在 GROMACS 中启用人工智能深度势，实现从头算精度的分子动力学模拟

Andong Hu、Luca Pennati、Stefano Markidis 和 Ivy Peng
瑞典皇家理工学院，瑞典斯德哥尔摩{andonghu, pennati, markidis, ivybopeng}@kth.se

arXiv:2602.02234v1 [cs.DC] 2 Feb 2026

摘要—先进的人工智能深度势能够以远低于第一性原理量子力学计算（例如密度泛函理论）的计算成本，提供从头算**精度的结果。本研究通过集成 DeePMD-kit，将人工智能深度势引入生产级分子动力学（MD）代码 GROMACS。DeePMD-kit 提供特定领域的深度学习（DL）模型，用于描述原子间势能和力场。具体而言，我们将 GROMACS 神经网络势与 DeePMD-kit 中的 C++/CUDA 后端相结合，从而支持跨多种 DP 模型系列和深度学习后端的人工智能深度势推理。我们在 GROMACS 中评估了两种近期的大原子模型架构：基于注意力机制的 DPA2 和基于图神经网络的 DPA3。评估使用了四种**从头算**精度的蛋白质水溶液基准体系（1YRF、1UBQ、3LZM、2PTC），并在 NVIDIA A100 和 GH200 GPU 上进行。结果表明，在 A100 和 GH200 GPU 上，DPA2 的吞吐量分别最高可比 DPA3 高出** $4.23\times$ **和** $3.18\times$ **。我们还对 DPA2 和 DPA3 的吞吐量、内存使用量及 GPU 内核级执行情况进行了表征研究，以作进一步比较。研究发现，内核启动开销和区域分解推理是生产级分子动力学模拟中人工智能深度势的主要优化重点。**

索引词—分子动力学、深度势、GRO- MACS、DeePMD-kit

I. 引言

分子动力学（MD）模拟 [1]–[3] 能够深入揭示复杂系统的原子级特性，而这些特性通常无法通过直接实验获取。它们是材料科学、生物化学、生命科学和药物发现领域至关重要的计算工具。尽管分子动力学模拟的目标是原子和分子系统及其相互作用，因此涉及量子力学，但绝大多数分子动力学模拟采用经典物理学方法，例如粒子推进器（即负责更新粒子坐标和速度的经典计算阶段），以及通过经验拟合量子规律得到的经典势能和伪势能。

采用经验力场（如 AM- BER [3], CHARMM [4] 和 GROMACS [2], ）的经典分子动力学可在大规模系统上高效扩展。然而，这类方法无法可靠地捕捉断键与成键、电荷转移、多体极化及相关量子力学效应。相比之下，基于密度泛函理论（DFT）的从头算 分子动力学（AIMD）能够计算原子间作用力，并达到近乎量子力学级别的精度。然而，其 $\mathcal{O}(N^3)$ 扩展性严重限制了可研究的系统规模和模拟时间尺度 [5]。

已提交预印本，待发表。

近年来，机器学习在原子间势方面取得进展，由此发展出第三种途径——人工智能深度势（DP） [6], 。这类模型基于量子力学参考数据中的势能进行训练，通过快速机器学习（ML）推理来计算能量和力，其精度接近从头算，而计算成本比密度泛函理论低几个数量级。这项能力使得人们能够对大型生物分子系统开展具有从头算精度的分子动力学模拟，并进行长时间轨迹模拟，同时兼容生产级分子动力学模拟工作流。例如，DeePMD-kit [7] 通过成熟且广泛使用的软件栈，提供先进的深度势模型，支持使用多种神经网络架构训练和部署深度势模型。

在本研究中，我们首先以当前先进的分子动力学代码 GROMACS 和 DeePMD-kit 为例进行案例研究，评估将 AI 深度势集成到生产级分子动力学代码中的机遇与挑战。我们发现，大多数推理技术栈以 Python 为核心，并依赖深度学习框架；而生产级分子动力学引擎对性能要求极高，是围绕经验力场构建的 C++/CUDA/MPI 代码。要打通这两个领域，需要稳定的 C/C++ 接口、协调一致的数据布局（例如邻居列表和周期性边界）、GPU执行流水线，以及原生支持高性能计算的构建系统。另一个挑战在于，人们尚未充分了解将深度势模型集成到端到端分子动力学工作流中的计算成本。例如，深度势在分子动力学工作流中的推理吞吐量、可扩展性、GPU利用率和内存使用量仍缺乏充分表征，因此性能瓶颈尚不明确。

通过在 GROMACS 中支持多个深度势模型系列和机器学习后端，我们为在生产级 GROMACS 工作流中实现从头算级别的力提供了一条稳定且向前兼容的路径。具体而言，我们首次将 DeePMD-kit 与 GROMACS 紧密集成，以支持经图形处理器加速的 AIMD 模拟工作流。我们的实现扩展了 GROMACS [2]神经网络势 (NNPot) 模块，使其能够部署最先进的人工智能深度势模型。我们使用两种近期的大原子模型 (LAM) 架构——DPA2 [8] 和 DPA 3 [9], ——在 NVIDIA A100 和 GH200 GPU 上，对四种蛋白质-水体系分子动力学基准进行了评估。随后，我们通过比较从头算级别的模拟与不包含量子力学效应的经典模拟，量化了引入量子力学效应的代价。此外，我们还通过在应用层和

GPU 内核层面进行性能分析，考察其计算强度、内存占用和内核启动行为，从而揭示在生产级分子动力学模拟中采用 DPA2 和 DPA3 模型的不同可行性。最后，我们重点指出了利用人工智能深度势实现可扩展、高保真 AIMD 模拟的优化重点。

我们将本文的贡献总结如下：

- 我们识别了在 GROMACS 的分子动力学模拟工作流中采用最先进的人工智能深度势模型所面临的机遇与挑战。
- 我们提出了一种设计策略，以实现人工智能深度势分子动力学模拟中的紧密系统集成与并行推理，并在最新的 GROMACS 版本中给出了实现（可从 https://github.com/HuXioAn/gromacs/tree/deepmd-oneModel 获取）。
- 我们使用 DPA2 和 DPA3 大原子模型，在 NVIDIA A100 和 GH200 GPU 上评估了四种溶剂化蛋白质（1YRF、1UBQ、3LZM、2PTC）。
- 我们首次对 GROMACS 中包含量子力学效应的深度人工智能分子动力学模拟进行了端到端性能表征，并概述了 DPA2 和 DPA3 分子动力学工作流的优化优先级。

II. 背景

在分子动力学模拟 [1]–[3],中，原子轨迹通过数值积分牛顿运动方程来计算，采用Verlet或蛙跳算法。作用在原子上的力计算为经典势能函数（$U(\mathbf{r})$）的负梯度，该函数通常分为短程和长程贡献。短程项包括成键贡献，如键、键角和二面角相互作用，以及非键范德华贡献（伦纳德-琼斯势）。长程静电相互作用用库仑势建模，通常使用粒子网格埃瓦尔德（PME）技术计算。库仑项分为短程快速衰减贡献和长程贡献。前者在实空间中通过成对原子相互作用计算，后者则通过三维快速傅里叶变换在傅里叶空间的笛卡尔网格上求解泊松方程来计算。所有非键短程相互作用仅在截断距离内的原子之间利用Verlet邻居列表计算，因此计算成本从二次降至线性。因此，包含 N 个原子和用于FFT的 N_g 个网格点的体系，其模拟成本按$\mathcal{O}(N) + \mathcal{O}(N_g \log N_g)$标度。

在经典分子动力学模拟中，$U(\mathbf{r})$ 的函数形式和相互作用参数由经验力场提供，例如 AMBER [3] 和 CHARMM [4],。这些力场通过将解析函数形式拟合到量子力学（QM）结果和实验数据而构建。相比之下，在从头算MD（AIMD）模拟中，力是在模拟过程中通过采用玻恩–奥本海默理论或卡尔-帕里内洛方法对电子结构进行量子力学处理而实时计算得出的。与经典分子动力学相比，AIMD模拟具有更高的准确性和保真度，因为它们能够捕捉键的断裂与形成、电子极化和电荷转移等现象。然而，AIMD 的计算成本随原子数以 $\mathcal{O}(N^3)$ 的比例增长，因此模拟规模仅限于数千个原子。

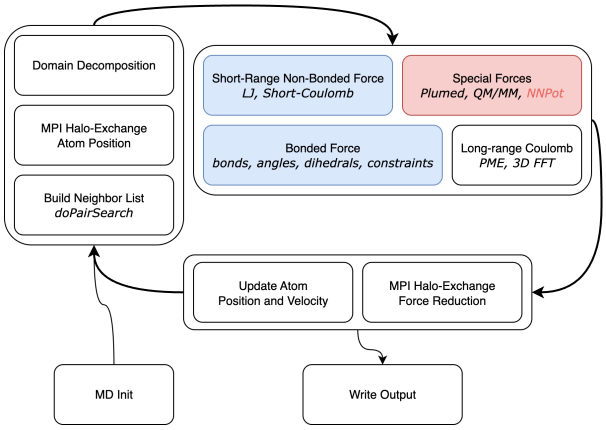


图 1：GROMACS 分子动力学模拟循环概览。

深度神经网络势和机器学习势（ML-Potentials） [11] 为解决这一精度与成本之间的权衡提供了方案。ML-Potentials 通过从大量包含能量和力的量子力学数据集中学习，来逼近 Born–Oppenheimer势能面 $U(\mathbf{r})$。与基于预定义函数形式的经验力场不同，ML-Potentials 依赖灵活的非线性函数逼近器，能够更好地捕捉复杂的相互作用和量子效应。为开发不同类别的 ML-Potentials，人们采用了多种用于编码原子结构的架构和方法，包括基于核的方法（如高斯近似势（GAP））、基组展开方法、深度神经网络以及图神经网络 [6], [9], [12]。训练完成后，ML-Potentials 能以仅按 $\mathcal{O}(N)$ 缩放的成本实现接近量子力学精度，为开展长时间、大规模且接近从头算 保真度的分子动力学运行带来了希望。

GROMACS [2] 是一款广泛使用的 C/C++开源分子动力学引擎，旨在从笔记本电脑到超级计算机等各种规模的模拟中实现高效运行。图 1 详细展示了该软件中分子动力学模拟主循环的结构。与其他经典分子动力学代码类似，GROMACS 使用经验力场计算标准的短程和长程相互作用，同时还可以通过专用接口在模拟中加入其他相互作用（图 1 中的特殊力 ）。GROMACS 通过一个名为神经网络势的力提供器 模块，引入了使用完整神经网络（NN）方法和混合神经网络方法运行分子动力学模拟的功能。在本研究中，我们选择以最新版 GROMACS 为基础，研究生产级分子动力学模拟工作流。

深度势分子动力学工具包 *(DeePMD-kit)* [13]是一个开源框架，旨在提供用于在分子动力学引擎中训练和部署的机器学习势能模型。DeePMD-kit 采用多后端架构，支持 NumPy、PyTorch、TensorFlow、JAX 和 PaddlePaddle。该框架将深度模型与执行后端解耦，

从而提高了与现有分子动力学代码集成的灵活性。

DeePMD-kit 支持标准深度神经网络势和图神经网络势。本文选择研究这两类中的两个近期深度势模型，即使用标准深度神经网络的 DPA2 模型，以及使用图神经网络的 DPA3 模型。

DPA2 和 DPA3 被称为大原子模型（LAMs） [8], [9]。LAMs 采用多任务方式训练：从一组数据集中训练一个统一的描述符，并为每个数据集训练相应的拟合网络 。这种方案降低了针对特定下游任务微调预训练大原子模型的成本。

DPA2、DPA3 和 DPA1 属于 DPA模型系列，这是 DeePMD团队开发的最新模型架构 [8], [9], [14]。DPA1 和 DPA2 利用注意力机制捕捉原子与其邻居原子之间的高维关系。原子及其邻居原子的信息通过多层感知机（MLP）嵌入网络编码为高维表征；在 DPA2 中，这一过程称为`repinit` 。随后，这些表征通过多个串联的自注意力层以残差方式更新；在 DPA2 中，这些层称为 `repformer` 。最后，经过训练的 MLP拟合网络可利用更新后的表征生成系统中每个原子的能量。DPA3 不使用注意力机制，而是利用图神经网络（GNN）和线图系列（LiGS）来捕捉并更新系统中的物理表征，例如键、键角和二面角。本文使用 DeePMD团队发布的 DPA2 和 DPA3 的最新版本`DPA-2.4-7M`和`DPA-3.1-3M` [9]。

III. 深度势 AIMD 模拟工作流

本节将介绍深度势 AIMD 模拟工作流的主要设计考量，并随后描述我们的实现。图 2 展示了耦合 GROMACS 和 DeePMD 的 AIMD 模拟工作流的总体架构。在此示例中，我们以蛋白质分子动力学模拟为测试案例，说明主要阶段和数据结构。主要数据结构包括水和蛋白质，以及它们的长程和短程力场。上层表示 GROMACS 内部的主要分子动力学运行。下层提供深度势模型描述符、算子，以及模型训练和推理功能。在这两个层级之间，虚线 C++ 接口和 NNPot 突出显示了一个耦合层，该层在两个层级之间交换原子和力场的信息。

整个分子动力学模拟工作流程由多个阶段组成。在实际进行分子动力学模拟之前，会先进行能量最小化（EM），以弛豫体系，确保初始结构在物理上合理。随后进行两步平衡过程：首先在恒容恒温（NVT）条件下稳定温度，然后在恒压恒温（NPT）条件下平衡体系的密度和压力。体系稳定后，便进行分子动力学模拟，以模拟原子随时间的实际运动。DeePMD 势能

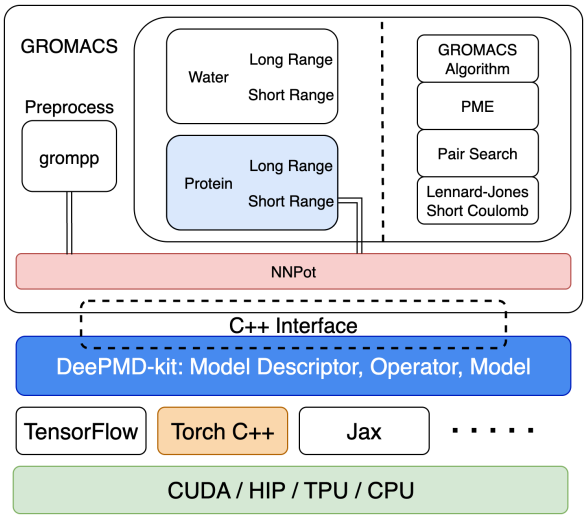


图 2：GROMACS 分子动力学架构及 DeePMD-kit 集成概览。

DeePMD 中的推理可参与所有这些阶段，因为这些阶段都需要计算作用于每个原子的力。

出于两个主要考虑，我们设计的 AIMD 工作流仅在最终分子动力学阶段使用深度势模型。首先，NVT 和 NPT 平衡阶段各自可能需要数万步。由于这些阶段仅为主要分子动力学运行提供起点，因此该阶段无需达到从头算质量，我们也就跳过了在此阶段使用深度势。其次，保留当前的预处理流程，可以使该工作流以相同的能量最小化和作为经典模拟与深度势模拟起点的平衡条件为基础。

AIMD 工作流旨在支持完整的深度势模拟或经典-深度势混合模拟，后者仅使用深度势模型模拟部分原子。具体而言，该工作流可以选择特定原子，使用深度势模型进行模拟。例如，在溶剂化蛋白质分子动力学模拟中，水和蛋白质原子可以根据具体应用场景，分别采用经典方法或 AIMD 方法处理。这一特性兼顾了 AIMD 方法的准确性与性能。

在 GROMACS 预处理阶段 `grompp`，即最终分子动力学运行阶段之前，NNPot 模块通过移除键和二面角，禁用 `protein` 组内的所有相互作用。此外，它还会将目标原子添加到排除列表中，以避免这些原子被纳入邻居列表，从而禁用非键短程相互作用。在分子动力学运行期间，NNPot 将蛋白质的原子数据传递给 DeePMD-kit 进行模型推理，以获得蛋白质原子之间的短程力。在图 2 所示的示例中，我们对蛋白质原子内部的相互作用使用深度势模型，但不对水原子使用。水原子的相互作用，包括蛋白质与水之间的相互作用，则由 GROMACS 中的经典算法提供。用于

深度势模型处理的原子组，可以使用 `mdp` 配置输入文件中的 NNPot 选项轻松完成选择。

我们设计的工作流旨在实现 GROMACS 与 DeePMD-kit 的紧密耦合。GROMACS 提供了用于对接机器学习模型的 NNPot 模块，但目前仅支持 PyTorch 模型。尽管 DeePMD-kit 为 GROMACS 2020.2 版本提供了代码补丁，通过将模型推理代码添加到 GROMACS 的核心函数 `do_力` 中，但这种集成需要对输入蛋白质进行侵入式预处理。综合考虑这两种方案，我们将 DeePMD-kit 推理集成为 GROMACS NNPot 模块的额外后端，并实现二者之间原子信息和力场的通信。

我们发现，在 AIMD 工作流中实现二者紧密耦合面临三项技术挑战。第一项挑战源于 DPA2 和 DPA3 模型内部的通信机制。与早期的深度势模型不同，早期模型只需获取 $Rcutoff$ 幽灵晕区域中的幽灵原子，而 DPA2 和 DPA3 则需要根据其内层 L 获取 $L \times Rcutoff$ 幽灵区域中的原子。我们通过扩大 GROMACS 子域的幽灵区域来解决这一问题，从而为 DPA 模型提供所需的原子信息。第二项挑战来自 GROMACS 中的非对称幽灵拓扑。由于经典方法是线性的，在两个子域的边界处，只有一侧会从另一子域获取幽灵原子，因为它们只需将两侧对同一原子的贡献相加。然而，深度势模型是非线性的，因此每个子域都需要完整的原子信息。我们通过在推理前进行通信，在每个子域中重建正确的拓扑和幽灵原子，从而解决了这一问题。第三项挑战是 GROMACS 神经网络势缺乏对并行推理的支持，因此无法获取幽灵原子信息。当前设计沿用了 NNPot 模块的设计：在模型推理前，将来自多个 GROMACS 子域的原子汇集到一个进程中，然后在该进程上执行深度势模型推理。虽然这简化了区域分解和幽灵原子的处理，但我们也承认，未来为了提高可扩展性，可能需要采用侵入性更强的区域分解方案。

我们的实现使用 DeePMD-kit 的 C++接口，并通过添加 DeePMD-kit 推理作为新的后端，扩展了 GROMACS NNPot 模块的功能。该实现支持在 GROMACS 分子动力学工作流中使用所有 DeePMD 模型系列（如 e2 a、DPA、DPA2、DPA3.....）及其后端（PyTorch、TensorFlow、JAX.....）。目前，NNPot 仅支持单个 MPI 进程上以 TorchScript 格式运行的 PyTorch 模型，且子域中的幽灵原子无法提供给模型。因此，我们为 NNPot 模块添加了新的推理后端，使用户能够在编译时选择原有的 PyTorch 后端或我们的 DeePMD 后端。我们还修改了 NNPot 接口，以便向推理运行时提供幽灵原子信息，这对未来实现并行推理至关重要。这些更改确保了长期可维护性和未来兼容性——正式承诺。

IV. 评估方法

我们使用四种蛋白质（1YRF、1UBQ、3LZM 和 2PTC）的四次分子动力学模拟，评估集成到 GROMACS 模拟工作流中的人工智能深度势模型所实现的从头算精度。这些蛋白质在大小、折叠拓扑和生物学相关性方面涵盖了广泛范围。它们的原子数分别为 582、1231、2643 和 4114。所有模拟均采用相同的 GROMACS 分子动力学模拟流程进行评估。

我们在表 Ia 中总结了模拟的关键参数。能量最小化阶段使用`steep` 积分器，且时间步数不固定。相反，模拟会持续运行，直到系统充分弛豫，且势能降低至所需阈值。另一方面，NVT、NPT 和分子动力学阶段使用`md`，并按照用户设定的模拟步数和时间运行。分子动力学期间所有短程相互作用的截断范围均设为 0.7 nm，以确保 DPA2 和 DPA3 模型配置具有安全的邻居列表裕度；这两个模型的截断范围为 0.6 nm。

在本研究中，我们在 GROMACS 模拟工作流程中采用了 DPA 模型系列中最新的两种深度势能模型，即 DPA2（`DPA-2.4-7M`）和 DPA3（`DPA-3.1-3M`）。DPA2 和 DPA3 是通过多任务方式预训练的基础模型，之后可以将其冻结，并分支出专用于不同下游任务的模型，正如第二节所述。为评估它们的跨系统泛化能力和可迁移性，在分子动力学实验中，我们将这两个 DPA 模型冻结为溶剂化_蛋白质_片段 分支。随后，我们测试了由 DPA2 模型衍生的`SPICE2` 分支，以展示模型架构和分支之间的差异。请注意，DPA2-溶剂化_蛋白质_片段和`DPA2-SPICE2`使用相同的描述符，但采用了基于不同数据集的不同拟合网络 。在本文余下部分中，我们将用溶剂化_蛋白质 作为溶剂化_蛋白质_片段的简称。

我们在两个配备不同 NVIDIA GPU 的测试平台上评估上述分子动力学模拟。第一个测试平台配备两块 A100-40GB GPU，我们在实验中使用其中一块。该平台配有一颗 AMD EPYC 7302P 16 核处理器作为 CPU，以及 256 GB DDR4主存。第二个测试平台配备一台 NVIDIA GH200 480 GB超级芯片。该 GH200 搭载一颗 72 核、基于ARM64的Grace CPU，配有 480 GB LPDDR5X内存；还搭载一块 Hopper 100 GPU（H100），配有 96 GB HBM3内存。CPU 与 GPU 通过缓存一致性 NVLink-C2C互连，可提供 900 GB/s 的带宽，也使利用 CPU-GPU统一内存进行模型推理成为可能 [15]。由于大多数机器学习推理工作负载都在 GPU 上执行，因此这一配置非常适合展示深度势能从头算分子动力学模拟工作流在当今基于GPU的高性能计算系统上的性能。

我们在表 Ib 中总结了软件环境信息。GROMACS 编译时启用了 CUDA 图形处理器支持。

所有分子动力学运行均使用单个 MPI 进程和 8 个 OpenMP 线程。在所有模型推理阶段，均使用 PyTorch 作为 DeePMD-kit 的后端，因为它是 DPA2 和 DPA3 模型的原生后端。

表 I：实验规格

(a) 测试案例参数

	EM	NVT 和 NPT	MD
Δt（fs）	-	2	1
步数	-	50K	10K
总时间（ps）	-	100	10
截断半径 (nm)	1.2	1.2	0.7
神经网络势	No	No	Yes
神经网络原子组	-	-	蛋白质

(b) 软件环境

	A100	GH200
GROMACS DeePMD-kit	2025.2 v3.1.0	2025.2 v3.1.0
PyTorch	2.7.1	2.5.1
CUDA	12.6	12.4
GCC	12.2	11.4
OS/Linux	CentOS8/4.18	RHEL9.4/5.14

在所有测试中，GROMACS 和 DeePMD 部分均以单精度浮点数作为主要数据精度。对于每个蛋白质基准，我们都会在不同蛋白质、模型和硬件测试平台的组合上运行多项测试。为了提供不考虑量子力学效应的分子动力学模拟参考，我们还进行经典分子动力学模拟，仅使用经验拟合量子规律来计算势能和赝势。由于某些基于 DPA3 的测试所需的 GPU 内存超过了特定硬件的可用内存，因此无法运行；更多详情见第 V 节。

我们使用 GROMACS 内置的计时与性能模块测量模拟性能，该模块报告模拟时间与实际挂钟时间之比，单位为 ns/day。这一指标常用于分子动力学基准测试，可清晰、一致地衡量模拟吞吐量；数值越高，表示性能越好。

使用命令行工具 `nvidia-smi` 每 10 秒追踪一次 GPU 内存使用量。报告的内存使用量反映了 GROMACS 分子动力学运行与 DeePMD 模型推理两者的总消耗。为了深入表征分子动力学工作流中的深度势模型，我们使用 Nsight Systems 和 Nsight Compute 对分子动力学模拟的前 200 步进行性能分析。此外，我们在 DeePMD-kit C++ 源代码中使用 NVTX 注释标记模型推理调用。随后捕获并可视化这些注释，以便精确识别相应的模拟步骤。

V. 评估结果

本节首先验证 AIMD 模拟工作流，然后从模拟吞吐量和可扩展性方面评估主要的机器学习分子动力学推理阶段。此外，我们还对两个图形处理器平台上的性能进行表征，包括 DPA2 和 DPA3 模型的内核执行和内存使用量。

我们通过在四项蛋白质基准测试中，将基于 DPA 的量子力学力场与经验模拟力场进行比较，评估 GROMACS 中达到从头算精度的模拟工作流程是否正确。图 3 展示了溶剂化蛋白质 1YRF 在启用 DPA2 和 DPA3 的量子力学效应后分别进行 10 ps 分子动力学运行的状态，并与经验模拟结果进行了比较。橙色球体表示原子及其相连的键，蓝色带状结构表示蛋白质的二级结构。从这三张快照中可以验证，蛋白质中心和下部附近的 α-螺旋二级结构（蓝色带状结构）在四种模拟设置中保持一致。此外，DPA2-SPICE2 和 DPA2 溶剂化蛋白质是通过冻结共享表示而获得的两个下游模型；该共享表示由多样化的通用数据集训练得到。通过考察这两个下游模型，我们验证了大原子模型的基本特性，例如跨系统泛化和可迁移性。

A. 深度势模型比较

我们使用两种大原子模型 DPA2 和 DPA3，在 A100 和 GH200 GPU 架构上评估四个蛋白质基准测试的模拟吞吐量。图 4 展示了两个 GPU 测试平台上的分子动力学模拟吞吐量，单位为 ns/day。对于 1YRF 蛋白质，DPA2 模型在 GH200 上实现了最高吞吐量 0.7 ns/day。随着蛋白质基准测试中的原子数从 582 个增加到 4,114 个，模拟吞吐量降至 0.2 ns/day。在 A100 图形处理器上，DPA2 模型的模拟吞吐量从 1YRF 蛋白质的 0.55 ns/day 降至最大蛋白质 2PTC 的 0.1 ns/day。请注意，由于 `DPA2-solvated_`蛋白质和 `DPA2-SPICE2`是源自同一 DPA2 模型的分支，因此它们具有相同的神经网络架构，吞吐量也完全相同。在两个 GPU 测试平台上，随着原子数增加，DPA2 模型的模拟吞吐量呈现近似线性的缩放趋势。这种线性扩展符合 DPA2 模型中机器学习推理的理论预期。与密度泛函理论（DFT）等第一性原理量子力学方法的三次方缩放相比，DPA 模型所实现的模拟吞吐量证实，深度势模型在达到从头算精度的分子动力学模拟中具有更优越的缩放特性。

与 DPA2 相比，基于图神经网络（GNN）的 DPA3 模型在所有蛋白质基准测试中的模拟吞吐量始终低于 DPA2。具体而言，在 A100 上，DPA2 的性能平均是 DPA3 的 $4.23\times$ 倍；在 GH200 上则是 $3.18\times$ 倍。与 DPA2 类似，随着测试蛋白质规模增大，DPA3 的模拟吞吐量也会下降，并呈现近似线性扩展。与 DPA2 不同，由于 GPU 显存溢出（OOM），DPA3 无法在 A100 上完成对 3LZM 和 2PTC 蛋白质的模拟，也无法在 GH200 上完成对 2PTC 蛋白质的模拟。我们在第 V-C 节中介绍了它们的 GPU 内存使用量特征。

采用 DPA2 和 DPA3 的分子动力学模拟，从较小的图形处理器架构迁移到较大的图形处理器架构时，呈现出不同的加速效果。尽管这两个测试平台的图形处理器峰值吞吐量从 19.5 FP32 太 FLOPS 提升至 60TFLOPS，但对于 1YRF 蛋白质基准，DPA2 的模拟吞吐量仅从 0.55 ns/day 提升至 0.7 ns/day。不过，DPA2 在其他蛋白质基准测试中的加速效果更显著，例如从 A100 切换到 GH200 时，$2.2\times$ 在 1UBQ、$2.9\times$在 3LZM 和 $2\times$ 在 2PTC。由于其他基准的原子数为 $2\times$-$7\times$ 1YRF 的原子数，因此

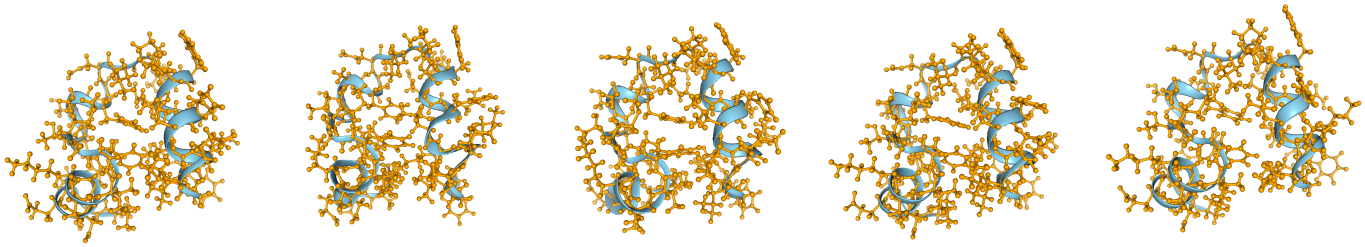


图 3：经过 10 ps 分子动力学模拟后 1YRF 蛋白质的快照。将初始结构（最左侧）与 DPA2 溶剂化蛋白质、DPA2-SPICE2、DPA3 溶剂化蛋白质以及经验 GROMACS 力场（最右侧）得到的轨迹进行比较。请注意，深度势模型与经验力场在局部原子位置上的偏差是预期之中的，因为深度势模型纳入了从头算级量子效应。

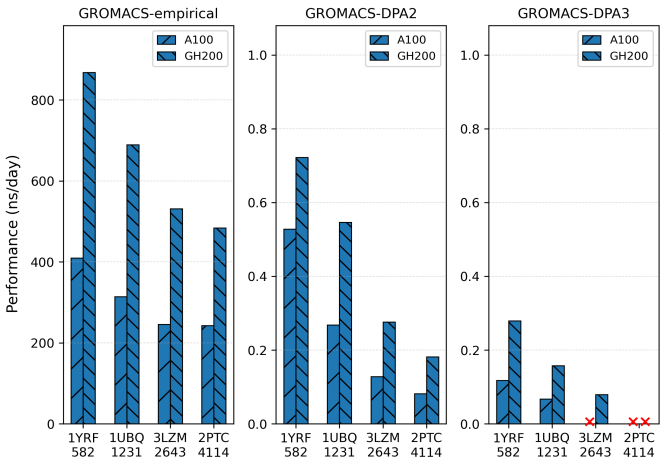


图 4：在 A100 和 GH200 上，对四种蛋白质模拟采用经典方法、从头算 结合 DPA2，以及从头算结合 DPA3 时的分子动力学模拟吞吐量。由于× 显存溢出，DPA3 的结果不可用。

它们的计算负载更高，因此能更充分地利用 GH200 增加的 GPU 容量。与 DPA2 不同，即使对于 1YRF 蛋白质，从 A100 切换到 GH200 也能使模拟吞吐量提升近 3×。DPA3 基于图神经网络架构，GH200 上 HBM3 内存带宽的提升可能对性能改善贡献更大。

为了解达到从头算精度所需的计算开销，我们还配置了经典分子动力学模拟，并在两个 GPU 测试平台上测量其吞吐量。经典分子动力学在 A100 测试平台上的吞吐量为 200 至 400 ns/day，在 GH200 测试平台上则为 500 至 900 ns/day。引入深度势后性能大幅下降，这表明尽管机器学习推理可以降低密度泛函理论中由 $\mathcal{O}(N^3)$ 带来的高计算成本，但与经验分子动力学模拟相比，从头算分子动力学模型的计算开销仍然更大。

表 II：使用 DPA2 和 DPA3 在 A100 上对 1UBQ 进行 200 步分子动力学模拟时，性能分析所得的主要 GPU 内核细分。

内核类型	DPA2		DPA3	
	时间（%）	启动次数	时间（%）	启动次数
GEMM	44.1	140k	8.0	210k
逐元素运算与索引	40.9	600k	75.5	1.7M
减少	7.5	120k	10.0	280k
其他	7.5	140k	6.5	410k
总计	**100**	**1.0M**	**100**	**2.6M**

要点 I： 基于深度势的从头算分子动力学模拟，其吞吐量大致随原子数量线性缩放，与密度泛函理论的三次方缩放不同。

要点 II： 在所有蛋白质基准测试中，*DPA2*深度势模型始终实现了比*DPA3*高3–4× 倍的模拟吞吐量。

B. 图形处理器上的深度势分子动力学特征分析

我们对图形处理器上的内核执行模式进行分析，以了解DPA2和DPA3的性能特征。DPA2和DPA3的分析轨迹表明，分子动力学模拟的运行时间主要由深度势模型推理决定。对于DPA2模型，图形处理器内核执行占总运行时间的96%，而DPA3为98.4%。表II列出了使用1UBQ蛋白质的分子动力学模拟时这两个模型的内核执行时间和核函数启动次数明细。

在 DPA3 中，逐元素和索引内核占运行时间的 75% 以上。仅索引内核（例如`indexFuncLargeIndex`和`indexSelectLargeIndex`）就占内核执行时间的 34% 以上。融合操作和散射/收集操作也占据了运行时间的很大一部分。这些内核主要负责在内存中重新排列、选择、散射或收集数据。因此，它们未能充分利用图形处理器的计算能力，却给内存子系统带来了更大压力。DPA3 中也包含与通用矩阵乘法相关的计算内核，但仅占运行时间的约 8%。DPA3 的这些内核特征与其图神经网络层架构

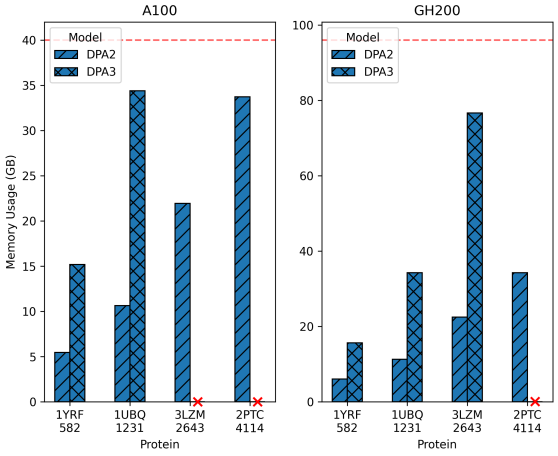


图 5：A100 和 GH200 上 DPA2 与 DPA3 模型的 GPU 内存使用量。红色虚线表示 GPU 容量。由于显存溢出，×标记的结果不可用。

相符。因此，DPA3 中的内存受限操作可能会受益于具有高内存带宽的硬件。DPA2 工作流中，密集矩阵乘法和索引操作的比例较为均衡。通用矩阵乘法核函数占运行时间的 44.1%，逐元素和索引内核则占 40.9%。通用矩阵乘法占主导的特性表明，DPA2 可能会受益于 Tensor Core 加速硬件。两个 DPA 工作流的核函数启动次数都显著更多：DPA3 为 2.61M，DPA2 为 1.01 M，这意味着启动开销增加，核函数融合效率降低。降低启动开销的一种常见优化方法是使用 CUDA Graphs。然而，分析结果表明，当前后端引擎并未使用 CUDA Graphs。

核函数特征凸显了两种模型中不同的优化重点。例如，核函数融合和 GEMM 调优是 DPA2 的主要优化机会。DPA3 则可以探索通过 CUDA Graphs 进行优化，以减少核函数启动开销，并融合较小的逐元素核函数，从而提高整体 GPU 占用率。鉴于深度势模型具有重复且可预测的特性，这些优化可能带来显著的性能提升。一个限制是深度势模型，尤其是 DPA系列，缺乏对域分解的支持。因此，推理时所有原子都需要在一块 GPU 上处理，从而限制了可扩展性和硬件可移植性。未来的工作应着重于实现跨 MPI 进程、借助幽灵原子进行分布式推理。

要点 III： *DPA2* 工作流在以 *GEMM* 为主的操作和索引操作之间较为均衡，而 *DPA3* 工作流则主要由受内存带宽限制的索引操作主导。

C. 图形处理器内存使用量

通过内核特征分析，我们发现 DPA2 和 DPA3 都存在大量受内存带宽限制的索引操作。结合 DPA3 频繁出现的显存溢出错误（第 V-A 节），我们决定进一步研究 GPU 内存对 DPA 分子动力学工作流的影响。图 5 展示了 DPA2 和 DPA3 在 A100 和 GH200 测试平台上的 GPU 内存使用量。两种 DPA 分子动力学工作流的 GPU 内存使用量都很高。对于包含 582 个原子的最小蛋白质基准，DPA2 工作流需要约 5 GB 的 GPU 内存。对于其他规模更大的基准，在 A100 图形处理器上，GPU 内存使用量依次增至 10、22，最终达到 34 GB。DPA2 的内存使用量在两个图形处理器测试平台上保持不变。

DPA3 的内存使用量始终显著高于 DPA2，在相同的蛋白质模拟中达到约 $3\times$ 。例如，对于最小的蛋白质基准，它使用 15 GB 内存；对于 1UBQ蛋白质，使用 35 GB 内存；对于 3LZM蛋白质，使用 70 GB 内存。由于内存占用更高，只有在 GH200 系统上才能使用 DPA3 模型运行 3LZM蛋白质。对于最大的结构 2PTC 蛋白质，预计 DPA3 的内存需求将超过 140 GB，因此超出 GH200 的 96 GB 容量，导致无法执行该实验。

两种 DPA 模型的内存需求都随原子数量线性增长。这符合预期，因为这两种模型的描述符都使用消息传递神经网络。在这两种模型的核心迭代过程中，邻居原子的表征会相互用作下一次更新的输入。这意味着，原子 i 的表征不仅会受到其邻居 j 的影响，还会受到 j 的邻居原子的影响，因为这些信息已在先前的更新中编码到 j 的表征中。换言之，原子的感受野可以扩展到邻居区域之外，以捕捉系统中更远距离的关系，但这可能会对计算和可扩展性产生负面影响。

高 GPU 内存使用量将成为生产级深度势能 AIMD 模拟的主要限制因素。这凸显了采用内存感知模型设计或针对更大系统的并行执行策略的必要性。这一发现表明，只有借助分布式内存系统，或利用大量 CPU 侧内存的异构内存系统（例如 NVIDIA GH200 [15] 和 AMD MI300A [16]），大规模模拟才切实可行。混合精度推理或激活检查点等未来优化也有望进一步缓解这些内存限制。

要点 IV： *DPA2* 和 *DPA3* 的内存使用量均随原子数量线性增长。*DPA3* 比 *DPA2* 需要 $3\times$ 更多 *GPU* 内存，因此限制了其在大型模拟中的可行性。

VI. 相关工作

GROMACS [17] 和 LAMMPS [1]等分子动力学软件利用高性能计算系统上 CPU 和图形处理器之间的多级并行来提升性能。包括 DeePMD 模型系列在内的深度神经网络势的出现，使多个领域的分子动力学模拟能够以相对较低的成本达到接近密度泛函理论的精度 [6], [18]。现在，若干分子动力学软件已支持深度势，可通过定制接口实现，也可通过与 DeePMD-kit 耦合实现

及类似框架。例如，DeePMD-kit 已与多个分子动力学软件包集成，包括 LAMMPS、i-PI 和 AMBER [19], ; 与此同时，许多软件也在开发各自的深度势接口。据我们所知，本研究首次通过 GROMACS 官方接口集成了最新版本的 GROMACS 和 DeePMD-kit，并首次对 DPA2 和 DPA3 模型在蛋白质分子动力学模拟中的 GPU 级性能进行了全面表征。

已有多项研究探索了深度势模型在各种高性能计算系统上的性能与可扩展性。Jia等 [20] 展示了在 NVIDIA V100 加速系统上使用深度势进行大规模分子动力学模拟，规模达到 1 亿个原子。Du等 [21] 在一个 4096 节点的 ARMv8 集群上实现了 5 亿个原子，并利用 ARM SVE 向量化指令集加速深度势推理。然而，大多数研究都集中于较早期的深度势模型，尚未探索新兴的基于注意力机制和图神经网络（GNN）的 DPA 模型系列。此外，迄今为止所研究的大多数基准系统仅限于水和铜等材料，而深度势在生物分子模拟中的性能仍未得到探索。本工作通过研究蛋白质体系中的基于注意力机制的 DPA2 和基于 GNN 的 DPA3 模型，填补了这些空白。

VII. 结论

在本工作中，我们将 DeePMD-kit 框架与 GROMACS 神经网络势相结合，从而在 GROMACS 中启用人工智能深度势。我们首先在多个蛋白质体系的图形处理器上验证了该工作流，结果表明力评估正确，且分子动力学轨迹稳定。随后，我们使用所提出的 AIMD 工作流，在四个溶剂化蛋白质分子动力学模拟中评估了两个最新的深度势大原子模型 DPA2 和 DPA3，以实现从头算质量。结果表明，在 A100 和 GH200 GPU 上，DPA2 的吞吐量最高分别比 DPA3 高出 $4.23\times$和 $3.18\times$。我们首次刻画了基于深度势、具备从头算质量的 GROMACS GPU 端到端分子动力学工作流。通过采用由同一预训练 DPA-2 基础模型衍生的两个下游模型，我们评估了大原子模型的基本性质，例如跨体系泛化能力和可迁移性。最后，我们确定内核启动开销和区域分解推理是分子动力学工作流中人工智能深度势的主要优化重点。

致谢

本研究得到了 SeRC（瑞典电子科学研究中心）SESSI 高效模拟软件计划的支持。

参考文献

[1] A. P. Thompson 等人，“LAMMPS——一种灵活的模拟工具，可用于原子、介观和连续尺度上的粒子材料建模”，《计算物理通讯》，第 271 卷，第 108171 页，2022 年。[2] M. J. Abraham 等人，“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”，《*SoftwareX*》，第 1–2 卷，第 19–25 页，2015 年。

[3] D. A. Case 等人，“Amber 生物分子模拟程序”，计算化学杂志，第 26 卷，第 1688 页，2005 年 16，第 1668 页，不详。[4]A. D. M. Jr. 等人，“ 用于蛋白质分子建模和动力学研究的全原子经验势”，《物理化学杂志 B》，第 102 卷，第 18 期，第 3586–3616 页，1998 年。 [在线]。获取地址：https://doi.org/10.1021/jp973084f [5] X. He、Y. Zhu、A. Epstein 和 Y. Mo，“从头算分子动力学模拟所得扩散性质的统计方差”，npj 计算材料，第 4 卷，第 1 期，第 18 页，2018 年。 [在线]。获取地址：https://doi.org/10.1038/s41524-018-0074-y [6] L. Zhang、J. Han、H. Wang、R. Car 和 W. E，“深度势分子动力学：具备量子力学精度的可扩展模型”，《物理评论快报》，第 120 卷，第 14 期，第 143001 页，2018 年 4 月。[7] H. Wang、L. Zhang、J. Han 和 W. E，“DeePMD-kit：用于多体势能表示和分子动力学的深度学习软件包”，《计算物理通讯》，第 228 卷，第 178–184 页，2018 年。 [在线]。获取地址：https://www.sciencedirect.com/science/article/pii/S0010465518300882 [8] D. Zhang 等人，“DPA-2：作为多任务学习器的大型原子模型”，npj 计算材料，第 10 卷，第 1 期，第 293 页，2024 年。 [在线]。获取地址：https://doi.org/10.1038/s41524-024-01493-2 [9] ——，“大型原子模型时代的图神经网络”，2025 年 6 月，arXiv:2506.01686 [physics]。 [在线]。获取地址：http://arxiv.org/abs/2506.01686 。R. Iftimie、P. Minary 和 M. E. Tuckerman，“从头算分子动力学：概念、近期进展与未来趋势”，美国国家科学院院刊，第 102 卷，第 6659 页，2005 年 19，第 6654 页，不详。[11] F. No´e、A. Tkatchenko、K.-R. M¨uller 和 C. Clementi，“分子模拟中的机器学习”，《物理化学年度评论》，第 71 卷，第 71 卷，2020 年，第 361–390 页，2020 年。[12] A. P. Bart´ok、M. C. Payne、R. Kondor 和 G. Cs´anyi，“高斯近似势：无需电子即可达到量子力学精度”，《物理评论快报》，第 104 卷，第 13 期，第 136403 页，2010 年 4 月，出版方：美国物理学会。 [在线]。获取地址：https://link.aps.org/doi/10.1103/PhysRevLett.104.136403 。/ [10] [13] J. Zeng 等人，“DeePMD-kit v：用于机器学习势的多后端框架”，《化学理论与计算杂志》，第 21 卷，第 9 期，第 4375–4385 页，2025 年 5 月。 [在线]。获取地址：https://doi.org/10.1021/acs.jctc.5c00340 . /acs.jctc. c D. Zhang 等人，“用于分子模拟的基于注意力机制的深度学习势模型预训练”，npj 计算材料，第 10 卷，第 1 期，第 94 页，2024 年。 [在线]。获取地址：https://doi.org/10.1038/s41524-024-01278-7 [14] [15] G. Schieffer、J. Wahlgren、J. Ren、J. Faj 和 I. Peng，“利用集成式 CPU-GPU 系统内存进行高性能计算：初探 Grace Hopper”，载于《第 53 届国际并行处理会议论文集》，丛书：ICPP ’24。ACM，2024 年，第 199–209 页。 [在线]。获取地址：https://doi.org/10.1145/3673038.3673110 . / 。J. Wahlgren 等人，“剖析 AMD MI300A APU 上的 CPU-GPU 统一物理内存”，载于《2025 IEEE 国际工作负载特征分析研讨会》，2025 年，第 368–380 页。[17] S. P´all 等人，“在 GROMACS 中实现分子动力学模拟的异构并行化与加速”，《化学物理学杂志》，第 153 卷，第 13 期，第 134110 页，2020 年 10 月。 [在线]。获取地址：https://doi.org/10.1063/5.0018516 . / 。J. Chang 等人，“使用 DPA-2 对生物分子进行高效、精确的力场优化”，2024 年 6 月，arXiv:2406.09817 [physics]。 [16] [18] 2406.09817 在线。获取地址：http://arxiv.org/abs/2406.09817 。J. Zeng 等人，“Deepmd-kit v：深度势模型软件包”，《化学物理学杂志》，第 159，5，054801，卷，第 期，第 页。08 2023 [在线]。获取地址：https://doi.org/10.1063/5.0155600 。/。 [19] W. Jia 等人，“利用机器学习将具备从头算精度的分子动力学扩展至 1 亿个原子”，载于《高性能计算、网络、存储与分析国际会议论文集》，丛书：SC ’20。IEEE 出版社，2020 年。[21] Q. Du 等人，“将深度学习分子动力学扩展至 4096 节点 ARMv8 集群上的 5 亿个原子”，载于《2025 IEEE 国际集群计算会议（CLUSTER）》，2025 年 9 月，第 1–12 页，ISSN：2168-9253。 [在线]。获取地址：https://ieeexplore.ieee.org/document/11186476 [20]