arXiv:2602.13140v1 [cs.LG] 13 Feb 2026

FlashSchNet：快速且准确的粗粒化神经网络分子动力学

Pingzhi Li[1], **Hongxuan Li**[1,2], **Zirui Liu**[3], **Xingcheng Lin**[2], **和 Tianlong Chen**[1]

[1]北卡罗来纳大学教堂山分校 [2] 北卡罗来纳州立大学 [3] 明尼苏达大学双城分校

图神经网络（GNN）势函数（如 SchNet）通过学习多体相互作用，提高了分子动力学（MD）模拟的准确性和可迁移性，但由于内核碎片化以及受内存带宽限制的流水线未能充分利用图形处理器，其速度仍慢于经典模型力场。我们表明，现有方法缺失的一项原则是使 GNN-MD 具备输入输出感知能力，仔细考虑图形处理器高带宽内存（HBM）与片上 SRAM 之间的数据读写。我们提出了 FlashSchNet，这是一种高效且准确的输入输出感知 SchNet风格图神经网络分子动力学框架，建立在四项技术之上：（1）*Flash* 径向基，将成对距离计算、高斯基展开和余弦包络融合为单次分块处理，在每个距离仅计算一次的同时，将其复用于所有基函数；（2）*Flash* 消息传递，融合截断、邻居收集、滤波器乘法和归约，避免在 HBM 中物化边张量；（3）*Flash*聚合，通过 CSR分段归约重新构造散射加法，将原子写入次数减少至原来的特征维度分之一，并在前向和反向传播中实现无竞争累加；（4）逐通道 *16* 位量化，利用 SchNet 多层感知机权重较低的逐通道动态范围，在几乎不损失准确性的情况下进一步提升吞吐量。在单个 NVIDIA RTX PRO 6000 上，FlashSchNet 在 64 个并行副本上对包含 269 个珠子的粗粒化（CG）蛋白质实现了**1000 ns/day** 的聚合模拟吞吐量（比 CGSchNet 基线快**6.5**× 倍，并将峰值内存降低**80%**），超越了经典力场（例如MARTINI），同时保持了 SchNet 级别的准确性和可迁移性。

代码： https://github.com/UNITES-Lab/flash-molecular-dynamics

1 引言

分子动力学（MD）模拟是计算化学、药物发现和材料科学中的核心工具，为探究分子运动的原子级分辨率提供了计算显微镜（Karplus & McCammon，2002年；Dror等人，2012；Hollingsworth & Dror，2018）。通过数值积分牛顿运动方程，MD能够生成时间分辨轨迹，将微观相互作用与宏观可观测量联系起来，从而实现热力学估算、构象探索以及对稀有事件的机制研究（Chodera & Noé，2014）。然而在实际应用中，经典模型MD始终面临一项权衡：经验力场速度快但属于近似，而Car–Parrinello等第一性原理分子动力学虽然更加精确，却要昂贵数个数量级（Car & Parrinello，1985）。即使采用MARTINI（Marrink等人，2007）等广泛使用的粗粒化模型，以牺牲原子级细节换取效率，对数百万至数十亿个时间步反复计算力仍是一个根本性瓶颈，限制了常规工作流程中可实现的时间尺度和系统规模（De Vivo等人，2016）。

受这一空白的启发，图神经网络（GNN）势能模型迅速发展，成为机器学习力场（MLFF）的主流类别之一。它们植根于几何深度学习原理（Bronstein 等人，

✉ 联系邮箱：{pingzhi, tianlong}@cs.unc.edu，xingcheng_lin@ncsu.edu，zrliu@umn.edu

预印本。审稿中。

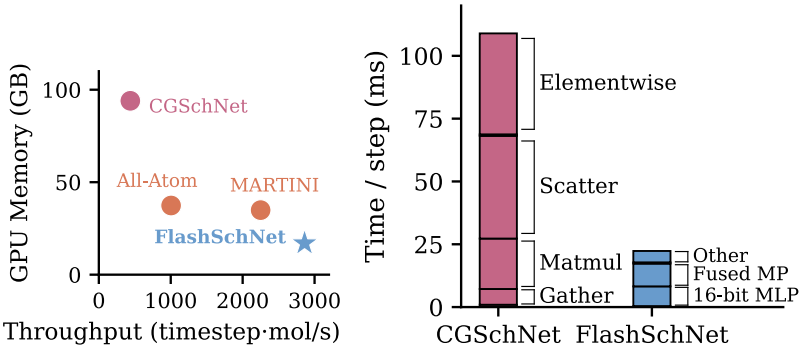

图 1：左：SchNet 式图神经网络分子动力学的内存与吞吐量权衡。与 CGSchNet 基线相比，FlashSchNet 在提升吞吐量 6× 的同时，实现了 5×的内存削减。右：各步骤耗时分解。FlashSchNet 通过融合内核和 16 位量化，消除了散射和逐元素运算造成的瓶颈。所有测试均在包含 269 个珠子的蛋白质（1ENH）上进行，共使用 64 个副本。

2021 年），这类模型将原子表示为节点，将局部相互作用表示为边，并利用消息传递以数据驱动且具有物理结构的方式捕捉多体效应（Gilmer 等人，2017 年）。SchNet（Schütt 等人，2017 年）以及后续的几何图神经网络（例如 DimeNet（Gasteiger 等人，2020年）和 E（3）-等变架构，例如 NequIP（Batzner 等人，2022 年）、Allegro（Musaelian 等人，2023）和 MACE（Batatia 等人，2022 年））已展现出很高的精度，以及在不同化学环境中的更强可迁移性，使机器学习势能模型能够以极低的计算成本逼近第一性原理精度。

然而，更高的准确率并未转化为更快的实际模拟速度。在 SchNet 式图神经网络分子动力学中，连续滤波器卷积（CFConv）会反复构建逐边特征（距离、径向基、截断函数），并应用小型多层感知机，随后对动态邻域执行散射式聚合。使用 PyTorch 和 JAX 等高级深度学习框架实现时，这些计算会被拆分为许多内核，并反复将中间边张量写入图形处理器的高带宽内存（HBM）；与此同时，聚合过程因原子更新争用而需要大量同步。因此，尽管名义浮点运算次数不高，该工作负载仍严重受内存带宽限制，且未能充分利用图形处理器。例如，CGSchNet（Charron 等人，2025年）在包含 269 个珠子的粗粒化（CG）蛋白质上运行 64 个并行副本时，模型浮点运算利用率（MFU）仅为 **2.5%** 模型浮点运算利用率（MFU）[1]。这些观察揭示了实用图神经网络分子动力学所缺失的一项原则：通过优化高带宽内存与片上 *SRAM* 之间的读写，使整个流程具备输入输出感知能力。

我们识别出 SchNet 式图神经网络分子动力学中的四个主要瓶颈，它们都源于内存读写：❶ 径向基展开 在不同内核中分别计算成对距离、高斯基函数值和余弦截断值，并将中间张量（距离、展开后的基函数、截断值）写入高带宽内存，尽管每个张量都只会被使用一次；❷ 消息传递 为截断掩码、邻居收集、滤波器乘法和散射聚合分别启动不同操作，在各阶段之间将大小为 $O(E \times F)$（边数 × 特征维度）的边张量写入高带宽内存；❸ 散射聚合 使用原子加法累积消息，在高邻域密度下会产生 $O(E \times F)$ 次冲突原子写入，导致操作串行化；❹滤波网络 为每条边反复加载多层感知机权重，使这些小型矩阵乘法严重受带宽限制。

我们提出：❶ *Flash* 径向基 将成对距离计算、高斯基展开和余弦包络融合为一次分块遍历，每个距离只计算一次，并在芯片上将其复用于所有基函数。❷ *Flash* 消息传递 将截断掩码、邻居收集、滤波器乘法，

[1]MFU 的计算方式是将实际达到的每秒万亿次浮点运算除以图形处理器的峰值每秒万亿次浮点运算。我们启用 TF32 张量核心，并报告在单张 NVIDIA RTX PRO 6000 上的测量结果。

和归约融合到一个内核中，从而避免在高带宽内存中物化中间边张量。 ❸ *Flash*聚合 通过 CSR 分段归约重新设计散射加法，将原子写入减少至原来的特征维度分之一，并在前向和反向传播中实现无竞争累加。 ❹ 逐通道 *16* 位量化 利用 SchNet 多层感知机权重较低的逐通道动态范围，在几乎不损失精度的情况下进一步提升吞吐量。如图 1所示，FlashSchNet实现了显著的端到端加速和内存节省。我们的贡献总结如下：

- 我们发现 IO 效率低下是 SchNet 式图神经网络分子动力学的关键瓶颈，并展示了如何利用模型固有结构——即利用图稀疏性实现无争用的压缩稀疏行聚合，并利用较低的逐通道动态范围进行 16 位权重量化——从算法层面减少内存流量。

- 我们将这些见解转化为四种内核级实现（即Flash 径向基、Flash 消息传递、Flash聚合和量化滤波网络），共同消除了中间张量物化，实现了端到端加速并节省了内存。

- 我们将这些技术整合到 FlashSchNet 中，相较于 CGSchNet 基线，实现了**6.5× 加速**和**80% 的内存削减**。据我们所知，这是首个性能超越 MARTINI 等经典粗粒化力场的 SchNet 式图神经网络分子动力学方法：它在单张 RTX PRO 6000 上，对包含 269 个珠子的粗粒化蛋白质进行 64 个并行副本的模拟，总吞吐量达到**1000 纳秒/天**，同时保留了学习势能的准确性和可迁移性。

2 相关工作

分子动力学模拟与机器学习力场。分子动力学（MD）模拟是研究分子系统的基础。传统力场（例如 AMBER（Wang 等，2004 年）、CHARMM（Brooks 等，2009））计算效率高，但由于函数形式固定，可迁移性有限。从头算分子动力学（Car & Parrinello，1985；Kühne 等，2020年）通过第一性原理计算实现了高精度，但其 $\mathcal{O}(N^3)$ 缩放限制了适用范围。机器学习力场（MLFF）弥合了精度与效率之间的差距。早期方法包括核方法，如 GAP（Bartók 等，2010）和 sGDML（Chmiela 等，2019），以及神经网络势能（Behler & Parrinello，2007）。基于图神经网络的机器学习力场，例如 SchNet（Schütt 等，2017 年）、DimeNet（Gasteiger 等，2020年）和 PhysNet（Unke & Meuwly，2019 年），直接作用于分子图，具有更强的泛化能力。 $E(3)$ 等变模型，包括 NequIP（Batzner 等，2022 年）、Allegro（Musaelian 等，2023）和 MACE（Batatia 等，2022 年），通过保持几何对称性实现了更高的数据效率，但计算成本也更高。近期的通用机器学习力场（Ju 等，2025年；Neumann 等，2024；Yang 等，2024）通过大规模训练增强了可迁移性，但推理成本仍是大规模模拟的瓶颈。

近期研究探索了粗粒化图神经网络力场，旨在提升可扩展性，同时保持物理保真度。Airas 和 Zhang (Airas 和 Zhang，2026年) 通过从蛋白质语言模型中提炼结构先验，提出了一种溶剂感知粗粒化势，重点关注二级结构和溶剂暴露。Majewski 等人 (Majewski 等人，2023) 开发了神经网络粗粒化力场，利用长时间的原子级轨迹，在多种蛋白质中重现蛋白质热力学。Charron 等人 (Charron 等人，2025年) 提出了一种可迁移粗粒化模型，可泛化至未见过的序列，并准确预测折叠能量景观和突变效应。尽管这些模型提升了物理保真度，且在某些情况下增强了泛化能力，但其推理仍受内存和 I/O 限制，难以扩展至长轨迹或大型生物分子系统。

高效图神经网络。一类研究通过在不同系统层级优化稀疏消息传递算子来提升图神经网络的效率，包括采用专用 CUDA 内核的图中心框架（Wang 等人，2019 年；Fey 和 Lenssen，2019 年）、根据图结构调整执行方式的运行时系统（Wang 等人，2021年）、融合算子并减少内核启动次数的编译器栈（Xie 等人，2022 年），以及采用兼容 CSR 设计的加速稀疏原语，如 SpMM 和采样稠密稠密矩阵乘法（Chen 等人，2020年；Huang 等人，2020年；Rahman 等人，2021年）。这些工作主要针对大型、基本静态图上的通用图神经网络负载，此类负载以稀疏线性代数运算为主。SchNet风格图神经网络的分子动力学则大不相同：动态邻居列表、带有逐边多层感知机的连续滤波卷积，以及力计算对高效反向传播的需求，使得反复生成边张量

表1：全文所用符号汇总。

符号	描述
N	原子或珠子的数量
E	邻居图中有向边的数量
D	隐藏特征维度
D_r	径向基维度
T	交互块数量
r_{cut}	邻居列表构建的截断半径
$\mathbf{r}_i \in \mathbb{R}^3$	原子 i 的位置
$\mathbf{x}_i^{(t)} \in \mathbb{R}^D$	第 t 层原子 i 的隐藏特征
$\mathrm{X}^{(t)} \in \mathbb{R}^{N\times D}$	所有原子的堆叠隐藏特征
源节点、目标节点 $\in \{1,\dots,N\}^E$	边的源和目标索引数组
$d_e \in \mathbb{R}$	边 e 的标量距离
$\mathbf{b}_e \in \mathbb{R}^{D_r}$	边的径向基向量 e
$\mathrm{B} \in \mathbb{R}^{E\times D_r}$	所有边上的堆叠径向基
$\mathbf{w}_e \in \mathbb{R}^D$	边的连续滤波器 e
$\mathrm{W} \in \mathbb{R}^{E\times D}$	所有边上的堆叠滤波器
$\mathbf{m}_e^{(t)} \in \mathbb{R}^D$	第 t层边 e 的消息
$\mathcal{E}$	总势能
$\epsilon_i \in \mathbb{R}$	来自原子 i的单原子能量贡献
$\mathbf{F}_i \in \mathbb{R}^3$	原子 i 上的力

物化和争用密集型散射加法成为关键瓶颈，因此需要采用输入输出感知融合和无争用的 CSR 风格聚合。

内存受限运行时优化。 现代图神经网络和机器学习势工作负载通常受内存限制，因为不规则的收集/散射操作与小型稠密内核交错执行，使吞吐量主要受数据移动而非浮点运算次数的制约。FlashAttention（Dao 等人，2022 年）是输入输出感知算法设计的典范，它会明确考量高带宽内存与静态随机存取存储器之间的数据流量，并通过分块和重新计算最大限度地实现片上复用。在图神经网络运行时中，高层框架通常会将消息构建和聚合拆分为多个内核执行，这些内核会将中间结果物化，并遭受原子操作争用（Gong 等人，2025年），因此需要采用融合技术来减少内存流量和启动开销（Liu 等人，2024）。在分子模拟领域，TorchMD-Net 2.0 通过优化邻居搜索和高效的力计算，对模拟技术栈进行工程优化，从而实现了显著的加速（Pelaez 等人，2024）。这些工作表明，要显著缩短实际运行时间，就需要进行端到端的输入输出感知重设计，协同优化特征构建、消息传递和聚合。

3 背景

我们在表1中总结了本文使用的符号。第3.1 节回顾分子动力学和力计算。第3.2 节介绍 SchNet 架构，并重点说明主导运行时间的算子。第3.3 节展示促使我们采用 IO 感知设计的硬件瓶颈。

3.1 分子动力学与力计算

分子动力学（MD）通过反复计算力 $\{\mathbf{r}_i\}_{i=1}^N$，并积分运动方程，模拟原子/珠子位置 $\{\mathbf{F}_i\}$ 的演化（Karplus& McCammon，2002年）。在基于能量的分子动力学中，力场定义了一个标量势能 $\mathcal{E}(\{\mathbf{r}_i\})$，力为：

$$\mathbf{F}_i = -\nabla_{\mathbf{r}_i}\mathcal{E} \in \mathbb{R}^3.$$

随后，时间积分器通过以下方式更新状态：

$$\mathbf{r}_i \leftarrow \mathbf{r}_i + \Delta t\,\mathbf{v}_i + \cdots, \qquad \mathbf{v}_i \leftarrow \mathbf{v}_i + \Delta t\,\mathbf{F}_i/m_i + \cdots,$$

其中 $\mathbf{v}_i$ 表示速度，m_i 表示质量，省略项取决于所选的恒温器积分器（例如朗之万）。因此，每个分子动力学步都需要 (i) 计算 $\mathcal{E}$，以及 (ii) 通过反向传播得到 $\mathbf{F}_i$，这使得端到端吞吐量主要受学习势能函数的前向 和反向 计算开销影响。

3.2 SchNet 模型

SchNet（Schütt 等人，2017 年）是一种连续滤波器消息传递网络，根据原子位置 E $\{\mathbf{r}_i\}_{i=1}^N$ 和类型 $\{Z_i\}_{i=1}^N$预测势能，并通过 $\mathbf{F}_i = -\nabla_{\mathbf{r}_i} E$获得力。模型维护原子级隐藏特征 $\mathbf{x}_i^{(t)} \in \mathbb{R}^D$ （堆叠为 $\mathbf{X}^{(t)} \in \mathbb{R}^{N\times D}$），并在径向截断所生成的稀疏邻居图上，迭代应用距离依赖相互作用。下面我们将按照 Charron 等人 （2025年）所采用的具体架构，介绍计算流程和主要算子。

构建邻居列表。给定位置后，SchNet 首先构建一个截断半径为r_{cut}的邻居列表，该列表由两个索引数组 src、dst $\in \{1,\dots,N\}^E$表示，用于索引 E有向边。每条边e 表示从源节点 j =src$[e]$ 到目标节点 i = dst$[e]$的一次相互作用。

距离和径向基。 对于每条边 e，SchNet 都会计算位移向量和标量距离

$$\mathbf{u}_e = \mathbf{r}_{\texttt{dst}[e]} - \mathbf{r}_{\texttt{src}[e]} \in \mathbb{R}^3, \qquad d_e = \|\mathbf{u}_e\|_2,$$

并将 d_e 展开为一个 D_r 维径向基向量 $\mathbf{b}_e = \text{RBF}(d_e) \in \mathbb{R}^{D_r}$，通常还会通过平滑截断包络进行调制。对所有边进行堆叠后得到 B $\in \mathbb{R}^{E\times D_r}$。

滤波器网络。 一个小型多层感知机将每个径向基向量映射为 D 维连续滤波器：

$$\mathbf{w}_e = \text{MLP}_{\text{filter}}(\mathbf{b}_e) \in \mathbb{R}^D,$$

生成堆叠滤波器 W $\in \mathbb{R}^{E\times D}$。由于这个多层感知机会针对每条边进行计算，因此生成的张量规模为 $O(E \times D)$，并构成内存流量的主要来源。

CFConv 消息传递与聚合。 连续滤波器卷积（CFConv）通过将源节点特征与学习得到的滤波器逐元素相乘来生成边消息，$\mathbf{m}_e^{(t)} = \mathbf{x}_{\texttt{src}[e]}^{(t)} \odot \mathbf{w}_e \in \mathbb{R}^D$，并将这些消息通过对入边求和聚合到目标节点上：$\mathbf{h}_i^{(t)} = \sum_{e:\ \texttt{dst}[e]=i} \mathbf{m}_e^{(t)} \in \mathbb{R}^D$。随后，带有残差连接的逐点更新网络生成$\mathbf{x}_i^{(t+1)}$；这一交互块会重复 T 次。标准实现通过 scatter_add 执行聚合；当多条边共享同一目标节点时，这会导致 $O(E \times D)$ 次原子写入，并产生严重争用。

能量读出。 经过 T 交互块后，输出多层感知机将原子特征映射为逐原子能量贡献：

$$\epsilon_i = \text{MLP}_{\text{out}}(\mathbf{x}_i^{(T)}) \in \mathbb{R}, \qquad E = \sum_{i=1}^{N} \epsilon_i,$$

可选择与先验能量项（例如，成键相互作用）组合，作为 $E \leftarrow E + E_{\text{prior}}$ (Charron 等人，2025年）。SchNet 包含多个多层感知机子模块，即多层感知机$_{\text{filter}}$、分块更新网络，以及多层感知机$_{\text{out}}$。由于反复加载权重，它们都受带宽限制。

通过自动微分求力。 在分子动力学中，力是通过对标量能量关于位置求导得到的：$\mathbf{F}_i = -\nabla_{\mathbf{r}_i} E \in \mathbb{R}^3$。这需要反向传播经过基于邻居列表索引的距离与径向基函数计算、所有多层感知机子模块以及聚合算子。关键在于，聚合的反向传播同样涉及散射式累积（此时针对源节点），因此前向与反向效率对于实际的分子动力学吞吐量都至关重要。

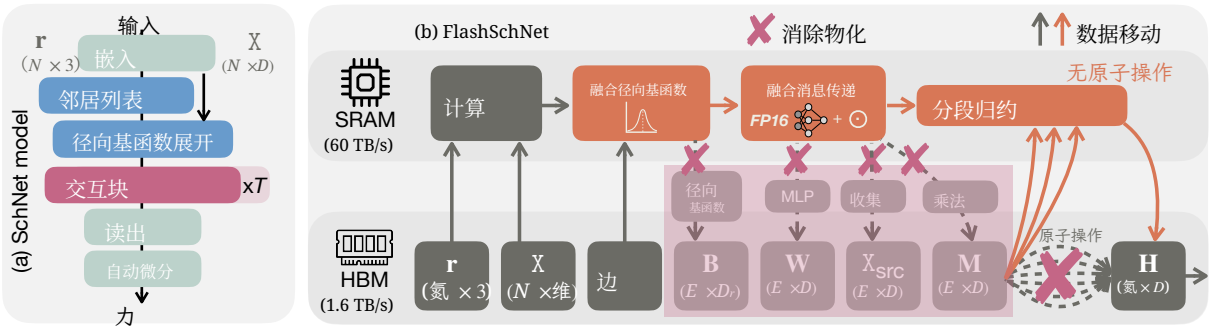

图2：（a）用于分子动力学的 SchNet 模型架构：原子位置$\mathbf{r}$ 和嵌入$\mathbf{X}$ 经由邻居列表构建、径向基展开和 T 交互块进行处理，随后通过能量读出和自动微分计算力。（b）FlashSchNet I/O 感知执行模型。基线流水线（下方，阴影部分）将中间边张量（$\mathbf{B} \in \mathbb{R}^{E\times D_r}$,$\mathbf{W}, \mathbf{X}_{\text{src}}, \mathbf{M} \in \mathbb{R}^{E\times D}$）物化到高带宽内存，并使用原子散射进行聚合。FlashSchNet（上方，橙色）将这些操作融合为三个内核，使中间结果保留在静态随机存取存储器中：融合径向基函数 在一次遍历中计算距离、高斯基展开和余弦包络；融合消息传递 结合半精度浮点数滤波器多层感知机、邻居收集和逐元素乘法；分段归约 以无争用的压缩稀疏行式累加取代原子散射加法。红色叉号表示已消除的高带宽内存物化。FlashSchNet 流水线将内存流量减少 $\sim E/N$，并消除了所有原子操作争用。

3.3 硬件性能挑战

我们聚焦于图形处理器。现代图形处理器可提供很高的峰值浮点运算次数，但常常受限于高带宽内存（HBM）与片上存储之间的内存流量。对于SchNet 式图神经网络分子动力学，占主导地位的算子是邻居图上稀疏的、基于索引的流水线，因为它们会反复物化大型边张量（例如，$\mathrm{B} \in \mathbb{R}^{E\times D_r}$ 和$\mathrm{W} \in \mathbb{R}^{E\times D}$），并执行相对于其HBM读写量而言算术强度较低的散射式归约。这使得运行时间主要受带宽限制，而碎片化的内核还会因反复加载和存储中间结果而进一步降低有效吞吐量。

边中间结果带来的高带宽内存流量。 SchNet流水线的核心会将边距离扩展为 $\mathrm{B} \in \mathbb{R}^{E\times D_r}$，并将滤波器扩展为 $\mathrm{W} \in \mathbb{R}^{E\times D}$，概念上形成行数为$\mathbf{m}_e^{(t)}$的边消息 $\mathbf{M}^{(t)} \in \mathbb{R}^{E\times D}$。即使每个元素的计算量不大，写入和重新读取这些边张量仍会导致每个块产生$O(E \cdot D_r)$和 $O(E \cdot D)$的高带宽内存流量；这一流量会随 T 个交互块而增加，并且在力计算所需的反向传播中再次增加。

图归约中的散射竞争。 CFConv中的聚合通常使用scatter_add，将$\mathbf{m}_e^{(t)}$累加到目标节点$\mathbf{h}_i^{(t)}$。这会执行 $O(E \cdot D)$次原子更新；当许多边共享同一目标节点（即局部度较高）时，并发原子操作会串行执行，显著降低吞吐量。此外，聚合的反向传播也需要进行散射累加，因此争用会同时影响前向和反向传播，直接限制分子动力学步的实际运行时间。

混合精度与张量核心。 图形处理器提供专用的张量核心，可在多种精度下加速矩阵乘法和融合多层感知机基本操作（例如 半精度浮点数）。在 SchNet 中，滤波器、更新和读出网络由多层感知机构成，其权重会被反复加载，因此对计算吞吐量和内存带宽都很敏感。使用半精度浮点数权重和激活值，同时将关键累加和力输出保留为 FP32，可以减少权重／激活值流量，并通过将这些多层感知机映射到张量核心上来提高计算吞吐量。

4 Flash-SchNet

本节介绍FlashSchNet，这是一种输入输出感知的 SchNet 式图神经网络分子动力学实现，可加速端到端的分子动力学步，包括前向能量评估 + 反向力计算。如图2所示，我们的设计针对第3节指出的主要瓶颈。从总体上看，FlashSchNet 沿用基线 CGSchNet（Charron 等人，2025年）的架构，但：(a) 融合单次使用边流水线，避免在高带宽内存中物化大型中间结果；(b) 以无争用的压缩稀疏行分段归约取代原子散射归约；(c) 对多层感知机子模块应用逐通道 16 位量化，在无损的前提下降低计算时间和内存流量。

FlashSchNet 的计算目标是 FlashSchNet。 考虑邻居图 $(\text{src}, \text{dst}) \in \{1, \dots, N\}^E$ 下索引为 t 的层中的一个交互块。对于边 e，定义位移 $\mathbf{u}_e = \mathbf{r}^{\texttt{dst}[e]} - \mathbf{r}^{\texttt{src}[e]}$、距离$d_e = \|\mathbf{u}_e\|_2$、径向基 $\mathbf{b}_e = \text{RBF}(d_e) \in \mathbb{R}^{D_r}$以及平滑截断包络 $C(d_e) \in \mathbb{R}$。CFConv 聚合可写为

$$\mathbf{h}_i^{(t)} = \sum_{e:\ \texttt{dst}[e]=i} \left(\mathbf{x}^{(t)}_{\texttt{src}[e]} \odot \mathbf{w}_e \right),$$

其中 $\mathbf{w}_e = \text{MLP}_{\text{filter}}\big(\mathbf{b}_e \cdot C(d_e)\big) \in \mathbb{R}^D$。基线实现通常会将 B $\in \mathbb{R}^{E \times D_r}$ 和 W $\in \mathbb{R}^{E \times D}$ 物化为高带宽内存中的中间结果，并使用 scatter_add 实现求和，从而导致大量内存流量和原子操作争用。FlashSchNet 在避免边张量物化并消除聚合路径上的原子操作的同时，计算相同的 $\mathbf{h}_i^{(t)}$ 。

4.1 SchNet 交互的输入输出感知重构

单次使用边流水线。 每条边的计算形成一条单次使用链：距离到径向基，再到滤波器多层感知机，最后到门控消息。我们将这条链视为一个流式算子并对其进行融合，使得 $\mathbf{u}_e$、d_e、$\mathbf{b}_e$ 以及多层感知机$_{\text{filter}}$ 内部的中间激活值能够在片上生成和消费。从概念上讲，我们用融合边算子 $\mathbf{h}_i^{(t)} = \sum_{e:\ \texttt{dst}[e]=i} \Psi\big(\mathbf{x}^{(t)}_{\texttt{src}[e]}, \mathbf{r}^{\texttt{src}[e]}, \mathbf{r}_{\texttt{dst}[e]}\big)$ 替代显式边张量，其中Ψ 封装了距离计算、径向基与包络评估、滤波器多层感知机以及门控。

基于力的模拟的精度约定。 力需要通过 $d_e = \|\mathbf{u}_e\|_2$ 以及截断函数和基函数进行梯度计算。我们将位置 $\mathbf{r}_i$、距离 d_e、能量累积 $\mathcal{E}$以及力输出 $\mathbf{F}_i$ 保留为 FP32，并对归约使用 FP32 累积。W16A16 应用于 SchNet 多层感知机子模块，详见第 4.4 节。

4.2 Flash 消息传递：融合边计算

融合前向算子。 对于每条边 e with (j, i) = (src[e],dst[e])，我们计算：

$$\mathbf{u}_e = \mathbf{r}_i - \mathbf{r}_j, \quad d_e = \|\mathbf{u}_e\|_2,$$
$$\widetilde{\mathbf{b}}_e = \text{RBF}(d_e) \cdot C(d_e), \quad \mathbf{w}_e = \text{MLP}_{\text{filter}}(\widetilde{\mathbf{b}}_e),$$

然后构造消息 $\mathbf{m}_e^{(t)} = \mathbf{x}_j^{(t)} \odot \mathbf{w}_e$ ，并将其直接输入聚合操作以计算 $\mathbf{h}_i^{(t)}$ 。这样就无需将 B 和 W 作为高带宽内存中的中间结果物化。

片上复用。 我们对边进行分块，并组织计算，使短时间窗口内会重复使用的值（例如，共享同一目标节点的边所对应的 $\mathbf{r}_i$ ）保存在寄存器或共享内存中。即使这会引入少量重新计算，也能减少全局内存流量。

4.3 Flashaggregation 分段归约

前向传播中的按目标分组的分段归约。为避免原子操作争用，我们按目标节点对边重新排序，并执行分段归约。令 dst_ptr $\in \{0, \dots, E\}^{N+1}$和 perm $\in \{1, \dots, E\}^E$定义按目标分组的布局，其中节点 i对应的边占据

$$p \in [\texttt{dst_ptr}[i],\ \texttt{dst_ptr}[i+1]).$$

Then

$$\mathbf{h}_i^{(t)} = \sum_{p=\texttt{dst_ptr}[i]}^{\texttt{dst_ptr}[i+1]-1} \mathbf{m}_{\texttt{perm}[p]}^{(t)}.$$

我们将每个目标分段的独占所有权分配给一个块，在寄存器中累加，并为每个特征通道仅写入一次。

反向传播中的源节点分组分段归约。相对于源特征的主要梯度如下：

$$\nabla \mathbf{x}_j^{(t)} = \sum_{e:\ \texttt{src}[e]=j} \nabla \mathbf{h}_{\texttt{dst}[e]}^{(t)} \odot \mathbf{w}_e.$$

我们通过构建源节点分组布局，并对 $\nabla \mathbf{x}_j^{(t)}$ 采用相同的独占所有权原则进行累加，从而避免原子操作争用。

动态邻居列表下的索引构建。邻居列表可能会在分子动力学步之间发生变化，因此当 (src,dst) 发生变化时，必须重建分组布局。我们分别使用 dst 和 src进行桶排序，构建目标节点分组索引和源节点分组索引，从而为每个节点生成连续边段，以支持独占所有权分段归约。我们在第 5节中，将此桶排序开销与整体加速比一并报告。

4.4 多层感知机子模块的 W16A16 混合精度

动机。 SchNet 滤波网络呈现出明显的通道级幅值结构。如图 3所示，权重幅值在不同输出通道间分布不均；这一模式在各个交互块中均保持一致，因此采用通道级量化，几乎无损地降低多层感知机计算和输入输出成本。

通道级量化。 我们对所有多层感知机子模块采用 W16A16（16 位权重、16 位激活值），包括多层感知机$_{\text{filter}}$、分块更新网络和读出网络多层感知机 $_{\text{out}}$。我们改进最优大脑压缩（Frantar 等人，2023）方法，为每个线性层计算逐通道量化尺度，以最小化量化损失。使用半精度浮点数权重和中间激活值后，多层感知机通用矩阵乘法可映射到张量核心，从而减少权重和激活值的存储开销。

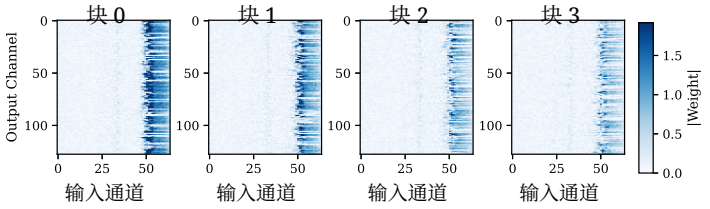

图3：滤波网络呈现出明显的通道级幅值分布，因此可采用通道量化实现无损加速。

相较于 FP32 的通信开销。我们将位置 $\mathbf{r}_i$、距离 d_e、能量累积 $\mathcal{E}$和力输出$\mathbf{F}_i$保留为 FP32，并按照第 4.1 节所述，在归约过程中使用 FP32 累积。

4.5 端到端集成

在每个分子动力学步中，我们构建邻居列表，在启用时更新分段归约索引，运行T 采用融合消息传递和分段归约的交互块，通过读出网络计算能量，并通过自动微分获得力 $\mathbf{F}_i = -\nabla_{\mathbf{r}_i}\mathcal{E}$。该配置支持在第5节中对融合、分段归约和 W16A16 进行受控消融实验。

成本分析。 FlashSchNet 避免在 HBM 中物化 B $\in \mathbb{R}^{E\times D_r}$ 和 W $\in \mathbb{R}^{E\times D}$，将每步的主要 IO 从 $\text{IO}^{\texttt{base}} = \Theta\big(T \cdot E(D_r + D)\big) + \Theta\big(T \cdot ED\big)$ 降至 $\text{IO}^{\texttt{flash}} = \Theta\big(T \cdot (ED + ND)\big)$，消除了径向基和滤波器物化，并以无争用的分段存储 $O(ND)$ 替代了争用严重的 $O(ED)$ 原子聚合。由于 $E \gg N$在典型的图神经网络分子动力学中（例如， 10^5 对比 10^2），总 IO 降低了 $\sim E/N$。16位量化进一步将多层感知机权重和激活值流量减少一半。

5 实证评估

5.1 端到端结果

实验设置。 我们根据FlashSchNet 在 Charron 等人的基准测试套件基础上，对五种快速折叠蛋白质进行评估（2025年）：Chignolin（CLN，10 个残基）、TRPcage（2JOF，20 个残基）、Homeodomain（1ENH，54 个残基）、Villin（1YRF，35 个残基）和 Alpha3D（2A3D，73 个残基）。所有模拟均在 300 K 下采用 64 个并行副本，并在单个 NVIDIA RTX PRO 6000 GPU 上使用 4 fs 的步长。我们将其与三种参考方法进行比较：CGSchNet（Charron 等人，2025年）（FP32 参考机器学习力场）、经典 MARTINI 力场（Marrink等人，2007）以及全原子模拟。结构保真度通过 Cα 均方根偏差、天然接触分数 Q和 GDT-TS 进行评估。吞吐量以每秒每摩尔的时间步数表示·mol/s（即，所有副本汇总后的每秒模拟步数）。更多详情见附录A。

折叠动力学得以保持。 为验证我们的优化保留了底层势能的物理保真度，我们对Chignolin、TRPcage和Villin各模拟16 ns。图4展示了均方根偏差与 Q 随模拟时间的演化。所有轨迹均表现出多次可逆折叠转变，且均方根偏差与 Q之间呈预期的负相关。Chignolin在折叠（$Q > 0.8$）与未折叠（$Q < 0.4$）状态之间表现出纳秒尺度的快速转变；TRPcage表现出动态涨落， Q 在0.4与0.9之间振荡；Villin在亚稳态盆地中的停留时间更长，并多次到达天然态（$Q > 0.85$）。这些结果证实 FlashSchNet 能够正确采样构象景观，且未引入数值伪影。

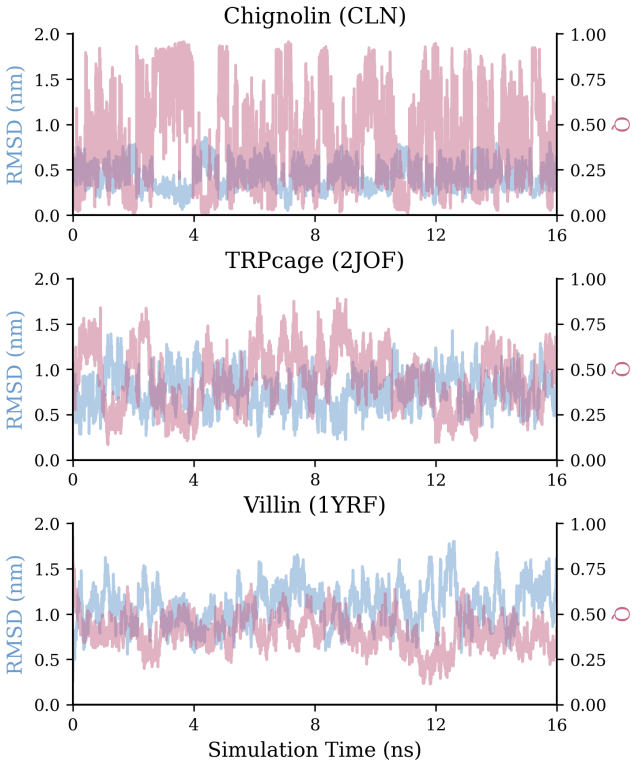

图 4：使用 FlashSchNet 模拟的三种快速折叠蛋白质的 Cα 均方根偏差与天然接触分数（Q）轨迹。这些图展示了多次可逆的折叠/去折叠事件，并呈现均方根偏差与 Q 之间预期的负相关。FlashSchNet 成功捕捉了 Chignolin 独特的折叠时间尺度（纳秒级转变），相比之下 TRPcage 和 Villin 的停留时间更长。

表 2：结构精度基准。FlashSchNet 保留了基线模型 CGSchNet 的高保真度，并显著优于经典模型 MARTINI。全原子模拟作为实验参考方法。

蛋白质	指标	机器学习力场		经典模型	参考方法
		FlashSchNet	CGSchNet	MARTINI	全原子
Chignolin	GDT-TS	0.90	0.90	0.66	1.00
	最大 Q	0.89	0.96	0.83	0.95
TRPcage	GDT-TS	0.72	0.72	0.64	0.88
	最大 Q	0.89	0.96	0.60	0.95
Villin	GDT-TS	0.74	0.78	0.46	0.88
	最大 Q	0.88	0.96	0.56	0.93

结构保真度与CGSchNet基线相当。 表2 使用GDT-TS和最大亚稳态 Q对结构准确性进行基准测试。FlashSchNet 在所有蛋白质上，GDT-TS评分与CGSchNet基线的差距保持在0.04以内，同时两种机器学习力场在稳定近天然结构方面均显著优于MARTINI。对于最大亚稳态，也呈现类似趋势，Q 即FlashSchNet持续达到 $Q \geq 0.88$，与CGSchNet相当，且显著高于MARTINI（$Q \approx 0.56$–0.83）。这些结果证实，FlashSchNet 在显著提高模拟速度的同时，保留了原始CGSchNet模型的物理准确性。

表 3：计算效率基准测试。评估的蛋白质包括 Chignolin (CLN)、TRPcage (2JOF)、同源结构域 (1ENH)、Villin (1YRF) 和 Alpha3D (2A3D)。报告的性能指标为速度（timestep·mol/s）和峰值内存（GB）。与经典模型基准相比，FlashSchNet 在单个 RTX PRO 6000 GPU 上展现出具有竞争力的吞吐量。

蛋白质体系	指标	MLFF		经典模型	参考方法
		FlashSchNet	CGSchNet	MARTINI	全原子
Chignolin	速度	**5222**	3578	2580	1437
	峰值内存	**3.7**	22.7	35.0	36.3
TRPcage	速度	**4938**	1729	2550	1419
	峰值内存	**8.8**	29.2	34.9	38.3
同源结构域	速度	**3095**	477	2250	1005
	峰值内存	**18.0**	92.5	34.9	37.4
Villin	速度	**3912**	1056	2340	1275
	峰值内存	**12.9**	94.2	35.0	47.9
Alpha3D	速度	**2610**	288	2160	861
	峰值内存	**22.4**	94.1	31.7	63.6

吞吐量达到与经典力场相当的水平。表3 总结了吞吐量和内存使用情况。在同源结构域（1ENH）系统中，FlashSchNet 的吞吐量约为**3000 步·mol/s**（即1000 ns/天），相较于 CGSchNet 基线（约 **6.5× 倍加速**，约为500 步·mol/s）。这使机器学习力场与经典势函数之间的差距基本消失，因为FlashSchNet 的性能已与 MARTINI 相当（约2900 步·mol/s），同时显著优于全原子模拟（约 1200 步·mol/s）。此外，FlashSchNet 通过消除大型中间结果的物化，将峰值内存从 92GB（CGSchNet）降至**18GB**（$> 80\%$ **的内存减少**）。这有望使得在商用硬件（例如单块 RTX 5090）上模拟大型系统成为可能。

5.2 消融实验结果

对动态图拓扑的鲁棒性。图神经网络分子动力学中的一项关键挑战是，邻居图会在整个模拟过程中不断演变，尤其是在构象转变期间。如图6所示，拉长的 1ENH 蛋白质在300k 步内展开，导致邻接矩阵从近似对角结构转变为稠密的非对角结构，同时边的数量增加。图5 显示，在这些条件下，CGSchNet 的吞吐量显著下降，这可能是因为边分布到更多目标节点时，散射竞争加剧（Gong 等人，2025年）。相比之下，FlashSchNet 通过不受边分布模式影响且无竞争的 CSR 分段归约，保持了稳定的吞吐量。对于涉及大幅构象变化的实际分子动力学工作流而言，这种鲁棒性至关重要。

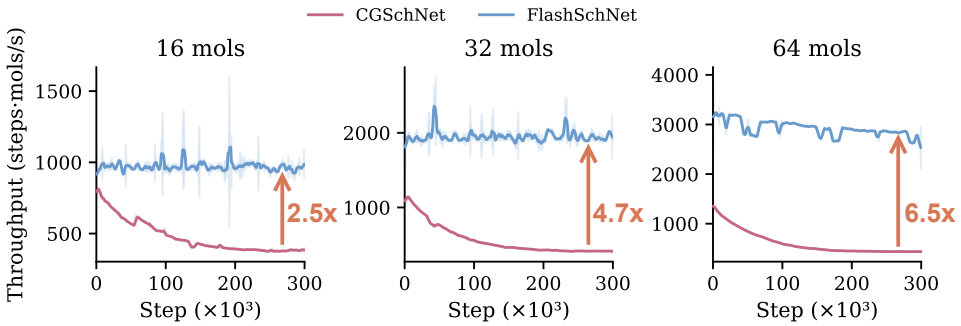

图5：在三种批大小（即 16、32和 64个并行副本）下，对1ENH蛋白进行300k步延展模拟时的逐步吞吐量比较。FlashSchNet 在模拟过程中即使图拓扑不断演变，也能保持稳定的吞吐量；而随着邻居图变得更稠密、对角线特征减弱，CGSchNet的性能会下降，如图 6所示。随着批大小增大，加速差距也随之扩大，在 6.5× 、64个副本时达到最大。

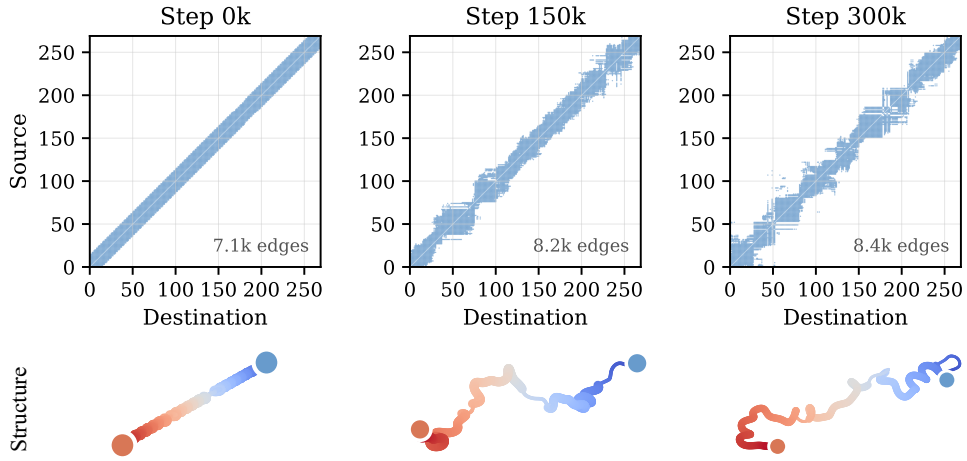

图6：1ENH延展模拟过程中图拓扑与蛋白质结构的演变。上：第0、150k和300k步的邻接矩阵，显示随着蛋白质展开，非对角线区域密度不断增加（例如 边数从7.1k增加到8.4k）。下：对应的三维结构按残基索引着色，展示了从紧凑的折叠态到延展构象的转变。

内存削减带来了更好的可扩展性。如图7所示，我们考察了在四个大小各异的蛋白质体系中，吞吐量如何随并行副本数量扩展：Chignolin（CLN，10 个残基）、Villin（1YRF，35 个残基）、同源结构域（1ENH，54 个残基）和 Alpha3D（2A3D，73 个残基）。在所有体系中，CGSchNet 都会在较小批大小下耗尽图形处理器内存（见插图），从而限制了其在副本交换等增强采样方法中的实用性；这类方法受益于大量并发

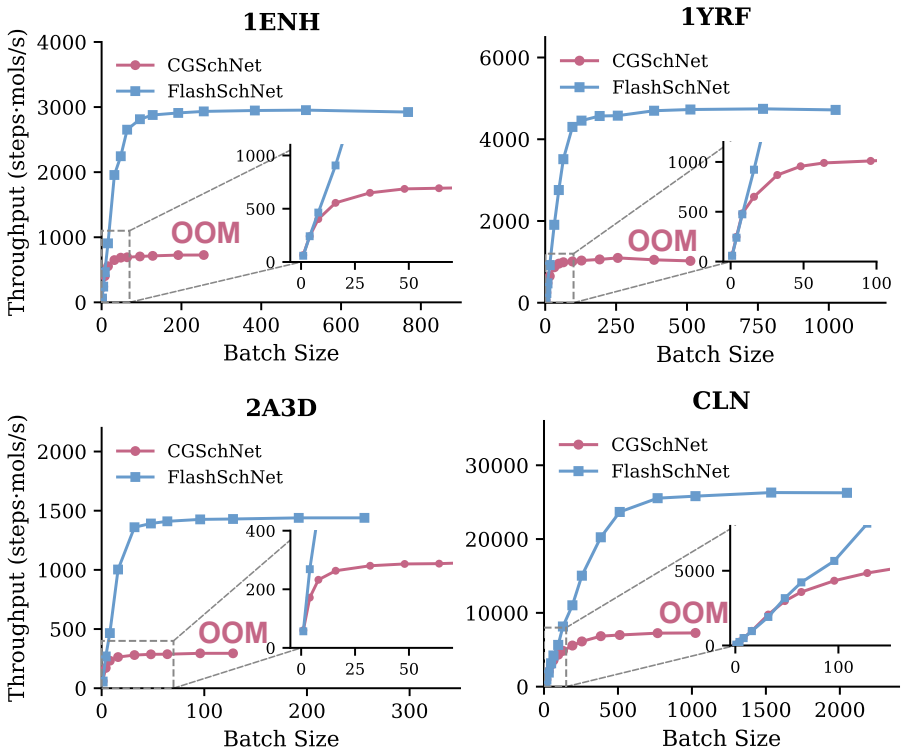

图7：单个 RTX PRO 6000 GPU 上，四种蛋白质体系的吞吐量随批大小的变化。达到饱和之前，FlashSchNet 始终能够扩展到比 CGSchNet 大 3-10× 倍的批大小，而 CGSchNet 在小得多的批大小下就会遇到内存不足（OOM）。我们的 IO感知设计带来的内存削减，对于需要大量并行轨迹的增强采样工作流至关重要。

轨迹。相比之下，FlashSchNet 可将批大小扩展至原来的 3-10× 倍，具体取决于体系大小，即从最大蛋白质 Alpha3D 的 256 个副本，到 Chignolin 的 2048 个副本。随着计算资源得到充分利用，吞吐量在逐渐饱和之前近似线性增长。对于需要通过大量独立轨迹实现统计收敛的工作流程而言，这种可扩展性尤为宝贵。

6 结论

我们提出了 FlashSchNet，这是一种 IO感知的 SchNet风格图神经网络分子动力学框架，旨在解决学习势能函数受内存限制的问题。通过确定高带宽内存流量而非浮点运算次数是主要瓶颈，我们开发了四种利用模型固有结构、从算法层面减少数据移动的技术。*Flash* 径向基将距离计算和基函数展开融合为单次分块处理。*Flash* 消息传递消除了中间边张量物化。*Flash*聚合通过采用压缩稀疏行分段归约重新构造散射加法，实现无竞争累加。逐通道 *16* 位量化利用较低的逐通道动态范围，进一步提升吞吐量。这些技术结合后，相较于 CGSchNet 基线实现了**6.5× 倍加速**和**80% 的内存削减**；在单个 RTX PRO 6000 GPU 上，对包含 269 个珠子的粗粒化蛋白质进行 64 个并行副本模拟时，整体吞吐量达到 1000 ns/day。据我们所知，FlashSchNet 是首个在墙钟效率上快于经典粗粒化力场（例如 MARTINI），同时又保留学习势能准确性和可迁移性的 SchNet 式图神经网络分子动力学方法。

影响声明

FlashSchNet通过输入输出感知的融合内核、无争用聚合和轻量化量化，提高了SchNet风格图神经网络分子动力学的效率并降低了内存占用。这些改进使在固定图形处理器预算下进行更多并发多副本模拟成为可能，从而提升统计效率

以及对稀有事件的覆盖范围。这有助于扩大计算化学、药物发现和材料科学领域的学术及工业用户对高精度学习型分子动力学的使用。通过提高利用率并减少冗余内存移动，这些技术还可能降低每模拟纳秒的能耗，但总体环境影响取决于效率提升是否会带来更多的总模拟量。

参考方法

Airas, J. 和 Zhang, B. 蛋白质语言模型的知识蒸馏可生成基础隐式溶剂模型。*arXiv* 预印本 *arXiv:2601.05388*，2026年。3

Bartók, A. P., Payne, M. C., Kondor, R. 和 Csányi, G. 高斯近似势：无需电子即可达到量子力学的精度。物理评论快报，104(13):136403，2010。3

Batatia, I., Kovacs, D. P., Simm, G., Ortner, C. 和 Csányi, G. MACE：用于快速、准确力场的高阶等变消息传递神经网络。神经信息处理系统进展，35:11423–11436，2022 年。2，3

Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E. 和 Kozinsky, B. 用于数据高效且准确的原子间势的 E (3)-等变图神经网络。自然通讯，13(1):2453，2022 年。2，3

Behler, J. 与 Parrinello, M. 高维势能面的广义神经网络表示。《物理评论快报》,98(14):146401,2007. 3

Best, R. B.、Hummer, G. 与 Eaton, W. A. 天然接触决定全原子模拟中的蛋白质折叠机制。《美国国家科学院院刊》, 110(44):17874–17879,2013.17

Bronstein, M. M.、Bruna, J.、Cohen, T. 与 Veličković, P. 几何深度学习：网格、群、图、测地线与规范。*arXiv* 预印本 *arXiv:2104.13478*, 2021年. 1

Brooks, B. R.、Brooks III, C. L.、Mackerell Jr, A. D.、Nilsson, L.、Petrella, R. J.、Roux, B.、Won, Y.、Archontis, G.、Bartels, C.、Boresch, S. 等。CHARMM：生物分子模拟程序。《计算化学杂志》，30(10):1545–1614，2009。3 Car, R. 与 Parrinello, M. 分子动力学与密度泛函理论的统一方法。《物理评论快报》，55(22):2471，1985。1, 3 Charron, N. E.、Bonneau, K.、Pasos-Trejo, A. S.、Guljas, A.、Chen, Y.、Musil, F.、Venturin, J.、Gusew, D.、Zaporozhets, I.、Krämer, A. 等。利用机器学习得到的可迁移粗粒化模型探索蛋白质景观。《自然·化学》，17(8):1284–1292，2025年。2, 3, 5, 7, 9, 17, 18 Chen, Z.、Yan, M.、Zhu, M.、Deng, L.、Li, G.、Li, S. 与 Xie, Y. FuseGNN：在 GPGPU 上加速图卷积神经网络训练。见《第39届计算机辅助设计国际会议论文集》，第1–9页，2020年。3 Chmiela, S.、Sauceda, H. E.、Poltavsky, I.、Müller, K.-R. 与 Tkatchenko, A. sGDML：利用机器学习构建精确且数据高效的分子力场。《计算机物理通讯》，240:38–45，2019 年。3 Chodera, J. D. 与 Noé, F. 生物分子构象动力学的马尔可夫状态模型。《结构生物学当前观点》，25:135–144，2014。1 Dao, T.、Fu, D.、Ermon, S.、Rudra, A. 与 Ré, C. FlashAttention：具有输入输出感知的快速且内存高效的精确注意力。神经信息处理系统进展，35:16344–16359，2022 年。4

De Vivo, M.、Masetti, M.、Bottegoni, G. 和 Cavalli, A. 分子动力学及相关方法在药物发现中的作用。《药物化学杂志》，59(9):4035–4061，2016 年。1

Dror, R. O.、Dirks, R. M.、Grossman, J.、Xu, H. 和 Shaw, D. E. 生物分子模拟：分子生物学的计算显微镜。《生物物理学年度评论》，41:429–452，2012 年。1Fey, M. 和 Lenssen, J. E. 使用 PyTorch Geometric 进行快速图表示学习。*arXiv* 预印本 *arXiv:1903.02428*，2019 年。3Frantar, E.、Singh, S. P. 和 Alistarh, D. 最优大脑压缩：一种实现精确训练后量化和剪枝的框架，2023。https://arxiv.org/abs/2208.11580。8Gasteiger, J.、Groß, J. 和 Günnemann, S. 面向分子图的方向性消息传递。*arXiv* 预印本 *arXiv:2003.03123*，2020年。2，3Gilmer, J.、Schoenholz, S. S.、Riley, P. F.、Vinyals, O. 和 Dahl, G. E. 面向量子化学的神经消息传递。载于国际机器学习会议，第 1263–1272 页。Pmlr，2017 年。2Gong, Y.、Tarafder, A. K.、Afrin, S. 和 Kumar, P. 识别和分析 {图神经网络} 系统中的陷阱。载于*2025 USENIX* 年度技术会议（*USENIX ATC 25*），第 1605–1624 页，2025年。4，11Hollingsworth, S. A. 和 Dror, R. O. 人人都适用的分子动力学模拟。《神经元》，99(6):1129–1143，2018 年。1

Huang, G., Dai, G., Wang, Y., and Yang, H. Ge-spmm：面向图神经网络的通用图形处理器稀疏矩阵-矩阵乘法。在*SC20*：高性能计算、网络、存储与分析国际会议，第 1–12 页。IEEE，2020年。3

Ju, S., You, J., Kim, G., Park, Y., An, H., and Han, S. 将预训练的通用机器学习原子间势应用于锂离子电池中液态电解质的物理化学模拟。数字发现，2025年。3Karplus, M. and McCammon, J. A. 生物分子的分子动力学模拟。自然结构生物学，9(9):646–652，2002年。1，4

Kühne, T. D., Iannuzzi, M., Del Ben, M., Rybkin, V. V., Seewald, P., Stein, F., Laino, T., Khaliullin, R. Z., Schütt, O., Schiffmann, F., 等。Cp2k：一种电子结构与分子动力学软件包——Quickstep：高效且准确的电子结构计算。化学物理学杂志，152(19)，2020年。3

Liu, J., Cai, Z., Chen, Z. 和 Wang, M. Df-gnn：面向图形处理器上注意力图神经网络的动态融合框架。*arXiv* 预印本 *arXiv:2411.16127*，2024。4

Majewski, M., Pérez, A., Thölke, P., Doerr, S., Charron, N. E., Giorgino, T., Husic, B. E., Clementi, C., Noé, F., and De Fabritiis, G. 蛋白质热力学粗粒化势的机器学习。自然通讯，14(1):5739，2023。3Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P., and De Vries, A. H. Martini 力场：用于生物分子模拟的粗粒化模型。物理化学杂志*B*，111(27): 7812–7824，2007。1，9Musaelian, A., Batzner, S., Johansson, A., Sun, L., Owen, C. J., Kornbluth, M., and Kozinsky, B. 为大规模原子尺度动力学学习局部等变表示。自然通讯，14(1):579，2023。2，3Neumann, M., Gin, J., Rhodes, B., Bennett, S., Li, Z., Choubisa, H., Hussey, A., and Godwin, J. Orb：一种快速、可扩展的神经网络势能。*arXiv* 预印本 *arXiv:2410.22570*，2024。3Pelaez, R. P., Simeon, G., Galvelis, R., Mirarchi, A., Eastman, P., Doerr, S., Tholke, P., Markland, T. E., and De Fabritiis, G. TorchMD-Net 2.0：用于分子模拟的快速神经网络势能。《化学理论与计算杂志》，20(10):4076–4087，2024。4

Rahman, M. K., Sujon, M. H., and Azad, A. Fusedmm：用于图嵌入和图神经网络的统一 sddmm-spmm 内核。载于《2021 年 IEEE 国际并行与分布式处理研讨会》（IPDPS），第 256–266 页。IEEE，2021 年。3 Schütt, K., Kindermans, P.-J., Sauceda Felix, H. E., Chmiela, S., Tkatchenko, A., and Müller, K.-R. SchNet：一种用于模拟量子相互作用的连续滤波卷积神经网络。神经信息处理系统进展，30，2017 年。2，3，5 Unke, O. T. and Meuwly, M. PhysNet：一种用于预测能量、力、偶极矩和部分电荷的神经网络。《化学理论与计算杂志》，15(6):3678–3693，2019 年。3 Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. 通用 AMBER 力场的开发与测试。《计算化学杂志》，25(9):1157–1174，2004 年。3 Wang, M., Zheng, D., Ye, Z., Gan, Q., Li, M., Song, X., Zhou, J., Ma, C., Yu, L., Gai, Y., 等。深度图学习库：一个以图为核心、高性能的图神经网络软件包。arXiv 预印本 arXiv:1909.01315，2019 年。3 Wang, Y., Feng, B., Li, G., Li, S., Deng, L., Xie, Y., and Ding, Y. {GNNAdvisor}：一种用于在 {图形处理器}上加速 {图神经网络} 的自适应高效运行时系统。载于《第 15 届 USENIX 操作系统设计与实现研讨会》（OSDI 21），第 515–531 页，2021年。3 Xie, Z., Wang, M., Ye, Z., Zhang, Z., and Fan, R. Graphiler：利用消息传递数据流图优化图神经网络。《机器学习与系统会议论文集》，4:515–528，2022 年。3 Yang, H., Hu, C., Zhou, Y., Liu, X., Shi, Y., Li, J., Li, G., Chen, Z., Chen, S., Zeni, C., 等。MatterSim：一种涵盖不同元素、温度和压力的深度学习原子模型。arXiv 预印本 arXiv:2405.04967，2024 年。3 Zemla, A. LGA：一种用于查找蛋白质结构三维相似性的方法。《核酸研究》，31(13): 3370–3374，2003 年。17 Zhang, Y. and Skolnick, J. 用于自动评估蛋白质结构模板质量的评分函数。《蛋白质：结构、功能与生物信息学》，57(4):702–710，2004 年。18

附录目录

A 评估指标与方案的详细说明 17

A.1 均方根偏差（RMSD）.17A.2 天然接触分数（Q）.17A.3 GDT-TS评分.17

A 评估指标和协议的详细信息

A.1 均方根偏差（RMSD）

Cα 均方根偏差（RMSD）用于量化采样构象与参考结构之间的几何偏差。给定 N Cα 查询结构中的原子位置 $\{\mathbf{r}_i\}_{i=1}^{N}$ 以及参考结构中对应的位置 $\{\mathbf{r}_i^{\mathrm{ref}}\}_{i=1}^{N}$，RMSD 定义为经刚体变换后可达到的最小欧几里得距离：

$$\mathrm{RMSD} = \min_{\mathbf{R}\in\mathrm{SO}(3),\mathbf{t}\in\mathbb{R}^3} \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{R}\mathbf{r}_i + \mathbf{t} - \mathbf{r}_i^{\mathrm{ref}}\|^2} \tag{A.1}$$

其中，$\mathbf{R}$ 表示旋转矩阵，$\mathbf{t}$ 表示使两个结构最佳对齐的平移向量。通常通过奇异值分解（SVD）计算这种叠合。RMSD 值越低，表示与天然态的结构保真度越高。

A.2 天然接触分数（Q）

天然接触分数（Q）可作为反应坐标，根据残基间距离量化结构相似性，反映拓扑保真度而非整体叠合（Best 等人，2013）。其定义如下：

$$Q = \frac{1}{N_c}\sum_{(i,j)\in\mathcal{C}} \frac{1}{1+\exp\left[\beta(r_{ij} - \lambda r_{ij}^0)\right]} \tag{A.2}$$

其中，$\mathcal{C}$ 表示天然接触对的集合，$N_c = |\mathcal{C}|$ 表示天然接触的总数，r_{ij} 表示查询结构中残基 i 和 j 之间的 Cα 距离，r_{ij}^0 表示参考天然结构中对应的距离。

天然接触定义。对于参考全原子结构中满足以下两个条件的任意残基对（i，j），我们将其定义为天然接触：(1) 序列间隔为 $|i-j|\geq 3$；(2) 重原子间距离小于 4.5 Å。这些已识别残基对的 Cα 距离构成参考集合 $\{r_{ij}^0\}$。

参数。按照（Charron 等人，2025年），我们设定 $\beta = 10$ 纳米$^{-1}$ 和 $\lambda = 1.5$。参数 β用于调节 sigmoid 函数的陡峭程度，而 λ 则考虑了天然距离附近的热涨落。这些超参数生成平滑的自由能面，能够清晰地区分类天然态（$Q\approx 1$）与未折叠构象（$Q < 0.5$）。

最大的亚稳态 Q。在模拟过程中，蛋白质会对 Q 进行概率分布采样。为评估力场稳定天然态的能力，我们计算Q的一维概率密度，应用 Savitzky-Golay 平滑，并识别最右侧的局部最大值（即具有最高 Q 值的稳定盆地）。该指标反映了模型所形成的折叠态的结构保真度。

A.3 GDT-TS评分

全局距离测试总分（GDT-TS）（Zemla，2003 年）通过识别能够在指定距离阈值内叠合的最大 Cα 原子子集来量化结构相似性。对于特定的截断值 d，设 P_d 表示经过最优叠合后，查询结构中与参考结构中对应位置的距离在 α Å 以内的 Cd 原子所占百分比。GDT-TS

计算公式如下：

$$\text{GDT-TS} = \frac{P_1 + P_2 + P_4 + P_8}{4} \tag{A.3}$$

其中 P_1、P_2、P_4 和 P_8 分别对应 1、2、4 和 8 Å 的截断值。与均方根偏差不同，GDT-TS 对局部高变异区域（例如环区）的敏感度较低，因而能稳健地衡量全局拓扑。所有 GDT-TS 计算均使用 TM-score 程序（Zhang 和 Skolnick, 2004 年）。

评估方案。 遵循（Charron 等人，2025年）中的方案，我们在均方根偏差与 Q 空间中构建二维自由能面，并采用 k-均值聚类，设定 100 个中心。我们找出最类天然的簇（定义为 Q 最高且均方根偏差最低），并从该能量盆地中随机抽取 10 个代表性结构。报告的 GDT-TS 为这些样本与实验参考结构比对所得分数的平均值。