arXiv:2301.04835v2 [physics.comp-ph] 4 Apr 2023

PySAGES：利用 GPU 加速的灵活高级采样方法

Pablo F. Zubieta Rico[a]、Ludwig Schneider[a]、Gustavo R. Pérez-Lemus[a]、Riccardo Alessandri[a]、Siva Dasetty[a]、Cintia A. Menéndez[a]、Yiheng Wu[a]、Yezhi Jin[a]、Yinan Xu[a]、Trung D. Nguyen[b]、John A. Parker[b]、Andrew L. Ferguson[a]、Jonathan K. Whitmer[c]、Juan J. de Pablo[a]

*a*芝加哥大学普利兹克分子工程学院，芝加哥南埃利斯大道*5640*号，伊利诺伊州芝加哥，邮编*60637*，美国 *b*芝加哥大学研究计算中心，芝加哥南德雷塞尔大道*6054*号，伊利诺伊州芝加哥，邮编*60637*，美国 *c*圣母大学化学与生物分子工程系，印第安纳州圣母镇纽兰德楼*250*号，邮编*46556*，美国

摘要

分子模拟是物理学、化学和生物学研究中的重要工具。通过提供对高级采样方法和技术的支持，使研究人员能够计算相关的底层自由能景观，模拟的能力可以得到大幅提升。因此，开发能够无缝适用于广泛复杂系统的软件至关重要。在以往为高级采样提供开源、由社区支持的软件的努力基础上，我们推出 PySAGES：一种以 Python 实现的高级通用系综模拟软件套件（SSAGES），为分子动力学模拟中的大规模并行增强采样方法应用提供完整的图形处理器支持，例如自适应偏置力、谐和偏置或前向通量采样。PySAGES 库提供直观的界面，便于管理系统配置、纳入新的集体变量，并实现复杂的基于自由能的采样方法，因而可作为开发和实现新兴模拟技术的通用平台。通过针对不同类别分子系统的清晰简洁示例，我们展示了这一新工具的功能、核心特性和计算性能。我们期望 PySAGES 能为科学界提供一个稳健且易于使用的平台，以加速模拟、改善采样，并便捷地估算各种材料和过程的自由能。

关键词：增强采样方法、GPU 加速

1. 引言

分子模拟广泛应用于众多科学与工程领域 [1]。随着分子模拟越来越多地用于发现新现象和解读复杂的实验测量，所研究系统的复杂性也随之增加。经典原子分子动力学（分子动力学）模拟通常受限于微秒级时间尺度和数十纳米的长度尺度。对于自由能景观崎岖的系统，这样的时间尺度可能不足以确保对相关相空间进行充分采样，因此必须采用先进方法来克服自由能垒。在这方面，识别恰当的集体变量（集体变量）十分有用，也日益普遍。这些变量通常是系统原子坐标的可微函数；随后，可以施加偏置来探索由这些集体变量定义的空间，从而克服能垒，并增强对热可及相空间的采样。

图形处理器或 TPU 等硬件加速器，以及专为快速分子动力学计算设计的专用硬件迅速发展 [2, 3], ，为研究人员提供了更多机会，得以对更大的系统进行更长时间的模拟。尤其是 GPU，提供了一种广泛可用的快速模拟方案；如今，已有多个软件包可在此类设备上进行分子动力学模拟，例如 HOOMD-蓝色[4], OpenMM [5], JAX MD [6, 7],LAMMPS [8], 和 Gromacs [9], 。

如上所述，增强采样方法旨在克服系统中分隔多个亚稳态的高能垒，同时便于计算相关热力学量关于不同集体变量的函数，例如自由能面（FES）。PLUMED [10], Colvars [11], 以及我们自己的 SSAGES 软件包 [12], 等多个库，都提供了开箱即用的解决方案，可用于执行增强采样分子动力学模拟。

在文献中已有的各种增强采样方法中，近期提出的一些方案依赖机器学习（ML）策略来近似自由能面及其梯度（广义力） [13, 14, 15, 16]。同样，用于识别与高方差或慢自由度（DOF）相关的有意义集体变量（CV）的算法，也基于深度神经网络 [17, 18, 19, 20, 21, 22]。这些

进展凸显了将机器学习框架无缝
w可与现有的分子动力学软件库配合使用。

截至目前，尚无解决方案能够将增强采样技术、硬件加速和机器学习框架结合起来，以促进增强采样分子动力学模拟在GPU上运行。尽管一些分子动力学库支持GPU并提供有限的增强采样方法 [5, 9, 23, 24, 25], ，但目前还没有任何软件包能够让用户在同一平台中，以与PLUMED 和SSAGES为CPU-基础分子动力学模拟所提供的相同后端无关方式，利用所有这些功能。

本文介绍 PySAGES，即用于高级通用系综模拟的 Python 套件。这是一款以 Python 编写、基于 JAX 的免费开源软件包，其设计理念遵循 SSAGES，使用户能够轻松地在中央处理器、GPU 和 TPU 上执行增强采样分子动力学模拟。目前，PySAGES 可与 HOOMD-blue、OpenMM、JAX MD 和 ASE 结合使用，并可通过 ASE 进一步与 CP2K、Quantum ESPRESSO、VASP 和 Gaussian 等软件结合使用。目前，PySAGES 提供以下增强采样方法：伞形采样、元动力学、良温元动力学、前向通量采样、字符串方法、自适应偏置力、人工神经网络采样、使用神经网络的自适应偏置力、组合力频率和谱自适应偏置力。PySAGES 还包含一些最常用的集体变量；更重要的是，只要能够用 JAX 提供的 NumPy [26] 界面表示，定义新的集体变量就相对简单。所有集体变量都可以通过 JAX 函数变换自动求导。PySAGES 具有高度模块化的特点，因此即使是在面向用户的脚本中，也能随着新方法的出现轻松实现这些方法。

在以下各节中，我们将概述 PySAGES 的设计与实现，并通过一系列示例展示其在不同应用领域中解决研究问题的灵活性。我们还将讨论其在 GPU 上的性能，并介绍一些关于如何通过未来开发以及社区参与和贡献来扩展和改进该软件包、以覆盖更多研究用例的展望。

2. 实现

我们先简要介绍 PySAGES 的核心分量、它们如何协同工作，以及与各个后端的通信如何使 PySAGES

在运行时对模拟施加偏置。图中概述了PySAGES 的执行工作流程，并展示了用户界面与模拟主要阶段及其与后端交互之间的对应关系图1。

为了在尽量减少对既有工作流干扰的同时提供统一的用户界面，PySAGES 只需用户将传统的后端脚本代码封装到模拟生成器函数中。此方法兼容PySAGES所支持的不同模拟后端之间各不相同的 Python 接口。图2展示了一个模拟生成器函数示例，以及如何修改传统的 OpenMM 脚本以执行增强采样分子动力学模拟。图2。

模拟开始时，会调用模拟生成器函数，实例化所需数量的模拟副本。随后，PySAGES 会针对每个副本查询粒子信息以及后端将使用的设备。此外，在这一初始阶段，PySAGES 还会通过 JAX 的 `gradtransform` 对集体变量进行自动微分，以便估算偏置力，并为用户声明的采样方法生成专用的初始化和更新例程。

与 SSAGES 一样，PySAGES 会将模拟信息封装到一个名为 `Snapshot` 的对象中。该对象以不依赖后端和设备的格式提供最重要的模拟信息，例如粒子位置、速度和力。为此，PySAGES 使用 DLPack [27]，让基于 C++ 的分子动力学库能够直接访问后端分配的缓冲区中的各种粒子属性内容，并尽可能避免复制数据。

模拟和采样方法的设置完成后，PySAGES 会将控制权交还给后端，后端随后运行指定数量的时间步，或直到满足其他停止条件。为实现信息的双向交换，PySAGES 会向后端添加一个类似力的对象或函数，并在时间积分过程中调用它。在此过程中，采样方法的状态会得到更新，计算出的偏置力也会添加到后端的合力中。

最后，采样方法收集的信息会被返回，可用于根据选定的集体变量计算自由能。与 SSAGES 不同，PySAGES 提供了易于使用的 `analyze` 界面，简化了模拟后分析流程，包括根据所选采样方法自动计算自由能。这项功能可以大幅减少从模拟中获取有价值见解所需的时间和精力。

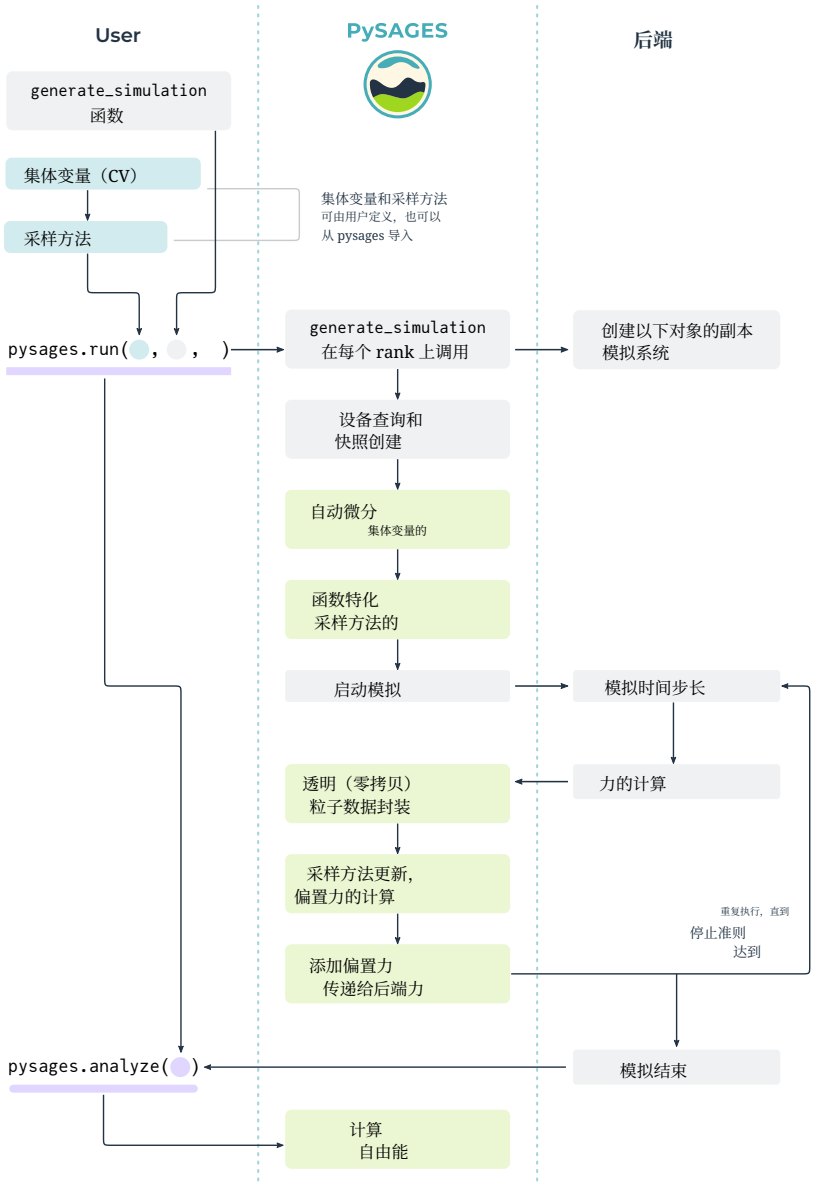

图1：PySAGES 模拟流程图。用户为模拟编写脚本，声明要使用的集体变量和采样方法。



图2：PySAGES Python接口的使用示例。借助 PySAGES，只需对代码进行少量修改，即可轻松扩展现有的分子动力学脚本，以执行增强采样分子动力学。通常，用户只需将定义模拟系统的代码封装到模拟生成器 函数中。

PySAGES 提供了一种简单的方式，可通过统一的前端利用不同的并行计算框架（包括 MPI）来运行增强采样模拟。这是通过 Python 的 `concurrent.futures`接口实现的。具体而言，对于 MPI 并行化，用户只需向 PySAGES 的 `run` 方法传入一个额外的 `MPIPoolExecutor`（来自 `mpi4py`）即可。如果用户选择 `UmbrellaSampling` 之类的方法，每个图像的工作负载就会分配到可用的 MPI 节点上。另一方面，对于大多数采样方法，并行化接口允许用户对同一系统运行多个副本，例如，以分析计算给定系统自由能时的不确定性。

为确保我们实现的可复现性和正确性，并遵循软件工程最佳实践，我们实现了一套全面的单元测试套件，并利用 GitHub 的持续集成服务。此外，我们还使用 `trunk.io` [28] 来遵循质量标准，并促进开发人员之间的协作。

2.1. 增强采样方法

虽然我们假设读者对增强采样方法有一定的基础了解，但在此我们仍对这些技术进行概述。对于希望进一步了解增强采样基本原理的读者，我们推荐阅读若干优秀的近期综述文章 [29, 30, 31,32, 33, 34, 35, 36, 37, 38]。此外，我们还讨论

了增强采样方法在 PySAGES 中的实现总体结构，并概述了该库中已提供的各种方法。

增强采样方法是一类模拟技术，通过操控常规分子动力学模拟，更有效地采样构型空间。在分子动力学中，集体变量 ξ 通常是所有粒子位置的函数 $\xi(\{r_i\})$。

对于给定的统计系综（例如正则 NVT系综），相应的自由能可写为 $A = -k_\mathrm{B}T\ln(Z)$，其中 A 是亥姆霍兹自由能，Z 是正则配分函数。为了明确自由能对 ξ 的依赖关系，我们写出配分函数：

$$Z(\xi) \propto \int \mathrm{d}^N r_i \; \delta(\hat{\xi}(\{r_i\}) - \xi) \; e^{-U(\{r_i\})/k_\mathrm{B}T} \tag{1}$$

对该配分函数进行归一化，可得集体变量子空间中构型出现的概率 $p(\xi) = Z(\xi)/(\int \mathrm{d}\xi \; Z(\xi))$。将该概率代入自由能表达式，得到：

$$A(\xi) = -k_\mathrm{B}T\ln(p(\xi)) + C \tag{2}$$

其中，C 是一个常数。

如果对自由能关于 ξ 求导，我们得到

$$\frac{dA(\xi)}{d\xi} = \frac{\int \mathrm{d}^N r_i \; \frac{dU}{d\xi} \; \delta(\hat{\xi}(\{r_i\}) - \xi) \; e^{-U(\{r_i\})/k_\mathrm{B}T}}{\int \mathrm{d}^N r_i \; \delta(\hat{\xi}(\{r_i\}) - \xi) \; e^{-U(\{r_i\})/k_\mathrm{B}T}} = \left\langle \frac{dU}{d\xi} \right\rangle_\xi, \tag{3}$$

其中 $\langle \ldots \rangle_\xi$ 表示条件平均。

基于集体变量的增强采样方法旨在以计算上可行的方式准确确定 $p(\xi)$ 或 $dA(\xi)/d\xi$，并据此恢复 $A(\xi)$。

在 PySAGES 中，采样方法的实现遵循 JAX 函数式编程模型。新方法通过将 `SamplingMethod` 类子类化来实现，并且必须定义一个 `build` 方法。此方法会返回两个方法：`initialize` 和 `update`，用于模拟偏置过程的一部分。对于熟悉 JAX MD 的读者，可以将它们视为 JAX MD 的 `simulate` 积分方法所返回的高层函数的类似物。`initialize` 方法会分配所有必要的辅助对象，并将其存储在 `State` 数据结构中；而 `update` 方法则利用任意时刻模拟中的信息来更新 `State`。

PySAGES 既允许在用于设置分子动力学模拟的 Python 脚本中无缝编写新方法，也开箱即用地提供了几种最重要的已知采样方法的实现。下面将列出这些方法并作简要介绍。

2.1.1. 谐和偏置

对相空间中某个特定区域进行采样的一种简单方法，是围绕某一点 ξ_0 对模拟施加谐和偏置。这会向哈密顿量添加一个二次势能项，当系统偏离目标点时，该项会使势能增加：$\mathcal{H}_b = \mathcal{H} + k/2(\xi - \xi_0)^2$, 其中$k > 0$为弹簧常数。通过将偏置后的概率分布除以已知的偏置权重 $p(\xi) = p_b(\xi)/e^{-k}/2(\xi-\xi_0)^2/k_\mathrm{B}T$，即可恢复无偏的概率分布 $p(\xi)$ 。

这种方法的缺点是，只能用于探索相空间中已知点附近的自由能景观。对于许多系统而言，这可能并不足够，因为其自由能景观十分复杂。

2.1.2. 伞形采样

伞形采样是一种传统技术，通过组合多个谐和偏置模拟来构建谐和偏置。它是一种广为人知的方法，可用于探索相空间中的已知路径，并获得沿该路径的自由能剖面 [39, 40]。通常，两个感兴趣点之间的路径由 N 相空间中的点 ξ_i来描述。在这些点中的每一个点上，都会进行一次谐和偏置模拟，然后将得到的出现频数直方图合并，以获得单一的自由能剖面。

在 PySAGES 中，我们为多维集体变量实现了伞形积分。该方法近似作用于偏置点的力，并对这些力进行积分，以求得自由能剖面 A（ξ），从而能够探索复杂的高维自由能景观。

2.1.3. 改进的字符串方法

当只知道端点而不知道路径本身时，可以使用改进的（基于样条的）字符串方法，求出这两个端点之间的平均自由能路径（MFEP） [41]。基于样条的字符串方法通过使用三次样条对 MFEP 进行插值，改进了原有的字符串方法。在此方法中，路径上的中间点根据记录到的作用于这些点的平均力移动，但移动方向仅限于垂直于路径轮廓的方向。这样可以确保路径上各点之间的距离保持不变。

此方法已得到广泛应用，并被证明是有效求出最小自由能路径（MFEP）的方式，可用于寻找相空间中两点之间的 [41]。

2.1.4. 自适应偏置力采样

自适应偏置力（ABF）采样方法是一种用于绘制复杂自由能景观的技术。无需预先了解系统的势能即可应用该方法，因为它会沿积分路径对各点的自由能导数进行实时估算。ABF 通过向系统施加额外的力来偏置原子的运动，并在模拟过程中持续更新偏置力的大小和方向。在长时间极限下，这会产生一个沿所关注过渡坐标不受平均力作用的哈密顿量，从而形成平坦的自由能面，使系统能够展现加速动力学，进而提供可靠的自由能估算 [42, 43]。与 SSAGES 类似，PySAGES 中 ABF 的实现基于 [43] 中描述的算法。

2.1.5. 元动力学

元动力学是另一种用于增强复杂系统采样的常用方法。在元动力学中 [44], , 沿一个或多个集体变量施加由高斯函数构成的偏置势。用户控制这些高斯函数的高度和宽度（σ）。模拟过程中，会按照用户定义的时间间隔逐步累积添加高斯偏置势。在标准元动力学中，高斯偏置势的高度固定不变。

相比之下，在良温元动力学（WTMD） [45] 模拟中，每个时间步都会根据预设的温度相关偏置因子调整高斯偏置势的高度。与标准元动力学相比，WTMD 中高斯高度的这种缩放能够更快收敛，因为它将探索的自由能范围限制在由偏置因子确定的范围内。

在 PySAGES 中，我们实现了标准元动力学和 WTMD。用户设置偏置因子的值后，即可启用 well-tempered 变体。为了提升计算性能，我们新增了可选支持，可将偏置势存储在预定义网格上。用户可根据需求，以牺牲精度为代价换取更快的模拟。

2.1.6. 前向通量采样

前向通量FFS）属于与上述方法不同的一类增强采样方法。在前文介绍的方法中，通过对系统施加偏置，计算相空间（A）某一区域到目标区域（B）的自由能变化。而在FFS中不添加偏置，而是通过高效筛选轨迹，使其从相空间中的 A 到 B，从而完成采样。由于不使用偏置，系统的内禀动力学得以保留，因此可以研究转变路径的动力学和微观信息 [46]。在PySAGES 中，我们实现了FFS [47, 48]的直接版本。

2.1.7. 人工神经网络采样

人工神经网络采样（ANN） [13]采用正则化神经网络，根据集体变量空间各区域的访问直方图直接近似自由能，并生成可避免振铃和边界伪影的偏置力 [13], 。这些伪影常见于元动力学或基函数采样等方法 [49]。通过对采样不足的区域进行插值，并将偏置外推至新的、尚未探索的区域，这种方法能够有效地快速适应各种自由能景观。

PySAGES 的实现提供了比 SSAGES 更灵活的网络正则化方法；SSAGES 使用的是贝叶斯正则化。

2.1.8. 使用神经网络进行力偏置

使用神经网络进行力偏置（FUNN） [14] 基于与 ANN 相同的理念，即依靠人工神经网络提供连续函数来对模拟施加偏置；但它并不使用集体变量空间访问情况的直方图，而是在模拟推进过程中，利用平均力的 ABF估计值训练网络并更新其参数。该方法具备 ABF 的所有特性，同时它所生成的广义平均力平滑近似，使系统自由能的收敛速度远快于 ABF。

2.1.9. 组合力频率采样

组合力频率采样（CFF）方法 [15]结合了基于广义力的技术（如自适应偏置力或 FUNN）的速度优势，以及基于频率的方法（如元动力学或人工神经网络）的优点。相较于早期基于力的方法，它有诸多改进，包括无需使用超参数来抑制早期估计值、实现自动化

力的积分以生成自由能，并提供自由能随时间变化的显式表达式，从而能够使用副本交换或重新加权。

原则上，通过使用直方图的稀疏存储，可以将该方法扩展到更高维度，而不会遇到内存限制。不过，PySAGES 尚未实现这种优化。

2.1.10. 谱自适应偏置力

谱自适应偏置力 [50] 是一种遵循与基于神经网络的采样方法相同原理的方法，即构建自由能的连续近似。然而，与FUNN 等方法不同，它通过拟合指数收敛的基函数展开来实现这一点，可视为基函数采样方法的一种推广。与后者不同，且类似于CFF，它能够恢复系统自由能的显式表达式。无论在运行时间还是收敛速度方面，这种方法都极其高效。

2.2. 集体变量

如前所述，增强采样计算通常涉及选择一个集体变量。对于给定系统，合适的集体变量可以只是两组原子质心之间的距离，也可以是复杂的专用量。

下面列出 PySAGES 中预定义的一组集体变量，并按使用时所需的原子坐标组数排序：

1. `TwoPointCV`。此子类适用于定义需要两组坐标的集体变量，其中包括 `Distance` 和 `Displacement`（向量）。 2. `ThreePointCV`。这是包含三组原子的集体变量子类，例如`Angle`。 3. `FourPointCV`。这是包含四组原子的集体变量子类，例如`DihedralAngle`。 4. `AxisCV`。这是投影到指定轴上的集体变量子类，其中包括 `Component` 和 `PrincipalMoment`。 5. `CollectiveVariable`所有集体变量的通用基类。在 PySAGES 中，直接继承此类且不属于前述类别的集体变量包括：`RingPhaseAngle`、`RingAmplitude`、`RadiusofGyration`、`Asphericity`、`Acylindricity`、`ShapeAnisotropy`、`RingPuckeringCoordinates` [51]（向量）。

在 PySAGES 中，我们为用户提供了一个简单的框架来定义集体变量，并使用 JAX 自动微分。为了说明这一点，我们

比较如何编写一个集体变量的计算，该集体变量用于测量两个原子组之间的向量在一条轴上的投影，这条轴穿过另外两个原子组；我们分别在 SSAGES 和PySAGES 中进行比较（见图3）。在PySAGES 中，梯度计算会自动完成，而在SSAGES 中则必须显式编写代码。

利用人工神经网络（例如自编码器）发现的数据驱动且可微的集体变量，[29, 21, 18, 52, 53]基于原子的任意特征化，原则上可以根据上述集体变量通用抽象类在 PySAGES 中实现。

下面的第二个示例展示了微分编程在 PySAGES 中声明集体变量的强大能力。

2.2.1. 案例研究：用于界面的集体变量

当两种不互溶的液体彼此接触时，其中一种液体的密度会逐渐变化。这个过渡区域就是液-液界面，其位置在许多研究中都至关重要（参见第 3.1.3 节）。然而，确定此类界面的位置并非易事，因为它通常会随着模拟的进行而波动。作为界面的代表性集体变量，我们可以采用密度梯度达到最大值处的位置。更正式地说，令 $\rho(x)$ 表示感兴趣液体在垂直轴坐标 x 处的密度。我们希望确定界面的位置：

$$I = \arg\max_x |\rho'(x)| \tag{4}$$

然而，在分子模拟中，密度函数 ρ（x）无法直接测量，因为原子的坐标是离散的。为了近似求得 ρ（x），我们将坐标划分为多个分箱，每个分箱的宽度为 δ，并创建直方图 p（x），记录落入位置 x 附近分箱中的原子数。换句话说，

$$p(x) = \sum_{i=1\ldots n} \left[\, |x_i - x| < \delta/2 \,\right] \tag{5}$$

其中，x_i 表示原子 i 的坐标。如上所述，p（x）不可微。因此，与其他研究一样，[54]，我们采用高斯核进行核密度估计，以修改 p（x）。修改后的 $\tilde{p}$（x）定义如下：

$$\tilde{p}(x) = \sum_{i=1\ldots n} \exp\left(-\frac{(x_i - x)^2}{2\sigma^2}\right) \tag{6}$$



图3：在 PySAGES 中编写集体变量的示例。左侧是用 SSAGES 编写的相同集体变量，右侧则是 PySAGES 版本。通常，用户只需将集体变量编写为 JAX 中的可微函数即可。

其中， σ 是一个超参数，用于决定高斯核的宽度。然后，密度梯度可近似为：

$$\tilde{p}\,'(x) = \frac{\tilde{p}(x+\delta/2) - \tilde{p}(x-\delta/2)}{\delta} \tag{7}$$

并将界面的位置计算为 $I = \arg\max_x |\tilde{p}\,'(x)|$。arg max 运算符也是不可微的。因此，我们用一个 somax 函数替代它，将原始输入转换为概率。将这些分箱记作 $m\ \ j = 1 \ldots m$，最后将界面的位置计算为：

$$I = \frac{\sum_j x_j \, \exp |\tilde{p}\,'(x_j)|}{\sum_j \exp |\tilde{p}\,'(x_j)|} \tag{8}$$

如集体变量的代码片段（见附录 A）所示，PySAGES 可以简洁而直接地实现复杂的集体变量，例如这个集体变量。

3. 结果与讨论

要评估像 PySAGES 这样的软件包，至少必须考虑两个因素：物理正确性和计算性能。

首先，为了评估 PySAGES中实现的增强采样方法是否正确，我们在附录 B.1中展示了丙氨酸二肽（ ϕ 和 ψ 的二面角ADP）的自由能景观。这个例子常用于对新的增强采样算法进行基准测试。类似地，我们还在附录 B.2中展示了丁烷的自由能随二面角变化的情况。结果表明，PySAGES 使用不同的方法和后端重现了预期的自由能景观。在第 3.1 节中，我们将进一步探究 PySAGES 在这些简单模型系统之外的适用性和正确性。

其次，我们在PySAGES 上展示了图形处理器上使用两种不同后端的性能，详见第 3.2 节。具体而言，我们将增强采样模拟的性能与纯分子动力学模拟以及其他增强采样实现的性能进行比较。

3.1. 使用 PySAGES 的增强采样示例应用

为展示 PySAGES 在不同情境下的多功能性和有效性，我们列举了若干示例，说明如何利用增强采样方法，在生物学、药物设计、材料工程、高分子物理和从头算模拟等不同领域获得宝贵见解。

这些示例展示了 PySAGES 在不同研究领域中的应用，以及不同增强采样方法和后端的实用性。

总的来说，这些示例证实，PySAGES 中实现的增强采样方法按预期运行，所得结果与现有文献一致。

3.1.1. 用于药物发现的蛋白质–配体复合物结构稳定性

高通量对接技术是药物先导化合物发现中广泛使用的计算方法。然而，这些技术受限于对蛋白质构象以及配体在对接区域内稳定性的信息不足 [55]。为了解决这一问题，人们开发了动态解离对接（DUCK）方法，通过计算破坏蛋白质-配体复合物中最重要的天然接触（氢键相互作用）所需的功，来评估配体结合的稳定性 [56]。事实证明，该方法与经典对接互为补充且彼此正交，因此两种技术可以并行用于药物发现 [57, 58]。然而，DUCK 与传统增强采样技术结合时，可能需要较长时间才能收敛 [56], 因此不适用于高通量药物发现流程。

在此，我们展示了如何将 PySAGES 与 OpenMM 结合，高效应用于药物发现：通过协同利用用户友好的界面、原生并行能力以及收敛速度快的新型增强采样方法，加速对配体数据库进行虚拟筛选。在此示例中，我们研究了 SARS-CoV-2 病毒的主蛋白酶（Mpro）（蛋白质数据库：7JU7 [59]），移除了配体，并选取单体 A 作为对接受体。使用 RDock [60]对 SMILES字符串 `CCCCOCC(=O)c1ccc(C)cc1N[C@H]1N[C@@H](c2cccnc2)CS1`进行了配体对接。我们使用评分最高的构象初始化系统，并采用 ff14SB [61], TIP3P [62], 和 GAFF [63]力场进行模拟。我们进行了 10 ns 的平衡过程，以确定配体与蛋白质之间最稳定的氢键。随后，使用该平衡过程的最后一帧初始化 PySAGES 中采用自适应偏置力、元动力学、FUNN、人工神经网络和谱自适应偏置力进行的增强采样计算。我们将这些方法与使用牵引分子动力学模拟的同一系统进行比较（见图 4b），后者使用 Amber20 [64] 。结果表明，采用谱自适应偏置力或 FUNN 等新型增强采样方法，可以将模拟时间缩短一个数量级。这能够大幅加快药物发现进程，并帮助更快地识别潜在的候选药物。

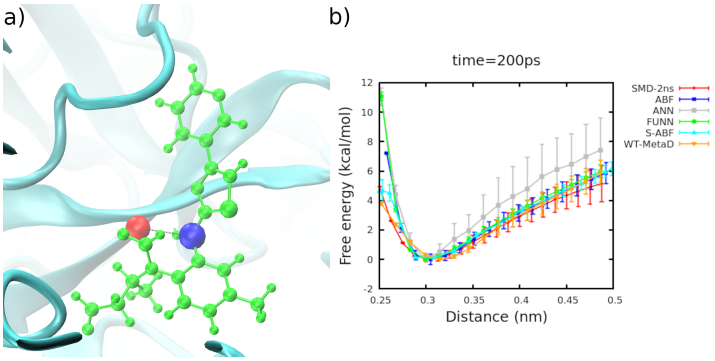

图 4：动态解离对接（DUCK）方法详解。对于经典对接获得的候选结合模式，使用 分子动力学 模拟进行短程运行，并从中选出最稳定的受体-配体天然接触。本例中，该接触是面板 a) 中突出显示的红色和蓝色原子之间的氢键。b) 比较 PYSAGES 中用于 DUCK 计算的不同方法；每种方法的结果均取 5 个不同副本的平均值。红色曲线为参考结果，即时长为 2 ns 的 牵引式 分子动力学 模拟。相比之下，PYSAGES 中的不同方法采用短 10 倍的模拟时长：只有 人工神经网络 [13]的性能逊于参考方法；谱系 自适应偏置力 [50]或 FUNN [14]表现最佳。

3.1.2. 二嵌段共聚物球形畴的裂变

现在，我们使用粗粒化模型研究二嵌段共聚物单个球形畴的裂变。我们采用了先前研究中发表的一种软粗粒化耗散粒子动力学（DPD）模型 [65, 66, 67]。该模型由 $n = 200$ 条链组成，每条链含有 $N = 256$ 个珠子，用以表示液态聚合物熔体。每条链中的前 $N_A = 16$ 个珠子为A型，其余 $N_B = 240$ 个为B型。

采用标准 DPD势，通过排斥参数 $A_{ii} = 5k_BT/\sigma^2$来实现不可压缩性。不过，在异类粒子之间施加更强的相互作用$A_{AB} = A_{ii} + \Delta A k_BT/\sigma^2$，其中 $\Delta A \in [0.1,\ 0.4]$，以产生排斥作用并引发微相分离。Flory-Huggins 参数 $\Delta A \propto \chi N > 0$ 可用于表征这种相分离。该非键合势的相互作用范围为 1σ，与将温度维持在 $T = 1k_BT = 1\epsilon$的 DPD恒温器范围相同。

此外，使用静止长度为零的谐弹簧力将珠子连接到聚合物链上，弹簧常数为 k =$16/3k_B/\sigma^2$，由此得到的平均键长为 $b_0 = 0.75\sigma$。该聚合物熔体的平衡相是体心立方（BCC）相，其中球形 A 液滴分散在 B 熔体中。[68] 然而，我们将聚合物限制在边长为 $L_0 = 10\sigma$ 的紧密立方模拟盒中，这导致形成一个单一的

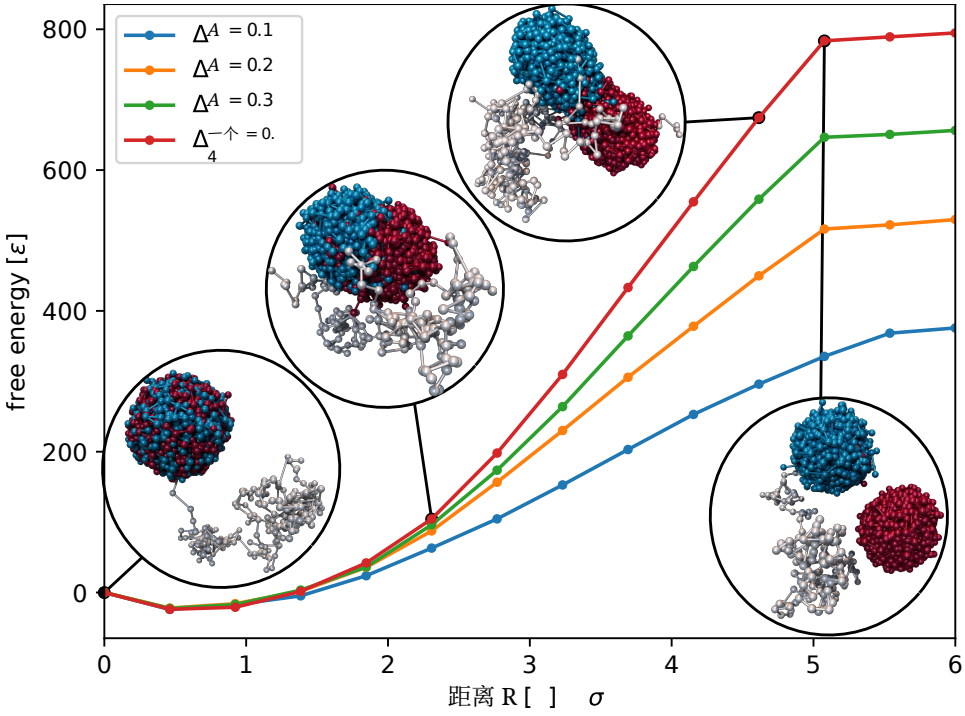

图5：球形嵌段共聚物畴裂变的自由能景观。形成球形畴的链端被分成两组（蓝色）和（红色），为清晰起见，其他链端未显示，仅显示一条链（灰色）。起初形成一个球形畴，但随着我们进一步约束蓝色组与红色组之间的质心距离，该畴首先伸长，随后完全分离。在分离过程中，自由能持续增加，并且异类之间的排斥作用越强，增幅越陡峭 ΔA。畴一旦分离，自由能便趋于平台。

B基质中的球形畴。我们以 $\Delta t = 10^{-3}\tau$ 为时间步长进行模拟，并将每次模拟平衡 $t = 1000\tau$，随后进行 $t = 1000\tau$ 的生产运行。关于该系统在有无 图形处理器时的性能讨论，PySAGES 可参见第 3.2.1 节。

在定义二嵌段共聚物系统后，下一步是在系统中定义一个集体变量 。在本例中，我们关注单个球形 A 畴裂变为两个大小相等的较小 A 畴。为此，我们将聚合物链分成两组：前 $n = 100$条链将形成第一个小畴（图图5中的蓝色），后 $n = 100$ 条链将形成第二个球形畴（图图5中的红色）。为了定义并确保两组彼此分离，我们将集体变量 定义为蓝色 A链尾与红色 A链尾的质心之间的距离， R。起初，在没有偏置的情况下，两组形成一个球形畴，蓝色和红色聚合物链尾充分混合，如 $R < 1\sigma$ 较小时的图5所示。

为研究球形畴的裂变，我们使用谐和偏置（见第2.1.1节）在两组之间施加分离距离 R_0。系统中的高密度$\sqrt{\mathcal{N}} = \frac{\rho_0}{N} R_{e0}^3 \approx 344$会导致波动较小，并抑制不利构象。因此，我们使用较高的弹簧力常数 $k_{CV} = 1500\epsilon/\sigma^2$，以促进分离。

我们研究了 $R \in [0, 6]\sigma$ 的分离过程，采用14个副本并使用伞形积分（见第2.1.2节）来确定自由能剖面，如图5所示。随着外部分离距离 R_0 <style id='15}>的增加，我们观察到单个畴如何分裂成两个畴。在较低的分离距离$R < 2\sigma$下，单个畴基本未发生变形，但两组在单个球形畴内部彼此分离。进一步增大分离距离会超出球形畴的尺寸，使该畴变形为细长的棒状形状。两组仍保持连接，以尽量减小AB界面。

当 4σ 与 5σ 之间的间距达到一定程度时，畴变形会变得非常剧烈，以至于在两组之间形成另一个AB界面、进而形成两个球形畴所需付出的代价，低于液滴畴变形及其拉长的AB界面所带来的熵惩罚。分离之后，自由能景观对间距不再敏感，因为两个畴之间没有相互作用。

分离的自由能剖面由异类组分间的排斥作用 $\chi N \propto \Delta A$ 控制。排斥作用越强，裂变扩大 AB 表面积所需的能量就越多。对于最强的

相互作用 $\Delta A = 0$。4ϵ，总自由能垒达到约 800ϵ；而对于最低的 $\Delta A = 0$。1ϵ，其仍低于 400ϵ。这两种能垒都比热涨落 $1k_BT = 1\epsilon$ 大几个数量级，因此预计不会发生自发分离，裂变只能通过增强采样进行研究。

值得注意的是，在最小间隔距离 $R_0 = 0$ 处，体系并未处于最低自由能态。强制完全混合并不有利，因为这两组分自然倾向于稍微分离，以优化链末端尾部的熵。

3.1.3. 水相界面中的液晶锚定

液晶（LC）像液体一样流动，却具有类似晶体的各向异性特性。由于液晶的锚定行为对水相界面处少量分子的浓度变化高度敏感，近年来已被用作界面分子传感器的原型 [69]。界面处存在的分子会显著改变液晶分子的自由能面，这取决于液晶分子相对于该界面的位置和取向。在此例中，我们重新考察了液晶的一些典型界面：4-氰基-4′-戊基联苯（5CB）与纯水及月桂基硫酸钠（SDS）的界面。对于 5CB 和水，此前的研究重点是获取 5CB 在水相界面的自由能面 [70]。在本研究中，我们在一个厚度为 16 nm、包含 1000 个处于向列相（300 K）的 5CB 分子的薄层上施加了混合锚定条件；该薄层与一个厚度为 3 nm、含有 62 个 SDS 分子的水层相互作用，SDS 位于其中一个界面处。所用力场为：用于 5CB 的统一原子力场 [71], 、用于水的 TIP3P [62] 、以及用于 SDS 的 GAFF [63] 和 Lipid 17。用于研究该系统的集体变量（CVs）为：位于两个界面处的一个 5CB 分子质心之间的距离（见附录 A），以及同一分子相对于盒子 z 轴的倾斜取向。纯水界面和含 SDS 界面的自由能面均展示于图 6。可以看到，纯水的自由能面在平行于表面的取向处呈现一个极小值，其形状与 [70] 中计算得到的结果相似。相反，SDS 的存在使界面处相同相对位置和取向对应的极小值转变为极大值（图 6 左上），并使极小值移至 5CB 垂直于界面的取向；这与实验观察一致，即水中存在 SDS 时，锚定方式会从平面锚定转变为垂面锚定。

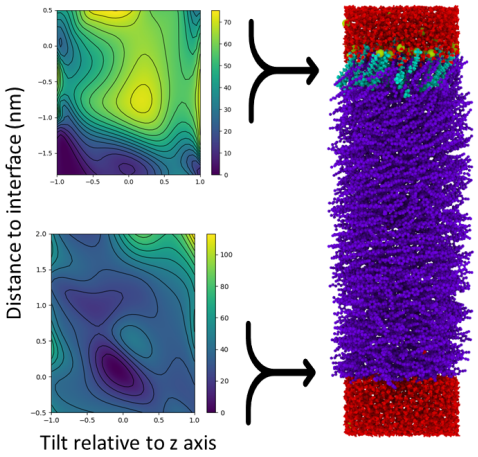

图 6：含有 SDS 和水的混合锚定板中 5CB 的自由能面。右侧：系统快照，其中水分子为红色、5CB 为紫色、SDS 为绿色，钠离子为黄色。左上：水–SDS 界面附近 5CB 分子的自由能面。左下：纯水界面附近 5CB 的自由能面。两个自由能面均使用 FUNN 方法，通过 PySAGES 和 OpenMM 得到。

3.1.4. 从头算增强采样模拟

在从头算多相催化模拟领域，捕捉动态效应和熵效应对于准确描述相关现象至关重要 [38]。经典力场无法准确捕捉催化过程中涉及的关键键断裂事件，因此有必要采用基于第一性原理计算的分子动力学模拟。鉴于反应事件通常受较大的自由能垒限制，增强采样技术是此类模拟不可或缺的一部分。将PySAGES 与ASE结合，可使用多种第一性原理计算器。

例如，我们使用 VASP 作为计算器，开展简单的 从头算增强采样模拟。集体变量 是钠原子与氯原子之间的间隔距离，采用 PBE 泛函 [72]，和谱系自适应偏置力作为增强采样方法（参见第 2.1.10 节）。结果如 图 7所示，自由能曲线沿 Na–Cl 距离方向的最低点对应真空中 Na 原子与 Cl 原子之间的平衡距离。

3.1.5. 使用机器学习力场进行增强采样

深度神经网络（NN）力场在保持第一性原理分子动力学精度的同时，计算成本可与经典分子动力学相近。

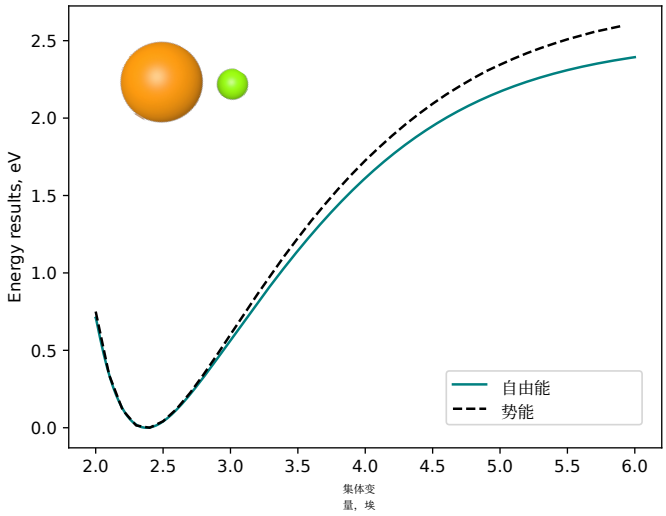

图7：使用 PySAGES 中的谱自适应偏置力，结合 ASE + VASP 计算的 Na–Cl 距离的自由能（$T = 300$ K）和势能。

通过 ASE 可以使用 NN 势，例如 DeepMD [73],和高斯近似势（GAP）。此外，JAX MD 可用于利用更通用的 NN 势，这些势可用于增强采样计算。将 PySAGES 与 ASE 或 JAX MD 结合，可通过使用 NN 力场并利用 PySAGES 提供的任一种增强采样方法高效采样稀有事件，从而用于主动学习，如参考文献 [74]所述，其中使用并行回火元动力学，在水中尿素分解过程中生成了准确的 NN 力场。

为了测试 PySAGES 处理不同 神经网络 力场的能力，我们选取了三个使用上述方法训练的不同系统。对于 Deep分子动力学，我们使用一个预训练的水模型，其中增强采样系统是在真空中的单个水分子，集体变量是分子的内角，采样方法为 自适应偏置力 (第 2.1.4 节）。图8 的结果表明，该自由能曲线的最低点约为 105 度，处于实验值的范围内。

接下来，在 图8b 中，采用了 GAP 势来模拟 Si–H 非晶混合物 [75]。在此情况下，使用了一个含有 244 个原子的系统，集体变量是混合物中三个 Si–Si–H 原子构成的键角。自由能的全局最小值与以下直方图一致：

无偏模拟的结果见 [75]。

最后，我们研究了硅晶体的图神经网络（GNN）模型 [76]，使用 PySAGES 和 JAX MD。在本例中，使用了一个包含 64 个原子的晶态硅系统，集体变量是晶体中的硅—硅距离。图8c 的结果表明，对于该模型，自由能最小值对应的硅—硅最近邻距离几乎与 2.35 埃这一实验值完全一致。

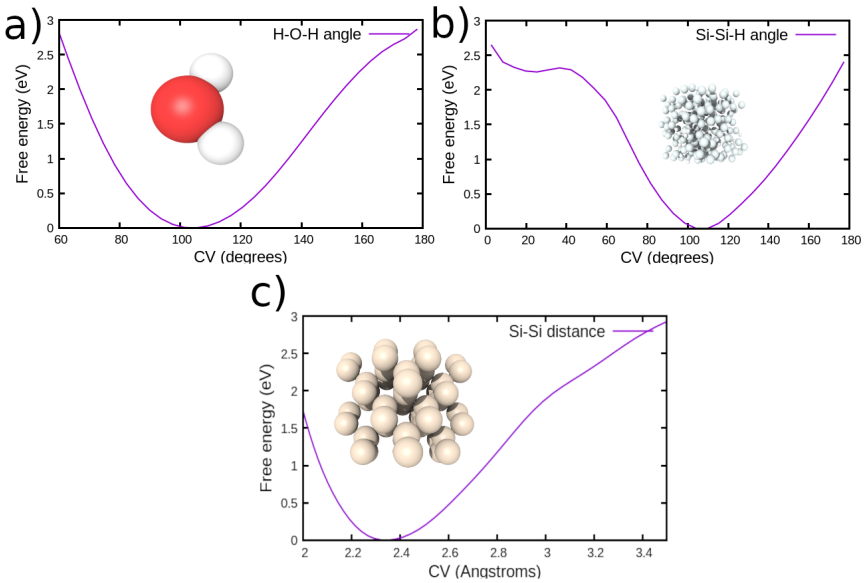

图8：自由能计算：a）使用 ASE 和 DeepMD 模型计算水分子内角；b）使用 ASE 和 GAP 模型计算硅—硅—氢键角；c）使用 JAX MD 和图神经网络模型计算硅—硅距离。

3.2. 性能

我们的分析表明，PySAGES 在 ~14至少比SSAGES 在配备 NVIDIA V100 GPU 的机器上快 15 倍。为获得这一估算值，我们沿两个球形聚合物域之间的质心距离进行伞形采样增强采样，以测量文中所述球形二嵌段共聚物共混物裂变的自由能景观（图5），该过程见第3.1.2节。为确保对不同库和分子动力学引擎版本的支持与兼容性，我们使用SSAGES v0.9.2- alpha 和 PySAGES v0.3.0 分别搭配 HOOMD-blue v2.6.0 和 HOOMD-blue v2.9.7来估算性能。

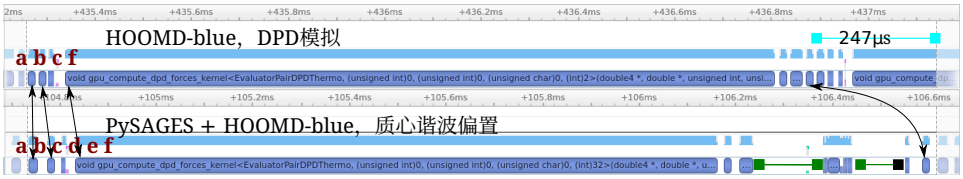

图9：该图展示了使用 NVIDIA Nsight Systems 在 NVIDIA V100 GPU 上记录的 1.8 毫秒剖面时间线片段。上排显示原生 HOOMD-blue 模拟的一个时间步，下排显示对质心集体变量施加谐和偏置的 PySAGES/HOOMD-blue 模拟。浅蓝色表示图形处理器活动，深蓝色表示各个 CUDA 计算内核。栗色字母标示两种模拟中相同的计算步骤：a）积分的前半步，b）计算键力，c）成对作用力，d）计算集体变量，e）向 HOOMD-blue 模拟添加谐波偏置力，f）第二个积分步骤。d）和 e）仅在 PySAGES 中执行，并在图形处理器上运行。我们观察到，在 PySAGES 通过 Python 协调 GPU–JAX/CuPy 的过程中，图形处理器存在空闲时间（绿色条），但请注意，即使在图形处理器内存中也没有内存复制。每个时间步额外用于集体变量偏置的时间为 247 µ秒（青绿色条）。

3.2.1. 图形处理器 利用率分析

PySAGES 旨在使用 图形处理器 执行模拟中的每个计算密集型步骤，并在 GPU 设备与主机CPU 内存之间实现零拷贝指令，其显式后端适用于 HOOMD-蓝色 [4]和 OpenMM[5], , 同时通过 JAX [77]为用户提供 Python 代码。本节中，我们通过考察两个示例系统（每个后端各一个）来研究 PySAGES 的计算效率。

对于 HOOMD-blue，我们考察了一个由高度粗粒化DPD二嵌段共聚物构成的系统，如第3.1.2节所述。模拟盒中共有 $nN = 51\,200$ 个粒子，密度为 $\rho = 51。2/\sigma^3$。我们使用搭载于Intel Xeon Gold 6248R CPU的NVIDIA V100 GPU @ 3.00GHz 进行基准测试。仅使用 HOOMD-blue v2.9.7 运行时，平均每秒时间步数（TPS）为 754，这符合 HOOMD-blue 在 GPU 上的预期高性能。

图9 展示了执行单个时间步期间的详细分析时间线。在1.8毫秒内，HOOMD-blue将大部分计算工作用于计算成对DPD作用力。可以注意到，HOOMD-blue经过设计，在一个时间步内图形处理器几乎没有空闲时间。加入PySAGES计算部分后，我们观察到新增了一部分计算，用于计算集体变量并将力添加到每个粒子上。这会导致图形处理器出现一小段空闲时间，因为执行过程还需要Python运行时界面与JAX进行交互。未来，我们计划异步执行集体变量的计算，与常规的

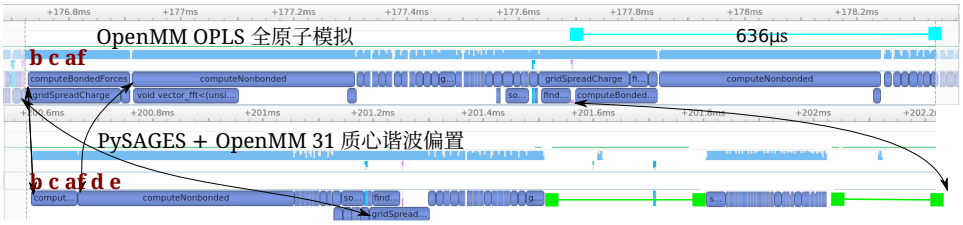

图10： 1.6毫秒的 OpenMM OPLS 模拟，模拟对象为 40,981 个以聚合物形式存在的粒子，并使用粒子网格 Ewald 方法（PME）对长程库仑力进行求和。颜色和标签与 图9相同。OpenMM 支持异步 图形处理器 内核执行，因此与 HOOMD-blue 相比，时间线的线性排序程度更低，但我们仍可识别出 集体变量 计算 d）和力偏置 e），以及 图形处理器 （绿色）的同步空闲阶段。总体而言，与 MM相比，OpenHOOMD-blue 的性能下降更为明显。

力计算并行进行，从而隐藏这段很短的、受CPU限制的图形处理器空闲时间。不过，我们测得额外计算造成的总延迟仅约为 247 µ秒。对于用户友好的集体变量定义而言，我们认为这是可以接受的开销。

为了连接集体变量空间中的多个点，我们可以使用伞形采样（见第 2.1.2 节）或改进弦方法（见第 2.1.3 节）等增强采样方法来计算 MFEP。这些高级采样方法的共同点是会对系统的多个副本进行模拟。借助 PySAGES，我们可以使用 Python 模块 `mpi4py` 及其 `MPIPoolExecutor` 轻松并行执行这些模拟。这使我们能够在多块 GPU 上执行模拟副本，即使这些 GPU 分布在不同的主机上也是如此。在我们的示例中，我们使用 14 个副本进行伞形积分，并使用了 7 块 Nvidia V100 GPU。使用单块 V100 GPU 执行模拟，所有副本共进行 $5 \cdot 10^5$ 个时间步，耗时 2 小时 59 分钟。理想情况下，使用 7 块 GPU 可将求解时间缩短至约 26 分钟。通过我们的 MPI 并行实现，求解时间为 28 分钟。同步开销以及最终分析等无法并行的部分共耗时 2 分钟，约占 9% 的开销。这种基于 MPI 的多 GPU 实现，使得在高性能计算（HPC）环境中能够自动高效地进行增强采样。

对于专为单副本模拟设计的增强采样方法，我们提供了一种实现，使多个副本能够并行运行，这称为尴尬并行计算。在这种情况下，内置分析会对多个副本的结果取平均值，并估算不确定性。

上一节中，我们通过 JAX 展示了 PySAGES 与 HOOMD-blue 之间快速的 GPU 互操作性。不过，PySAGES 的理念是独立于模拟后端开发增强采样方法，因此这里我们展示了使用 OpenMM 也能实现类似的性能。由于 OpenMM 专注于全原子模拟，我们使用 BigSMILES [78] 记法 `{[$]CC([$])(C)C(OCC(O)CSC1=CC=C(F)C(F)=C1)=O}`和 OPLS-AA 力场 [79, 80]（包括通过粒子网格 Ewald 方法（PME）计算的长程库仑力）模拟聚合物的全原子模型。我们模拟了由 31 个大分子组成的 40 聚体体相体系，总计包含 40 981 个原子。作为概念验证，我们计算了每条聚合物链的质心，并通过 PySAGES 对其施加谐和偏置。作为性能指标，我们评估了在与上文 HOOMD-blue 示例相同的硬件配置上运行的每日纳秒数（NS/DAY）。对于无偏置的纯 OpenMM 模拟，我们达到 ≈ 136 NS/DAY 的性能。对于使用 PySAGES 偏置的模拟，我们达到 ≈ 75 NS/DAY 的性能，相当于约 50% 的偏置开销。图10展示了与 HOOMD-blue 类似的时间序列分析。

值得注意的是，与 HOOMD-blue 相比，OpenMM 的执行模型更多地采用独立内核的并行执行，这也改变了执行顺序。因此，相同的 CPU 同步对执行过程造成的影响比在 HOOMD-blue 中更大。此外，对于该系统，单个时间步的执行速度比 HOOMD-blue 更快，因此同步开销也更加明显。在这种情况下，PySAGES 和 OpenMM 的并行化预计会带来更大的性能优势。我们还注意到，计算质心并对全部 31 条聚合物链施加偏置，比前一个示例中计算单个集体变量的成本更高。这些因素共同解释了 PySAGES 在这项 OpenMM 模拟中的较高开销，但总体性能良好，而且显著优于那些需要在 CPU 上进行集体变量计算的替代实现。

4. 结论

我们介绍了 PySAGES，这是一个用于分子动力学模拟中增强采样的库。它使用户能够使用多种增强采样方法和集体变量，并可通过基于 Python 和 JAX 的简洁界面实现新方法和新变量。

我们通过多个示例应用展示了如何使用 PySAGES，涵盖药物设计、材料工程、

高分子物理和从头算分子动力学模拟。我们希望这些示例能让读者了解该库的灵活性，以及它以高性能方式解决各类问题的潜力。

正如我们的分析所展示的，对于大型问题，即使后端已在图形处理器上执行计算，PySAGES 进行偏置模拟的速度仍可比 SSAGES 等库快一个数量级以上。

不过，与任何新开发的软件一样，PySAGES 仍在开发中，我们也在持续改进它。近期，我们计划添加重启功能，让用户能够更灵活地运行长时间模拟。此外，我们计划优化 PySAGES 侧的计算，使其能够与后端的力计算完全异步运行，从而进一步提升当前性能。我们也欢迎社区参与 PySAGES 的开发，无论是提出新功能建议、报告错误，还是贡献代码。

总的来说，我们认为，PySAGES 为有兴趣在多个领域开展分子模拟和从头算模拟的研究人员提供了实用工具。它拥有易于使用的框架，可用于定义和使用采样方法及集体变量，并且在 GPU 设备上具有出色的性能。

展望未来，我们期待 PySAGES 能够实现完全端到端可微的自由能计算。这将为力场和材料设计带来新的可能性，推动这些领域取得重大进展。

代码可用性

代码可在以下位置获取：PySAGES GitHub 仓库 https://github.com/SSAGESLabs/PySAGES。

致谢

本研究获得美国能源部基础能源科学、材料科学与工程部通过中西部计算材料综合中心（MICCoM）提供的支持。L. S. 感谢芝加哥大学埃里克和温迪·施密特科学人工智能博士后奖学金的支持。R.A. 获荷兰研究理事会（荷兰研究理事会 Rubicon 计划 019.202EN.028）支持。作者还感谢芝加哥大学研究计算中心提供计算资源。

利益冲突声明

A.L.F. 是 Evozyne, Inc. 的联合创始人和顾问，也是美国专利申请16/887,710和17/642,582、美国临时专利申请62/853,919、62/900,420、63/314,898和63/479,378，以及国际专利申请PCT/US2020/035206和PCT/US2020/050466的共同发明人。

附录A：用于计算到界面距离的集体变量

集体变量的实现，见集体变量所述，参见第 2.2.1 节，即一组原子与由另一组原子定义的界面之间的距离。

```
class DistanceToInterface(TwoPointCV):
def __init__(self, indices, axis, sigma, scope, bins=100, coeff=1):
super().__init__(indices)self.axis = axisself.sigma = sigma
self.scope = scopeself.bins = binsself.coeff = coeff@property
def function(self):return lambda r1, r2: distance_to_interface(r1, r2,
 axis=self.axis,sigma=self.sigma, scope=self.scope,bins=self.bins,
 coeff=self.coeff)
def distance_to_interface(p1, p2, axis, sigma, scope, bins, coeff):
mobile_axis = barycenter(p1)[axis]
positions_axis = p2.flatten()[axis::3]
centers = np.linspace(scope[0], scope[1], bins)
centers = np.expand_dims(centers, 1)
positions_axis = np.expand_dims(positions_axis, 0)
diff = positions_axis - centers
mass = np.exp(-0.5 * (diff / sigma) ** 2)mass = np.sum(mass, axis=1)
mass_diff = np.abs(mass[1:] - mass[:-1])centers = np.squeeze(centers)
centers_mean = (centers[1:] + centers[:-1]) / 2
probability = nn.softmax(mass_diff * coeff)
interface = np.sum(probability * centers_mean)
return mobile axis - interface_
```

附录 B. 基准测试系统

在以下各节中，我们将展示丙氨酸二肽和丁烷基准测试系统的自由能计算结果。这些系统增强采样模拟所选用的所有参数的详细信息汇总于附录 B.3。

附录 *B.1.* 丙氨酸二肽

第一个测试系统是在真空中的丙氨酸二肽（图 B.11），这是文献中常用的增强采样方法基准系统。

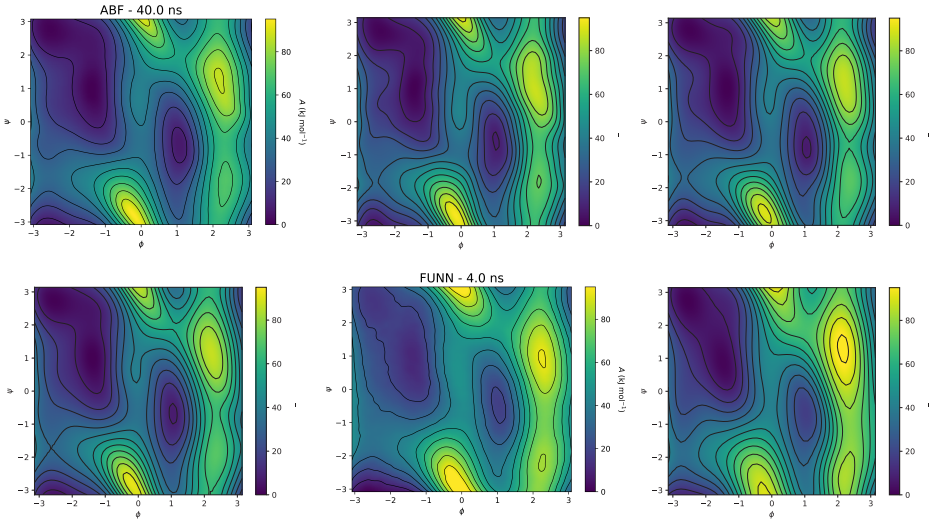

图 B.11：使用不同增强采样方法（ABF、元动力学、谱自适应偏置力、人工神经网络、FUNN、CFF），通过 PySAGES 和 OpenMM 获得的真空中丙氨酸二肽（Amber ff99SB [81]）自由能景观，该景观以二面角 ϕ 和 ψ 为变量。每个面板还标示了自由能收敛所需的模拟时长。较长的 ABF 模拟结果代表基准真值。

附录 *B.2.* 丁烷

作为第二个测试系统，我们计算丁烷分子（真空中）沿 C-C-C-C 二面角的自由能曲线，ϕ_{CCCC}，见图 B.12。

附录 *B.3* 示例系统详情

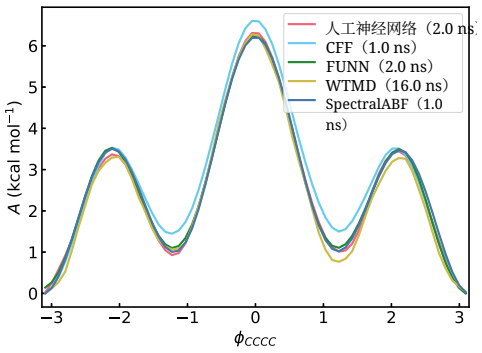

图 B.12：使用不同增强采样方法获得的丁烷分子二面角自由能剖面（使用 OPLS- 基力场 [79]），采用 PySAGES 和HOOMD-blue： 人工神经网络、CFF、FUNN、Spectral 自适应偏置力、WTMD。图例还标明了模拟时长。较长的 自适应偏置力模拟代表基准真值。

表 B.1：各示例的参数和方法详情。除元动力学外，对于所有方法，我们在 ADP 的每个集体变量方向上使用了包含 50 个点的网格，在丁烷的集体变量方向上使用了包含 64 个点的网格。

N(自适应偏置力) = 在计算自适应偏置力的完整平均值之前使用的阈值参数。

ADP =丙氨酸二肽

系统	后端	CV	方法	设置	Fig.
ADP	OpenMM	ϕ和 ψ	ABF ANN CFF FUNN 元动力学 谱自适应偏置力	$N = 500$（默认） 拓扑结构 = (8,8) 拓扑结构 = (14,) 拓扑结构 = (14,) $\sigma = 0.35$ 弧度 h = 1.2 千焦/摩尔 步长 = 500 —	B.11
丁烷	HOOMD-blue	ϕ_{CCCC}	ANN CFF FUNN WTMD 谱自适应偏置力	拓扑结构 = (8,8) 拓扑结构 = (8,) 拓扑结构 = (8,) $\sigma = 0.10$ 弧度 h = 0.01 千焦/摩尔 步幅 = 50 $\Delta T = 5000$ —	B.12

参考文献

1. nobelprize.org, 《2013年诺贝尔化学奖》（https://nobelprize.org/prizes/chemistry/2013/summary/，访问于2022年11月）。2. D. E. Shaw、M. M. Denero、R.O. Dror、J.S. Kuskin、R.H. Larson、J.K. Salmon、C. Young、B. Batson、K. J.Bowers、J. C. Chao 等, 《Anton：一台用于分子动力学模拟的专用机器》, 《ACM 通讯》51 (7) (2008年) 91–97。3. D. E. Shaw、J. Grossman、J. A. Bank、B. Batson、J. A.Butts、J.C. Chao、M. M. Denero、R. O. Dror、A.Even、C. H. Fenton 等, 《Anton 2：提升专用分子动力学超级计算机在性能和可编程性方面的标准》, 载于：SC'14：国际高性能计算、网络、存储与分析会议论文集，IEEE，2014年，第41–53页。4. J. A. Anderson、J. Glaser、S. C.Glotzer, 《HOOMD-blue：用于高性能分子动力学和硬粒子蒙特卡罗模拟的 Python 软件包》, 《计算材料科学》173 (2020年) 109363。5. P. Eastman、J. Swails、J.D.Chodera、R. T.McGibbon、Y. Zhao、K. A. Beauchamp、L.-P. Wang、A. C. Simmonett、M. P.Harrigan、C. D.Stern、R. P. Wiewiora、B. R. Brooks、V. S.Pande, 《OpenMM 7：快速开发分子动力学高性能算法》, 《PLOS计算生物学》13 (7) (2017年) 1–17。6. S. Schoenholz、E. D. Cubuk, 《JAX, M.D.：可微物理框架》, 载于《神经信息处理系统进展》第33卷，2020年，第11428–11441页。7. S. S. Schoenholz、E. D. Cubuk, 《JAX, M.D.：可微物理框架》, 《统计力学杂志：理论与实验》2021 (12) (2021年) 124016。8. A. P. Thompson、H. M. Aktulga、R. Berger、D.S. Bolintineanu、W. M. Brown、P. S. Crozier、P. J. in 't Veld、A. Kohlmeyer、S. G. Moore、T. D. Nguyen、R. Shan、M. J. Stevens、J. Tranchida、C. Trott、S. J. Plimpton, 《LAMMPS ——一种灵活的模拟工具，用于对原子、介观和连续介质尺度的粒子材料进行建模》, 《计算物理通讯》271 (2022年) 108171。

9. M. J. Abraham, T. Murtola, R. Schulz, S.Páll,J. C. Smith, B. Hess, E. Lindahl, Gromacs：通过从笔记本电脑到超级计算机的多级并行实现高性能分子模拟，So- wareX 1（2015年）19–25。10. G. A. Tribello, M. Bonomi,D.Branduardi, C. Camilloni, G.Bussi, PLUMED 2：为老牌工具增添新功能，《计算机物理通讯》185（2）（2014年）604–613。11. G. Fiorin, M. L. Klein, J. Hénin, 使用集体变量驱动分子动力学模拟，《分子物理》111（22-23）（2013年）3345–3362。12. H. Si dky, Y. J. Colón, J. Helerich, B. J. Sikora, C. Bezik,W.Chu, F. Giberti, A. Z. Guo, X. Jiang, J. Lequieu, J. Li, J. Moller,M.J. Quevil- lon, M. Rahimi, H. Ramezani-Dakhel, V.S. Rathee, D. R. Reid,E.Sev- gen, V. Thapar, M. A. Webb, J.K. Whitmer, J. J. de Pablo, SSAGES：用于先进广义系综模拟的软件套件，《化学物理学杂志》148（4）（2018年）044104。13. H. Sidky, J. K. Whitmer, 利用人工神经网络学习自由能景观,《化学物理学杂志》148（10）（2018年）104111。14. A. Z. Guo, E. Sevgen, H. Sidky, J.K. Whitmer, J. A. Hubbell, J.J. de Pablo, 利用神经网络进行力偏置的自适应增强采样，《化学物理学杂志》148（13）（2018年）134108。15. E. Sevgen, A. Z. Guo, H. Sidky, J.K. Whitmer, J. J. de Pablo, 结合力频率采样，模拟具有崎岖自由能景观的系统，《化学理论与计算杂志》16（3）（2020年）1448–1455。16. D. Wang, Y. Wang, J. Chang,L. Zhang, H.Wang, et al., 利用自适应强化动力学高效采样高维自由能景观，《自然·计算科学》2（1）（2022年）20–29。17. C. R. Schwantes, V. S. Pande, 改进马尔可夫状态模型的构建，揭示 NTL9 折叠中的许多非天然相互作用，《化学理论与计算杂志》9（4）（2013年）2000–2009。

18. W. Chen, A. L. Ferguson, 利用自编码器进行分子增强采样：在线发现集体变量并加速自由能景观探索，《计算化学杂志》39（25）（2018年）2079–2102。19. A. Mardt, L. Pasquali, H. Wu,F. Noé, VAMPnets：用于分子动力学学习的深度学习方法，《自然-通讯》9（1）（2018年）1–11。20. W. Chen, H. Sidky, A. L. Ferguson, 时间滞后自编码器在动力系统慢模态发现中的能力与局限，《化学物理学杂志》151（6）（2019年）064123。21. W. Chen, H. Sidky, A. L. Ferguson, 利用无状态可逆 VAMPnets 非线性发现分子慢模态，《化学物理学杂志》150（21）（2019年）214114。22. H. Sidky, W. Chen, A. L. Ferguson, 分子潜在空间模拟器，《化学科学》11（35）（2020年）9459–9467。23. T.-S. Lee, D. S. Cerutti, D.Mermelstein, C. Lin,S. LeGrand,T. J.Giese, A. Roitberg, D. A. Case, R. C.Walker, D. M. York, 图形处理器加速的分子动力学和 Amber18 中的自由能方法：性能提升与新功能，《化学信息与建模杂志》58（10）（2018年）2043–2050。24. J. C. Phillips, D. J. Hardy, J. D. Maia,J. E. Stone, J. V. Ribeiro, R. C. Bernardi, R. Buch, G. Fiorin,J. Hénin, W. Jiang,et al., 基于 CPU 和图形处理器架构的可扩展分子动力学，使用 NAMD，《化学物理学杂志》153（4）（2020年）044130。25. C. Kobayashi, J. Jung, Y. Matsunaga, T.Mori, T. Ando, K. Tamura, M. Kamiya, Y. Sugita, GENESIS 1.1：一种支持多计算平台的增强采样算法混合并行分子动力学模拟器（2017年）。26. C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P.Virta- nen, D. Cournapeau, E. Wieser, J. Taylor, S.Berg, N.J. Smith,R.Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk,M.Brett, A. Haldane, J.F. del Río, M. Wiebe, P. Peterson, P.Gérard-Marchant,K.Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T. E. Oliphant, 使用 NumPy 进行数组编程，《自然》585（7825）（2020年）357–362。27. DLPack（https://github.com/dmlc/dlpack，访问于2022年11月）。

28. trunk.io，Trunk.IO（https://trunk.io，访问于2022年9月）。29. H. Sidky、W. Chen、A. L.Ferguson，机器学习在集体变量发现和生物分子模拟增强采样中的应用，《分子物理》118（5）（2020）e1737742。30. N. E. Jackson、M. A. Webb、J. J. de Pablo，面向多尺度材料设计的机器学习最新进展，《化学工程现状》23（2019）106–114。31. P. Tiwary、A. v. d. Walle，加速分子动力学增强采样方法综述，《纳米力学多尺度材料建模》（2016）195–221。32. A.-h. Wang、Z.-c. Zhang、G.-h. Li，生物分子增强采样分子动力学模拟进展，《中国化学物理学报》32（3）（2019）277。33. A. Mitsutake、Y. Mori、Y. Okamoto，增强采样算法，《生物分子模拟》（2013）153–195。34. Y. Miao、J. A. McCammon，生物分子自由能计算的无约束增强采样：综述，《分子模拟》42（13）（2016）1046–1055。35. Y. I. Yang、Q. Shao、J. Zhang、L. Yang、Y.Q. Gao，分子动力学中的增强采样，《化学物理学杂志》151（7）（2019）070902。36. C. Abrams、G. Bussi，分子动力学中的增强采样：使用元动力学、副本交换和温度加速，《熵》16（1）（2014）163–199。37. V. Limongelli，2020年的配体结合自由能与动力学计算，WIREs Computational Molecular Science 10（4）（2020）e1455。38. G. Piccini、M.-S. Lee、S. F. Yuk、D. Zhang、G. Collinge、L. Kollias、M.-T. Nguyen、V.-A. Glezakou、R. Rousseau，多相催化中结合增强采样的从头算分子动力学，《催化科学与技术》12（2022）12–37。39. J. Kästner，两个或更多反应坐标中的伞形积分，《化学物理学杂志》131（3）（2009）034109。

40. J. Kästner，伞形采样，《Wiley跨学科评论：计算分子科学》1（6）（2011）932–942。41. E. Weinan、W. Ren、E. Vanden-Eijnden，用于计算势垒跨越事件中最小能量路径的简化改进弦方法，《化学物理学杂志》126（16）（2007）164103。42. J. Comer、J. C. Gumbart、J. Hénin、T. Lelièvre、A. Pohorille、C. Chipot，自适应偏置力方法：你一直想了解却不敢提问的一切，《物理化学杂志B》119（3）（2015）1129–1151。43. E. Darve、D. Rodríguez-Gómez、A. Pohorille，用于标量和向量自由能计算的自适应偏置力方法，《化学物理学杂志》128（14）（2008）144120。44. A. Laio、M. Parrinello，逃离自由能极小值，《美国国家科学院院刊》99（20）（2002）12562–12566。45. A. Barducci、G. Bussi、M. Parrinello，良定义温度元动力学：平滑收敛且可调的自由能方法，《物理评论快报》100（2）（2008）020603。46. S. Hussain、A. Haji-Akbari，使用前向通量采样研究稀有事件：近期突破与未来展望，《化学物理学杂志》152（6）（2020）060901。47. R. J. Allen、P. B. Warren、P. R. ten Wolde，生化网络中稀有切换事件的采样，《物理评论快报》94（2005）018104。48. R. J. Allen、D. Frenkel、P. R. ten Wolde，模拟稀有事件……

平衡或非平衡随机系统，《化学物理学杂志》124 (2) (2006) 024102。49. J. K. Whitmer、C.-c. Chiu、A. A. Joshi、J. J. de Pablo，基函数采样：材料性质计算的新范式，《物理评论快报》113 (19) (2014年) 190602。50. P. F. Zubieta Rico、J. J. de Pablo，自由能景观的索伯列夫采样

景观，arXiv（2022年）。 arXiv:2202.01876。51. D. Cremer, J. A. Pople，环折叠坐标的一般定义，《美国化学会志》97（6）（1975）1354–1358。

52. J. M. L. Ribeiro, P. Bravo, Y. Wang, P. Tiwary，针对增强采样的重加权自编码变分贝叶斯（RAVE），《化学物理学杂志》149（7）（2018年）072301。53. C. Wehmeyer, F. Noé，时间滞后自编码器：用于分子动力学的慢集体变量的深度学习，《化

-

化学物理学 148 (24) (2018) 241703. 54. W. Wang, Z. Wu, R. Gómez-Bombarelli, 使用可微分模拟学习成对势（2022）。 arXiv:2209.07679. 55. A. Sethi, K. Joshi, K. Sasikala, M. Alvala，现代药物发现中的分子对接：原理与最新应用，载于：V. Gaitonde, P. Karmakar, A. Trivedi（编），《药物发现与开发》，第2卷，IntechOpen，2019年，第3章，第1–21页。56. S. Ruiz-Carmona, P. Schmidtke, F. J. Luque, L. Baker, N. Matassova, B. Davis, S. Roughley, J. Murray, R. Hubbard, X. Barril，动态解离与准结合态作为药物发现工具，《自然·化学》9 (3) (2017) 1755–4349。57. M. Majewski, X. Barril，结构稳定性可预测蛋白质–配体复合物的结合模式，《化学信息与建模杂志》60 (3) (2020) 1644–1651。58. M. Rachman, D. Bajusz, A. Hetényi, A. Scarpino, B. Merő, A. Egyed, L. Buday, X. Barril, G. M. Keserű，通过动态解离发现一种新型激酶铰链结合片段，RSC 医药化学 11 (2020) 552–558。59. N. Drayman, J. K. DeMarco, K. A. Jones, S.-A. Azizi, H. M. Froggatt, K. Tan, N. I. Maltseva, S. Chen, V. Nicolaescu, S. Dvorkin, K. Furlong, R. S. Kathayat, M. R. Firpo, V. Mastrodomenico, E. A. Bruce, M. M. Schmidt, R. Jedrzejczak, M. A. Munoz-Alia, B. Schuster, V. Nair, K. yeon Han, A. O'Brien, A. Tomatsidou, B. Meyer, M. Vignuzzi, D. Missiakas, J. W. Botten, C. B. Brooke, H. Lee, S. C. Baker, B. C. Mounce, N. S. Heaton, W. E. Severson, K. E. Palmer, B. C. Dickinson, A. Joachimiak, G. Randall, S. Tay，马赛替尼是一种广谱冠状病毒3CL 抑制剂，可阻断SARS-CoV-2的复制，《科学》373 (6557) (2021) 931–936。60. S. Ruiz-Carmona, D. Alvarez-Garcia, N. Foloppe, A. B. Garmendia-Doval, S. Juhos, P. Schmidtke, X. Barril, R. E. Hubbard, S. D. Morley,

RDock：一种快速、灵活的开源程序，用于将配体对接到蛋白质和核酸，《PLOS计算生物学》10 (4) (2014年) 1–7。61. J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, C. Simmerling，ff14SB：提高蛋白质侧链和主链参数的准确性，《化学理论与计算杂志》11 (8) (2015年) 3696–3713。62. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein，液态水模拟中简单势函数的比较，《化学物理学杂志》79 (2) (1983) 926–935。63. J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case，通用 Amber 力场的开发与测试，《计算化学杂志》25 (9) (2004) 1157–1174。64. D. A. Case, K. Belfon, I. Y. Ben-Shalom, S. R. Brozell, D. S. Cerutti, T. E. Cheatham, III, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, G. Giambasu, M. K. Gilson, H. Gohlke, A. W. Goetz, R. Harris, S. Izadi, S. A. Izmailov, K. Kasavajhala, A. Kovalenko, R. Krasny, T. Kurtzman, T. S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, V. Man, K. M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, A. Onufriev, F. Pan, S. Pantano, R. Qi, D. R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. L. Simmerling, N. R. Skrynnikov, J. Smith, J. Swails, R. C. Walker, J. Wang, L. Wilson, R. M. Wolf, X. Wu, Y. Xiong, Y. Xue, D. M. York, P. A. Kollman，Amber 2020（2020年）。65. L. Schneider, M. Heck, M. Wilhelm, M. Müller，剪切流中层状取向之间的转变，《大分子》51 (12) (2018) 4642–4659。66. L. Schneider, M. Müller，对称二嵌段共聚物的流变学，《计算材料科学》169 (2019) 109107。67. L. Schneider, G. Lichtenberg, D. Vega, M. Müller，圆柱限域中的对称二嵌段共聚物：形成手性形貌的一种途径？《ACS应用材料与界面》12 (44) (2020) 50077–50095。68. M. W. Matsen，嵌段共聚物熔体的标准高斯模型，《物理学杂志：凝聚态》14 (2) (2001) R21。69. Z. Wang, T. Xu, A. Noel, Y.-C. Chen, T. Liu，液晶在生物传感中的应用，《软物质》17 (2021) 4675–4702。

70. H. Ramezani-Dakhel、M. Sadati、M. Rahimi、A. Ramirez-Hernandez、B. Roux、J. J. de Pablo, 《理解水相界面处液晶的原子尺度行为》, 《化学理论与计算杂志》13 (1) (2017) 237–244。71. G. Tiberio、L. Muccioli、R. Berardi、C. Zannoni, 《迈向计算机模拟液晶：通过分子动力学模拟获得 n-氰基联苯的真实相变温度和物理性质》, ChemPhysChem 10 (1) (2009) 125–136。72. J. P. Perdew、K. Burke、M. Ernzerhof, 《广义梯度近似简明介绍》 [《物理评论快报》77, 3865 (1996)],《物理评论快报》78 (1997) 1396–1396。73. L. Zhang、J. Han、H. Wang、R. Car、W. E, 《深度势分子动力学：一种具有量子力学精度的可扩展模型》, 《物理评论快报》120 (2018) 143001。74. M. Yang、L. Bonati、D. Polino、M. Parrinello, 《使用元动力学构建反应事件的神经网络势：水中尿素分解案例》, 《今日催化》387 (2022) 143–149, CASALE SA 百年：关于催化过程的科学视角。75. D. Unruh、R. V. Meidanshahi、S. M. Goodnick、G. Csányi、G. T. Zimányi, 《非晶态 si : H 的高斯近似势》, 《物理评论材料》6 (2022) 065603。76. E. D. Cubuk、B. D. Malone、B. Onat、A. Waterland、E. Kaxiras, 《基于神经网络的经验势中的表示》, 《化学物理学杂志》147 (2) (2017) 024104。77. R. Frostig、M. J. Johnson、C. Leary, 《通过高级跟踪编译机器学习程序》, 《机器学习系统》4 (9) (2018)。78. T.-S. Lin、C. W. Coley、H. Mochigase、H. K. Beech、W. Wang、Z. Wang、E. Woods、S. L. Craig、J. A. Johnson、J. A. Kalow 等, 《BigSMILES：一种用于描述大分子的结构化线性表示法》, 《美国化学会中央科学》5 (9) (2019) 1523–1531。79. W. L. Jorgensen、D. S. Maxwell、J. Tirado-Rives, 《OPLS 全原子力场在构象能和有机液体性质方面的开发与测试》, 《美国化学会志》118 (45) (1996) 11225–11236。

80. L. Schneider、M. Schwarting、J.Mysona、H. Liang、M. Han、P. M. Rauscher、J. M. Ting、S. Venkatram、R. B. Ross、K.J. Schmidt、B. Blaiszik、I. Foster、J. J. de Pablo，《针对小分子性质的计算机模拟主动学习》，Mol. Syst. Des. Eng. (2022)。81. V. Hornak、R. Abel、A. Okur、B.Strockbine、A. Roitberg、C. Simmerling，《多种 AMBER 力场的比较及其开发

f

改进的蛋白质主链参数，《蛋白质：结构、功能与生物信息学》65（3）（2006）712–725。