而提供更好地逼近目标体系自由能的可能途径.

鉴于生成式模型在语言图像绘画等方面的巨大成功^[131], Janson 等^[132] 基于生成对抗模型和 transformer 架构训练的构象系综生成神经网络成功产生了训练数据集中没有的内秉无序蛋白 (IDP) 构象, 该过程与分子模拟直接采样相比所用计算代价非常小, 不过正确性依然有待进一步在更多体系中验证.

5 结 论

综上所述, 变分方法处理分子体系自由能地貌图目前已经有了较多不同视角的尝试, 但还都限于在较为简单的体系探索, 和其他理论上不甚严格的刻画分子体系自由能地貌图的神经网络方法相比也还没有展示出明显的系统优势. 比如使用变分的 VAMPNets^[67] 和使用回归^[119] 两种方法在丙氨酸二肽体系中就没有明显的表现差异. 不过变分更严格的理论基础有可能会让误差控制更加容易, 也很可能会在将来较大分子体系的应用和进一步发展体现出更多的优势. 从理论方法的角度, 现有的这些不同变分目标函数都是为了更好地逼近分子体系自由能地貌图的准确描述, 如何将它们集成并能够依据应用需求灵活选择关注视角显然是个有价值的任务. 当前的变分和自编码器模型中还有很多需要人工调节和尝试的环节, 最为突出的就是目前的所有方法都不能通过自主学习优化获得自编码器中间低维隐空间的适当维度. 另外变分计算本身原则上也可以在神经网络中数值实现, 从而有可能增加灵活性和可泛化能力, 不过目前尚没有见到这类尝试, 有可能是个有价值的发展方向.

从应用的角度, 目前最迫切需要解决的问题可能是将这些变分构建向更大更复杂分子体系的延伸. 从自由能地貌图构象空间的层次来看, 超过两个时空尺度的体系显然会带来更多挑战, 在同一个自由能地貌图时空尺度层次上, 多个亚稳态之间过渡路径交汇的可能性和准确处理也有待解决. 这些问题的可靠处理在较大的复合体分子机器的

理解中很有必要.

目前大模型的应用如火如荼^[133], 不过在 AI 的科学应用领域尚没有发力. 主要原因之一是作为通用大模型训练素材的语音图像材料非常丰富, 而特定科学领域的数据一般都不够丰富或者很多都难以理解. 不过这些模型集成多模态的能力显然对 AI 在广泛科学应用中和特定的复杂分子体系中都有参考价值. 已有的这些变分构建方法, 还有未来可能出现的其他新颖构建, 很可能在将来被统一到一个多目标大模型中.

参考文献

- [1] Thomas C, Tampe R 2020 *Annu. Rev. Biochem.* **89** 605
- [2] Jiang F, Doudna J A 2017 *Annu. Rev. Biophys.* **46** 505
- [3] Latorraca N R, Venkatakrishnan A J, Dror R O 2017 *Chem. Rev.* **117** 139
- [4] Wei G, Xi W, Nussinov R, Ma B 2016 *Chem. Rev.* **116** 6516
- [5] Dignon G L, Best R B, Mittal J 2020 *Annu. Rev. Phys. Chem.* **71** 53
- [6] Choi J M, Holehouse A S, Pappu R V 2020 *Annu. Rev. Biophys.* **49** 107
- [7] Spöner J, Bussi G, Krepl M, et al. 2018 *Chem. Rev.* **118** 4177
- [8] Bussi G, Laio A 2020 *Nat. Rev. Phys.* **2** 200
- [9] Mobley D L, Gilson M K 2017 *Annu. Rev. Biophys.* **46** 531
- [10] Rodnina M V, Beringer M, Wintermeyer W 2007 *Trends Biochem. Sci.* **32** 20
- [11] Bernardi R C, Melo M C R, Schulten K 2015 *Biochim. Biophys. Acta* **1850** 872
- [12] Sugita Y, Okamoto Y 1999 *Chem. Phys. Lett.* **314** 141
- [13] Faraldo-Gomez J D, Roux B 2007 *J. Comput. Chem.* **28** 1634
- [14] Laio A, Parrinello M 2002 *Proc. Natl. Acad. Sci. U. S. A.* **99** 12562
- [15] Barducci A, Bussi G, Parrinello M 2008 *Phys. Rev. Lett.* **100** 020603
- [16] Maragliano L, Vanden-Eijnden E 2006 *Chem. Phys. Lett.* **426** 168
- [17] Abrams J B, Tuckerman M E 2008 *J. Phys. Chem. B* **112** 15742
- [18] Darve E, Rodriguez-Gomez D, Pohorille A 2008 *J. Chem. Phys.* **128** 144120
- [19] Torrie G M, Valleau J P 1977 *J. Comput. Phys.* **23** 187
- [20] Carter E A, Ciccotti G, Hynes J T, Kapral R 1989 *Chem. Phys. Lett.* **156** 472
- [21] Sprik M, Ciccotti G 1998 *J. Chem. Phys.* **109** 7737
- [22] Zwanzig R W 1954 *J. Chem. Phys.* **22** 1420
- [23] Kirkwood J G 1935 *J. Chem. Phys.* **3** 300
- [24] Oberhofer H, Dellago C, Geissler P L 2005 *J. Phys. Chem. B* **109** 6902
- [25] Chen M, Cuendet M A, Tuckerman M E 2012 *J. Chem. Phys.* **137** 024102
- [26] Lesage A, Lelievre T, Stoltz G, Henin J 2017 *J. Phys. Chem. B* **121** 3676
- [27] Tribello G A, Gasparotto P 2019 *Front. Mol. Biosci.* **6** 46
- [28] Comer J, Gumbart J C, Henin J, Lelievre T, Pohorille A, Chipot C 2015 *J. Phys. Chem. B* **119** 1129
- [29] Darve E, Pohorille A 2001 *J. Chem. Phys.* **115** 9169
- [30] Huber T, Torda A E, van Gunsteren W F 1994 *J. Comput. Aided. Mol. Des.* **8** 695
- [31] Wang F, Landau D P 2001 *Phys. Rev. Lett.* **86** 2050
- [32] Valsson O, Tiwary P, Parrinello M 2016 *Annu. Rev. Phys. Chem.* **67** 159
- [33] Husic B E, Pande V S 2018 *J. Am. Chem. Soc.* **140** 2386
- [34] Dellago C, Bolhuis P G, Csajka F S, Chandler D 1998 *J. Chem. Phys.* **108** 1964
- [35] Bolhuis P G, Chandler D, Dellago C, Geissler P L 2002 *Annu. Rev. Phys. Chem.* **53** 291
- [36] van Erp T S, Moroni D, Bolhuis P G 2003 *J. Chem. Phys.* **118** 7762
- [37] Moroni D, Bolhuis P G, van Erp T S 2004 *J. Chem. Phys.* **120** 4055
- [38] Hummer G 2004 *J. Chem. Phys.* **120** 516
- [39] Bolhuis P G, Swenson D W H 2021 *Front. Data Comput.* **4** 2000237
- [40] E W, Vanden-Eijnden E 2010 *Annu. Rev. Phys. Chem.* **61** 391
- [41] Sarich M, Banisch R, Hartmann C, Schütte C 2013 *Entropy* **16** 258
- [42] Cybenko G 1989 *Math. Control Signal Syst.* **2** 303
- [43] Leshno M, Lin V Y, Pinkus A, Schocken S 1993 *Neural Netw.* **6** 861
- [44] Zhou D X 2020 *Appl. Comput. Harmon. Anal.* **48** 787
- [45] Alzubaidi L, Zhang J, Humaidi A J, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaria J, Fadhel M A, Al-Amidie M, Farhan L 2021 *J. Big Data* **8** 53
- [46] He K, Zhang X, Ren S, Sun J 2016 *IEEE Conference on Computer Vision and Pattern Recognition* Las Vegas, USA, 27–30 June, 2016 pp770–778
- [47] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, Polosukhin I 2017 *Advances in Neural Information Processing Systems* Long Beach, USA, December 4–9, 2017
- [48] Ho J, Jain A, Abbeel P 2020 *Advances in Neural Information Processing Systems* Virtual pp6840–6851
- [49] Baydin A G, Pearlmutter B A, Radul A A, Siskind J M 2018 *J. Mach. Learn. Res.* **18** 1
- [50] Rumelhart D, Hinton G, Williams R 1986 *Nature* **323** 533
- [51] Michelucci U 2022 arXiv: 1312.6114 [stat. ML]
- [52] Kingma D P, Welling M 2019 arXiv: 1906.02691 [cs. LG]
- [53] Waterfall J J, Casey F P, Gutenkunst R N, Brown K S, Myers C R, Brouwer P W, Elser V, Sethna J P 2006 *Phys. Rev. Lett.* **97** 150601
- [54] Rumelhart D E, Hinton G E, Williams R J (Anderson J A, Rosenfeld E, ed) 1988 *Neurocomputing* (Vol. 1) (Cambridge: The MIT Press) pp696–700
- [55] Arfken G B, Weber H J, Harris F E 2011 *Mathematical Methods for Physicists: A Comprehensive Guide* (Cambridge: Academic Press)
- [56] Blei D M, Kucukelbir A, McAuliffe J D 2017 *J. Am. Stat. Assoc.* **112** 859
- [57] Ganguly A, Earp S W 2021 arXiv: 2108.13083 [cs. LG]
- [58] Marquardt D W 1963 *J. Soc. Ind. Appl. Math.* **11** 431
- [59] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L 2019 *Advances in Neural Information Processing Systems* pp8026–8037
- [60] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M 2016 *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* Savannah, GA, USA, November 2–4, 2016 pp265–283
- [61] Ma Y, Yu D, Wu T, Wang H 2019 *Front. Data Comput.* **1** 105

- [62] Hadji I, Wildes R P 2018 [arXiv: 1803.08834 \[cs. CV\]](#)
- [63] Ghorbani M, Prasad S, Klauda J B, Brooks B R 2022 *J. Chem. Phys.* **156** 184103
- [64] Mardt A, Hempel T, Clementi C, Noe F 2022 *Nat. Commun.* **13** 7101
- [65] Perez-Hernandez G, Paul F, Giorgino T, De Fabritiis G, Noe F 2013 *J. Chem. Phys.* **139** 015102
- [66] Wu H, Noé F 2019 *J. Nonlinear Sci.* **30** 23
- [67] Mardt A, Pasquali L, Wu H, Noe F 2018 *Nat. Commun.* **9** 5
- [68] Kleiman D E, Shukla D 2023 *J. Chem. Theory Comput.* **19** 4377
- [69] Chen H, Roux B, Chipot C 2023 *J. Chem. Theory Comput.* **19** 4414
- [70] Schütte C, Fischer A, Huisinga W, Deuffhard P 1999 *J. Comput. Phys.* **151** 146
- [71] He Z, Chipot C, Roux B 2022 *J. Phys. Chem. Lett.* **13** 9263
- [72] Bonati L, Zhang Y Y, Parrinello M 2019 *Proc. Natl. Acad. Sci. U. S. A.* **116** 17641
- [73] Bittracher A, Mollenhauer M, Koltai P, Schütte C 2023 *Multiscale Model. Simul.* **21** 449
- [74] Wang Y, Ribeiro J M L, Tiwary P 2019 *Nat. Commun.* **10** 3573
- [75] Beyerle E R, Mehdi S, Tiwary P 2022 *J. Phys. Chem. B* **126** 3950
- [76] Zhang J, Lei Y K, Yang Y I, Gao Y Q 2020 *J. Chem. Phys.* **153** 174115
- [77] Kingma D P, Welling M 2013 [arXiv: 1312.6114 \[stat. ML\]](#)
- [78] Tiwary P, Berne B J 2016 *Proc. Natl. Acad. Sci. U. S. A.* **113** 2839
- [79] Wu H, Paul F, Wehmeyer C, Noe F 2016 *Proc. Natl. Acad. Sci. U. S. A.* **113** E3221
- [80] Wu H, Mey A S, Rosta E, Noé F 2014 *J. Chem. Phys.* **141** 214106
- [81] Chodera J D, Swope W C, Noé F, Prinz J H, Shirts M R, Pande V S 2011 *J. Chem. Phys.* **134** 244107
- [82] Prinz J H, Chodera J D, Pande V S, Swope W C, Smith J C, Noe F 2011 *J. Chem. Phys.* **134** 244108
- [83] Rosta E, Hummer G 2015 *J. Chem. Theory Comput.* **11** 276
- [84] Mey A S, Wu H, Noé F 2014 *Phys. Rev. X* **4** 041018
- [85] Hinrichs N S, Pande V S 2007 *J. Chem. Phys.* **126** 244101
- [86] Noe F 2008 *J. Chem. Phys.* **128** 244103
- [87] Chodera J D, Noé F 2010 *J. Chem. Phys.* **133** 265
- [88] Schütt K, Kindermans P J, Saucedo Felix H E, Chmiela S, Tkatchenko A, Müller K R 2017 *Advances in Neural Information Processing Systems* Long Beach, ACM, USA, 2017 pp991–1001
- [89] Husic B E, Charron N E, Lemm D, Wang J, Perez A, Majewski M, Kramer A, Chen Y, Olsson S, de Fabritiis G, Noe F, Clementi C 2020 *J. Chem. Phys.* **153** 194101
- [90] Battaglia P W, Hamrick J B, Bapst V, et al. 2018 [arXiv: 1806.01261 \[stat. ML\]](#)
- [91] Kipf T N, Welling M 2016 [arXiv: 1609.02907 \[cs. LG\]](#)
- [92] Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y 2017 [arXiv: 1710.10903 \[stat. ML\]](#)
- [93] Ghorbani M, Prasad S, Klauda J B, Brooks B R 2022 [arXiv:2201.04609 \[physics.comp-ph\]](#)
- [94] Hempel T, Del Razo M J, Lee C T, Taylor B C, Amaro R E, Noe F 2021 *Proc. Natl. Acad. Sci. U. S. A.* **118** e2105230118
- [95] Maragliano L, Fischer A, Vanden-Eijnden E, Ciccotti G 2006 *J. Chem. Phys.* **125** 24106
- [96] Pan A C, Sezer D, Roux B 2008 *J. Phys. Chem. B* **112** 3432
- [97] Weinan E, Ren W, Vanden-Eijnden E 2005 *Chem. Phys. Lett.* **413** 242
- [98] Branduardi D, Gervasio F L, Parrinello M 2007 *J. Chem. Phys.* **126** 054103
- [99] Leines G D, Ensing B 2012 *Phys. Rev. Lett.* **109** 020601
- [100] Invernizzi M, Parrinello M 2020 *J. Phys. Chem. Lett.* **11** 2731
- [101] Berezhkovskii A, Szabo A 2005 *J. Chem. Phys.* **122** 14503
- [102] Langer J S 1969 *Ann. Phys.* **54** 258
- [103] Valsson O, Parrinello M 2014 *Phys. Rev. Lett.* **113** 090601
- [104] Bilonis I, Koutsourelakis P S 2012 *J. Comput. Phys.* **231** 3849
- [105] Dempster A P, Laird N M, Rubin D B 2018 *J. R. Stat. Soc. B* **39** 1
- [106] Bonati L, Piccini G, Parrinello M 2021 *Proc. Natl. Acad. Sci. U.S.A.* **118** e2113533118
- [107] Tishby N, Pereira F C, Bialek W 2000 [arXiv: physics/0004057 \[physics.data-an\]](#)
- [108] Still S 2014 *Entropy* **16** 968
- [109] Song Y, Kingma D P 2021 [arXiv: 2101.03288 \[cs. LG\]](#)
- [110] Arjovsky M, Chintala S, Bottou L 2017 *International Conference on Machine Learning* Sydney pp214–223
- [111] Huang Y P, Xia Y, Yang L, Wei J, Yang Y I, Gao Y Q 2021 *Chin. J. Chem.* **40** 160
- [112] Ribeiro J M L, Bravo P, Wang Y, Tiwary P 2018 *J. Chem. Phys.* **149** 072301
- [113] Chen M 2021 *Eur. Phys. J. B* **94** 211
- [114] Qiu Y, O'Connor M S, Xue M, Liu B, Huang X 2023 *J. Chem. Theory Comput.* **19** 4728
- [115] Monroe J I, Shen V K 2022 *J. Chem. Theory Comput.* **18** 3622
- [116] Ma A, Dinner A R 2005 *J. Phys. Chem. B* **109** 6769
- [117] Chen W, Ferguson A L 2018 *J. Comput. Chem.* **39** 2079
- [118] Chen H, Liu H, Feng H, Fu H, Cai W, Shao X, Chipot C 2022 *J. Chem. Inf. Model.* **62** 1
- [119] Wehmeyer C, Noe F 2018 *J. Chem. Phys.* **148** 241703
- [120] Williams M O, Kevrekidis I G, Rowley C W 2015 *J. Nonlinear Sci.* **25** 1307
- [121] Mezić I 2005 *Nonlinear Dyn.* **41** 309
- [122] H. Tu J, W. Rowley C, M. Luchtenburg D, L. Brunton S, Nathan Kutz J 2014 *J. Comput. Dynam.* **1** 391
- [123] Zhang J, Chen M 2018 *Phys. Rev. Lett.* **121** 010601
- [124] Rydzewski J, Valsson O 2021 *J. Phys. Chem. A* **125** 6286
- [125] Belkacemi Z, Gkeka P, Lelievre T, Stoltz G 2022 *J. Chem. Theory Comput.* **18** 59
- [126] Kikutsuji T, Mori Y, Okazaki K I, Mori T, Kim K, Matubayasi N 2022 *J. Chem. Phys.* **156** 154108
- [127] Sun L, Vandermause J, Batzner S, Xie Y, Clark D, Chen W, Kozinsky B 2022 *J Chem Theory Comput* **18** 2341
- [128] Wang Y, Lamim Ribeiro J M, Tiwary P 2020 *Curr. Opin. Struct. Biol.* **61** 139
- [129] Jung H, Covino R, Arjun A, Leitold C, Dellago C, Bolhuis P G, Hummer G 2023 *Nat. Comput. Sci.* **3** 334
- [130] Zhao L, Wang L 2023 *Chin. Phys. Lett.* **40** 120201
- [131] Wu T, He S, Liu J, Sun S, Liu K, Han Q L, Tang Y 2023 *IEEE-CAA J. Automatica Sin.* **10** 1122
- [132] Janson G, Valdes-Garcia G, Heo L, Feig M 2023 *Nat. Commun.* **14** 774
- [133] Naveed H, Ullah Khan A, Qiu S, Saqib M, Anwar S, Usman M, Akhtar N, Barnes N, Mian A 2023 [arXiv: 2307.06435 \[cs. CL\]](#)

SPECIAL TOPIC—Machine learning in biomolecular simulations

Variational analysis and AI algorithm implementation of free energy landscapes of molecular system^{*}

Du Bo-Chuan¹⁾ Tian Pu^{1)2)†}¹⁾ (*School of Life Sciences, Jilin University, Changchun 130012, China*)²⁾ (*School of Artificial Intelligence, Jilin University, Changchun 130012, China*)

(Received 14 November 2023; revised manuscript received 18 January 2024)

Abstract

Accurate description of the free energy landscape (FES) is the basis for understanding complex molecular systems, and for further realizing molecular design, manufacture and industrialization. Major challenges include multiple metastable states, which usually are separated by high potential barriers and are not linearly separable, and may exist at multiple levels of time and spatial scales. Consequently FES is not suitable for analytical analysis and brute force simulation. To address these challenges, many enhanced sampling methods have been developed. However, utility of them usually involves many empirical choices, which hinders research advancement, and also makes error control very unimportant. Although variational calculus has been widely applied and achieved great success in physics, engineering and statistics, its application in complex molecular systems has just begun with the development of neural networks. This brief review is to summarize the background, major developments, current limitations, and prospects of applying variation in this field. It is hoped to facilitate the AI algorithm development for complex molecular systems in general, and to promote the further methodological development in this line of research in particular.

Keywords: variation, neural networks, complex molecular system, free energy landscape**PACS:** 87.80.-y, 87.15.A-**DOI:** 10.7498/aps.73.20231800

^{*} Project supported by the Interdisciplinary Integration and Innovation Project of JLU, China (Grant No. JLUXXJC2021ZZ05).

[†] Corresponding author. E-mail: tianpu@jlu.edu.cn



分子体系自由能地貌图的变分分析及AI算法实现

杜泊船 田圃

Variational analysis and AI algorithm implementation of free energy landscapes of molecular system

Du Bo-Chuan Tian Pu

引用信息 Citation: *Acta Physica Sinica*, 73, 068702 (2024) DOI: 10.7498/aps.73.20231800

在线阅读 View online: <https://doi.org/10.7498/aps.73.20231800>

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

基于忆容器件的神经形态计算研究进展

Research progress of neuromorphic computation based on memcapacitors

物理学报. 2021, 70(7): 078701 <https://doi.org/10.7498/aps.70.20201632>

基于机器学习构建的环三亚甲基三硝胺晶体势

Energetic potential of hexogen constructed by machine learning

物理学报. 2020, 69(23): 238702 <https://doi.org/10.7498/aps.69.20200690>

基于深度学习的联合变换相关器光学图像加密系统去噪方法

In depth learning based method of denoising joint transform correlator optical image encryption system

物理学报. 2020, 69(24): 244204 <https://doi.org/10.7498/aps.69.20200805>

等离子环境中带电体能量的Collin变分

Collin variational study of charged conductors' energy in plasma environment

物理学报. 2018, 67(20): 209401 <https://doi.org/10.7498/aps.67.20180504>

电磁诱导透明暗孤子的耗散变分束缚分析

Dissipative variational analysis for bounded dark solitons of electromagnetically induced transparency

物理学报. 2018, 67(6): 064207 <https://doi.org/10.7498/aps.67.20172567>

基于物理信息神经网络的光斑质心计算

Calculation of spot centroid based on physical informed neural networks

物理学报. 2022, 71(20): 200601 <https://doi.org/10.7498/aps.71.20220670>