arXiv:2603.06567v1 [cs.LG] 6 Mar 2026

可扩展的基于注意力的机器学习原子间势配方：借助全对全节点注意力实现长程精度

Eric Qu[2,*], **Brandon M. Wood**[1], **Aditi S. Krishnapriyan**[2,3,†], **Zachary W. Ulissi**[1,†]

[1]Meta 的 FAIR， [2]加州大学伯克利分校， [3]劳伦斯伯克利国家实验室

*工作完成于 Meta，†通讯作者

机器学习原子间势（MLIPs）发展迅速，许多顶尖模型依赖于强基于物理的归纳偏置。然而，当模型扩展到生物分子和电解质等更大系统时，它们难以准确捕捉长程（LR）相互作用，导致当前方法依赖显式的基于物理的项或组件。在本工作中，我们提出 AllScAIP，一种直接、基于注意力且能量守恒的机器学习原子间势模型，可扩展至 O(1 亿) 训练样本。它使用数据驱动的全对全节点注意力组件来应对长程挑战。大量消融实验表明，在低数据/小模型情形下，归纳偏置可提高样本效率。然而，随着数据和模型规模扩大，这些收益减弱甚至逆转，而全对全注意力对于捕捉长程相互作用仍然至关重要。我们的模型在分子系统上达到最先进水平的能量/力精度，并在多项基于物理的评估（OMol25）中表现优异，同时在材料（OMat24）和催化剂（OC20）上具有竞争力。此外，它能够进行稳定的长时间尺度分子动力学模拟，准确恢复实验可观测量，包括密度和汽化热预测。

日期： 2026年3月9日
通讯方式： A.S.K. aditik1@berkeley.edu,Z.W.U. zulissi@meta.com**代码：** https://github.com/facebookresearch/fairchem

Meta

1 引言

机器学习原子间势（MLIP）从头算数据中学习替代势能面，以近乎达到密度泛函理论精度且成本实用的方式开展分子动力学模拟（Unke et al., 2021）。该领域已从基于描述符的神经网络势（Behler 和 Parrinello, 2007）发展到不变消息传递网络（Schütt et al., 2017; Gasteiger et al., 2021, 2022）、具有高阶方向特征的 SE(3) 等变架构（Batzner et al., 2022; Passaro and Zitnick, 2023; Fu 等，2025），以及近期出现的基于注意力的模型（Orbital-Materials, 2024; Qu and Krishnapriyan, 2024; Mazitov et al., 2025）。在这些方法中，设计的主要差异在于其归纳偏置：对称性（平移、置换和旋转不变性或
等变性）、局域性（截断和邻居模板）、光滑性/正则化、架构约束，以及显式的物理结构（能量守恒梯度）。这些选择体现了两种视角：一种强调采用更丰富的归纳偏置，预先编码几何和物理学知识，从而在小型数据集上提高样本效率；另一种则将某些归纳偏置视为可随数据量和参数规模增长而逐步减少的脚手架，让模型自行发现特征，同时优先考虑训练效率和稳定性。

随着机器学习原子间势的能力不断增强，评估对象已转向规模更大、复杂度更高的系统。在生物分子和电解质系统等场景中，长程相互作用对于描述细微而重要的相互作用至关重要。静电作用、诱导／极化、色散以及电子结构变化等长程效应，会在较长的长度尺度上耦合原子。然而，适用于超大规模数据集且最具可扩展性的机器学习原子间势，采用的是消息

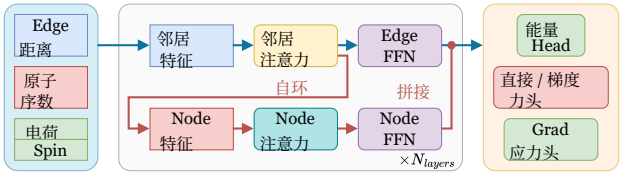


图1 AllScAIP 模型设计。简洁的骨干设计可实现高效的尺度扩展。

传递网络，并基于带有距离截断的局部半径图构造。解决这些局限的一种常见方法，是加入明确的基于物理的归纳偏置（Unke等人，2021年）：预测逐原子电荷（或多极矩），并通过 Ewald／PME 或 FMM 计算库仑项（Loche等人，2025年；Cheng，2025年；Kim等人，2025年）；加入极化或电荷平衡求解器（Ko等人，2021）；纳入解析或学习得到的色散（例如多体范德华相互作用）（Sauer等人，2025）；或引入连续介质／弹性修正（Gong等人，2025）。这些策略已在多个有针对性的小规模数据集上得到广泛验证。然而，随着领域转向开发能够在涵盖众多不同系统的大规模异质数据集上实现高精度的模型（这在一定程度上得益于大规模数据集的发布（Levine等人，2025年）），如何开发能够有效扩展至这一场景的方法，仍是一项尚待解决的挑战。

我们假设，若干归纳偏置可以在缩放过程中学习——特别是旋转对称性（Qu和Krishnapriyan，2024年；Mazitov等人，2025年）、通过半径图结构实现的局域性（Kreiman等人（2025年），以及长程相互作用；而另一些归纳偏置则更难学习，可能需要通过架构或损失函数进行编码（例如，通过基于梯度的力实现能量守恒）。基于这一观点，我们开发了一种概念简单且可扩展的注意力机器学习原子间势，它包含两个操作阶段（如图1所示）：邻域自注意力（基于EScAIP模型（Qu和Krishnapriyan，2024年）），在固定局部模板上运行以解析局部信息；随后是全连接节点自注意力，用于全局混合信息，使信号能够遍历整个图。这两个阶段都使用现成的多头自注意力 CUDA 内核；这些内核经过精心设计和优化，适用于计算机视觉和语言领域中广泛使用的 AI/ML 应用（Lefaudeux等人，2022年）。为检验这一假设，我们在方案中加入另外两个要素，以提供架构本身无法直接捕获的特征：勒让德角度编码（*LAE*），一种紧凑且具备旋转感知能力的边编码，可为邻域注意力提供高阶方向信息；以及欧几里得旋转位置编码（*ERoPE*）（基于Frank等人（2024年）），一种各向同性纯距离编码，可将距离信息注入节点注意力。这些要素作为归纳偏置，可以明确地构建到模型架构中。我们在不同模型规模和数据规模下对这些组件进行消融实验，以检验随着模型容量和数据规模增大，这类几何归纳偏置是否确实能够被学习。

在不同数据集和规模下，消融实验支持“归纳偏置可在规模化训练中学习”的假设。在低数据量、小模型的条件下，每个组成部分都发挥作用：加入 LAE 可提供方向/角度信号，从而降低力误差；欧几里得 RoPE 利用距离信息改善能量预测；全局节点注意力则通过无需深层堆叠即可实现多跳通信，带来最大幅度的提升。随着模型容量和数据集规模同时增大，情况发生了变化：几何编码的边际收益减小，有时甚至转为负值，这表明角向和径向特征可以通过大规模端到端训练来学习。相比之下，全连接节点注意力阶段始终是提升长程能力最持久的来源。

总之，我们的结果支持一种先验较少的可扩展机器学习原子间势方案。只要数据和参数充足，包括旋转等变性、高阶方向特征乃至长程相互作用在内的多种归纳偏置，大多都可以学习得到。由此得到的模型 AllScAIP（**全**连接 **可**扩展 **注**意力 **原**子间**势**），在大型分子数据集 Open Molecules 2025（迄今最大、最多样化的分子数据集）上，在能量/力精度和相关的基于物理的评估方面均达到最先进水平（Levine等人，2025年）。在 OMol25 上训练的 AllScAIP 能够进行稳定的长时间尺度分子动力学模拟，并准确再现实验可观测量，包括密度和汽化热预测。AllScAIP 在材料（OMat24）（Barroso-Luque等人，2024）和催化剂系统（OC20）（Tran et al.，2023年）上的表现也具有竞争力。综上，这些结果表明，数据驱动的机器学习原子间势发展路径或可与其他方法相竞争：优先采用可扩展组件，尽量减少先验，让规模化训练处理其余问题。

2 相关工作

机器学习原子间势。神经网络原子间势（NNIPs）取得了显著进展。这类机器学习模型经过训练，可根据原子序数和位置等系统描述符预测能量和每个原子的力（Yuan et al., 2025）。我们将当前方法分为两类：（1）使用等变节点特征（具有更多归纳偏置）的模型，包括 NeuqIP（Batzner et al., 2022）、MACE（Batatia et al., 2022）、SCN（Zitnick et al., 2022）、eSCN（Passaro and Zitnick, 2023）、Equiformer（Liao and Smidt, 2022；Liao et al., 2024）、eSEN（Fu 等，2025）、UMA（Wood 等人，2025）；（2）使用标量节点特征的模型，这类模型仅保证基本对称性，例如旋转和平移不变性：SchNet（Schütt et al., 2017）、DimNet（Gasteiger et al., 2020）、GemNet（Gasteiger et al.,

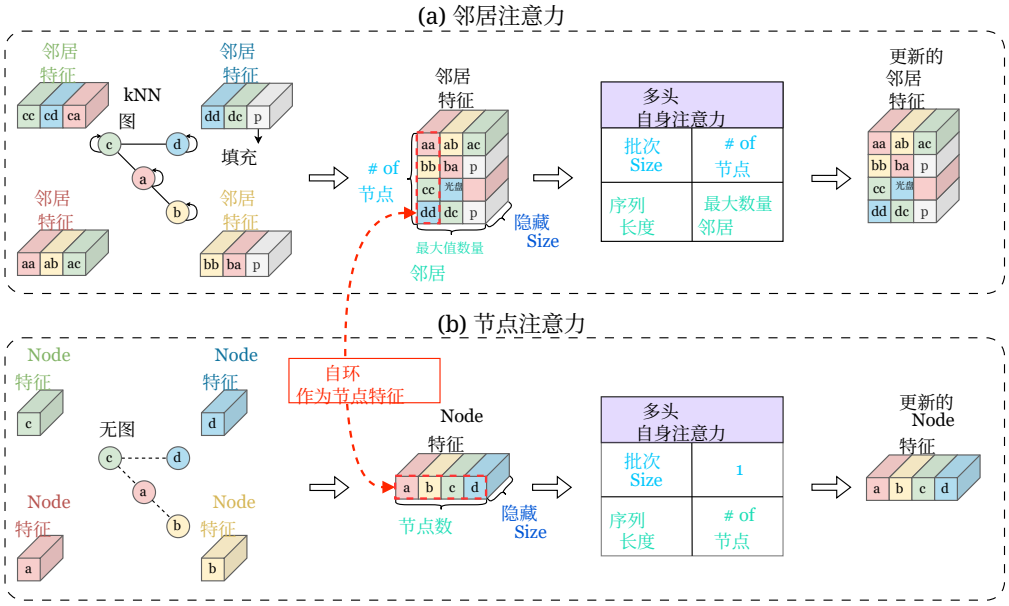


图 2 AllScAIP 中所用注意力操作的示意图。 (a) 邻居注意力。(b) 节点注意力。

2021，2022），EScAIP (QuandKrishnapriyan,2024年), MindScAIP(Liu等人,2026年), OrbNet(Orbital-Materials，2024），PET-MAD (Mazitov等人,2025年)。

机器学习原子间势中的长程相互作用。 为使局部的基于图神经网络的机器学习势能够建模长程相互作用，主要范式是注入显式的长程物理：模型预测电荷、多极矩或代理电荷密度，并用 Ewald/PME/FMM 计算静电作用，包括 PhysNet（Unke 和 Meuwly，2019年）、4G-HDNNP（Ko 等人，2021年）、AIMNet（Zubatyuk 等人，2019年）、BAMBOO（Gong 等人，2024年）、LODE（Grisafi 和 Ceriotti，2019年）以及 DPLR（Zhang 等人，2022年）。近期的“潜变量”方法通过学习一个隐藏的逐原子变量，并直接施加类 Ewald 的长程能量（LES），从而避免对标签的需求（Cheng，2025年）。色散同样已通过机器学习多体范德华相互作用加以纳入（Tu 等人，2023年）。

3 方法

3.1 注意力操作

我们将注意力视为一种现成操作，并利用经过高度优化的 CUDA 内核（Lefaudeux 等人，2022）。两个阶段之间唯一的区别在于我们如何打包标记：局部（邻居）注意力针对每个节点的固定邻居列表进行运算，其复杂度为 $\mathcal{O}(Nk)$；而全对全节点注意力则混合每个图中的所有节点，计算成本为 $\mathcal{O}(N^2)$。

邻域自注意力。遵循 EScAIP 模型（Qu 和 Krishnapriyan，2024），每个中心原子会聚合至多 k 个邻居及一个自身标记，从而得到形状为$(\#\text{节点}) \times (k{+}1) \times d_{\text{model}}$的张量，并将其输入标准的多头自注意力（MHSA）（图2 (a)）。我们在同一模板上执行两次方向性传递：中心→邻居（向外）以及邻居→中心（向内）。平滑的基于距离的包络函数会进行填充/掩码处理，使远距离原子对的权重逐渐衰减。

全连接节点自注意力。 局部邻域能够解析精细几何结构，但长程相互作用需要全局通信。因此，我们通过将节点打包为 $1\times(\#\text{节点})\times d_{\text{model}}$，在节点流上应用多头自注意力（见图 2（b）），并使用相同的算子。在实践中，这一阶段补充了局部处理流程：邻域注意力负责精细的各向异性相互作用；节点注意力则能在单个步骤中实现多跳耦合。

3.2 几何编码

在邻域注意力和节点注意力中，注意力矩阵的每个条目都对应一个几何对：在邻域情形下是边—边相互作用，在全局情形下是节点—节点相互作用。我们利用了这一点

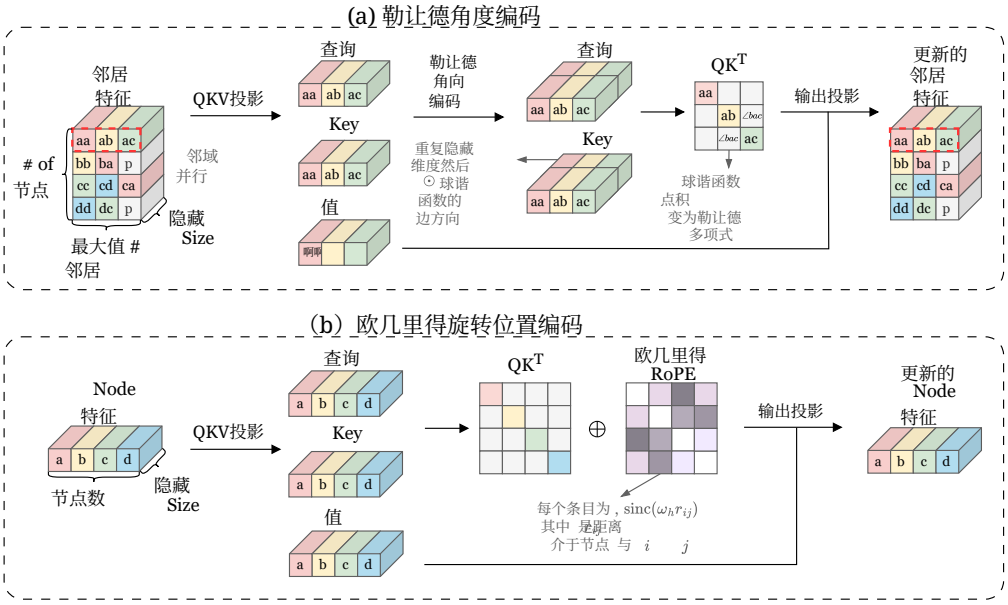


图 3 AllScAIP 中使用的几何编码示意图。（a）勒让德角度编码（LAE）；（b）欧几里得旋转位置编码。

将几何编码注入注意力 logits 中。勒让德角度编码提供方向（角度）信号，欧几里得 RoPE 提供径向（距离）信号；这些信号并未明确存在于原始 Q/K/V 流中。原则上，足够大的模型可以从坐标中学习这些信号。至于应将其作为输入（归纳偏置），还是让模型直接从数据/尺度扩展中学习，这是一个重要的选择。

勒让德角度编码（LAE）。该编码用于向邻域注意力 logits 注入角向和高阶方向特征。每条有向边 $i \to j$ 都会被赋予一个紧凑的角度编码，该编码由在单位方向 $\hat{\mathbf{r}}_{ij}$ 上计算的实球谐函数构成。对于阶数 $\ell = 0, \ldots, L$，我们计算 $\mathbf{Y}_\ell(\hat{\mathbf{r}}_{ij}) \in \mathbb{R}^{2\ell+1}$，将每个阶块重复 r_ℓ 次，并沿 ℓ 拼接，从而得到维度为 $d_{h_r} = \sum_\ell r_\ell$ 的 $\gamma_{ij} \in \mathbb{R}^{d_{h_r}}$（$2\ell+1$）。我们选择 $\{r_\ell\}$，使其总和与每头维度相匹配：$\sum_{\ell=0}^{L} r_\ell = d_h$。对于头 h，查询/键相应地重复 $2\ell+1$ 次，并与 γ_{ij} 逐元素调制，

$$\tilde{q}_{i,h} = \bar{q}_{i,h} \odot \gamma_{ij}, \quad \tilde{k}_{j,h} = \bar{k}_{j,h} \odot \gamma_{ij},$$

并应用通常的缩放点积，$a_{ij,\,h} = (\tilde{q}_{i,\,h}^{\top} \tilde{k}_{j,\,h}) / \sqrt{d_h}$，随后应用标准注意力。根据球谐函数加法定理，每个阶数内的内积对应于（相差一个常数因子）$P_\ell(\cos\theta)$，因此 LAE 以线性成本提供多阶角度结构。LAE 具备旋转感知能力、参数开销小，并且能够与现成的注意力内核集成（图3 （a））。

欧几里得旋转位置编码（ERoPE）。为了将距离信息注入节点注意力 logits，我们添加各向同性径向偏置。遵循 Frank 等人（2024）的方法，对于每一对 $(i,\ j)$，其间距为 r_{ij}，我们通过对单位球面上所有方向的平面波相位取平均，构建纯径向编码，从而得到各向同性核 $K_\omega(r) = \text{sinc}(\omega r) = \sin(\omega r) / (\omega r)$。我们选取一组 M 对数间隔频率 $\{\omega_k\}_{k=1}^{M} \subset [\omega_{\min}, \omega_{\max}]$（在各注意力头之间共享），计算 $\mathbf{s}_{ij} = [\text{sinc}(\omega_1 r_{ij}) \ldots\ldots \text{sinc}(\omega_M r_{ij})]^{\top}$。每个注意力头 h 使用学习到的权重向量 $\mathbf{w}_h \in \mathbb{R}^M$ 对它们进行混合，以生成加性对数几率偏置 $b_{ij,\,h} = \mathbf{w}_h^{\top} \mathbf{s}_{ij}$，最终的 logit 为

$$a_{ij,h} = \frac{q_{i,h}^{\top} k_{j,h}}{\sqrt{d_h}} + b_{ij,h}.$$

由于 $b_{ij,h}$ 仅取决于 r_{ij}，因此该偏置在构造上具有平移不变性和旋转不变性。小型频率组仅增加 $\mathcal{O}(M)$ 每对的计算量，可按批次预计算，并提供平滑且频谱丰富的径向先验，同时不改变 softmax 机制（图3 (b)）。

表 1 AllScAIP 模型中的归纳偏置

归纳偏置	状态	满足方式／位置
平移不变性	**已强制实施**	仅使用相对几何信息（距离/方向）。
置换等变性	**已强制满足**	对具有共享权重的集合应用注意力。
广延性（可加性）	**已强制满足**	能量和周期性边界条件距离的求和聚合。
局部性先验	**强制实施**	邻域注意力中的平滑截断距离 k近邻图。
能量守恒	**强制实施**	基于梯度的力： $F = -\mathrm{d}E/\mathrm{d}x$；可微k近邻图。
旋转等变性	可学习	非硬编码；可在缩放下学习。
长程效应	可学习	全对全节点注意力提供无限的感受野。
径向/距离先验	可选（软）	**欧几里得 RoPE**：节点注意力中的基于距离的编码。
高阶方向特征	可选（软）	**LAE**：紧凑的球谐编码在邻域注意力中调制 Q/K（角度/方向）。

3.3 AllScAIP 模型

图 1展示了 AllScAIP 的完整架构。输入由原子序数、边距离径向基函数以及可选的电荷/自旋初始化。每个模块首先对邻居流应用邻域注意力（可选地加入 LAE 编码）。随后是 RMSNorm、残差路径和边前馈神经网络。接着，该模块将自环邻居特征作为节点特征，并对节点流执行全对全节点注意力（可选地使用欧几里得 RoPE），然后依次进行 RMSNorm、残差路径和节点前馈神经网络。更新后的节点特征与邻居特征拼接后传递至下一层。经过 N_{layers}个模块后，节点流输入能量头；力通过直接监督或能量梯度（能量守恒选项）进行监督，并在需要时加入基于梯度的应力头。

3.4 归纳偏置

我们区分 **通过架构编码的** 硬先验、**可选**作为特征提供的软先验，以及**可学习的**、交由规模效应习得的物理结构。我们的目标是构建一个先验轻量的模型，只保留那些难以高精度学习的先验（局域性、守恒性、置换/平移对称性），同时让大型模型和数据集学习其余内容。表 1列出了我们考量的归纳偏置。简言之，我们强制施加核心对称性（平移、置换）、局域性、能量守恒和广延性；可选地注入少量几何特征；并让模型随着规模扩大而学习旋转和长程相互作用。能量守恒的表述遵循 Liu等人 (2026年) 提出的可微 kNN 算法。我们在 § 5.2 中验证这些性质。

4 消融实验

模型组件消融实验。 我们首先在固定模型容量和数据规模的条件下，分别隔离各架构组件的贡献。我们使用中型配置（AllScAIP-md，约 8500 万个参数），在 OMol25 4M划分（Levine等人，2025年）上训练 80 个轮次，并采用直接力监督。我们分别启用或关闭邻域自注意力（NeiAtt；始终启用）、勒让德角度编码（LAE）、全对全节点注意力（NodeAtt）和欧氏径向偏置（ERoPE）。我们在验证集上报告总计及四个数据划分——生物分子、电解质、金属配合物和中性有机物——的每原子能量平均绝对误差（mEV）和力平均绝对误差（meV/Å）（表 2，顶部的 4M 部分）。

移除 LAE 会导致所有数据划分上的力性能持续下降，而对能量的影响较小。这与 LAE 作为方向/角度特征的作用相符：加入方向信息有助于模型更好地拟合向量目标。相比之下，关闭 ERoPE 主要会损害能量预测，对力的影响则较轻微，这与 ERoPE 仅依赖距离的各向同性偏置作为标量预测的径向先验相一致。最后，消融全对全节点注意力会降低能量和力的预测表现，在生物分子中尤为明显，因为这类系统规模更大，长程效应也更强。完整配置在所有数据划分上都取得了最佳结果，证实该模块能提高低数据量场景下的样本效率。

表 2 数据和模型规模扩展下的组件消融。我们报告每原子能量平均绝对误差（毫电子伏）和力平均绝对误差（毫电子伏/埃），数值越低越好。每种大小下，采用 **编码**（邻域注意力 +LAE +节点注意力 +ERoPE）的构型为基准；其他行根据相对于该基准的表现着色（绿色 = 表示更好，红色 = 表示更差）。吞吐量列报告推理速度（在一台 H200 上对 1000 个原子进行计算，单位为纳秒/天）。我们报告 OMol25 4M 划分（80 轮次）和完整 102M 数据集（10 轮次）的结果[1]

						OMol 4M（80 轮次）									
		消融实验				生物分子		电解质		金属配合物		中性有机物		总计	
Size	吞吐量	NeiAtt	LAE	NodeAtt	ERoPE	E ↓	F ↓	E ↓	F ↓	E ↓	F ↓	E ↓	F ↓	E ↓	F ↓
35M	2.279	✓	✓	✓	✓	0.38	5.57	1.27	8.93	2.85	33.52	1.89	13.52	1.46	9.35
	7.623	✓		✓		0.5	6.32	1.51	9.81	2.96	34.56	2.23	14.51	1.75	10.21
85M	1.124	✓	✓	✓	✓	0.23	3.51	1.05	6.38	2.46	27.83	1.21	8.93	1.16	6.88
	3.333	✓		✓	✓	0.29	4.27	1.17	7.22	2.67	29.79	1.53	10.53	1.22	7.6
	1.392	✓	✓	✓		0.29	3.78	1.16	6.67	2.58	28.26	1.36	9.03	1.3	7.1
	4.014	✓		✓		0.32	4.24	1.17	7.16	2.64	29.19	1.57	10.12	1.25	7.55
	1.281	✓	✓			0.52	4.58	1.39	7.42	2.81	29.86	1.43	10.01	1.56	7.93
	4.327	✓				0.55	4.96	1.45	7.92	2.85	30.78	1.57	10.77	1.72	8.48
Size	**吞吐量**	NeiAtt	LAE	NodeAtt	ERoPE	**OMol 102M（10 轮次）**									
35M	2.279	✓	✓	✓	✓	0.21	3.85	0.75	6.01	2.09	26.02	1.01	8.54	0.85	6.61
	7.623	✓		✓		0.29	4.47	0.81	6.73	2.12	26.86	1.08	9.17	0.98	7.22
85M	1.124	✓	✓	✓	✓	0.15	2.81	0.53	4.59	1.83	21.93	0.73	6.03	0.67	5.1
	3.333	✓		✓	✓	0.2	3.16	0.55	4.99	1.83	22.48	0.79	6.47	0.73	5.51
	1.392	✓	✓	✓		0.2	2.93	0.58	4.76	1.8	21.59	0.76	5.94	0.72	5.13
	4.014	✓		✓		0.15	2.91	0.52	4.7	1.83	22.31	0.72	6.29	0.64	5.23
	4.327	✓				0.4	3.83	0.79	5.57	2.01	23.91	0.85	7.05	0.96	6.15

数据与模型规模扩展消融实验。 为了量化归纳偏置与尺度之间的相互作用，我们考察两个维度：在模型容量固定时改变数据规模（OMol 4M →OMol 102M），以及在数据规模固定时改变模型大小（35M → 85M）。

在容量固定的情况下，**数据缩放**（4M → 102M）会使各个领域的能量误差和力误差均有所下降。在许多情况下，在相同数据规模下，移除固定几何编码后，模型表现与完整模型持平或略有提升，这表明随着数据量增加，角向/径向信号可以直接学习。全对全节点注意力始终发挥作用，说明全局混合阶段是实现长程相互作用的一种可靠机制。

在数据规模固定时进行**模型规模扩展**（35M → 85M），绝对误差也会随着模型容量增大而下降。在 4M 数据规模下，几何编码仍有帮助，但其优势会随着模型规模增大而减弱；在 102M 数据规模下，中型模型采用固定编码时表现持平或略逊一筹，但全对全注意力仍带来可测量的优势。简言之，随着数据和参数规模扩大，固定几何编码的边际价值会降低，而全局混合的架构优势仍能持续发挥作用。

5 结果

5.1 推理效率与系统规模扩展

我们测试 AllScAIP 模型的效率和内存随规模扩展的表现。图 4 展示了在单张 H200 141G GPU 上（关闭图生成）原子数·纳秒/天随系统规模的变化，并呈现出斜率变化。在较小 N 规模区间，速度主要受邻域注意力影响（$\mathcal{O}(Nk)$），原子数·纳秒/天几乎保持平稳，或随 N线性变化；一旦全对全节点注意力占据主导（$\mathcal{O}(N^2)$），其数值便大致按 $1/N$ 下降（双对数图中的斜率为 -1）。虚线标出了这一实测转折点。总体而言，这些曲线显示出可预测的尺度扩展规律，并且存在明显的区间切换；与基线方法相比，我们的模型仍然高效。我们的方法面向约 10^3-10^5 个原子的规模，这也是许多有趣科学问题所在的领域（生物分子、电解质、软物质、介观晶粒），同时 GPU 仍能保持高效。此外，我们还在图 9中提供了原始吞吐量随系统规模的变化，并在图 8中提供了各组件（邻域注意力/前馈网络、节点注意力/前馈网络）的细分数据和实际运行时间测量。

[1]我们注意到，FAIRChem 代码库中存在一个已确认的 错误，导致 AllScAIP 和所有基线方法的每原子平均绝对误差指标看起来始终较差。我们已在附录中附上未归一化的总能量平均绝对误差 A.4。

表3 对称性与守恒性检查。周期性边界条件（PBC）和非周期性边界条件（非PBC）系统的广延性误差（越低越好）、通过随机旋转下的力余弦相似度评估旋转等变性（越高越好），以及 MD22 大分子在微正则系综分子动力学模拟中的能量漂移（越低越好）。

模型	训练集	PBC 能量 Δ [毫电子伏] ↓	非周期性边界条件 能量 Δ [毫电子伏] ↓	随机旋转微正则系综分子动力学能量漂移 余弦相似度 ↑	[毫电子伏 / 原子 / ps] ↓
UMA-M-1p1	UMA-459M	2.41×10^{-2}	3.83×10^{-3}	0.9999	1.5×10^{-3}
AllScAIP-md	NA（随机初始化）	3.62×10^{-2}	2.49×10^{-3}	0.6827	-
AllScAIP-md	OMol-102M	4.18×10^{-2}	4.73×10^{-3}	0.9999	4.3×10^{-3}

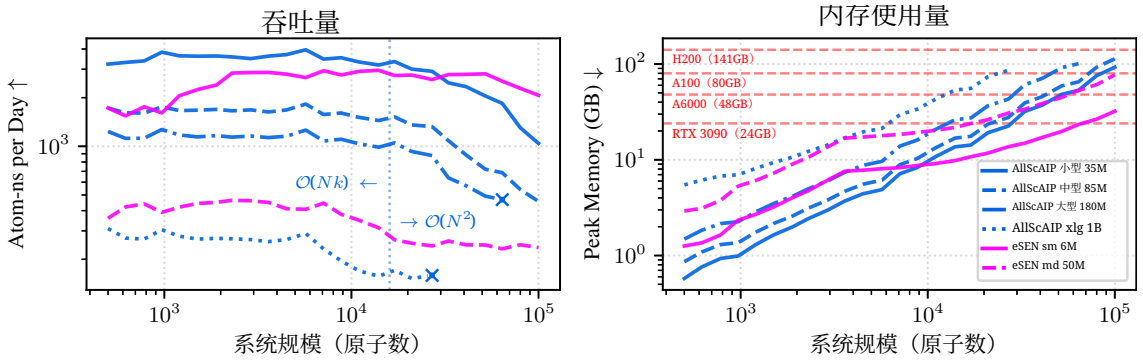


图 4 吞吐量和内存与系统规模的关系。左：原子时间（原子数 × 纳秒/天，越高越好）与系统规模的关系。结果是在单张 H200 141G 上测得，且关闭了图生成。我们报告了 AllScAIP 的四种模型规模（35M/85M/180M/1B）以及 eSEN 基线方法。竖直虚线大致表示节点注意力的$\mathcal{O}(N^2)$超过邻域注意力的 $\mathcal{O}(Nk)$的临界点，其中 k 为邻居数量的最大值。右：显存使用量与系统规模的关系。水平虚线表示常见 GPU 的显存容量。

5.2 对称性与能量守恒检查

我们验证了 AllScAIP 满足关键的归纳偏置性质，并将其与一个在多种化学体系上训练的强基线模型 UMA 进行比较（Wood 等人，2025年）（表 3）。

广延性。我们在周期性和非周期性设置下均测试了广延性：

(i) **周期性边界条件下超胞倍增：**给定一个周期性结构，我们通过沿一个晶格矢量平移晶胞并重新计算能量来构建一个 $2\times$ 超胞；广延性要求 $E_{2\times}\approx 2E_{1\times}$。我们报告了 $E_\Delta = |E_{2\times} - 2E_{1\times}|$。

(ii) **真空复制：**对于非周期性系统，我们复制该构型，并将副本平移一个较大的偏移量 R，然后评估 $E(R)$；由于 $R\to\infty$ 这些片段之间不存在相互作用，因此 $E(R)\to 2E_{\text{single}}$。在实践中，我们使用 $R = 1000$ 埃，并报告偏差 $E_\Delta = |E(R) - 2E_{\text{single}}|$。

由于该架构在注意力中仅使用相对几何特征，且不使用绝对位置编码，因此在这两项测试中都表现出近乎完美的可加性，证实学得的长程耦合不会引入非物理行为。

旋转等变性。我们从 OMol 测试结构中抽取 1000 个，随机生成一个旋转 $R\in$SO(3)，并对力进行两次评估：在原始坐标上计算$\mathbf{F}^{(1)}$，以及在旋转后的原子位置 $R\mathbf{X}$上计算$\mathbf{F}^{(2)}$。随后，我们对第一个预测结果 $R\mathbf{F}^{(1)}$进行旋转，并计算其与$\mathbf{F}^{(2)}$的逐原子余弦相似度，再对所有原子和结构取平均。基于 OMol-102M 训练的 AllScAIP 达到 0.9999 的余弦相似度，与 UMA 不相上下，而随机初始化的模型所得数值则低得多。这表明旋转等变性是学到的。

能量守恒。我们遵循Liu等人(2026年提出的可微 kNN 算法，以实现能量守恒。遵循Fu 等人(2025年的做法，我们对七种大型 MD22 分子进行 100 ps 的 NVE 分子动力学模拟（Chmiela 等人，2023年），并报告能量漂移（eV/原子/ps）。AllScAIP 的能量漂移较小，与 UMA 模型的缩放相当。

表4 OMol25 验证结果。 生物分子、电解质、金属配合物和中性有机物的每原子能量平均绝对误差（毫电子伏）和力平均绝对误差（毫电子伏/埃）。AllScAIP 各变体在两个数据划分上的总体能量误差均最低，且力误差表现出竞争力。

		生物分子		电解质		金属配合物		中性有机物		总计	
数据集	**模型**	E ↓	F ↓	E ↓	F ↓	E ↓	F ↓	E ↓	F ↓	E ↓	F ↓
All	eSEN-sm-d.	0.67	6.30	1.24	9.41	2.53	33.08	1.23	13.84	1.49	9.92
	eSEN-sm-cons.	0.59	4.61	1.01	8.08	2.30	28.86	0.84	11.11	1.27	8.25
	eSEN-md-d.	0.34	**2.61**	0.69	**4.40**	**1.73**	**19.99**	**0.59**	**5.63**	0.84	**4.76**
	GemNet-OC	**0.15**	3.88	0.56	5.98	1.83	25.12	0.86	10.38	0.66	6.52
	AllScAIP-sm-ft-cons.	0.22	3.84	0.75	6.01	2.09	26.02	1.00	8.55	0.85	6.61
	AllScAIP-md-d.	**0.15**	2.91	**0.52**	4.70	1.83	22.31	0.72	6.29	**0.64**	5.24
	AllScAIP-md-cons.	0.22	3.23	0.53	5.29	1.86	22.40	**0.59**	7.55	0.67	5.78
4M	eSEN-sm-d.	0.88	8.12	1.93	12.64	3.37	40.44	2.16	20.17	2.19	13.01
	eSEN-sm-cons.	0.86	6.17	1.61	11.16	2.72	35.33	1.50	16.92	1.89	11.10
	eSEN-md-d.	0.47	3.38	1.18	**6.51**	2.53	**27.31**	1.21	**9.26**	1.32	**6.78**
	GemNet-OC	0.25	5.20	1.04	8.42	2.66	32.76	1.64	15.59	1.13	8.98
	AllScAIP-sm-ft-cons.	0.27	4.06	0.91	7.81	2.26	32.10	1.11	12.73	1.04	8.19
	AllScAIP-md-ft-cons.	**0.20**	**3.01**	**0.75**	6.79	**2.06**	29.41	**0.78**	9.45	**0.90**	7.67

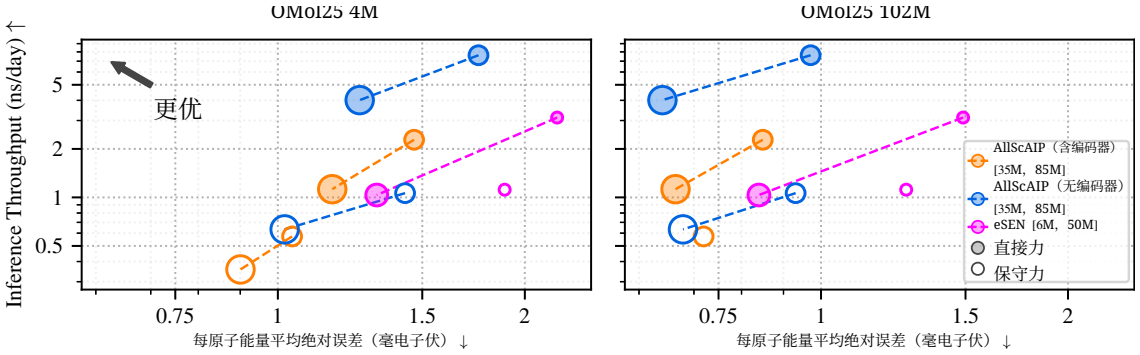


图 5 OMol25 能量误差与效率吞吐量（推理）的关系。 每原子能量平均绝对误差（毫电子伏，↓）与吞吐量（纳秒/天，↑）的关系，展示了带编码和不带编码的 35M/85M 模型；并与 eSEN（Fu 等人，2025年）基线方法进行了比较。保守模型以空心标记表示。左：在 OMol25 4M 数据上训练的模型（80 轮次）。右：OMol25 102M（12 轮次）。我们的模型在 4M 规模下位于帕累托前沿；在 102M 规模下，带编码与不带编码模型之间的差距缩小，甚至出现反转，表明这些归纳偏置在大规模下可能并非必要。与 eSEN 相比，直接力 AllScAIP 模型和保守力 AllScAIP 模型之间更大的速度差距，是因为 AllScAIP 中使用的可微k近邻图构建是一项新引入的操作，尚未得到优化（Liu等人，2026年）。

5.3 开放分子(OMol25)

设置。 我们在 OMol25（Levine等人，2025年）数据集上按照以下设置训练 AllScAIP：

a. **4M划分：** (i) 进行 80 轮直接力训练（**–d.**）；(ii) 先进行 50 轮直接力训练，再切换为基于梯度的能量/力头并继续训练 30 轮的保守微调（**–ft-cons.**）。b. **完整 102M：** (i) 进行 10 轮直接力训练（**–d.**）；(ii) 进行 10 轮直接力训练，然后进行 $+2$ 轮保守微调（**–ft-cons.**）；(iii) 从头开始进行完全保守训练（**–cons.**）。

所有模型均在 V100 32G 上以完整 fp32 精度训练。

能量与力的精度。 在 OMol25 上，我们的模型处于精度–速度帕累托前沿（图 5），能量误差达到最先进水平，同时力误差也具有竞争力（表 4）。在 4M划分上，经过保守微调的中型模型整体能量表现最佳，且力误差接近最强基线；直接力模型则以能量略微增加为代价，换取更好的力预测。在完整 102M 划分上，中型直接力模型的整体能量精度最佳。四个数据划分均呈现相同趋势，在**生物分子**（OMol 中规模最大的系统）上的提升尤为显著，因为长程相互作用在这些系统中最为关键。

OMol25 测试评估。 我们在完整的 OMol25 评估套件上进行基准测试（表 5），并根据模型在七个类别中的平均排名进行排序。在完整的 102M 数据划分上训练的 AllScAIP-md-d. 按平均排名位居首位

表5 OMol25测试评估的评估结果。 结果以能量平均绝对误差 [毫电子伏] 报告（越低越好）。模型按各类别的平均排名排序，平均排名越低表示整体表现越好。

模型	能量一致性	训练集	Avg. Rank	配体口袋 相互作用能 ↓	配体应变 应变能 ↓	构象 ΔE能量 ↓	质子化 ΔE能量 ↓	距离缩放 ΔE能量 ↓	电离能/电子亲和能 ΔE能量 ↓	自旋能隙 ΔE能量 ↓
AllScAIP-md-d.		OMol-102M	3.43	29.496	4.149	4.985	16.104	10.082	143.035	350.825
AllScAIP-md-cons.	✓	OMol-102M	4.14	50.532	4.103	3.874	17.092	36.145	148.976	317.400
eSEN-md-d.		OMol-102M	4.14	64.246	3.113	3.566	14.988	109.767	145.776	303.883
UMA-M-1p1	✓	UMA-459M	4.86	76.778	2.962	2.532	18.457	138.094	136.386	335.058
UMA-S-1p2	✓	UMA-520M	5.00	66.604	4.365	3.738	18.914	32.174	176.939	246.565
AllScAIP-sm-ft-cons.	✓	OMol-102M	6.29	58.450	4.770	4.106	27.515	15.473	168.858	341.708
AllScAIP-md-ft-cons.	✓	OMol-4M	6.43	46.727	4.839	5.172	19.399	16.473	169.104	367.133
AllScAIP-sm-ft-cons.	✓	OMol-4M	8.71	78.260	4.559	5.343	24.614	23.492	198.899	401.358
GemNet-OC-r12		OMol-102M	9.71	19.373	8.381	12.049	31.645	77.074	177.457	373.831
eSEN-sm-cons.	✓	OMol-102M	10.57	147.261	4.656	4.487	21.770	196.964	222.827	391.469
GemNet-OC-r6		OMol-102M	10.86	47.022	9.554	11.479	36.925	80.683	199.721	370.975
UMA-S-1p1	✓	UMA-459M	11.14	127.723	4.856	5.170	27.969	194.688	206.872	369.181
eSEN-md-d.		OMol-4M	11.14	98.337	4.508	5.369	24.975	111.063	240.212	430.753
GemNet-OC-r12		OMol-4M	12.86	36.221	12.489	19.797	50.773	99.995	237.682	490.463
mace-omol-L-0	✓	OMol-102M	14.29	297.404	7.954	6.782	25.107	244.834	338.795	438.891
eSEN-sm-cons.	✓	OMol-4M	14.43	243.883	5.850	6.811	35.754	232.839	293.120	478.242
GemNet-OC-r6		OMol-4M	15.00	96.043	13.270	23.355	52.160	83.296	329.993	559.080

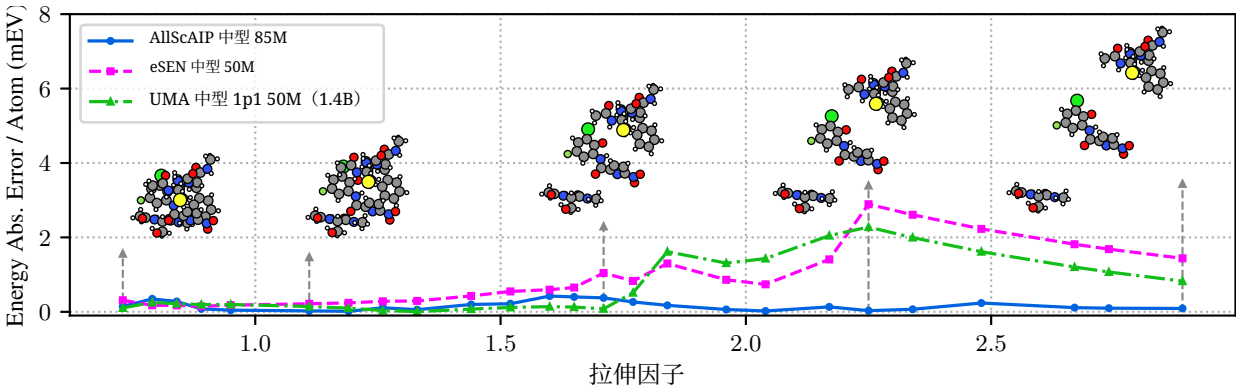


图6 OMol25距离尺度评估。将一组分子通过标量“拉伸因子”进行均匀压缩和拉伸，以探测长程能力（横轴；<1：压缩，>1：拉伸），并报告每原子能量绝对误差（毫电子伏；↓）。AllScAIP-md-cons. 在整个范围内始终保持误差低且平稳，而 eSEN 和 UMA md-1p1 在拉伸时误差急剧增大，表明其长程能力较差。插图展示了虚线标示的选定因子对应的几何结构示例。

。具体而言，我们的模型在距离缩放（LR）测试中表现出色（与第二名模型相比，误差降低约 90%）：当分子均匀压缩或拉伸时，AllScAIP 的能量误差始终较低且平稳，而 eSEN 和 UMA 在大幅拉伸条件下会明显退化（图 6）。这些结果表明，所提出的架构能够迁移到训练分布之外，在保持其余基准测试（配体相互作用/应变、构象、质子化、电离能/电子亲和能以及自旋能隙）竞争力的同时，提供稳健的长程行为。评估详情见 Levine等人 (2025年)。

分子动力学模拟及与实验的比较。为测试 AllScAIP 在运行分子动力学并根据模拟计算宏观可观测量时的表现，我们使用预训练模型直接进行了 NPT模拟（未针对分子动力学进行微调）。应力通过位移对应的能量梯度计算得出（训练期间未对其进行训练）。我们采用 1 fs 的时间步长进行了 300 ps 的 NPT分子动力学模拟，随后根据最后 150 ps 的数据（平衡后）计算密度和汽化焓。

按照Kim等人（2025年），我们对39种分子液体体系在298开尔文下进行了NPT分子动力学模拟，并对相应的孤立分子进行了NVT分子动力学模拟，以提取汽化焓所需的势能（图7）。我们使用在OMol25上训练的AllScAIP与eSEN进行比较，并使用在SPICE上训练的模型（附录C）与MACELES进行比较。与MACELES相比，AllScAIP取得了更低的平均绝对误差和 R^2。eSEN的结果显示出系统性的过度预测（压缩），而AllScAIP消除了这一偏差。

总体而言，分子动力学模拟结果表明，数据驱动的全对全节点注意力无需显式长程项即可产生符合实际的体相行为，这与我们的长程消融实验和距离缩放测试结果一致。关于使用 MD22（Chmiela 等人，2023年）数据集进行分子动力学模拟的更多结果，详见附录 D。

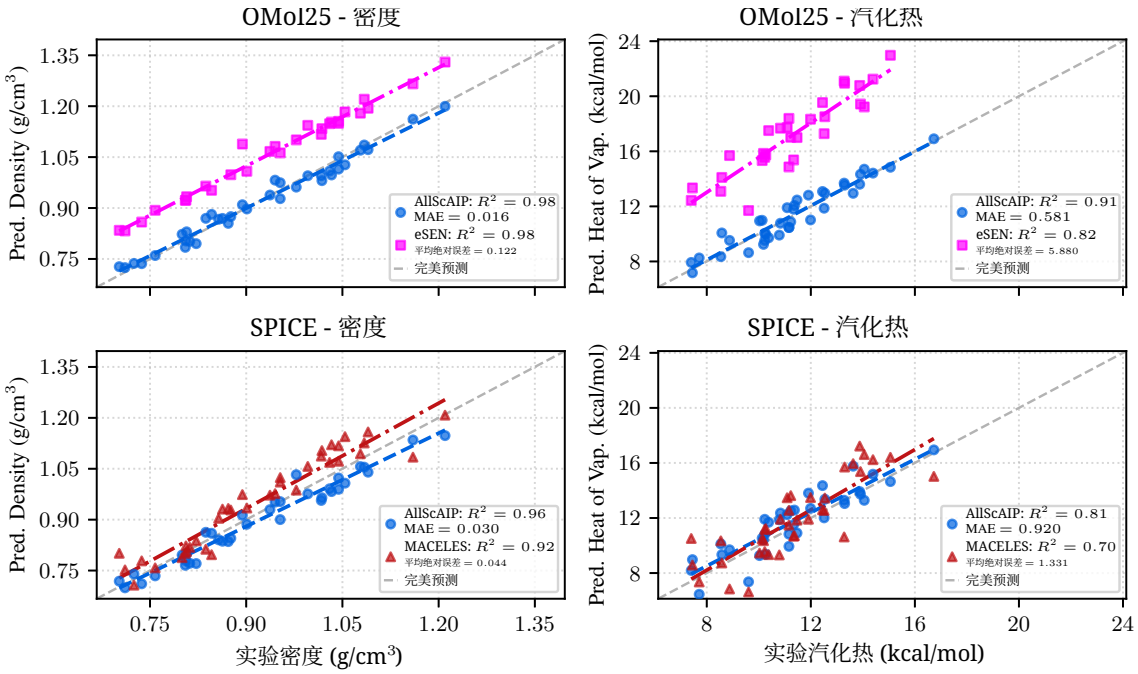


图7 分子动力学模拟与实验密度和汽化焓。 对39种不同的分子液体进行NPT分子动力学，并对孤立分子进行NVT分子动力学，时长均为300 ps。左：预测密度与实验密度的对比。右：预测汽化焓与实验汽化焓的对比。上：在OMol25上训练的模型。下：在SPICE上训练的模型（请注意，以不同密度泛函理论水平计算的SPICE结构也是OMol25的一个子集）。在这两项性质上，AllScAIP的平均绝对误差均低于MACELES和eSEN，且 R^2 更高。

5.4 开放材料与开放催化

AllScAIP 在开放材料（OMat24）（Barroso-Luque等人，2024年）和开放催化（OC20）（Chanussot等人，2021）上实现了具有竞争力的性能。这些是大规模材料数据集，其性质与 OMol25 截然不同。详情请参阅附录 B。

6 讨论

尺度扩展与归纳偏置的作用。我们的消融实验表明一条简单的准则：先扩展规模，再考虑偏置。方向/径向编码（LAE、欧几里得 RoPE）能提升低数据量/小模型情形下的数据效率，但随着数据量和参数规模的扩大，其边际收益会降低。相比之下，全对全节点注意力所提供的长程耦合架构能力，在我们测试的所有规模下都依然有益。我们还注意到，另有研究对图作为归纳偏置提出了质疑（Kreiman等人，2025年）。从实践角度看，这意味着应优先采用可扩展且表达能力强的组件，因为数据和算力只会不断增加；此类组件会随规模扩大而持续受益，而固定的归纳特征可能会限制灵活性。

机器学习原子间势的数据驱动路径与受约束路径。另一类研究通过 SE(3) 等变球谐（SH）不可约表示来编码方向。这种方法提供了丰富的先验方向基，但实际应用中它是带限的：在 L 阶处截断后，非常尖锐或特定于数据集的角特征（例如狭窄的氢键锥、π 堆叠取向等）需要更大的 L。这种表示能力强大，却也较为僵化：其容量受所选球谐阶数和耦合规则限制。我们的路径有所不同：我们保留标量主干，并采用数据驱动的方式。随着规模扩大，模型可以在需要之处分配容量，而不受球谐分解的约束，从而兼顾硬件效率与灵活性。

局限性与机遇。 效率研究表明，复杂度从局部 $\mathcal{O}(Nk)$ 明显转变为 $\mathcal{O}(N^2)$。我们认为，$\mathcal{O}(N^2)$ 是实现长程精度所需付出的、可控的代价，尤其是考虑到大语言模型实现超长上下文所采用的工程方案：激活检查点、键值缓存压缩和分页；其中大多数方案都可以直接应用于原子上的多头自注意力。除系统层面的工作外，还有几种建模途径可以进一步推迟进入 $\mathcal{O}(N^2)$ 阶段：采用跨池注意力的分层节点池、线性时间注意力，以及仅将部分节点路由至全局混合的混合专家模型。

致谢

本研究部分获得了 Research Corporation for Science Advancement 和阿诺德与梅布尔·贝克曼基金会提供的 Scialog 资助，项目编号为 #SA-AUT-2024-015b。我们感谢 Ishan Amin 提出最初的想法，并就 LAE 编码展开讨论。我们感谢加州大学伯克利分校 / 劳伦斯伯克利国家实验室的 Tobias Kreiman、Sam Blau、Ryan Liu、Ritwik Gupta、Sanjeev Raja、Nithin Chalapathi、Yue Jian、Yiheng Du、Michael Psenka、Boyu Qie、Kai Nelson 和 Rasmus Hoegh 提供的宝贵讨论。我们感谢 FAIR 化学团队的 C. Lawrence Zitnick、Misko Dzamba、Meng Gao、Muhammed Shuaibi、Daniel S. Levine、Luis Barroso-Luque 等人提出的宝贵讨论和意见。

参考文献

Luis Barroso-Luque, Muhammed Shuaibi, Xiang Fu, Brandon M. Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, C. Lawrence Zitnick, 和 Zachary W. Ulissi。开放材料 2024（OMat24）无机材料数据集与模型，2024年。Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, 和 Gábor Csányi。MACE：用于快速、准确力场的高阶等变消息传递神经网络。*Advances in Neural Information Processing Systems*，35：11423–11436，2022年。Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, 和 Boris Kozinsky。用于高效数据利用和准确原子间势的 E (3) 等变图神经网络。《自然-通讯》，13(1)：2453，2022年。Jörg Behler 和 Michele Parrinello。高维势能面的广义神经网络表示。《物理评论快报》，98(14)：146401，2007年。Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, 和 Zachary Ulissi。开放催化 2020（OC20）数据集与社区挑战赛。《*ACS* 催化》，11(10)：6059–6072，2021年。doi：10.1021/acscatal.0c04525。Bingqing Cheng。用于机器学习长程相互作用的潜在埃瓦尔德求和。《*npj* 计算材料》，11(1)：80，2025年。Stefan Chmiela, Valentin Vassilev-Galindo, Oliver T Unke, Adil Kabylda, Huziel E Sauceda, Alexandre Tkatchenko, 和 Klaus-Robert Müller。适用于数百原子分子的准确全局机器学习力场。*Science Advances*，9(2)：eadf0873，2023年。Peter Eastman, Pavan Kumar Behara, David L Dotson, Raimondas Galvelis, John E Herr, Josh T Horton, Yuezhi Mao, John D Chodera, Benjamin P Pritchard, Yuanqing Wang 等。用于训练机器学习势的类药分子和肽数据集 SPICE。《科学数据》，10(1)：11，2023年。J Thorben Frank, Stefan Chmiela, Klaus-Robert MÃžller, 和 Oliver T Unke。欧几里得快速注意力：以线性成本实现机器学习全局原子表示。*arXiv* 预印本 *arXiv:2412.08541*，2024年。Xiang Fu, Brandon M Wood, Luis Barroso-Luque, Daniel S Levine, Meng Gao, Misko Dzamba, 和 C Lawrence Zitnick。学习具有光滑性和表达能力的原子间势，用于物理性质预测。*arXiv* 预印本 *arXiv:2502.12147*，2025年。Johannes Gasteiger, Janek Groß, 和 Stephan Günnemann。分子图的方向性消息传递。发表于国际学习表征会议，2020年。https://openreview.net/forum?id=B1eWbxStPH。Johannes Gasteiger, Florian Becker, 和 Stephan Günnemann。GemNet：面向分子的通用方向性图神经网络。*Advances in Neural Information Processing Systems*，34：6790–6802，2021年。Johannes Gasteiger, Muhammed Shuaibi, Anuroop Sriram, Stephan Günnemann, Zachary Ward Ulissi, C Lawrence Zitnick, 和 Abhishek Das。GemNet-OC：为大型、多样化的分子模拟数据集开发图神经网络。《机器学习研究汇刊》，2022年。Sheng Gong, Yumin Zhang, Zhenliang Mu, Zhichen Pu, Hongyi Wang, Zhiao Yu, Mengyi Chen, Tianze Zheng, Zhi Wang, Lifei Chen 等。用于液态电解质开发的预测性机器学习力场框架。*arXiv* 预印本 *arXiv:2404.07181*，2024年。

Sheng Gong, Yumin Zhang, Zhenliang Mu, Zhichen Pu, Hongyi Wang, Xu Han, Zhiao Yu, Mengyi Chen, Tianze Zheng, Zhi Wang 等。用于液态电解质开发的预测性机器学习力场框架。《自然-机器智能》，第 1–10 页，2025年。Andrea Grisafi 和 Michele Ceriotti。将长程物理学融入原子尺度机器学习。《The Journal of Chemical Physics》，151(20)，2019年。Dongjin Kim, Xiaoyu Wang, Peichen Zhong, Daniel S King, Theo Jaffrelot Inizan, 和 Bingqing Cheng。用于机器学习原子间势的通用长程静电作用增强框架。arXiv 预印本 arXiv:2507.14302，2025年。Tsz Wai Ko, Jonas A. Finkler, Stefan Goedecker, 和 Jörg Behler。第四代高维神经网络势：实现准确静电作用，包括非局域电荷转移。《自然-通讯》，12：398，2021年。doi：10.1038/s41467-020-20427-2。Tobias Kreiman, Yutong Bai, Fadi Atieh, Elizabeth Weaver, Eric Qu, 和 Aditi S Krishnapriyan。Transformer 无需图先验即可发现分子结构。arXiv 预印本 arXiv:2510.02259，2025年。Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, 和 Grigory Sizov。xformers：一个模块化且易于定制的 Transformer 建模库。https://github.com/facebookresearch/xformers，2022年。Daniel S Levine, Muhammed Shuaibi, Evan Walter Clark Spotte-Smith, Michael G Taylor, Muhammad R Hasyim, Kyle Michel, Ilyes Batatia, Gábor Csányi, Misko Dzamba, Peter Eastman 等。开放分子 2025（OMol25）数据集、评估与模型。arXiv 预印本 arXiv:2505.08762，2025年。Yi-Lun Liao 和 Tess Smidt。Equiformer：用于三维原子图的等变图注意力 Transformer。发表于第十一届国际学习表征会议，2022年。Yi-Lun Liao, Brandon M Wood, Abhishek Das, 和 Tess Smidt。EquiformerV2：改进的等变 Transformer，可扩展至高阶表示。发表于第十二届国际学习表征会议，2024年。https://openreview.net/forum?id=mCOBKZmrzD。Ryan Liu, Eric Qu, Tobias Kreiman, Samuel M Blau, 和 Aditi S Krishnapriyan。从评估到设计：利用势能面光滑性指标指导机器学习原子间势架构。arXiv 预印本 arXiv:2602.04861，2026年。Philip Loche, Kevin K Huguenin-Dumittan, Melika Honarmand, Qianjun Xu, Egor Rumiantsev, Wei Bin How, Marcel F Langer, 和 Michele Ceriotti。适用于原子机器学习的快速、灵活长程模型。《The Journal of Chemical Physics》，162(14)，2025年。Arslan Mazitov, Filippo Bigi, Matthias Kellner, Paolo Pegolo, Davide Tisi, Guillaume Fraux, Sergey Pozdnyakov, Philip Loche, 和 Michele Ceriotti。PET-MAD：用于先进材料建模的轻量级通用原子间势。《自然-通讯》，16(1)：10653，2025年。Orbital-Materials。Orbital Materials 的 Orb 力场模型：Orbital Materials 的 Orb 力场模型。https: // github.com/orbital-materials/orb-models，2024年。访问日期：2024-10-29。Saro Passaro 和 C Lawrence Zitnick。将 SO(3) 卷积简化为 SO(2)，以实现高效等变 GNN。发表于国际机器学习会议，第 27420–27438 页。P MLR，2023年。Eric Qu 和 Aditi Krishnapriyan。可扩展性的重要性：提升神经网络原子间势在不同化学领域的速度与准确性。Advances in Neural Information Processing Systems，37：139030–139053，2024年。Mikkel Ohm Sauer, Peder Meisner Lyngby, 和 Kristian Sommer Thygesen。适用于二维范德华材料的色散校正机器学习势。《物理评论·材料》，9(7)：074007，2025年。Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, 和 Klaus-Robert Müller。SchNet：用于量子相互作用建模的连续滤波卷积神经网络。Advances in Neural Information Processing Systems，30，2017年。Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon M Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi 等。开放催化 2022（OC22）数据集及氧化物电催化剂相关挑战。《ACS 催化》，13(5)：3066–3084，2023年。

Nguyen Thien Phuc Tu、Nazanin Rezajooei、Erin R. Johnson和Christopher N. Rowley。一种对长程色散进行严格处理的神经网络势。*Digital Discovery*，2:718–727，2023年。doi: 10.1039/D2DD00150K。Oliver T. Unke和Markus Meuwly。PhysNet：一种用于预测能量、力、偶极矩和部分电荷的神经网络。*Journal of Chemical Theory and Computation*，15(6):3678–3693，2019年。doi: 10.1021/acs.jctc.9b00181。Oliver T Unke、Stefan Chmiela、Huziel E Sauceda、Michael Gastegger、Igor Poltavsky、Kristof T Schütt、Alexandre Tkatchenko和Klaus-Robert Müller。机器学习力场。*Chemical Reviews*，121(16):10142–10186，2021年。Brandon M Wood、Misko Dzamba、Xiang Fu、Meng Gao、Muhammed Shuaibi、Luis Barroso-Luque、Kareem Abdelmaqsoud、Vahe Gharakhanyan、John R Kitchin、Daniel S Levine等。UMA：一系列原子通用模型。*arXiv*预印本 *arXiv:2506.23971*，2025年。Eric C-Y Yuan、Yunsheng Liu、Junmin Chen、Peichen Zhong、Sanjeev Raja、Tobias Kreiman、Santiago Vargas、Wenbin Xu、Martin Head-Gordon、Chao Yang等。面向化学与材料原子模拟的基础模型。*arXiv*预印本 *arXiv:2503.10538*，2025年。Linfeng Zhang、Han Wang、Maria Carolina Muniz、Athanassios Z. Panagiotopoulos、Roberto Car和Weinan E。一种包含长程静电相互作用的深度势模型。*The Journal of Chemical Physics*，156(12):124107，2022年。doi: 10.1063/5.0083669。Larry Zitnick、Abhishek Das、Adeesh Kolluru、Janice Lan、Muhammed Shuaibi、Anuroop Sriram、Zachary Ulissi和Brandon Wood。用于建模原子相互作用的球面通道。*Advances in Neural Information Processing Systems*，35:8054–8067，2022年。Roman Zubatyuk、Justin S Smith、Jerzy Leszczynski和Olexandr Isayev。利用原子在分子中神经网络对化学性质进行准确且可迁移的多任务预测。*Science Advances*，5(8):eaav6490，2019年。

附录

A OMol25的补充结果

A.1 效率

我们在图8中报告了按组件与系统规模划分的运行时间细分。该剖析解释了吞吐量–规模曲线中观察到的斜率变化。对于适中的N，计算按 $\mathcal{O}(Nk)$ 缩放，并由邻域注意力驱动。超过交叉点后（180M更早，35M更晚），全对全节点注意力 $\mathcal{O}(N^2)$ 组件占据主导。我们注意到，性能分析器会增加时间和内存开销，并且需要禁用 `torch.compile`，因此这可能与实际性能有所不同。

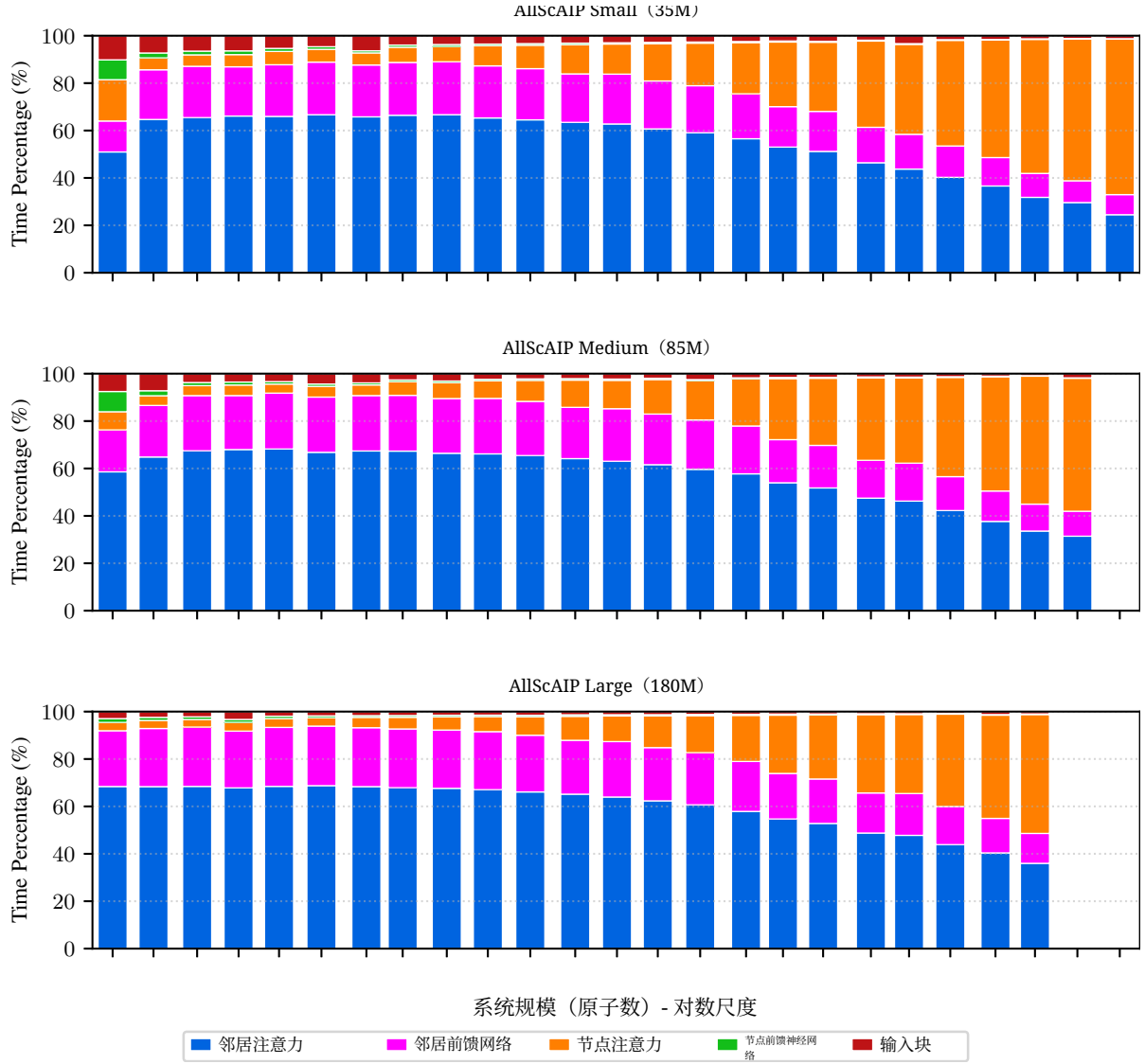


图 8 按组件划分的运行时间细分与系统规模的关系。堆叠条形图显示了在单个 H200 上（关闭图构建）AllScAIP 各模块实际推理时间所占的百分比，涵盖三种模型规模（上：35M，中：85M，下：180M）。颜色分别表示：邻居注意力（蓝色，$\mathcal{O}(Nk)$）、邻居前馈网络（洋红色）、节点注意力（橙色，$\mathcal{O}(N^2)$）、节点前馈神经网络（绿色）和输入/嵌入（红色）。在小规模系统中，运行时间主要消耗在局部模块（邻居注意力）上。随着原子数量增长至数万，全局节点注意力的开销不断增加，并最终占据运行时间的大部分；模型规模越大，交叉点出现得越早。

我们还提供了原始吞吐量与系统规模的关系图，以便进一步比较（图9 ）：

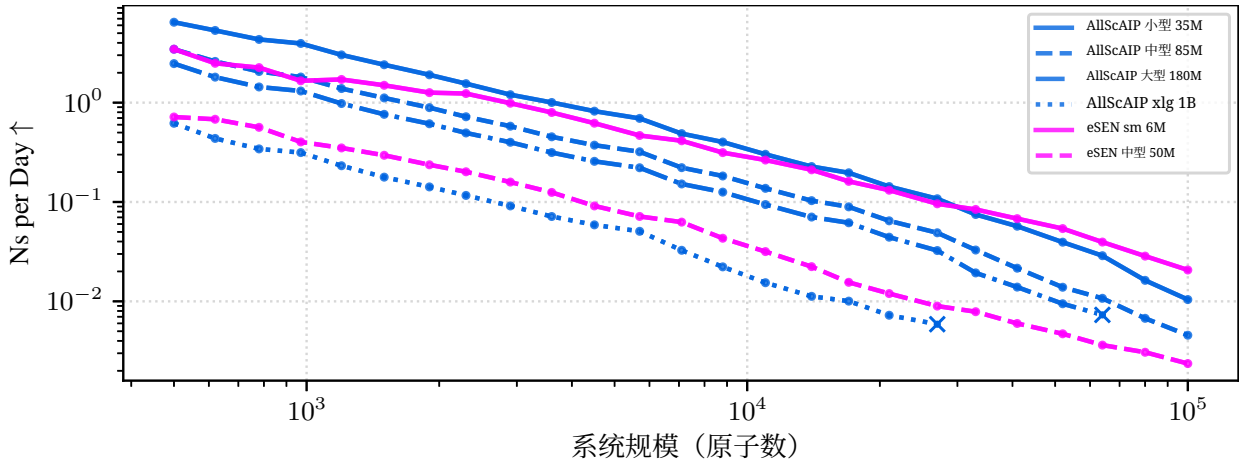


图 9 吞吐量与系统规模的关系。吞吐量（纳秒/天，越高越好）在单张 H200 141G 上测量，且关闭了图生成。折线展示了 AllScAIP 和 eSEN 基线方法的四种模型规模（35M/85M/180M/1B）。

A.2 OMol25 长程验证

为了进一步研究不同模型的长程误差，我们按系统直径和总电荷对 OMol25 验证集进行分箱。我们观察到，AllScAIP 的能量和力平均绝对误差仍然较低，而对于直径较大、电荷较高的系统，eSEN 的误差则显著增大（图 10）。

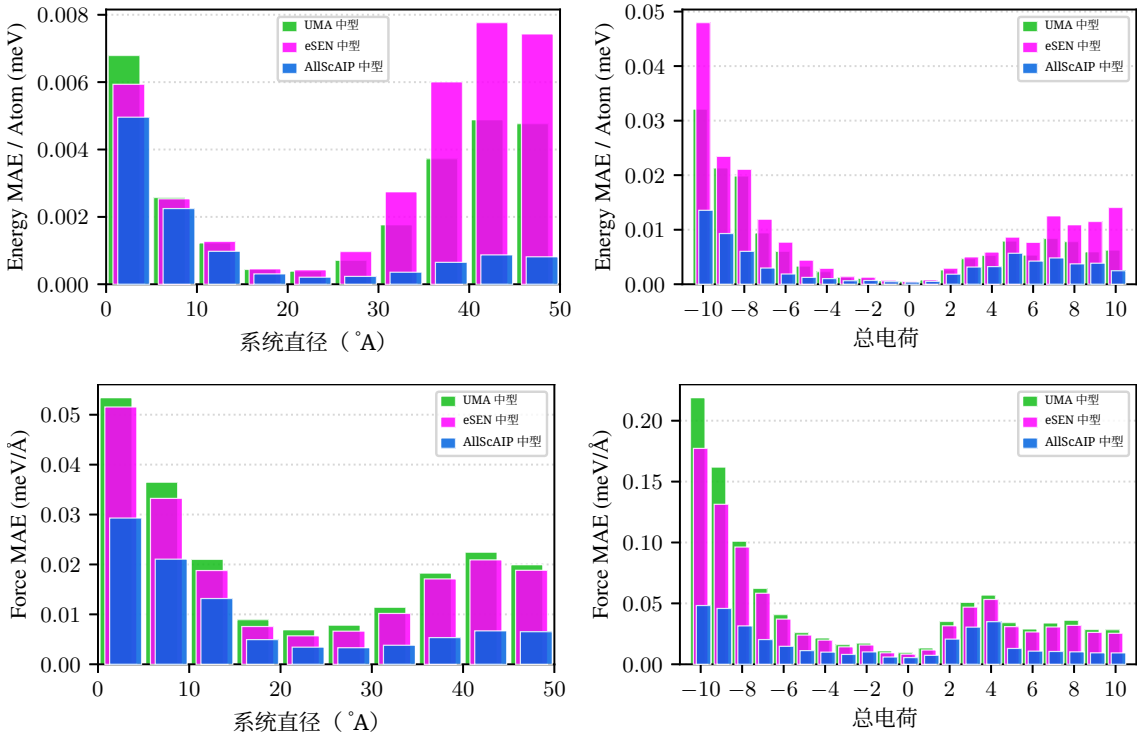


图 10 OMol25 预测误差与系统直径（埃）及总电荷的关系。对整个 OMol25 验证集，按（左）系统直径（埃）和（右）总电荷分组统计每原子能量平均绝对误差（毫电子伏，↓，上图）和力平均绝对误差（毫电子伏/埃，↓，下图）。柱形图对比 AllScAIP-md、UMA-md 和 eSEN-md。AllScAIP 在各分组中的误差始终较低且更稳定，而对于系统直径较大、总电荷较高的体系，eSEN 和 UMA 的误差会增大。

A.3 全对全节点注意力的可视化

为了具体展示全对全节点注意力的长程行为，我们可视化了一个有代表性的大型生物分子（图 10，左上），并从最后一个节点注意力层中提取注意力权重。

设置。 对于所选结构，我们计算成对距离 r_{ij}，并收集每个注意力头 h的 Softmax 后的注意力矩阵 $A^{(h)} \in \mathbb{R}^{N\times N}$。距离以 $\Delta r = 0.5$埃的区间宽度在 $[0,25]$埃范围内分箱。

平均注意力与距离。 对于每个注意力头，我们计算每个距离区间内的注意力平均值（先逐行平均，再跨行平均），并绘制所得曲线。图 10 （右上）展示了所有注意力头（细线）以及注意力头平均值（粗线）。一条竖线标示用于构建半径图的局部邻居截断距离（6埃）。我们一致观察到，部分注意力头在较大间距处分配了高于基线（1/N）的质量，而其他注意力头仍保持局部性，从而使平坦的注意力头平均值接近均匀水平。

距离和注意力热图。 成对距离矩阵证实了该体系的接触结构，而注意力头平均注意力矩阵则显示出垂直条带（“枢纽”）：即从许多其他原子（包括远距离原子）接收注意力的原子。我们发现，部分枢纽与化学特征显著的基团（例如带电或极性位点）重合。

解读。 综合来看，逐头距离分布和热图表明，节点注意力层中存在注意力头专门化和长程混合：部分注意力头能够跨越数十埃传递信息，与局部邻域注意力相辅相成。

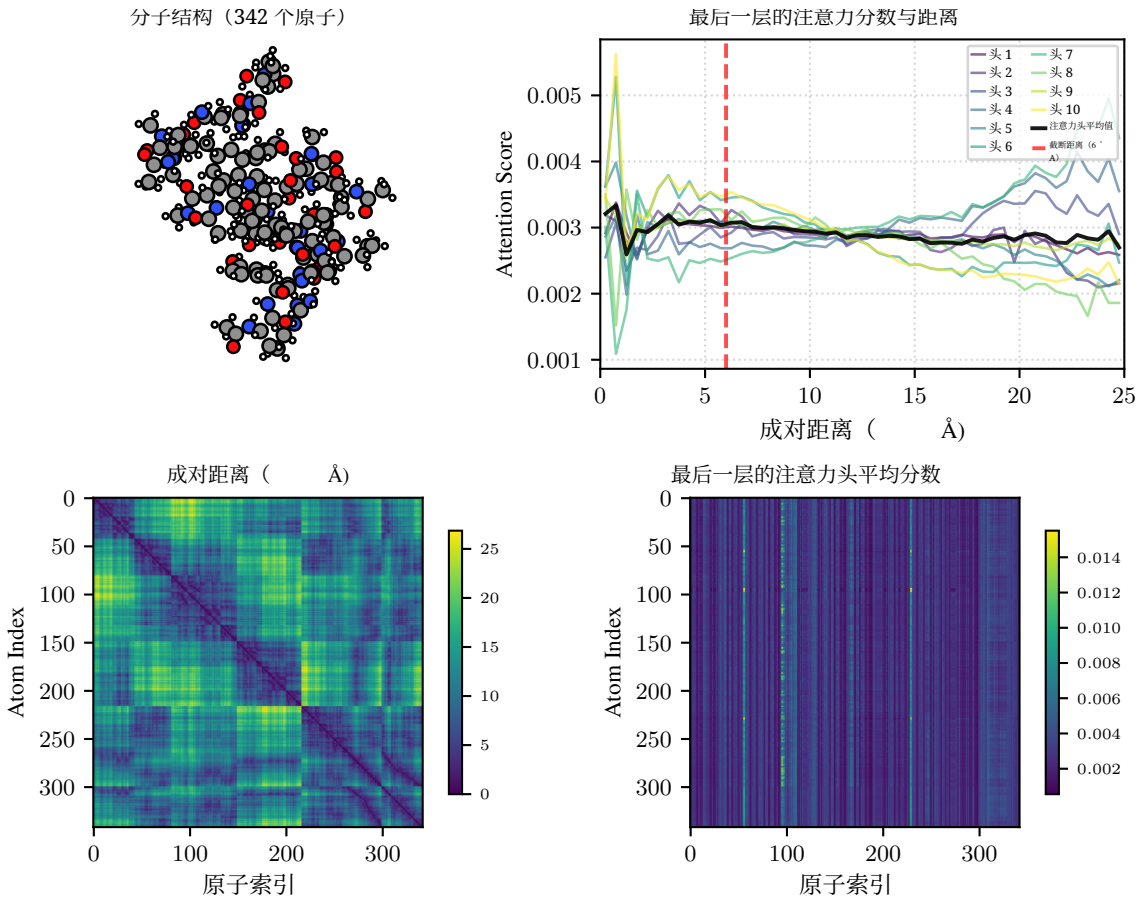


图 11 全对全节点注意力的解读。OMol 验证集中的一个蛋白质口袋示例，含有 342 个原子（左上）。右上：最后一个节点注意力层中，各注意力头的注意力权重随成对距离的变化；细线表示单个注意力头 ＝，粗黑线表示注意力头平均值 ＝。红色虚线标示局部邻居截断距离（6 埃）。左下：成对距离矩阵。右下：注意力头平均注意力矩阵。若干注意力头会将显著权重分配给远超局部截断距离的原子对；注意力热图还显示，某些列形成“枢纽”，在许多行中吸引长程注意力。

表 6 数据和模型规模扩展下的组件消融。我们报告总能量平均绝对误差（毫电子伏）和力平均绝对误差（毫电子伏/埃），数值越低越好。每种大小下，**采用编码**（邻域注意力 + LAE + 节点注意力 + ERoPE）的构型为参考；其他行的颜色表示相对于该参考构型的变化（绿色 = 更好，红色 = 更差）。吞吐量列报告推理速度（在一张 H200 上处理 1000 个原子的纳秒/天）。我们报告 OMol25 4M 划分（80 轮次）和完整的 102M（10 轮次）的结果。

						OMol 4M（80 轮次）									
		消融实验				生物分子		电解质		金属配合物		中性有机物		总计	
Size	**吞吐量**	NeiAtt	LAE	NodeAtt	ERoPE	E ↓	F ↓	E ↓	F ↓	E ↓	F ↓	E ↓	F ↓	E ↓	F ↓
35M	2.279	✓	✓	✓	✓	56.9	5.57	89.3	8.93	166.2	33.52	50.3	13.52	85.2	9.35
	7.623	✓		✓		73	6.32	112.3	9.81	174	34.56	58.2	14.51	102.9	10.21
85M	1.124	✓	✓	✓	✓	34.1	3.51	62.2	6.38	139.6	27.83	31.4	8.93	59.7	6.88
	3.333	✓		✓	✓	43.6	4.27	70.8	7.22	153.7	29.79	40.5	10.53	68.1	7.6
	1.392	✓	✓	✓		43.9	3.78	74.2	6.67	145.2	28.26	36.4	9.03	70.1	7.1
	4.014	✓		✓		48.3	4.24	78.4	7.16	150.6	29.19	40.8	10.12	75.2	7.55
	1.281	✓	✓			85.8	4.58	92.9	7.42	158.6	29.86	38.8	10.01	90.3	7.93
	4.327	✓				90.4	4.96	98.7	7.92	164.4	30.78	42.7	10.77	95.4	8.48
Size	**吞吐量**	NeiAtt	LAE	NodeAtt	ERoPE	**OMol 102M（10 轮次）**									
35M	2.279	✓	✓	✓	✓	33.2	3.85	48.3	6.01	122.3	26.02	25.8	8.54	48.6	6.61
	7.623	✓		✓		43	4.47	59.2	6.73	124	26.86	29.3	9.17	57.9	7.22
85M	1.124	✓	✓	✓	✓	23.7	2.81	36.3	4.59	105.3	21.93	19.7	6.03	37.3	5.1
	3.333	✓		✓	✓	30.8	3.16	41.8	4.99	106.4	22.48	21.5	6.47	42.7	5.51
	1.392	✓	✓	✓		30.2	2.93	41	4.76	104.3	21.59	20.5	5.94	41.7	5.13
	4.014	✓		✓		22.4	2.91	34.6	4.7	105.7	22.31	18.7	6.29	36.1	5.23
	4.327	✓				63.9	3.83	58.4	5.57	116.2	23.91	22.9	7.05	60.4	6.15

表7 OMol25 验证结果。生物分子、电解质、金属配合物和中性有机物的总能量平均绝对误差（毫电子伏）和力平均绝对误差（毫电子伏/埃）。

		生物分子		电解质		金属配合物		中性有机物		总计	
数据集	**模型**	E ↓	F ↓	E ↓	F ↓	E ↓	F ↓	E ↓	F ↓	E ↓	F ↓
All	eSEN-sm-d.	96.92	6.30	88.20	9.41	145.45	33.08	34.24	13.84	89.34	9.92
	eSEN-sm-cons.	86.72	4.61	72.51	8.07	132.71	28.86	22.98	11.11	75.98	8.25
	eSEN-md-d.	50.72	**2.61**	47.37	**4.39**	**97.55**	**19.99**	16.47	**5.62**	49.21	**4.76**
	GemNet-OC	23.35	3.88	**30.73**	5.98	103.10	25.12	21.22	10.38	**33.85**	6.52
	AllScAIP-sm-ft-cons.	31.94	3.34	39.70	6.06	123.69	25.94	18.77	8.97	42.48	6.41
	AllScAIP-md-d.	**22.25**	2.92	33.68	4.84	105.59	22.36	18.57	6.30	35.18	5.25
	AllScAIP-md-cons.	35.86	3.23	37.59	5.46	107.92	22.64	**15.74**	7.52	40.63	5.79
4M	eSEN-sm-d.	127.43	8.12	134.78	12.64	192.77	40.44	59.84	20.17	129.77	13.01
	eSEN-sm-cons.	125.62	6.17	116.65	11.16	156.48	35.33	42.00	16.92	114.36	11.09
	eSEN-md-d。	69.98	3.38	77.82	**6.50**	142.48	**27.31**	33.27	**9.26**	76.56	**6.78**
	GemNet-OC	39.58	5.20	56.32	8.42	148.50	32.76	40.98	15.59	57.82	8.98
	AllScAIP-sm-ft-cons。	43.82	4.08	62.09	7.98	130.94	32.24	29.35	12.72	59.93	8.26
	AllScAIP-md-ft-cons。	**28.27**	**3.62**	**43.85**	7.21	**123.35**	30.68	**23.11**	10.54	**44.40**	7.51

A.4 总能量平均绝对误差结果

我们展示了消融结果（表 6 以及总能量平均绝对误差指标上的验证误差（表 7）。

B OC20 和 OMat24 结果

设置。我们在 OC20（2.4 亿）和 OMat24（1 亿）数据集上训练 AllScAIP。在 OC20 上，我们先进行 3 轮直接力训练，然后进行 1 轮保守微调。直接力训练期间启用 fp16 自动混合精度。在 OMat24 上，我们进行 6 轮直接力训练 + 3 轮保守微调，并采用全 fp32 训练。

结果。我们使用验证集上的总能量平均绝对误差（ID 和 OOD-Both）评估 OC20，并在 OMat24 验证集上进行评估（表8）。我们将结果与 UMA（Wood 等人，2025年）进行比较；UMA 使用了更多数据（459M）进行训练，且计算预算更大。在 OC20 上，中型模型（AllScAIP-md-d.）表现出有竞争力的准确度，在 Val/ID 和 Val/OOD-Both 两项指标上均优于 UMA-S。分布外/分布内能量比（≈1.6）与 UMA-M 相当

这表明其泛化能力良好。在 OMat24 上，中等保守模型的表现优于 UMA-S。我们没有投入大量精力对模型在这两个数据集上的表现进行调优。

表 8 OC20 验证集（总能量）和 OMat24 验证结果。

	OC20（总能量）						OMat24			
	域内验证集			双重分布外验证集			Val			
模型	能量	力	F. Cos.	能量	力	F. Cos.	电子伏特/原子	力	应力	F. Cos.
UMA-S	63.6	24.1	0.63	107.0	29.2	0.65	11.3	57.1	2.9	0.98
UMA-M	43.1	15.8	0.73	70.0	19.2	0.75	10.0	47.3	2.7	0.99
UMA-L	32.6	12.0	0.77	49.8	14.5	0.79	9.7	43.5	2.5	0.99
eqV2-S							11	49.2	2.4	0.985
eqV2-M							10	44.8	2.3	0.986
eqV2-L							9.6	43.1	2.3	0.987
AllScAIP-sm-d.	61.1	18.3	0.67	92.8	22.9	0.68	12.7	60.1	-	0.98
AllScAIP-sm-ft-cons.	72.3	22.5	0.64	120.1	27.3	0.64	12.4	56.0	2.8	0.98
AllScAIP-md-d.	59.3	17.6	0.69	92.2	21.8	0.70	11.4	56.6	-	0.98
AllScAIP-md-ft-cons.							10.7	54.3	2.7	0.98

C SPICE 结果

设置。 为了与其他长程机器学习原子间势（例如 LES（Kim等人，2025年）进行比较，我们在 SPICE 数据集上使用 MACE-OFF 划分（AllScAIP-sm-cons.）训练了一个小型（3400万参数）的能量守恒变体（Eastman 等人，2023年），遵循 Kim等人 （2025年）的方法（我们注意到，SPICE 数据集已在 OMol25（Levine等人，2025年）中以更高理论水平纳入并重新计算）。我们在 H200 142GB 上以完整 fp32 精度训练 300 轮。

结果。 我们在七个留出测试子集上进行评估：PubChem、DES370K 单体/二聚体、二肽、溶剂化氨基酸、水和 QMugs。我们报告每原子能量平均绝对误差（毫电子伏；越低越好）和力平均绝对误差（毫电子伏/埃；越低越好）。表 9 将我们的结果与 MACE/MACELES（Kim等人，2025年）、EScAIP（Qu 和 Krishnapriyan，2024年）以及 eSEN （ Fu 等人，2025 年）进行了比较。我们发现，AllScAIP-sm-cons 在各数据划分上总体能量误差和力误差均达到最低。这些结果表明，该模型配方适用于不同的数据集设置，并且其性能优于其他长程机器学习原子间势设计，例如 MACELES（Kim等人，2025年）。

表 9 SPICE 测试结果。不同 SPICE 数据划分上的每原子能量平均绝对误差（毫电子伏，↓）和力平均绝对误差（毫电子伏/埃，↓）。AllScAIP 的总体能量和力误差最低。

	MACE-L		MACE-M		MACELES-M		EScAIP		eSEN		AllScAIP	
数据集	**E**	**F**	**E**	**F**	**E**	**F**	**E**	**F**	**E**	**F**	**E**	**F**
PubChem	0.88	14.75	0.91	20.57	0.71	17.24	0.53	5.86	0.15	4.21	**0.13**	**3.21**
DES370K M.	0.59	6.58	0.63	9.36	0.54	7.65	0.41	3.48	0.13	1.24	**0.07**	**0.97**
DES370K D.	0.54	6.62	0.58	9.02	0.47	7.48	0.38	2.18	0.15	2.12	**0.06**	**0.95**
二肽	0.42	10.19	0.52	14.27	0.52	11.46	0.31	5.12	0.25	3.68	**0.06**	**1.54**
溶剂化氨基酸	0.98	19.43	1.21	23.26	0.82	18.37	0.61	11.52	0.25	**3.68**	**0.13**	4.86
水	0.83	13.57	0.76	15.27	0.69	12.54	0.72	10.31	0.15	2.50	**0.08**	**2.12**
QMugs	0.45	16.93	0.69	23.58	0.76	20.11	0.41	8.74	0.12	3.78	**0.11**	**2.73**

D 补充分子动力学模拟

D.1 MD22上的结果

为评估分布偏移下的 MD 结构保真度，我们对七种 MD22 分子（零样本 NVT 模拟（Chmiela 等人，2023年）使用经 OMol 训练的模型进行（图 12）。我们计算了径向分布

$h(r)$，并将生产阶段的数据与使用 PBE+MBD 生成的参考轨迹进行比较。PBE+MBD 的理论精度低于 OMol，因此预计不会与我们的训练目标一致。在各个体系中，AllScAIP 重现了与 eSEN-sm 和 UMA-md 相同的峰值位置和整体形状，表明其平衡结构相当，并且分子间距离不存在明显偏差。

E 详细模型配置

表 10 AllScAIP 模型配置

模型大小	小型	中型	大型	超大型
隐藏_size	512	640	1024	2048
num_层	6	10	8	12
注意力_num_注意力	8	10	16	32
注意力_name	内存_高效			
激活函数	gelu			
归一化	RMSNorm			
node_方向_扩展_size	10			
edge_方向_扩展_size	6			
edge_距离_扩展_size	512			
attn_num_freq	32			
ffn_隐藏_层_乘数	2			
输出_隐藏_层_乘数	2			
max_num_元素	110			
max_批次_size	96			
knn_soft	True			
knn_Sigmoid_缩放	0.2			
knn_lse_缩放	0.1			
knn_use_low_mem	True			
距离_函数	高斯			
use_包络	True			
mlp_丢弃法	0.0			
注意力_丢弃法	0.0			

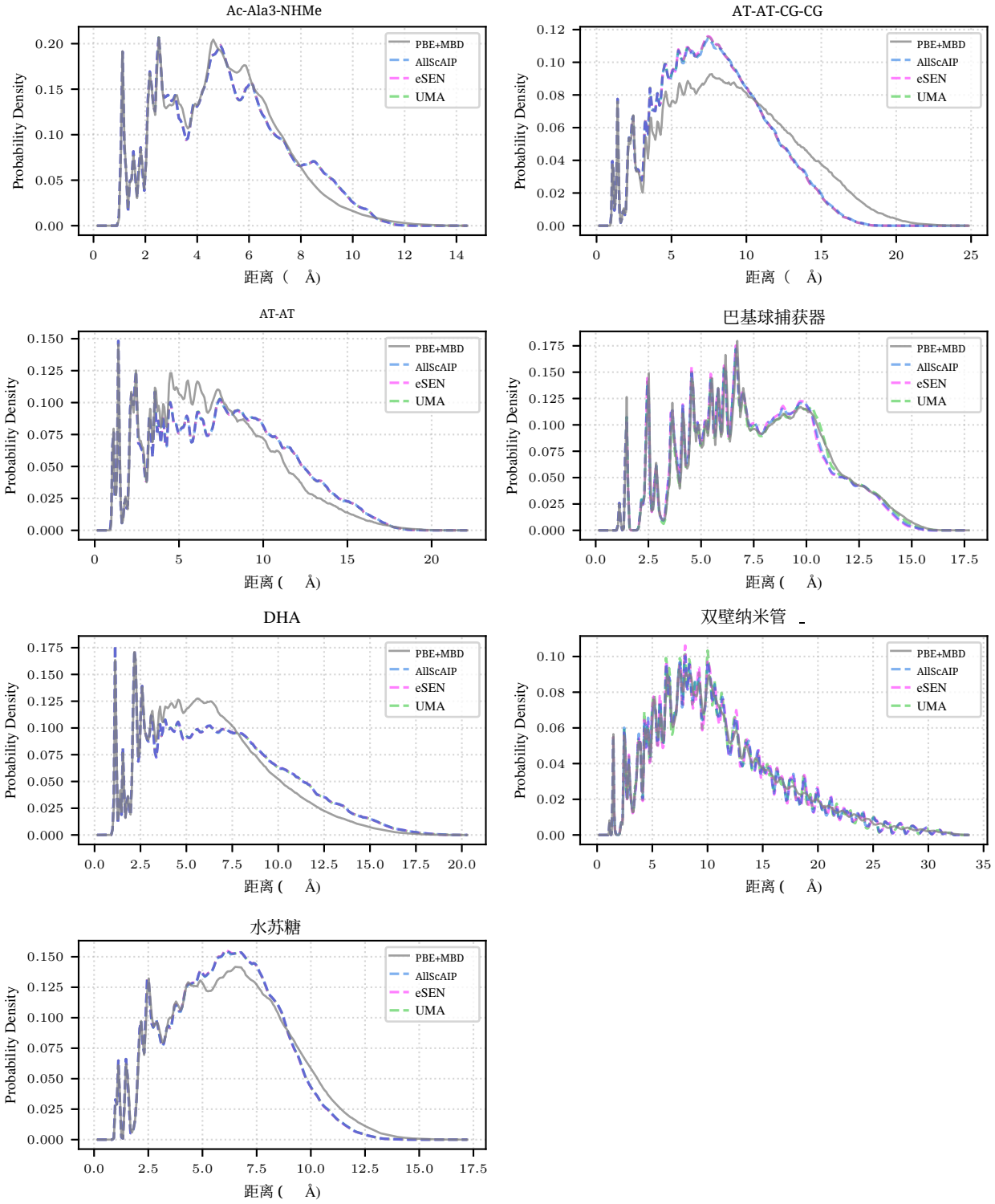


图12 MD22上的零样本MD：径向分布 $h(r)$。 使用基于OMol训练的AllScAIP在七个MD22系统上生成的NVT轨迹。曲线显示模拟得到的 $h(r)$，并与PBE+MBD参考计算进行比较。尽管参考计算所用的方法与OMol不同，AllScAIP在峰值位置和整体形状上均与eSEN/UMA一致，表明其结构统计特性相匹配。