arXiv:2601.08486v1 [physics.comp-ph] 13 Jan 2026

多任务微调使原子模型具备稳健的分布外泛化能力

Chengqian Zhang[1,2], Duo Zhang[1,2], Anyang Peng[1], Mingyu Guo[3,4], Yuzhi Zhang[3], Lei Wang[5], Guolin Ke[3], Linfeng Zhang[1,3], Tiejun Li[2, 6, 7†], Han Wang[8, 9†]

[1]AI for Science Institute，中国北京；[2]北京大学数据科学研究中心，中国北京；[3]深势科技，中国北京；[4]中山大学化学学院，中国广州；[5]中国科学院物理研究所北京凝聚态物理国家研究中心，中国北京；[6]北京大学LMAM与数学科学学院，中国北京；[7]北京大学机器学习研究中心，中国北京；[8]应用物理与计算数学研究所计算物理国家重点实验室，中国北京；[9]北京大学工学院HEDPS、CAPT，中国北京；

[†]通讯作者：tieli@pku.edu.cn wang han@iapcm.ac.cn,

摘要

精准的从头分子与材料设计，需要能够泛化到已知范围之外的结构–性质模型。尽管预训练原子级模型经过微调后可达到较高的分布内准确率，但其在分布外（OOD）条件下的可靠性仍不明确。我们发现了下游适配中的一种关键失效模式：标准微调会导致表征坍塌，抹去预训练得到的化学与结构先验，并严重降低分布外性能。为解决这一局限，我们提出多任务微调（MFT），将下游性质预测与从预训练继承而来的、具有物理依据的力场目标联合优化。这种方法既能保留重要的化学先验，又能实现任务特定适配。在分子与材料基准上，MFT 始终能够提升分布外泛化，使其接近由分布内准确率所限定的理论上限，同时优于标准微调、从头训练以及最先进的任务特定模型。这些结果表明，安全适配是大型原子级模型的核心要求，并使 MFT 成为推动稳健分子与材料发现的一种实用且数据高效的途径。

1. 引言

从头设计具有卓越性质的分子和材料仍然极具挑战性 [60]。这一工作的一项关键内容是准确建模结构–性质关系。这些关系支持两种互补策略：(i) 对数据库进行高通量虚拟筛选，以识别具有目标性质的候选物 [17], ；以及 (ii) 进行条件生成建模，生成满足指定性质约束的新结构 [10, 5]。原子级机器学习（ML）模型学习从三维原子结构（坐标和元素种类）到性质的映射，与仅基于化学式、一维字符串表征（例如简化分子线性输入规范，SMILES）或二维平面图表示的模型相比，始终具有更高的预测准确度 [85, 3]。因此，用于性质预测的原子级机器学习模型已成为分子与材料设计中不可或缺的工具。

从头设计旨在发现具有迄今未有记载的性质、化学组成或构型特征的分子和材料，因此需要能够可靠泛化到分布外（OOD）分子和材料的模型。然而，近期的基准研究凸显出分布外泛化能力显著下降，这表现为明显的

性质预测中，机器学习模型在分布内任务（ID；样本取自训练数据分布）与分布外任务之间存在性能差距。在有机分子领域，BOOM基准 [3] 根据目标属性值显著高于或低于训练数据覆盖范围的分子定义分布外划分，并揭示即使是表现最优的模型，其平均分布外误差也可达到分布内性能对应误差的3倍。Shi makawa 等人 [66] 进一步表明，在基于性质和基于结构的分布外设置下，机器学习性质预测模型的性能都会严重下降，尤其是在训练数据有限的任务中。对于无机晶体材料，分布外基准无论是识别特征空间或性质空间中密度最低的区域 [48] ，还是在组成空间或聚类特征空间中采用留一法协议 [48, 41], ，均一致表明，最先进水平的图神经网络（GNN）在训练域之外进行评估时，准确率会显著下降。总体而言，这些发现凸显了一个关键且尚未解决的挑战：在原子尺度性质预测中，无论是分子系统还是材料系统，也几乎不受分布外情形具体定义的影响，分布外泛化能力严重下降都是普遍存在的现象。

通过在大规模数据集上进行预训练，将先验化学知识编码到表征（特征）空间中，再针对下游性质预测任务进行微调，已被证明能够有效提升模型的泛化能力。已有研究表明，坐标去噪和原子类型掩码等自监督预训练策略能够学习稳健的表征，从而显著提升下游性质预测性能 [81, 85, 68]。类似地，基于力场任务的监督式预训练 [7, 32, 56, 78, 74, 8],——即在由第一性原理计算（如密度泛函理论（DFT））标注的分子或材料构型数据集上进行训练 [6, 11, 39], ——已被证明能够编码符合化学直觉的表征 [82], ，从而在针对材料和分子性质预测进行微调时大幅提升准确率 [67, 78, 34]。然而，利用此类先验化学知识所取得的性能提升，几乎完全是在分布内设置下进行评估的。针对性质预测任务微调预训练模型能否可靠地提升分布外性能这一根本问题，目前仍未得到充分探索。

预训练视觉和语言模型通常表现出强大的零样本性能 [13, 26], ，而针对特定任务的微调可能会通过表征坍塌损害分布外泛化，即在下游适配过程中通用先验知识受到扭曲 [2, 54]。相比之下，保留预训练表征（例如采用线性探测）能够带来更优的分布外鲁棒性；随着预训练表征质量的提升 [37]。

在这项工作中，我们证明，对预训练原子级模型进行标准微调（FT）（图1（b））会导致分布外泛化显著退化：微调模型的分布外性能远差于其分布内性能（图2（a））。我们进一步表明，这种退化伴随着预训练过程中习得的化学先验知识的丧失，表明发生了表征坍塌（图2（e）（h））。尽管线性探测（LP）模型（图1（c））的分布外性能略优于 FT，但其分布内泛化能力显著下降（图2（b））。为解决这一问题，我们提出多任务微调（MFT），联合优化下游性质预测任务和上游力场预训练任务（图1（d））。我们使用大型原子模型 DPA-3.1-3M 对 MFT 进行了系统评估，该模型 [83], 在涵盖材料和分子系统的多样化力场数据集上进行了预训练 [1]。采用 MFT 后，相对于从头训练所取得的分布外泛化提升，与相应的分布内提升相当，且与 FT 达到的分布内性能相当（图2（c））。这一提升还伴随着表征坍塌的明显缓解（图2（f）（i））。在绝对准确率方面，MFT DPA-3.1-3M 模型在分子（表3）和材料（表4）基准测试中均取得最佳分布外性能，其中也包括专为性质预测架构设计的基准测试，展现了所提方法的鲁棒性，以及其在真实世界分子和材料设计应用中的实用性与成熟度。

2. 结果

2.1. MFT概述

我们在图中概述了本研究采用的不同下游适配策略。1。每种下游适配设置均与从头训练（Scratch）基线进行比较；在该基线中，骨干网络和任务头都采用随机初始化，如图所示。1(a)。在微调过程中，预训练骨干网络会被迁移，并与随机初始化的任务专用预测头相结合，同时端到端更新所有模型参数，如图所示。1(b)。此外，我们还研究了LP模型的性能，即冻结预训练模型的骨干网络，仅微调任务专用头，如图所示。1(c)。这种设置将表征限制为保持固定，因此不会受到分布漂移的影响。

在本研究中，我们提出了 MFT，如图1(d) 所示。与微调类似，预训练模型的骨干网络会迁移到下游任务中。然而，在 MFT 中，骨干网络连接到两个任务头：一个

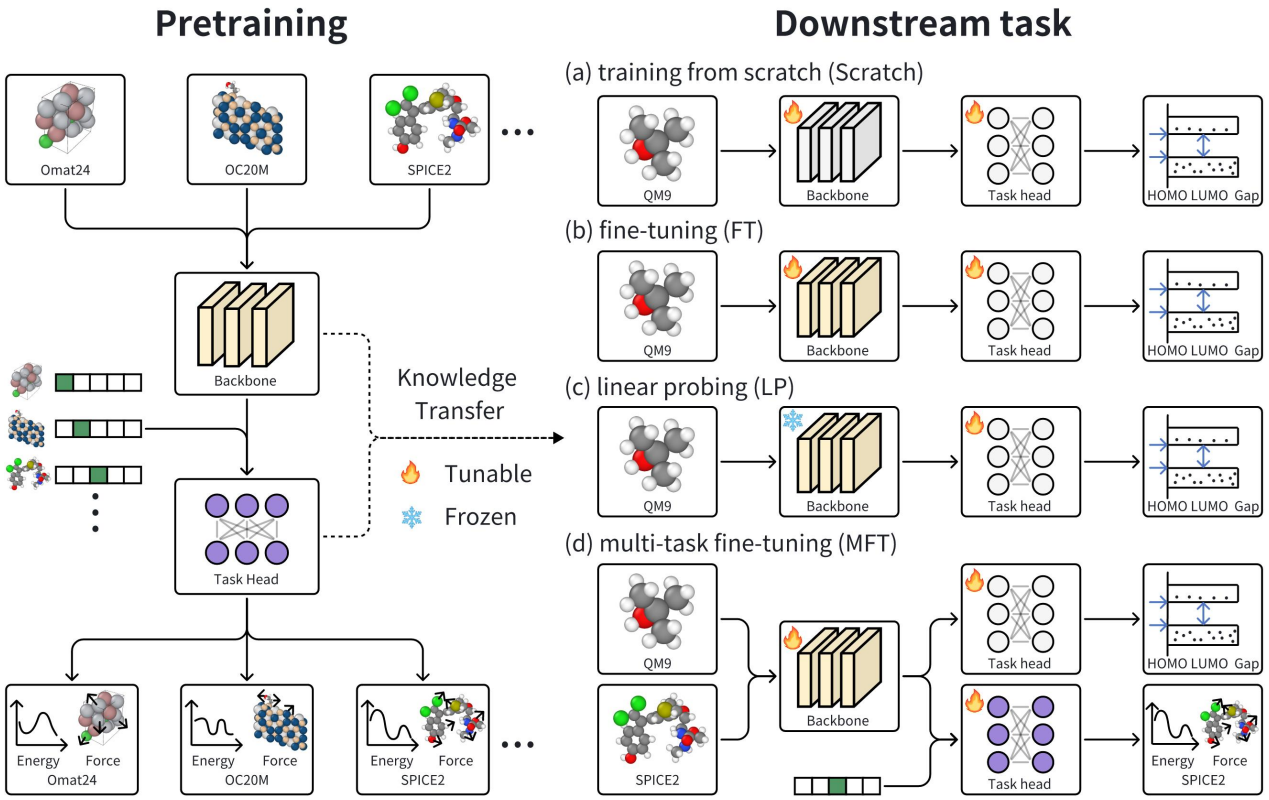


图 1：预训练与下游适配策略示意图。左侧面板展示预训练阶段。右侧对比了四种下游适配策略：(a) 从头训练（Scratch），即骨干网络和任务专用预测头均随机初始化，并进行端到端优化；(b) 微调（FT），即骨干网络从预训练模型初始化，并联合更新所有参数；(c) 线性探测（LP），即冻结预训练骨干网络，仅优化随机初始化的任务专用预测头；(d) 多任务微调（MFT），即联合优化预训练骨干网络和两个预测头：用于性质预测的任务专用预测头，以及从预训练阶段继承的辅助力场头。

为下游性质预测随机初始化，另一个则继承自预训练期间使用的力场预测头。两个任务头与共享骨干网络一起，针对下游性质目标和辅助力场目标进行联合优化。辅助预训练任务可以选择与目标下游领域高度契合的任务。例如，在针对分子性质预测进行微调时，SPICE2 数据集 [20],专为训练有机分子的机器学习原子间势而设计，可用作辅助力场任务。在进行材料性质预测时，OMat24 数据集 [6], 旨在用于通用材料机器学习势，并特别强调结构和组成多样性，可选作相应的力场任务。这种联合优化策略促使骨干网络学习一种表征空间，使其既能针对特定的下游性质预测任务达到最优，又能保留在训练过程中获得的广泛化学和构型先验知识预训练期间。表1总结了这些方法之间的比较，而第4节全面介绍了每种方法。

我们研究了预训练图神经网络模型 DPA-3.1-3M 的微调行为 [83] 在 QM9 分子性质预测任务上的表现。图中总结的结果2表明，分布内泛化与分布外泛化之间的性能差距与学习到的表征空间中的表征坍塌密切相关。微调预训练 JMP 模型时也观察到了类似的表征坍塌现象 [67], , 完整结果见补充材料第B节。这些观察表明，表征坍塌是微调预训练原子模型时普遍存在的一种失效模式，而非特定模型架构或预训练设置造成的假象。

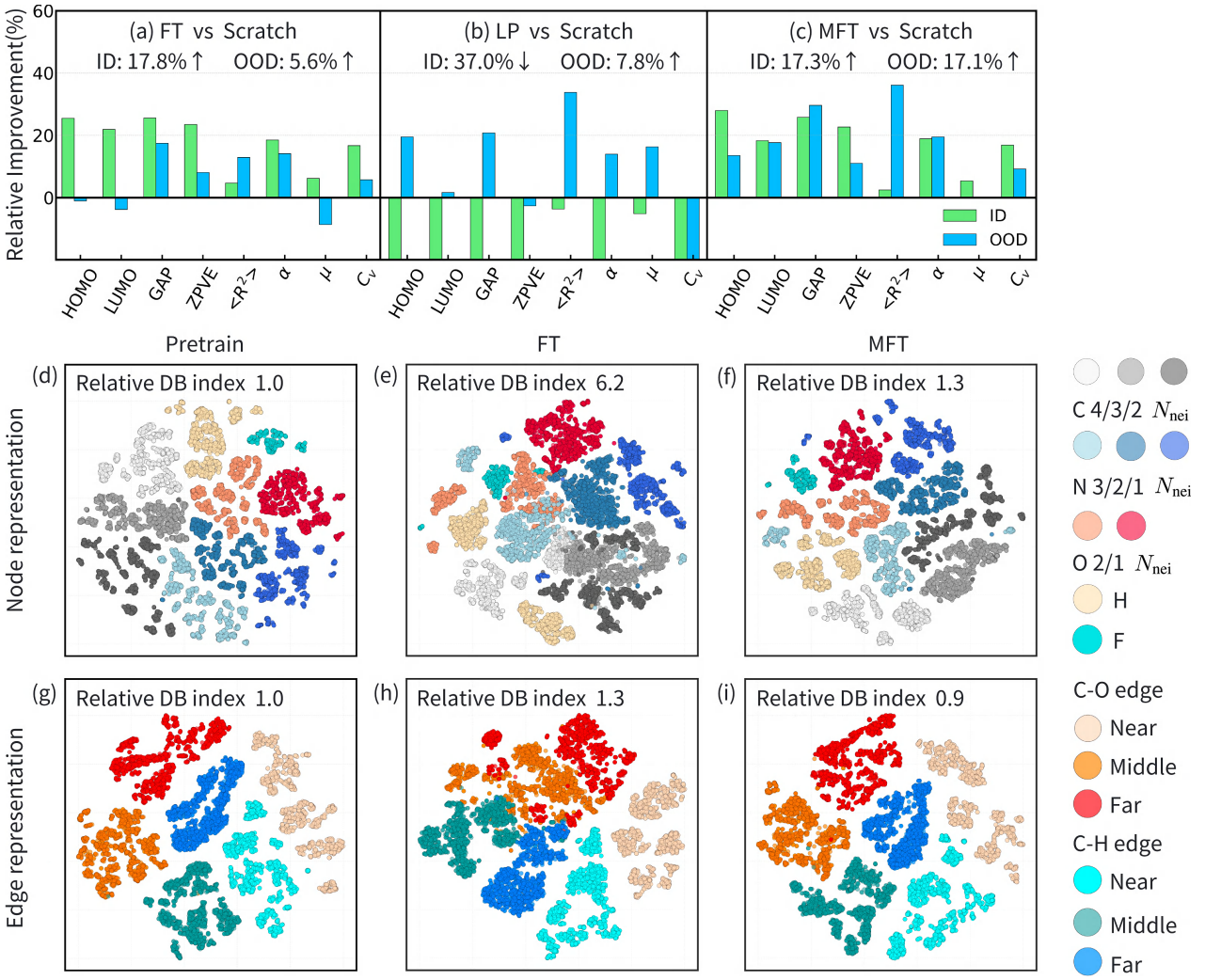


图 2：不同下游适配策略下表征坍塌的分析。(a–c) 微调（FT）、线性探测（LP）和多任务微调（MFT）相较于从头训练（Scratch）基线的相对性能提升。(d–f) 和 (g–i) 分别为原子（节点）表征和边表示的 t-SNE 可视化结果，投影至二维空间。原子表征聚类和边表示聚类的清晰度与分离度通过戴维斯–鲍尔丁（DB）指数进行量化，并以预训练模型为参照进行归一化。相对 DB 指数值越大，表明簇内聚性和可分性退化越严重，为预训练表征空间中的坍塌提供了定量证据。

2.2. MFT 弥合泛化差距

数据划分。QM9 数据集被划分为训练集、验证集、分布内测试集和分布外测试集。训练集、验证集和分布内测试集取自同一分布，而分布外样本则从性质分布的尾部选取（通过核密度估计得到的概率质量最低 10%），代表极端或罕见的分子性质。该划分方案遵循既有的分子性质预测分布外基准，即 BOOM [3]。

*FT*下分布内性能显著提升，但分布外性能退化。如图 2(a)所示，与从头训练相比，FT使分布内准确率平均提高了17.8%，但分布外准确率仅提高了5.6%。在八项性质预测任务中，分布外性能的相对提升始终小于相应的分布内性能提升，唯一的例外是电子空间范围（$\langle R^2 \rangle$）。值得注意的是，对于偶极矩（μ）和最低未占分子轨道（LUMO）能量，微调模型在分布外情形下的表现分别比从头训练基线低8.7%和3.8%。

表 1：下游适配策略对比。该表重点展示了从头训练（Scratch）、微调（FT）、线性探测（LP）和我们提出的多任务微调（MFT）在骨干网络初始化、更新参数以及是否包含辅助训练任务方面的差异。

方法	骨干网络初始化	更新骨干网络	更新任务头	辅助 Task
Scratch	随机	√	√	×
FT	预训练	√	√	×
LP	预训练	×	√	×
MFT（我们的方法）	预训练	√	√	√

*LP*增强分布外性能，但导致分布内性能退化。如图 2(b)所示，与从头训练的模型相比，LP的分布外准确率提高了7.8%，略高于FT带来的提升。相反，与从头训练基线相比，分布内泛化性下降了37%，表明冻结骨干网络严重限制了模型的表达能力。这种现象并不意外，因为下游性质预测任务所预测的目标量与力场学习中的目标量有很大差异。具体而言，性质预测模型旨在推断电子结构性质（最高占据分子轨道（HOMO）、LUMO、GAP和偶极矩）以及热力学性质（例如热容），而力场模型则通过训练来预测基态能量及其相对于原子坐标的导数（力）。这种预测范式的根本转变使得表征空间必须进行适应。

*MFT*弥合了泛化差距。 有效的微调策略必须使表征能够适应下游任务的需求，同时保留预训练期间获取的化学和构型先验知识，而本文提出的MFT策略满足了这些要求。如图所示 2(c)，采用MFT时，分布外泛化相对于从头训练的提升幅度达到17.1%，与相应的分布内提升幅度（17.3%）非常接近，也与微调（FT）达到的分布内性能（17.8%）相当。在表现最佳的情形（$\langle R^2 \rangle$）中，分布外泛化提升了36.1%；而在最差的情形（μ）中，分布外性能仍与从头训练的模型相当。从这个意义上说，MFT没有明显的泛化差距，其分布外泛化性接近由分布内泛化性所给出的理论上界。

2.3. MFT 缓解表示坍塌

预训练原子表示。图2（d）展示了预训练模型的原子（节点）表征。不同化学物种的原子以不同颜色表示，而同一物种的原子则根据其共价邻居数量进一步区分（详情请参见补充材料第D.1 节）。如图2（d）所示，预训练模型的原子表征与化学直觉高度一致：具有相同化学物种和相同共价环境的原子形成彼此明显分离的簇，与不同物种或不同配位环境的原子清晰区分开来。

原子表示坍塌。然而，微调后，不同物种的表征之间出现了显著重叠，如图 2(e) 所示。例如，氮原子和氧原子的表征开始混杂。此外，对局部共价环境的分辨能力显著降低。共价邻居数量不同的碳原子彼此混杂，有两个共价邻居的氧原子分裂成受氢和氟影响的多个簇，而氢原子则按与其形成单键的碳原子簇分开。表示聚类的清晰度和可分性可以通过戴维斯–博尔丁（DB）指数进行定量评估 [18], 其中，数值越低表示簇越清晰、越紧密。如图 2(e) 所示，微调后，原子表征的 DB 指数相较于预训练表征增加了 6.2 倍，表明簇可分性大幅模糊并丧失。综上，这些观察结果表明，微调后，预训练过程中学习到的、在化学身份和局部共价环境方面具有化学意义的先验表征发生了坍塌。

MFT 缓解了原子表示的坍塌。MFT 在分布外表现上的提升，伴随着原子表示坍塌的明显缓解。如图所示，2(f)，对应相同化学物种和共价环境的原子表示明显分离为不同的簇。仅观察到少数例外：具有三个共价邻居的氮原子受氢影响，被分为两个簇。定量衡量簇间分离程度的 DB 指数仅增加了 30%；相比之下，标准微调中，DB 指数相较于预训练原子表示的 DB 指数增加了 520%。

预训练边表示。作为一种图神经网络，DPA-3.1-3M 使用边表示来编码预设截断半径内的原子对关系。图 2(g) 展示了碳–氢（C–H）和碳–氧（C–O）原子对的边表示，这两种边类型在下游任务中十分常见。根据成对欧氏距离，这些边进一步分为近、中、远三类：近距离对应形成共价键的原子对，而中距离和远距离则分别表示其特征在有或没有角度信息的情况下更新的非键合原子对。预训练模型呈现出一

表 2：使用 1-瓦瑟斯坦距离（$\mathcal{W}_{\text{Pretrain}\to\text{FT}}$, $\mathcal{W}_{\text{Pretrain}\to\text{MFT}}$）量化预训练与微调后表征之间的分布偏移。报告了 QM9 LUMO 预测任务在训练集、分布内集和分布外集上的指标。

	训练集	ID	OOD
$\mathcal{W}_{\text{Pretrain}\to\text{FT}}$	3.76	3.75	6.54
$\mathcal{W}_{\text{Pretrain}\to\text{MFT}}$	1.51	1.49	1.97

高度有序的边表示，边根据化学身份和原子间距离明显聚类，这表明预训练模型编码了强大的化学与几何先验。

边表示的坍塌。相比之下，微调导致这种潜在组织结构明显退化，如图2（h）所示。尽管近距离边由于强共价信号而仍相对紧密，但中距离和远距离边的表征逐渐重叠：C–O 边表示开始与 C–H表示混合，中距离和远距离子类也变得纠缠不清。构型分辨率的这种损失体现为 DB 指数增加了 1.3 倍。总体而言，观察到的坍塌表明，模型区分长程空间相关性的能力受到损害，这说明微调优先进行下游标签拟合，却牺牲了对具有化学和几何意义的表征的保留。

MFT 可缓解边表示的坍塌。所提出的 MFT 方法能够有效缓解边表示的坍塌。如图 2(i) 所示，在 MFT 下，对应不同成对类型和原子间距离的边表示仍能很好地区分开来，其 DB 指数甚至比预训练边表示低 10%。这一表现与微调形成鲜明对比：微调会使 DB 指数增加 30%。

2.4. MFT 缓解了预训练表征与微调表征之间的分布偏移

*FT*导致显著的表征漂移。 我们使用1-瓦瑟斯坦距离（记作 $\mathcal{W}$）对预训练模型与微调模型的表征分布之间的相似性（或等价地，漂移程度）进行形式化量化。如表2所示，FT前后训练分布和ID测试分布的漂移距离非常接近，这符合预期，因为这些样本来自相同的底层分布，因此经历了相似的变化。相比之下，OOD测试分布呈现出显著更大的漂移，这表明FT对OOD样本的表征变化远强于对ID样本的影响。这种显著的分布偏移清楚地表明，FT导致了过拟合。

MFT 可缓解表征漂移。MFT 显著缩小了预训练表征与微调表征之间的分布偏移。表 2所示，在 MFT 框架下，训练集和分布内测试集的表征分布漂移分别为 1.51 和 1.49，远小于微调下的相应数值（分别为 3.76 和 3.75）。分布外分布漂移为 1.97，仅略高于训练集／分布内测试集的漂移，且明显小于微调导致的分布外分布漂移（6.54）。

然而，完全消除表征漂移会适得其反。如第2.2节所示，通过冻结骨干网络来维持零表征漂移的线性探测，分布内性能较差。这是因为模型无法弥合力场预训练与下游性质预测之间的范式转变。最佳微调策略必须把握好平衡：既要使表征适应特定任务的要求，又要保留预训练过程中学到的基础化学先验。我们的结果表明，MFT 成功实现了这种平衡。

2.5. MFT 优于最先进水平的分子性质预测器

如表 3所示，我们在 BOOM QM9 基准上将所提出的 MFT 策略与最先进水平（SOTA）的性质预测基线方法进行比较。MFT 模型（记为 DPA3-MFT）与应用于同一预训练骨干网络 DPA-3.1-3M 的标准微调（DPA3-FT）和线性探测（DPA3-LP）进行比较；图 1 展示了这些设置的示意图。为确保公平比较，我们还使用相同的模型架构从头训练模型（DPA3-Scratch）。DPA3-FT 和 DPA3-MFT 在分布内场景中均表现出明显优势：它们在全部八项任务中始终取得最佳分布内性能，证实从预训练模型进行微调优于最先进水平的任务特定从头训练，且 MFT 策略不会损害分布内准确率。按照 BOOM [3], 采用性质外推划分评估的分布外结果，凸显了 MFT 方法的主要优势。DPA3-MFT 在八项任务中的六项取得最佳分布外性能，在另一项任务中取得第二佳表现。相比之下，DPA3-FT 在泛化至分布外数据时往往会出现性能权衡，这可能是因为预训练期间学得的表征中所编码的化学先验发生了坍塌，导致分布外误差增大。DPA3-MFT 通过明确缓解这一效应，显著提升了模型在分布偏移下的稳健性。

表 3：QM9 BOOM 基准中八项分子性质预测任务的均方根误差（RMSE）。我们将方法 DPA3-Scratch（从头训练）、DPA3-LP（线性探测）、DPA3-FT（标准微调）和 DPA3-MFT（多任务微调）与 ChemBERTa [15], Graphormer [80], EGNN [61], ET [70], MoLFormer [58],RT [9], Geoformer [72], Chemprop [27], MACE [36], ModernBERT [73], 和 GotenNet [4]进行比较。最佳和次佳结果分别以**粗体**和下划线标示。黑色文本表示最佳分布内（ID）性能，而蓝色文本表示最佳分布外（OOD）性能。所有 DPA3 结果均为三次独立训练运行的平均值。

模型	#参数	拆分	HOMO	LUMO	GAP	ZPVE	$\langle R^2 \rangle$	α	μ	C_v
ChemBERTa	83M	ID	0.0070	0.0093	0.0104	0.00390	51.7	1.254	0.713	0.489
		OOD	0.0244	0.0256	0.0309	0.02253	302.6	6.328	2.723	2.765
Graphormer	47.1M	ID	0.0040	0.0042	0.0055	0.00024	33.4	0.431	0.626	0.180
		OOD	0.0236	0.0256	0.0275	0.01990	314.1	7.356	2.232	3.667
EGNN	0.2M	ID	0.0048	0.0052	0.0069	0.00103	19.6	0.566	0.481	0.267
		OOD	0.0212	0.0236	0.0312	0.00583	181.3	5.659	2.446	2.079
ET	6.8M	ID	0.0027	0.0031	0.0043	0.00057	28.2	0.490	0.368	0.160
		OOD	0.0220	0.0236	0.0271	0.01710	298.0	6.568	2.257	3.405
MoLFormer	48M	ID	0.0050	0.0052	0.0064	0.00106	40.1	1.047	0.602	0.785
		OOD	0.0236	0.0256	0.0275	0.01990	314.1	7.356	2.232	3.667
RT	27M	ID	0.0090	0.0102	0.0133	0.00289	68.2	2.264	1.104	0.654
		OOD	539.74	584.88	0.0339	27.6906	25458	69968	2.719	11435
Geoformer	51M	ID	0.0027	0.0028	0.0046	0.00030	10.9	0.326	0.847	0.124
		OOD	0.0157	0.0186	0.0240	0.00557	63.8	4.201	2.544	1.354
Chemprop	0.2M	ID	0.0041	0.0046	0.0058	0.00188	35.8	0.866	0.545	0.340
		OOD	0.0189	0.0179	0.0269	0.01277	233.7	4.850	2.304	2.118
MACE	3.9M	ID	0.0150	0.0135	0.0181	0.00180	9.8	0.322	0.430	0.134
		OOD	0.0339	0.0247	0.0409	0.00210	68.3	1.543	2.228	0.229
ModernBERT	111M	ID	0.0064	0.0076	0.0095	0.00073	40.1	0.870	0.698	0.407
		OOD	0.0216	0.0232	0.0324	0.00170	228.7	2.489	2.657	0.611
GotenNet	9.2M	ID	0.0052	0.0039	0.0088	0.00043	11.8	0.553	0.319	0.197
		OOD	0.0126	0.0118	0.0229	0.00053	16.7	1.825	2.173	0.302
DPA3-Scratch	3.3M	ID	0.0030 (0.0000)	0.0029 (0.0000)	0.0045 (0.0000)	0.00013 (0.00000)	8.5 (0.3)	0.365 (0.006)	0.294 (0.003)	0.116 (0.001)
		OOD	0.0155 (0.0004)	0.0130 (0.0002)	0.0244 (0.0003)	0.00029 (0.00000)	21.4 (0.8)	1.907 (0.019)	2.087 (0.003)	0.200 (0.006)
DPA3-LP	3.3M	ID	0.0046 (0.0000)	0.0054 (0.0000)	0.0077 (0.0000)	0.00016 (0.00000)	8.8 (0.0)	0.458 (0.000)	0.309 (0.000)	0.149 (0.000)
		OOD	**0.0125** (0.0001)	0.0128 (0.0002)	0.0193 (0.0001)	0.00029 (0.00000)	14.2 (0.0)	1.641 (0.010)	**1.747** (0.011)	0.281 (0.002)
DPA3-FT	3.3M	ID	0.0022 (0.0000)	**0.0023** (0.0000)	**0.0034** (0.0000)	**0.00010** (0.00000)	**8.1** (0.0)	0.297 (0.001)	**0.276** (0.001)	**0.097** (0.001)
		OOD	0.0156 (0.0001)	0.0135 (0.0003)	0.0201 (0.0002)	0.00026 (0.00000)	18.7 (2.1)	1.639 (0.022)	2.267 (0.018)	0.188 (0.002)
DPA3-MFT	3.3M	ID	**0.0021** (0.0000)	0.0024 (0.0000)	**0.0034** (0.0000)	**0.00010** (0.00000)	8.3 (0.0)	**0.296** (0.001)	0.278 (0.001)	**0.097** (0.001)
		OOD	0.0134 (0.0002)	**0.0107** (0.0002)	**0.0171** (0.0002)	**0.00025** (0.00000)	**13.7** (0.1)	**1.536** (0.005)	2.087 (0.008)	**0.181** (0.001)

2.6. MFT 优于最先进水平的材料性质预测器

为评估所提出的 MFT 策略在无机材料性质预测任务中的稳健性，我们进行了全面的基准测试，其中包括一个分布内随机划分，以及根据结构和性质标准构建的五种分布外划分 [48]。这些划分策略应用于 MatBench 的三个数据集 [19]：折射率（介电常数）、剪切模量（弹性）和钙钛矿形成能（钙钛矿）。三个数据集的摘要见表7。分布外划分策略利用轨道场矩阵（OFM） [53] 表示来编码电子结构和轨道相互作用，从而得到可进行物理解释的特征空间，以供后续聚类和密度估计使用 [31]。具体而言，五种分布外划分方法包括留一簇法（LOCO），该方法在 OFM 特征空间中通过 k-均值聚类将数据集划分为 50 个折。此外，四种稀疏目标策略（SparseXsingle、SparseXcluster、SparseYsingle 和 SparseYcluster）使用高斯核密度估计，识别特征空间（X）或性质空间（Y）中密度最低的样本。随后，使用 k-均值聚类将这些低密度样本进一步分为 50 个簇，并通过以下方式构建分布外测试集：从每个簇中选择一个代表点（single），或选择代表点及其 10 个最近邻（cluster），由此得到 50 个分布外折。对于每种划分方法，均以留一法生成 50 个分布外测试集：每次留出一个折作为分布外测试集，其余数据用于训练和验证。五种分布外划分策略的更多详情见补充材料第C.2节。

如表 4所示，DPA3-MFT 在评估的 18 项任务中有 13 项取得最佳表现，并在另外 4 项任务中排名第二。几乎在所有情况下，DPA3-MFT 的表现都优于标准微调基线（DPA3-FT），

表 4：MatBench 的介电常数、弹性和钙钛矿数据集上不同图神经网络模型的平均绝对误差（MAE）。评估包括对参考文献中介绍的五种不同分布外（OOD）设置进行 50 折交叉验证。 [48]（分布外划分策略的详细说明见第 C.2 节），以及用于分布内（ID）比较的 5 折随机划分。我们将 DPA3-FT（标准微调）和 DPA3-MFT（多任务微调）与 CGCNN [77], MEGNet [12], SchNet [62, 63], DimeNet++ [24]、ALIGNN [16], DeeperGATGNN [49], coGN [59],和 coNGN [59]进行基准比较。最佳和次佳结果分别以 **粗体** 和下划线标示。

分布外划分	数据集	CGCNN	MEGNet	SchNet	DimeNet++	ALIGNN	DeeperGATGNN	coGN	coNGN	DPA3-FT	DPA3-MFT
LOCO	介电常数	0.5144	2.6830	2.7074	2.7720	0.8592	0.5911	0.4984	0.4983	0.5305	**0.4690**
	弹性	**0.0585**	1.4468	1.4065	1.4242	0.0974	0.1173	0.1017	0.1019	0.0806	0.0774
	钙钛矿	0.0651	1.4654	1.4644	1.4666	0.0386	0.0365	0.0631	0.0632	0.0350	**0.0341**
SparseXcluster	介电常数	0.6006	2.8044	2.8368	2.9378	0.6953	1.4056	0.5241	0.5242	0.6069	**0.4806**
	弹性	**0.0499**	1.4113	1.3455	1.3562	0.0834	0.1109	0.1416	0.1417	0.0754	0.0689
	钙钛矿	0.0839	1.4485	1.4509	1.4567	0.0468	0.0464	0.1306	0.1306	**0.0384**	0.0403
SparseXsingle	介电常数	0.9888	3.8384	3.9767	3.8629	1.5115	1.5755	0.5519	**0.5518**	1.5335	1.2463
	弹性	0.0895	1.3099	1.2363	1.3214	0.0853	0.1140	0.2919	0.2918	0.0783	**0.0782**
	钙钛矿	0.0689	1.5005	1.5133	1.5248	0.0457	0.0373	0.2199	0.2200	0.0370	**0.0358**
SparseYcluster	介电常数	0.5254	2.5918	2.5988	2.5947	0.4359	0.3959	0.4673	0.4674	0.3354	**0.2688**
	弹性	0.0752	1.5659	1.5592	1.5454	0.0631	0.0858	0.0823	0.0823	0.0529	**0.0527**
	钙钛矿	0.0740	1.4685	1.4736	1.4732	0.0341	0.0333	0.0925	0.0926	0.0322	**0.0300**
SparseYsingle	介电常数	0.4777	2.5706	2.5566	2.5866	0.2513	0.2733	0.3589	0.3589	0.2008	**0.1790**
	弹性	0.0840	1.4491	1.4855	1.4828	0.0450	0.0807	0.1852	0.1853	0.0477	**0.0435**
	钙钛矿	0.0768	1.5048	1.5173	1.5148	0.0243	0.0259	0.1884	0.1886	0.0235	**0.0212**
随机	介电常数	0.5988	0.3391	0.3277	0.3400	0.3449	0.3355	0.3088	0.3142	0.3082	**0.2639**
	弹性	0.0895	0.0871	0.0796	0.0792	0.0715	0.0903	0.0689	0.0670	0.0627	**0.0609**
	钙钛矿	0.0452	0.0352	0.0342	0.0376	0.0288	0.0288	**0.0269**	0.0290	0.0272	0.0270

平均提升 7.8%；唯一的例外是钙钛矿 SparseXcluster 划分。与其他最先进模型相比，DPA3-MFT 始终保持性能优势，尽管某些竞争方法得益于任务特定架构设计。值得注意的是，在所有基于性质离群值的划分策略（SparseYsingle 和 SparseYcluster）中，DPA3-MFT 的分布外性能均居首位。例外出现在基于结构的分布外划分中，包括 LOCO 和 SparseXcluster 下的弹性任务，以及 SparseXsingle 下的介电任务。特别是，DPA3-MFT 在介电 SparseXsingle 划分上的表现相对较差，这是因为如图 6 所示，50个折中的一个折出现了异常高的平均绝对误差，从而不成比例地抬高了平均测试误差。MFT 带来的最显著提升出现在介电数据集上，DPA3-MFT 的表现较 DPA3-FT 提高了 16.0%（相比之下，弹性和钙钛矿任务分别提高了 4.1% 和 3.1%）。这可能是因为钙钛矿和弹性任务分别预测钙钛矿形成能和剪切模量（后者与能量的二阶导数密切相关），这些预测目标与预训练期间使用的能量与力预测目标天然一致。相比之下，介电（折射率）任务考察的是电子结构性质，与力场预训练之间存在更大的任务领域偏移。因此，成功适应需要进行更大幅度的表征更新，这使得保留预训练期间获得的具有化学意义的先验尤为关键；因此，MFT 在这一范围内尤其有效。

2.7. MFT 提升数据效率和少样本分布外泛化

为评估下游数据量对模型性能的影响，我们使用 QM9 带隙（GAP）预测任务和 MatBench 折射率（介电常数）预测任务，研究了 MFT 的数据效率。如图 3(a-b) 所示，我们将 MFT 与两种基线方法进行比较：从头训练和微调。在这些实验中，通过对原始训练数据进行随机下采样，将训练集规模缩减至 GAP 任务仅 114 个数据点、介电常数任务仅 14 个数据点。我们绘制了分布内和分布外测试集的均方根误差（RMSE）或平均绝对误差（MAE）随训练数据量变化的曲线，训练数据量跨度达三个数量级。实现细节见补充材料第 D.4节。

图 3(a) 展示了 QM9 GAP 预测任务的结果。在分布内准确率方面，在所有数据规模下，微调和 MFT 的表现始终优于从头训练。MFT 相较于微调保持着虽小但稳定的优势，不过当训练数据量达到完整数据集的 50% 时，这一差距有所缩小。相比之下，在分布外准确率方面，MFT 在所有训练样本量下均显著优于从头训练和微调基线，展现出强大的少样本泛化能力。值得注意的是，仅使用 1/128的训练数据，MFT 就能达到与使用完整数据集从头训练的模型相当的分布外准确率。

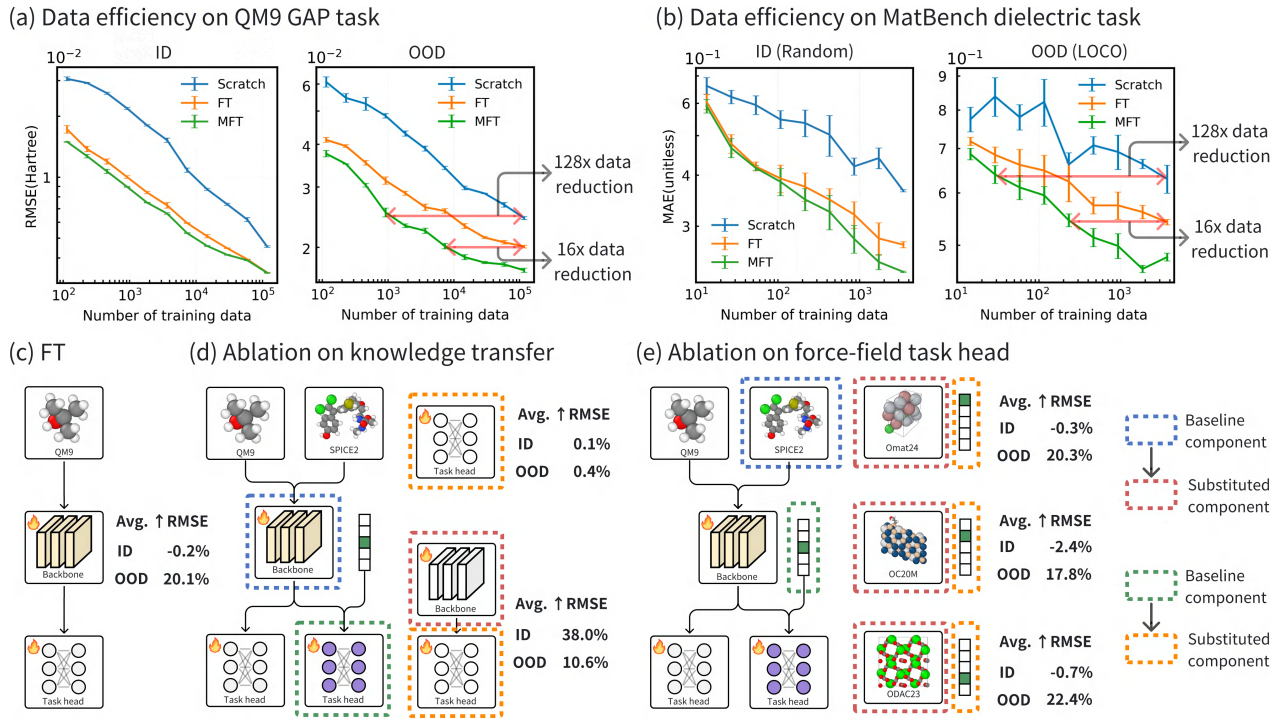


图 3：多任务微调（MFT）的数据效率和消融研究。(a) QM9 带隙（GAP）预测任务的数据效率。(b) MatBench 介电性质预测任务的数据效率。在 (a,b) 中，将 MFT 与从头训练（Scratch）和标准微调（FT）进行比较；展示了分布内（ID）和分布外（OOD）测试均方根误差或平均绝对误差随训练集大小的变化。(c–e) MFT 在 QM9 HOMO、LUMO 和 GAP 预测任务上的消融研究。(c) 微调基线。(d) 对迁移的预训练组件进行消融。(e) 对辅助力场任务选择进行消融。“Avg.↑RMSE”表示相对于基线多任务微调配置，HOMO、LUMO 和 GAP 预测任务的平均 RMSE 增幅。

此外，MFT 仅需 1/16 的训练数据，就能达到使用完整数据集训练的微调模型的分布外性能。

如图 3(b) 所示，MatBench 介电性质预测结果进一步验证了这些发现。在分布内准确率方面，在所有数据规模范围内，MFT 和标准微调均持续优于从头训练。尽管 MFT 的表现优于微调，但在极低数据量范围内，两者之间的差距有所缩小。在分布外性能方面，在所有训练集大小下，MFT 均显著优于微调和从头训练。值得注意的是，MFT 展现出卓越的数据效率，仅使用完整数据集的 1/128 和 1/16，其分布外准确率便可分别媲美使用完整数据集训练的从头训练模型和微调模型。这些趋势与 GAP 任务的观察结果高度一致。

鉴于分子与材料性质预测所需的标注数据通常较为稀缺，且生成成本高昂，例如，准确计算带隙往往需要耗费大量计算资源的杂化交换-相关泛函，因此，这种在广泛数据规模范围内展现出的稳健性能凸显了 MFT 在现实分子与材料设计应用中的实用价值。

2.8. 消融研究

我们针对 QM9 的最高占据分子轨道、LUMO 和 GAP 预测任务开展了消融实验，以区分知识迁移的贡献，并评估辅助任务选择的影响。基线多任务微调配置采用 SPICE2 作为辅助力场任务，并迁移预训练骨干网络和力场任务头。我们还考察了不迁移预训练任务头的变体，以及既不迁移预训练骨干网络也不迁移预训练任务头的变体。此外，我们还研究了多任务微调性能对辅助力场任务选择的敏感性，具体做法是将 SPICE2 替换为来自其他研究领域的数据集，包括 OMat24（无机材料）、OC20M（多相催化）和 ODAC23（金属有机框架，MOFs）。

如图所示。3(d)，迁移力场任务

头并非绝对必要：将任务头替换为随机初始化的任务头后，均方根误差仅略微增加，分布内增加 0.1%，分布外增加 0.4%。相比之下，若既不迁移预训练骨干网络，也不迁移任务头，性能就会显著下降；相较于基线，分布内均方根误差增加 38.0%，分布外增加 10.6%。值得注意的是，即使没有显式的知识迁移，联合多任务训练所达到的分布外性能仍优于微调，如图 3(c) 所示。这凸显了多任务框架在使下游性质预测与力场训练任务所编码的化学与构型先验相一致方面的内在优势。

我们进一步研究了辅助力场任务与下游性质预测任务之间的领域对齐所产生的影响。当辅助领域与目标任务存在显著差异时，会观察到明显的负迁移效应。结果汇总于图 3(e)。尽管分布内性能相对稳定，但在全部三个任务中，分布外性能均显著下降；分别使用 OMat24、OC20M 和 ODAC23 作为辅助任务时，平均均方根误差分别增加了 20.3%、17.8% 和 22.4%。值得注意的是，使用 OMat24 和 ODAC23 获得的分布外性能甚至低于微调。相比之下，OC20M 的分布外性能略优于微调，这很可能归因于催化吸附物与 QM9 中的小型有机分子之间存在部分结构重叠，这表明有限的结构相似性仍能促进分布外泛化。这些观察结果与文献中的既有发现一致：知识迁移并非总是有益的，在不匹配的领域间不加区分地迁移可能会导致负迁移 [14], , 即不相关的辅助数据会降低下游模型性能。

3. 讨论

预训练原子级模型为分子与材料性质预测提供了颇具前景的统一基础；然而，其实际影响在很大程度上取决于能否可靠地泛化到训练域之外。在这项工作中，我们发现并系统分析了预训练原子模型微调过程中一种关键但尚未得到充分探索的失效模式——表征坍塌。这一现象会导致预训练过程中获得的化学与几何意义先验退化，进而显著削弱分布外（OOD）泛化性能。

为应对这一挑战，我们提出多任务微调（MFT）。该方法通过联合优化下游性质预测目标与从预训练中继承的辅助力场任务，在保留通用化学与构型先验的同时，实现任务特定适配，从而完成训练。在广泛的分布外设置下，MFT 始终能够缓解表征坍塌，并在不牺牲分布内准确性的情况下，实现最先进的分布外性能。此外，MFT 出色的数据效率凸显了其对现实世界分子与材料发现的实际意义；在这些场景中，标注数据获取成本高昂，而超出已知范围的外推是常态，而非例外。

我们的分析从机制层面揭示了标准微调为何失效，以及 MFT 为何成功。我们表明，分布外性能与表征相对于预训练模型的漂移程度密切相关；有效的适应需要在可塑性（拟合下游目标所需）与稳定性（保留预训练知识所需）之间谨慎权衡。线性探测能够保证稳定性，但表达能力不足；而标准微调虽具备更强的表达能力，却要以表征坍塌为代价。相比之下，MFT 处于更有利的平衡点，能够在具有物理意义的辅助目标引导下，实现可控的表征适应。

我们预计，这些见解将为分子和材料发现领域下一代原子模型的设计提供指导；在这些领域，可靠地外推至此前尚未探索的化学领域至关重要。通过实现无需遗忘的微调，所提出的框架推动原子级基础模型朝着履行其使命迈进，使其成为数据高效、具有预测能力的科学发现可靠引擎。

4. 方法

本节介绍本文采用的预训练流程和微调策略，包括标准微调和多任务微调。

4.1. 预训练

在本研究中，大型原子级模型 DPA-3.1-3M [83] 在 OpenLAM-v1 数据集 [1], 上进行预训练。该数据集 [1], 由使用密度泛函理论（密度泛函理论）计算能量和原子力并据此标注的数据集组成 [28, 33]。OpenLAM-v1 数据集涵盖了广泛的体系，包括大型材料数据集（如 OMat）24 [6],、分子数据集（如 SPICE）2 [20], , 以及社区贡献的其他公开数据集。OpenLAM-v1 中包含的数据集完整列表见参考文献 [52]。我们将该数据集中的数据集记为 $\{D_m\}_{m=1}^{M}$，其中 OpenLAM-v1 数据集总共包含 $M = 31$ 个数据集。每个数据集 $D_m = \{(x_{mk},\ E_{mk},\ F_{mk})\}_{k=1}^{K_m}$ 包含 K_m 个构型，每个数据点由三维原子结构

组成 x_{mk}，其通过密度泛函理论计算得到的基态势能 E_{mk}，以及相应的原子力 F_{mk}。形式上，每个结构表示为 $x_{mk} = (R_{mk},\ Z_{mk})$，其中$R_{mk} \in \mathbb{R}^{N_{mk}\times 3}$ 表示 N_{mk} 个原子的笛卡尔坐标，$Z_{mk} \in \mathbb{N}^{N_{mk}}$ 表示它们的原子序数。对于具有周期性边界条件的结构，结构表示 x_{mk}中还包含模拟晶胞张量，并且标签中包含相应的、通过密度泛函理论计算得到的维里张量。为简化表述，本文余下部分不再明确讨论晶胞张量或维里张量。

用于学习原子表示的图神经网络骨干网络记为 f_θ，其可训练参数为 θ；特定于任务的拟合网络记为 g_ϕ，其可训练参数为 ϕ。预测的势能为

$$\hat{E}_{mk} = g_\phi\big(f_\theta(x_{mk}),\, c_m\big), \tag{1}$$

其中，c_m 表示数据集 D_m 的独热编码。原子力通过对预测势能关于原子坐标求负梯度得到：

$$\hat{F}_{mk} = -\nabla_{R_{mk}} \hat{E}_{mk}. \tag{2}$$

经验预训练目标定义为

$$\mathcal{L}_{\mathrm{pre}}(\theta,\phi) = \sum_{m\in\mathcal{B}_c}\sum_{k\in\mathcal{B}_m} \lambda_E \mathcal{L}_E(\hat{E}_{mk}, E_{mk}) + \lambda_F \mathcal{L}_F(\hat{F}_{mk}, F_{mk}), \tag{3}$$

其中，$\mathcal{B}_c$ 表示从数据集集合中随机抽取的数据集小批量，$\mathcal{B}_m$ 表示从数据集 D_m中随机抽取的构型小批量。这里，$\mathcal{L}_E$和 $\mathcal{L}_F$是误差度量，用于量化预测值与密度泛函理论参考值之间的差异，其中预测值和参考值分别为能量和力；而 λ_E 和 λ_F 是加权系数，用于平衡两个损失项的贡献。因此，预训练任务定义为最小化式（3）中的损失：

$$\text{(Pretraining)} \qquad \theta_0, \phi_0 = \arg\min_{\theta,\phi} \mathcal{L}_{\mathrm{pre}}(\theta,\phi). \tag{4}$$

骨干网络和任务头的预训练参数分别记为 θ_0 和 ϕ_0。

4.2. 微调

对于下游属性预测任务，我们考虑一个标注数据集 $\mathcal{D} = \{(x_k,\ y_k)\}_{k=1}^{K}$，其中 $y_k \in \mathbb{R}^d$ 表示我们关注的分子或材料性质（例如带隙或介电常数）。目标性质的维度记为 $d \geq 1$，其中 $d > 1$ 允许同时预测多个性质。在本研究中，

我们将研究限定在 $d = 1$ 的情形，即每次只预测单一性质。我们将任务专用预测头记为 h_ψ。由此得到的性质预测模型定义为

$$\hat{y}_k = h_\psi(f_\theta(x_k))\,. \tag{5}$$

性质预测的监督损失定义为

$$\mathcal{L}_{\mathrm{task}}(\theta,\psi) = \sum_{k\in\mathcal{B}} \mathcal{L}_P(\hat{y}_k, y_k), \tag{6}$$

其中，$\mathcal{B}$ 表示从数据集 $\mathcal{D}$中采样得到的一个小批量，而 $\mathcal{L}_P$ 是用于量化预测属性值与参考属性值之间差异的误差指标。下文将给出图1中示意展示的四种训练策略的数学公式。

从头训练。作为基线，所有模型参数 $\{\theta,\psi\}$均采用随机初始化，并通过最小化式（6）中的损失来训练性质预测模型：

$$\text{(Scratch)} \qquad \min_{\theta,\psi}\ \mathcal{L}_{\mathrm{task}}(\theta,\psi) \tag{7}$$

这种训练策略作为对照实验，用于量化预训练带来的实际性能提升。从头训练的模型性能通常受下游数据集 $\mathcal{D}$规模的限制，因为在缺乏有信息量的先验表示时，图神经网络在小规模标注数据上进行优化容易过拟合。在实践中，会根据验证性能采用早停来缓解过拟合。

线性探测（*LP*）。在 LP 中，预训练骨干网络的表征被视为冻结的特征提取模块，且仅优化任务专用预测头的参数：

$$\text{(LP)} \qquad \begin{aligned} &\min_{\psi}\ \mathcal{L}_{\mathrm{task}}(\theta_0,\psi),\\ &\text{Initialization:}\quad \psi \leftarrow \mathrm{rand.}, \end{aligned} \tag{8}$$

其中，$\psi \leftarrow \mathrm{rand}$。表示任务头参数$\psi$ 在训练前随机初始化。LP 在计算上效率较高，并保留了预训练期间学到的鲁棒表征。然而，冻结骨干网络也会限制模型的表达能力，以及其调整表征以适应下游任务的能力；当这些下游任务的预测目标与预训练任务相比存在显著差异时，这种限制尤为明显。

微调（*FT*）。在标准微调中，骨干网络参数使用预训练权重 θ_0进行初始化，而任务特定的头部参数则进行随机初始化。随后，使用下游任务目标对所有参数 $\{\theta,\psi\}$ 进行端到端更新：

$$\text{(FT)} \qquad \begin{aligned} &\min_{\theta,\psi}\ \mathcal{L}_{\mathrm{task}}(\theta,\psi),\\ &\text{Initialization:}\quad \theta \leftarrow \theta_0,\ \psi \leftarrow \mathrm{rand}. \end{aligned} \tag{9}$$

尽管这种优化方式能够使表征适应下游任务，并更好地拟合目标标签 y_i，但正如第2.2节所述，它往往会导致表征坍塌。

多任务微调（*MFT*）。MFT 在微调过程中将原始预训练任务（能量与力预测）重新引入，作为辅助目标。当没有关于下游数据集的先验知识时，可以将所有预训练数据集纳入辅助目标。或者，对于分子性质预测任务，可以选择 SPICE2 等分子预训练数据集；对于材料性质预测任务，则可以使用 OMat24 等材料预训练数据集。我们将选定的预训练数据集记为 D_{m^*}，并将正则化损失定义为

$$\begin{aligned}&\mathcal{L}_{\mathrm{reg}}(\theta,\phi)\\&=\sum_{k\in\mathcal{B}_{m^*}}\lambda_E\mathcal{L}_E(\hat{E}_{m^*k},E_{m^*k})+\lambda_F\mathcal{L}_F(\hat{F}_{m^*k},F_{m^*k}),\end{aligned}\tag{10}$$

其中，$\mathcal{B}_{m^*}$ 表示从数据集 D_{m^*}中抽取的一批构型。MFT 总目标定义为

$$\mathcal{L}_{\mathrm{MFT}}(\theta,\psi,\phi)=\mathcal{L}_{\mathrm{task}}(\theta,\psi)+\lambda_{\mathrm{reg}}\mathcal{L}_{\mathrm{reg}}(\theta,\phi),\tag{11}$$

其中，λ_{reg} 控制正则化损失相对于下游任务损失的权重。

优化问题表示为

$$\begin{aligned}&\text{(MFT)}\quad\min_{\theta,\psi,\phi}\ \mathcal{L}_{\mathrm{MFT}}(\theta,\psi,\phi),\\&\text{Initialization}:\quad\theta\leftarrow\theta_0,\ \psi\leftarrow\text{rand.},\ \phi\leftarrow\phi_0,\end{aligned}\tag{12}$$

其中，所有参数 $\{\theta,\psi,\phi\}$ 均进行联合优化。骨干网络参数 θ 和力场任务头参数 ϕ 从预训练值初始化，而下游任务头参数 ψ 则随机初始化。如第 2.2 节所示，这种 MFT 方案通过保留预训练期间学到的化学与构型先验，有效缓解了表征坍塌。

致谢

本研究所使用的计算资源由 AIforScienceInstitute 提供。Chengqian Zhang、Duo Zhang、Yuzhi Zhang、Guolin Ke 和 Han Wang 的工作得到了国家重点研发计划（项目编号：2022YFA1004300）的支持。Han Wang 还得到了国家自然科学基金（项目编号：12525113 和 12561160120）的资助。Tiejun Li 感谢国家重点研发计划（项目编号：2021YFA1003301）以及国家自然科学基金（项目编号：12288101）的支持。

数据可用性

DPA3 模型的多任务微调、标准微调和线性探测的训练与推理代码可在 DeePMD-kit 仓库中获取（https://github.com/deepmodeling/deepmd-kit）。用于复现本文所报告结果和图表的脚本可在以下地址获取：https://github.com/Chengqian-Zhang/Multitask-finetuning。

参考文献

[1]openlam-v1数据集，2025。https://www.aissquare.com/datasets/detail?name=OpenLAM-TrainingSet-v1&id=308&pageType=数据集。2,10[2] Armen Aghajanyan、Akshat Shrivastava、Anchit Gupta、Naman Goyal、Luke Zettlemoyer 和 Sonal Gupta。通过减少表征坍塌改进微调。载于国际学习表征会议，2021年。2,17

[3] Evan RAntoniuk、ShehtabZaman、Tal Ben-Nun、Peggy Li、James Diffenderfer、Busra Sahin、Obadiah Hersh Smolenski、Everett Grethel、Tim Hsu、Anna Hiszpan- ski、Kenneth Chiu、Bhavya Kailkhura 和 Brian Van Essen。BOOM：机器学习模型的分布外分子性质预测基准。在《第三十九届神经信息处理系统年会数据集与基准赛道》，2025年。1，2，4，6，17，19

[4] Sarp Aykent 和 Tian Xia。GotenNet：重新思考高效的三维等变图神经网络。在《第十三届国际学习表征会议》，2025。7，22

[5]Fan Bao、Min Zhao、Zhongkai Hao、Peiyao Li、Chongxuan Li 和 Jun Zhu。用于逆向分子设计的等变能量引导 SDE。在《第十一届国际学习表征会议》，2023年。1

[6] 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）无机材料数据集和模型。arXiv 预印本 arXiv:2410.12771，2024年。2，3，10

[7] Ilyes Batatia、Philipp Benner、Yuan Chiang、AlinM Elena、D´avid P Kov´acs、Janosh Riebesell、Xavier R

Advincula、Mark Asta、Matthew Avaylon、William J Baldwin 等。《原子尺度材料化学基础模型》。《化学物理学杂志》，163(18)，2025年。2[8] Anton Bochkarev、Yury Lysogorskiy 和 Ralf Drautz。《超越等变消息传递的半局域相互作用图原子簇展开》。《物理评论X》，14(2):021036，2024年。2[9] Jannis Born 和 Matteo Manica。《回归 Transformer 实现分子语言建模中的序列回归与生成并行处理》。《Nature Machine Intelligence》，5(4):432–444，2023年。7[10] Zhendong Cao、Xiaoshan Luo、Jian Lv 和 Lei Wang。《利用空间群信息生成晶体材料的 Transformer》。《科学通报》，2025。1[11] Lowik Chanussot、Abhishek Das、Siddharth Goyal、Thibaut Lavril、Muhammed Shuaibi、Morgane Riviere、Kevin Tran、Javier Heras-Domingo、Caleb Ho、Weihua Hu 等。《Open Catalyst 2020（OC20）数据集及社区挑战》。《ACS催化》，11(10):6059–6072，2021年。2[12] Chi Chen、Weike Ye、Yunxing Zuo、Chen Zheng 和 Shyue Ping Ong。《作为分子与晶体通用机器学习框架的图网络》。《材料化学》，31:3564–3572，2019。8[13] Xinlei Chen、Haoqi Fan、Ross Girshick 和 Kaiming He。《利用动量对比学习改进基线方法》，2020年。2, 17[14] Xinyang Chen、Sinan Wang、Bo Fu、Mingsheng Long 和 Jianmin Wang。《灾难性遗忘遭遇负迁移：用于安全迁移学习的批次谱收缩》。《神经信息处理系统进展》，32，2019。10[15] Seyone Chithrananda、Gabriel Grand 和 Bharath Ramsundar。《ChemBERTa：用于分子性质预测的大规模自监督预训练》，2020年。7[16] Kamal Choudhary 和 Brian DeCost。《用于提升材料性质预测效果的原子尺度线图神经网络》。《npj Computational Materials》，7(1):185，2021年。8[17] Stefano Curtarolo、Gus LW Hart、Marco Buongiorno Nardelli、Natalio Mingo、Stefano Sanvito 和 Ohad Levy。《通往计算材料设计的高通量之路》。《自然·材料》，12(3):191–201，2013。1[18] David L. Davies 和 Donald W. Bouldin。《一种簇分离度量》。《IEEE模式分析与机器智能汇刊》，PAMI-1(2):224–227，1979。5, 21[19] Alexander Dunn、Qi Wang、Alex Ganose、Daniel Dopp 和 Anubhav Jain。材料性质基准测试预测方法：MatBench测试集与automatminer参考算法。《npj Computational Materials》，6:138，2020年。7，20[20] Peter Eastman、Benjamin P Pritchard、John D Chodera和Thomas E Markland。肉豆蔻与香料：生物分子机器学习的模型与数据。《化学理论与计算杂志》，20(19):8583–8593，2024年。3，10[21]Utku Evci、Vincent Dumoulin、Hugo Larochelle和Michael C Mozer。Head2Toe：利用中间表征实现更好的迁移学习。载于Kamalika Chaudhuri、Stefanie Jegelka、Le Song、Csaba Szepesvari、Gang Niu和Sivan Sabato（编），《第39届国际机器学习会议论文集》，《机器学习研究论文集》第162卷，第6009–6033页。PMLR，2022年7月17日至23日。17[22] Shikun Feng、Yuyan Ni、Yanyan Lan、Zhi-Ming Ma和Wei-Ying Ma。用于三维分子预训练的分数阶去噪。载于《第40届国际机器学习会议论文集》，《机器学习研究论文集》第202卷，第9938–9961页。PMLR，2023年7月23日至29日。17[23] R´emi Flamary、Nicolas Courty、Alexandre Gramfort、Mokhtar Z. Alaya、Aur´elie Boisbunon、Stanislas Cham- bon、Laetitia Chapel、Adrien Corenflos、Kilian Fa- tras、Nemo Fournier、L´eo Gautheron、Nathalie T.H. Gayraud、Hicham Janati、Alain Rakotomamonjy、Iev- gen Redko、Antoine Rolet、Antony Schutz、Vivien Seguy、Danica J. Sutherland、Romain Tavenard、Alexander Tong和Titouan Vayer。POT：Python最优传输。《机器学习研究杂志》，22(78):1–8，2021年。21[24] Johannes Gasteiger、Shankari Giri、Johannes T. Mar- graf和Stephan G¨unnemann。面向非平衡分子的快速且具不确定性感知的方向性消息传递。载于分子机器学习研讨会，NeurIPS，2020年。8[25] Weifeng Ge和Yizhou Yu。从富者那里借取宝藏：通过选择性联合微调实现深度迁移学习。载于IEEE计算机视觉与模式识别会议（CVPR），2017年7月。17[26] Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Peiyi Wang、Qihao Zhu、Runxin Xu、Ruoyu Zhang、Shirong Ma、Xiao Bi等。DeepSeek-R1通过强化学习激励大语言模型进行推理。《Nature》，645(8081):633–638，2025。2，17[27] Esther Heid、Kevin P. Greenman、Yunsie Chung、Shih- Cheng Li、David E. Graff、Florence H. Vermeire，

Haoyang Wu、William H. Green和Charles J. McGill。Chemprop：用于化学性质预测的机器学习软件包。《化学信息与建模杂志》，64(1):9–17，2023年。7[28] Pierre Hohenberg和Walter Kohn。非均匀电子气。《物理评论》，136(3B):B864，1964年。10[29] Jeremy Howard和Sebastian Ruder。用于文本分类的通用语言模型微调，2018年。17[30] Shuyi Jia、Akaash R. Parthasarathy、Rui Feng、Guojing Cong、Chao Zhang和Victor Fung。用于材料性质预测的图神经网络基于导数的预训练。《Digital Discovery》，3:586–593，2024年。17[31] Mohammadreza Karamad、Rishikesh Magar、Yuting Shi、Samira Siahrostami、Ian D Gates和Amir Barati Farimani。用于材料性质预测的轨道图卷积神经网络。《物理评论：材料》，4(9):093801，2020年。7[32] Jaesun Kim、Jisu Kim、Jaehoon Kim、Jiho Lee、Yutack Park、Youngho Kang和Seungwu Han。用于高保真机器学习原子间势的数据高效多保真度训练。《美国化学会志》，147(1):1042–1054，2024年。2[33]Walter Kohn和Lu Jeu Sham。包含交换与关联效应的自洽方程。《物理评论》，140(4A):A1133，1965年。10[34] Lingyu Kong、Nima Shoghi、Guoxiang Hu、Pan Li和Victor Fung。Mattertune：用于微调原子基础模型、加速材料模拟与发现的一体化用户友好平台。《Digital Discovery》，4(4):2253–2262，2025。2，17[35] Simon Kornblith、Jonathon Shlens和Quoc V. Le。更好的ImageNet模型是否具有更好的迁移效果？载于《IEEE/CVF计算机视觉与模式识别会议论文集》（CVPR），2019年6月。17[36] D´avid P´eter Kov´acs、Ilyes Batatia、Eszter S´ara Arany和G´abor Cs´anyi。MACE力场架构评估：从药物化学到材料科学。《化学物理学杂志》，159(4):044118，2023年7月。7[37] Ananya Kumar、Aditi Raghunathan、Robbie Jones、Tengyu Ma和Percy Liang。微调可能扭曲预训练特征，并在分布外表现不佳。载于《第十届国际学习表征会议》，2022年。2，17[38] Yoonho Lee、Annie S Chen、Fahim Tajwar、Ananya Kumar、Huaxiu Yao、Percy Liang和Chelsea Finn。精细微调改善了对分布偏移的适应。载于《第十一届国际学习表征会议》，2023年。17

[39] Daniel S Levine, Muhammed Shuaibi, Evan Wal- ter Clark Spotte-Smith, Michael G Taylor, Muham- mad R Hasyim, Kyle Michel, Ilyes Batatia, G´abor Cs´anyi, Misko Dzamba, Peter Eastman 等。Open Molecules 2025（OMol25）数据集、评估与模型。arXiv 预印本 arXiv:2505.08762，2025年。2[40] Hao Li、Pratik Chaudhari、Hao Yang、Michael Lam、Avinash Ravichandran、Rahul Bhotika 和 Stefano Soatto。重新思考微调的超参数。载于国际学习表征会议，2020年。17[41] Kangming Li、Andre Niyongabo Rubungo、Xiangyun Lei、Daniel Persaud、Kamal Choudhary、Brian DeCost、Adji Bousso Dieng 和 Jason Hattrick-Simpers。探究材料机器学习中的分布外泛化。Communications Materials，6(1):9，2025年。2，17[42] Xingjian Li、Haoyi Xiong、Hanchao Wang、Yuxuan Rao、Liping Liu 和 Jun Huan。Delta：利用注意力特征图实现卷积网络深度学习迁移。载于国际学习表征会议，2019年。17[43] Yi-Lun Liao、Tess Smidt、Muhammed Shuaibi 和 Abhishek Das。将去噪泛化到非平衡结构可提升等变力场。Transactions on Machine Learning Research，2024年。17[44] Shikun Liu、Deyu Zou、Nima Shoghi、Victor Fung、Kai Liu 和 Pan Li。RoFT-Mol：分子图基础模型稳健微调基准测试。arXiv 预印本 arXiv:2509.00614，2025年。17[45] Xixian Liu、Rui Jiao、Zhiyuan Liu、Yurou Liu、Yang Liu、Ziheng Lu、Wenbing Huang、Yang Zhang 和 Yixin Cao。学习三维各向异性噪声分布可改进分子力场。载于第三十九届神经信息处理系统年会，2025年。17[46] Ziquan Liu、Yi Xu、Yuanhong Xu、Qi Qian、Hao Li、Xi- angyang Ji、Antoni Chan 和 Rong Jin。通过更好地利用预训练数据改进微调。神经信息处理系统进展，35:32568–32581，2022年。17[47] Aviv Netanyahu、Abhishek Gupta、Max Simchowitz、Kaiqing Zhang 和 Pulkit Agrawal。学习外推：一种传导式方法。载于第十一届国际学习表征会议，2023年。17[48] Sadman Sadeed O mee、Nihang Fu、Rongzhi Dong、Ming Hu 和 Jianjun Hu。基于结构的分布外（OOD）材料性质预测：一项基准研究。npj Computational Materials，10:144，2024年。2，7，8，17，20

[49] Sadman Sadeed Omee、Steph-Yves Louis、Nihang Fu、Lai Wei、Sourin Dey、Rongzhi Dong、Qinyang Li 和 Jianjun Hu。用于高性能材料性质预测的可扩展深层图神经网络。Patterns，3(5):100491，2022年。 8[50] Long Ouyang、Jeffrey Wu、Xu Jiang、Diogo Almeida、Carroll Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray 等。训练语言模型遵循人类反馈指令。神经信息处理系统进展，35:27730–27744，2022年。 17[51] F. Pedregosa、G. Varoquaux、A. Gramfort、V. Michel、B. Thirion、O. Grisel、M. Blondel、P. Prettenhofer、R. Weiss、V. Dubourg、J. Vanderplas、A. Passos、D. Cour- napeau、M. Brucher、M. Perrot 和 E. Duchesnay。Scikit-learn：Python 机器学习。机器学习研究杂志，12:2825–2830，2011年。 21[52] Anyang Peng、Chun Cai、Mingyu Guo、Duo Zhang、Chengqian Zhang、Antoine Loew、Linfeng Zhang 和 Han Wang。LAMBench：大型原子模型基准。arXiv 预印本 arXiv:2504.19578，2025年。 10[53]Tien Lam Pham、Hiori Kino、Kiyoyuki Terakura、Takashi Miyake、Koji Tsuda、Ichigaku Takigawa 和 Hieu Chi Dam。机器学习揭示材料中的轨道相互作用。先进材料科学与技术，18(1):756，2017年。 7，20[54] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。通过自然语言监督学习可迁移的视觉模型。载于 Marina Meila 和 Tong Zhang 编，《第38届国际机器学习会议论文集》，第139卷《机器学习研究论文集》，第8748–8763页。PMLR，2021年7月18日至24日。2，17[55] Raghunathan Ramakrishnan、Pavlo O Dral、Matthias Rupp 和 O Anatole von Lilienfeld。134,000个分子的量子化学结构与性质。Scientific Data，1，2014年。 19[56]Benjamin Rhodes、S ander Vandenhaute、Vaidotas ˇSimkus、James Gin、Jonathan Godwin、Tim Duignan 和 Mark Neumann。Orb-v3：大规模原子模拟。arXiv 预印本 arXiv:2504.06231，2025年。 2[57] Youngmin Ro 和 Jin Young Choi。AutoLR：深度网络微调中的逐层剪枝与学习率自动调优。AAAI人工智能会议论文集，35(3):2486–2494，2021年5月。 17[58] Jerret Ross、Brian Belgodere、Vijil Chenthamarak- shan、Inkit Padhi、Youssef Mroueh 和 Payel Das。大规模化学语言表征能够捕捉分子结构和性质。Nature Machine Intelligence, 4(12):1256–1264，2022。 7[59] Robin Ruff、Patrick Reiser、Jan St¨uhmer 和 Pascal Friederich。面向晶体结构的连通性优化嵌套线图网络。Digital Discovery, 3(3):594–601, 2024。 8[60] Benjamin Sanchez-Lengeling 和 Al´an Aspuru-Guzik。利用机器学习进行逆向分子设计：用于物质工程的生成模型。Science, 361(6400):360–365, 2018。 1[61] Victor Garcia Satorras、Emiel Hoogeboom 和 Max Welling。E(n)等变图神经网络。载于《机器学习研究论文集》第40届国际机器学习会议论文集，页码 9323–9332。PMLR，2021。 7[62] Kristof Sch¨utt、Pieter-Jan Kindermans、Huziel Enoc Sauceda Felix、Stefan Chmiela、Alexandre Tkatchenko 和 Klaus-Robert M¨uller。SchNet：一种用于量子相互作用建模的连续滤波卷积神经网络。神经信息处理系统进展，30，2017。 8[63] Kristof T Sch¨utt、Farhad Arbabzadah、Stefan Chmiela、Klaus R M¨uller 和 Alexandre Tkatchenko。深度张量神经网络带来的量子化学洞见。Nature Communications, 8(1):1–8, 2017。 8[64]Nofit Segal、Aviv Netanyahu、Kevin P Green-man、Pulkit Agrawal 和 Rafael G´omez-Bombarelli。已知的未知：材料和分子中的分布外性质预测。npj Computational Materials, 11(1):345, 2025。 17[65] Zhiqiang Shen、Zechun Liu、Jie Qin、Marios Sav-vides 和 Kwang-Ting Cheng。少即是多：重新审视小样本学习的微调策略。AAAI人工智能会议论文集，35(11):9594–9602，2021年5月。 17[66] Hajime Shimakawa、Akiko Kumada 和 Masahiro Sato。利用量子力学辅助机器学习对小数据分子性质进行外推预测。npj Computational Materials, 10(1):11, 2024。 2, 17[67]Nima Shoghi、Adeesh Kolluru、John R. Kitchin、Zachary Ward Ulissi、C. Lawrence Zitnick 和 Bran-don M Wood。从分子到材料：预训练用于原子性质预测的通用模型。载于《国际学习表征会议》第十二届会议，2024。 2, 3, 17, 22[68] Yang Shu、Xingzhuo Guo、Jialong Wu、Ximei Wang、Jianmin Wang 和 Mingsheng Long。CLIPood：使CLIP能够泛化到分布外数据。载于 Andreas Krause、E mma Brunskill、Kyunghyun Cho、Barbara

Engelhardt、Sivan Sabato 和 Jonathan Scarlett，编者，《机器学习研究论文集》第40届国际机器学习会议论文集，第202卷，页码 31716–31731。PMLR，2023年7月23日至29日。2[69] Ava P Soleimany、Alexander Amini、Samuel Goldman、Daniela Rus、Sangeeta N Bhatia 和 Connor W Coley。用于引导分子性质预测与发现的证据深度学习。ACS Central Science, 7(8):1356–1367, 2021。 17[70] Philipp Th¨olke 和 Gianni De Fabritiis。用于神经网络分子势的等变 Transformer。载于《国际学习表征会议》，2022。7，22[71] Laurens van der Maaten 和 Geoffrey Hinton。使用 t-SNE 可视化数据。机器学习研究杂志，9(86):2579–2605，2008。 21[72] Yusong Wang、Shaoning Li、Tong Wang、Bin Shao、Nanning Zheng 和 Tie-Yan Liu。采用原子间位置编码的几何 Transformer。载于《神经信息处理系统进展》，2023年。 7[73] Benjamin Warner、Antoine Chaffin、Benjamin Clavi´e、Orion Weller、Oskar Hallstr¨om、Said Taghadouini、Alexis Gallagher、Raja Biswas、Faisal Ladhak、Tom Aarsen、Nathan Cooper、Griffin Adams、Jeremy Howard 和 Iacopo Poli。更智能、更出色、更快速、更长上下文：一种用于快速、节省内存的长上下文微调与推理的现代双向编码器，2024年。7[74] Brandon M Wood、Misko Dzamba、Xiang Fu、Meng Gao、Muhammed Shuaibi、Luis Barroso-Luque、Ka-reem Abdelmaqsoud、Vahe Gharakhanyan、John R. Kitchin、Daniel S. Levine、Kyle Michel、Anuroop Sri-ram、Taco Cohen、Abhishek Das、Sushree Jagriti Sahoo、Ammar Rizvi、Zachary Ward Ulissi 和 C. Lawrence Zitnick。UMA：一系列原子通用模型。载于《第三十九届神经信息处理系统年会》，2025年。2[75]Mitchell Wortsman、Gabriel Ilharco、Jong Wook Kim、Mike Li、Simon Kornblith、Rebecca Roelofs、Raphael Gontijo Lopes、Hannaneh Hajishirzi、Ali Farhadi、Hongseok Namkoong 和 Ludwig Schmidt。零样本模型的鲁棒微调。载于《IEEE/CVF计算机视觉与模式识别会议（CVPR）》论文集，页码 7959–7971，2022年6月。 17[76]Fang Wu、Shuting Jin、Siyuan Li 和 Stan Z. Li。受指导者启发的机器学习：实现鲁棒的分子性质预测。载于《第三十八届神经信息处理系统年会》，2024年。 17

[77] Tian Xie 和 Jeffrey C. Grossman。用于准确且可解释地预测材料性质的晶体图卷积神经网络。《物理评论快报》，120:145301，2018年4月。8[78] Han Yang、Chenxi Hu、Yichi Zhou、Xixian Liu、Yu Shi、Jielan Li、Guanzhi Li、Zekun Chen、Shuizhou Chen、Claudio Zeni 等。MatterSim：一种适用于不同元素、温度和压力的深度学习原子模型。arXiv 预印本 arXiv:2405.04967，2024年。2, 17[79] Tianzhixi Yin、Gihan Panapitiya、Elizabeth D Coda 和 Emily G Saldanha。评估基于不确定性的主动学习对加速分子性质预测泛化的作用。《化学信息学杂志》，15(1):105，2023年。17[80] Chengxuan Ying、Tianle Cai、Shengjie Luo、Shuxin Zheng、Guolin Ke、Di He、Yanming Shen 和 Tie-Yan Liu。Transformer 在图表示方面的表现真的很差吗？载于《神经信息处理系统进展》，2021年。7[81]Sheheryar Zaidi、Michael Schaarschmidt、James Martens、Hyunjik Kim、Yee Whye Teh、Alvaro Sanchez-Gonzalez、Peter Battaglia、Razvan Pascanu 和 Jonathan Godwin。通过去噪进行预训练，以预测分子性质。载于《国际学习表征会议》，2023年。2, 17[82] Duo Zhang、Xinzijian Liu、Xiangyu Zhang、Chengqian Zhang、Chun Cai、Hangrui Bi、Yiming Du、Xuejian Qin、Anyang Peng、Jiameng Huang 等。DPA-2：作为多任务学习器的大型原子模型。npj Computational Materials，10(1):293，2024年。2[83] Duo Zhang、Anyang Peng、Chun Cai、Wentao Li、Yuanchang Zhou、Jinzhe Zeng、Mingyu Guo、Chengqian Zhang、Bowen Li、Hong Jiang、Tong Zhu、Weile Jia、Linfeng Zhang 和 Han Wang。面向大型原子级模型时代的图神经网络。arXiv 预印本 arXiv:2506.01686，2025。2, 3, 10, 21 [84] Linfeng Zhang、Mohan Chen、Xifan Wu、Han Wang、Weinan E 和 Roberto Car。用于绝缘体介电响应的深度神经网络。《物理评论B》，102(4):041121，2020年。22[85] Gengmo Zhou、Zhifeng Gao、Qiankun Ding、Hang Zheng、Hongteng Xu、Zhewei Wei、Linfeng Zhang 和 Guolin Ke。Uni-Mol：通用三维分子表征学习框架。载于《第十一届国际学习表征会议》，2023年。1, 2, 17

A. 相关工作

用于性质预测的预训练与微调。在预训练阶段，模型从大规模数据集中学习原子表示，随后将其迁移至下游任务。目前，三维原子系统的预训练主要有两种范式。第一种是自监督预训练，通常侧重于针对平衡结构的去噪目标；在这些结构中，逐原子力接近于零 [81, 85, 22]。近期，这一研究方向已扩展至非平衡构型，通过将原子力信息明确纳入学习目标来实现 [43, 45]。第二种范式是监督式预训练，即直接在标注有能量、力以及（某些情况下）应力张量的大规模分子或材料数据集上训练模型 [78, 30, 67, 34]。

在微调阶段，上述大多数研究采用标准微调范式：骨干网络参数由预训练模型初始化，并与额外的任务专用预测头联合优化。JMP [67] 进一步采用逐层学习率衰减（LLRD） [29] 来缓解过拟合。尽管如此，在本研究中，我们仍观察到 JMP-L 出现了明显的表征坍塌，这凸显了下游任务中安全适配所面临的持续挑战。RoFt-Mol [44] 对现有微调策略在广泛的分子性质预测任务上的表现进行了系统评估，并表明，最优微调机制在很大程度上取决于预训练目标和下游任务类型。与这些研究不同，我们专门致力于提升预训练三维原子模型的分布外泛化能力。这一问题在现实的分子与材料设计场景中至关重要，却尚未得到充分探索。

计算机视觉和 *NLP* 中的表征坍塌。 在计算机视觉（CV）和自然语言处理（NLP）领域，基础模型展现出了卓越的零样本能力和迁移能力 [13, 54, 26]。然而，越来越多的研究表明，将这些预训练模型适配到特定下游任务，往往会导致表征坍塌，并伴随分布外泛化能力显著下降 [2, 37, 75]。为缓解这一问题，研究者提出了多种策略，以保留预训练知识并减少微调过程中的过拟合。这些策略包括基于优化的约束，例如采用低于预训练阶段的学习率 [35, 40] ，或采用逐层学习率调度 [57, 65], ；也包括架构约束，即冻结部分模型参数，仅微调选定组件 [38, 21]。另一些方法则通过在预训练权重与微调权重之间进行插值 [75], ，或显式惩罚潜在表示流形中的偏移 [42]，来更新整个模型。此外，沿用预训练阶段的辅助训练目标，已被证明有助于提升微调鲁棒性和泛化能力 [25, 46, 50]。我们提出的 MFT 遵循这一范式，但将辅助目标建立在具有物理意义的力场监督之上，因此特别适用于分子与材料性质预测。

分子与材料设计中的分布外泛化。 分布外泛化是分子与材料发现中的一项核心挑战，但尚未得到充分探索。近期的大规模基准测试表明，与分布内评估相比，最先进的分子性质预测器在分布外设置下的性能会大幅下降 [3, 66], ，无机材料预测中也观察到了同样的趋势 [48, 41]。此前应对这一局限的工作主要集中于利用不确定性量化（UQ）评估预测的可靠性 [69], ，以及借助不确定性量化引导的主动学习生成新的训练数据，从而提升模型的泛化能力 [79]。其他方法则通过半监督学习利用大规模未标注数据 [76] ，或调整双线性传导框架，以增强模型在分子和材料领域的分布外鲁棒性 [64, 47]。尽管取得了这些进展，在微调大型预训练原子模型以进行下游性质预测的背景下，分布外泛化问题仍鲜有研究。尤其是，微调过程中分布外性能的系统性下降、这一现象与表征坍塌之间的联系，以及有原则的缓解策略，仍在很大程度上未得到探索。

B. JMP-L 的表征坍塌

在本节中，我们研究微调预训练 JMP-L 模型时的表征坍塌 [67]。我们采用与 DPA-3.1-3M 完全相同的 QM9 数据集划分，具体如第 2.2 节所述。与从头训练相比，微调使分布内准确率平均提高了 16.9%，但分布外准确率仅提高了 9.7%，如图 4 所示。在八项性质预测任务中，分布外的相对提升始终小于相应的分布内提升，唯一例外是电子空间范围（$\langle R^2 \rangle$），这与图 2（b）中展示的 DPA-3.1-3M 观察结果一致。值得注意的是，对于各向同性极化率（α）和偶极矩（μ），微调模型在分布外情形下分别比从头训练基线低 3.6% 和 13.6%。值得注意的是，JMP-L 的分布外提升是

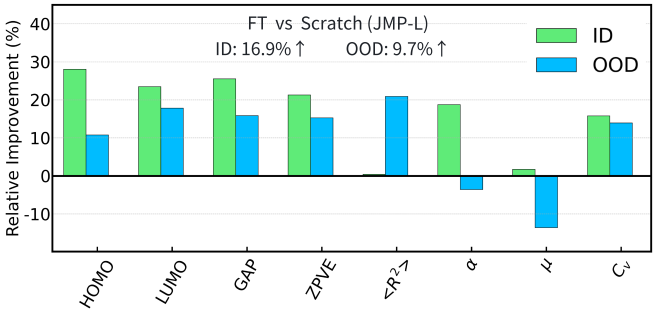


图 4：JMP-L 表征坍塌分析。与从头训练基线相比，微调（FT）带来的相对性能提升。

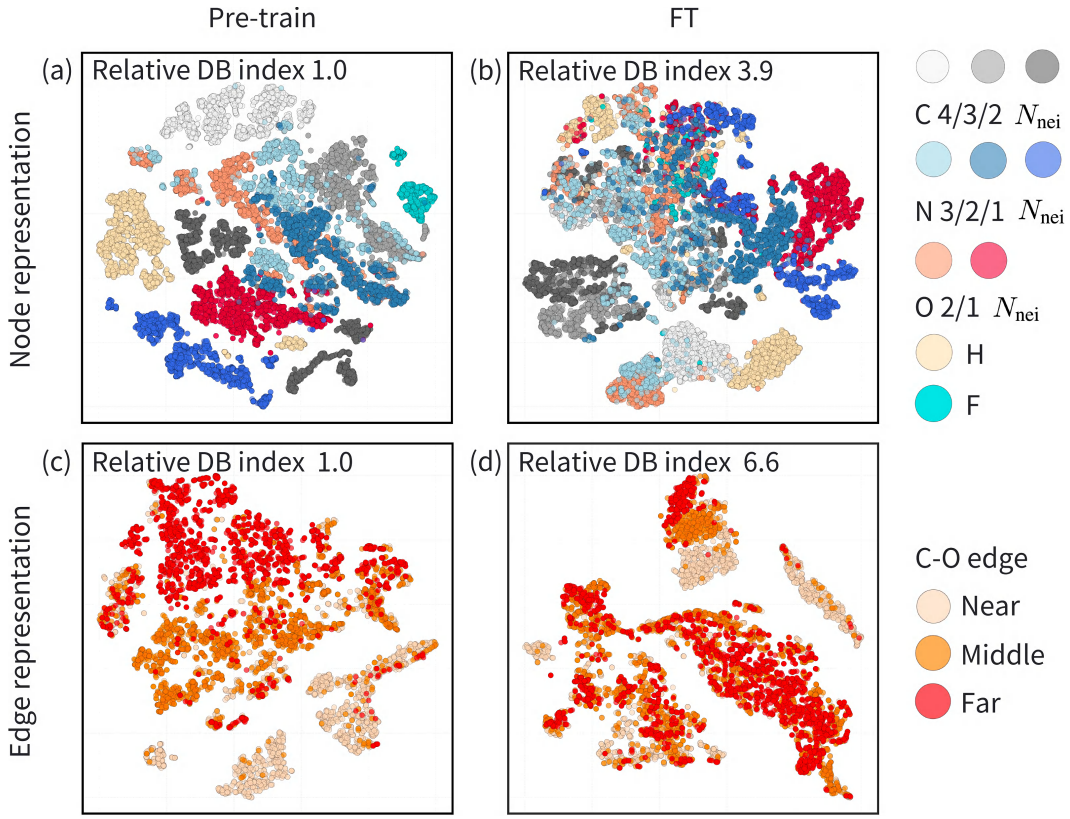


图 5：JMP-L 在 QM9 LUMO 任务中最后一个骨干层的原子（节点）和边表示的 t-SNE 可视化，降维至二维空间。Davies–Bouldin（DB）指数值以预训练模型基线为基准进行归一化。

表 5：使用 1-瓦瑟斯坦距离（$\mathcal{W}_{\text{Pretrain}\to\text{FT}}$, $\mathcal{W}_{\text{Pretrain}\to\text{MFT}}$）量化 JMP-L 预训练与微调表征之间的分布偏移。报告了 QM9 LUMO数据集在训练集、分布内和分布外数据集上的指标。

	训练	ID	OOD
$\mathcal{W}_{\text{Pretrain}\to\text{FT}}$	7.54	7.12	8.58

与 DPA-3.1-3M 相比，这种现象是合理的，因为 JMP-L 同时采用了逐层学习率衰减（LLRD）策略和低于预训练阶段的学习率，以缓解过拟合。然而，我们在 JMP-L 中仍然观察到表征坍塌和分布外性能退化现象，这表明下游任务的安全适配面临挑战。

图 5(a) 展示了预训练 JMP-L 模型最后一个骨干层的原子（节点）表征。颜色沿用图 2 中 DPA-3.1-3M 的配色。如图 5(a) 所示，同种元素且共价环境相同的原子往往形成簇，不过其聚类效果不如图 2 所示的 DPA-3.1-3M，例如氧和氮的表征之间存在一些重叠。然而，微调后，不同元素之间以及同种元素不同局部共价环境之间都出现了显著重叠，如图 5(b) 所示。几乎所有具有不同局部环境的不同元素都混杂在一起，Davies-Bouldin（DB）指数相较于预训练表征增加了 3.9 倍，表明簇可分性显著模糊并丧失。

图 5(c) 展示了碳-氧（C-O）对的边表示，并根据成对欧氏距离将其进一步划分为近、中、远三类。如图 5(c) 所示，我们可以在预训练表征空间中区分这三类边：远边聚集在左上方，近边聚集在右下方，尽管中边与远边之间存在一定混杂。然而，微调导致这种潜在组织结构明显退化，如图 5(d) 所示。中边和远边的表征大幅重叠。构型分辨率的这种损失定量体现为 DB 指数增加了 6.6 倍。

此外，我们使用 1-瓦瑟斯坦距离（$\mathcal{W}$）量化 JMP-L 模型微调前后表示分布之间的漂移。如表 5 所示，与 DPA-3.1-3M 类似，训练分布和分布内测试分布的漂移距离非常接近。然而，分布内测试分布的漂移（7.12）远大于 DPA-3.1-3M（3.75），这意味着微调后表征发生了显著变化。此外，分布外测试分布的漂移大于训练分布和分布内测试分布，导致过拟合。

C. 基准测试详情

表 6：QM9 基准测试属性的信息

性质	样本	单位	描述
α	130k	a_0^3	各向同性极化率
C_v	130k	$\frac{cal}{mol\ K}$	热容
HOMO	130k	Hartree	最高占据分子轨道能量
LUMO	130k	Hartree	最低未占分子轨道能量
GAP	130k	Hartree	HOMO-LUMO 能隙
μ	130k	德拜	偶极矩
$\langle R^2 \rangle$	130k	a_0^2	电子空间范围
ZPVE	130k	Hartree	零点振动能

表7：MatBench 基准测试属性信息

性质	样本	单位	描述
介电常数	4764	无量纲	折射率
弹性	10987	log10(GPa)	剪切模量
钙钛矿	18928	电子伏特/晶胞	形成能

C.1. QM9 分布外数据划分详情

对于小型分子数据集 QM9 [55], ，我们使用与 BOOM 相同的数据划分方式 [3]。训练集、分布内测试集和分布外测试集划分基于数据集的底层性质分布生成。对于表 6 中的八种分子性质，通过拟合核密度估计器（KDE）构建分布外划分

使用高斯核拟合属性值，以确定每个分子的概率密度。概率分数最低的 10% 分子被划分到 QM9 数据集的分布外划分中，从而有效分离出典型分子性质分布尾部的样本。不同于通过固定截断值进行传统划分的方法，这种基于密度的方法能够捕捉常规分布中的低概率样本，而这些分布可能是非单峰的或较为复杂。随后，从剩余分子中随机抽取 10% 的样本，生成分布内测试集，其余数据全部保留用于训练。

C.2. MatBench 的分布外数据划分详情

以下 MatBench 数据划分策略 [19], 最初由参考文献 [48], 提出，用于评估模型在特征空间和性质空间不同分布外区域中的无机材料上的表现。这些方法依赖于轨道场矩阵（OFM） [53], ，该矩阵通过编码电子结构和轨道相互作用来表示材料，从而提供具有物理可解释性的特征空间。

留一簇法（*LOCO*）。此方法使用基于 OFM 特征的 k-means 算法将数据集划分为 50 个簇。模型评估以迭代方式进行，每次将一个簇作为测试集，其余簇用于训练和验证。这样可以确保模型在与训练数据结构迥异的材料组上接受测试。

结构密度最低的单点目标（*SparseXsingle*）。此方法通过使用 t 分布随机邻域嵌入（t-SNE）将 1024 维 OFM 特征投影到二维空间，从而识别结构独特的材料。随后对该二维流形应用高斯核密度估计（KDE），找出密度最低的 500 个样本。接着通过 k-means 将这些样本分为 50 个簇，并从每个簇中提取一个代表性样本，组成包含 50 个样本的测试集。

属性密度最低的单点目标（*SparseYsingle*）。该方法关注属性分布中的离群值，使用高斯核密度估计来估算目标属性值（y 值）的密度。选取属性密度最低的 500 个样本，并通过 k-均值将其划分为 50 个簇。随后从每个簇中选取一个样本，从而得到一个代表稀有或极端属性区间的测试集。

结构密度最低的簇目标（*SparseXcluster*）。该方法将目标组织成簇，而非单个点，从而扩展了 SparseXsingle 的逻辑。确定 50 个具有代表性的稀疏样本后，将每个样本的 N 个最近邻纳入测试集。这些最近邻依据 OFM 特征空间中的欧几里得距离确定，并通过约束确保每个样本不会被分配到多个簇。

属性密度最低的簇目标（*SparseYcluster*）。类似地，该方法基于 SparseYsingle 方法构建。确定来自低密度属性区域的 50 个样本后，测试集将扩展至包含它们在 OFM 特征空间中的 N 个最近邻。这样便可评估模型在材料空间中与稀有属性值相关的局部区域上的表现。

此外，表4 中使用的5折随机划分来源于参考文献 [19]。

C.3. MatBench的详细结果

DPA3-FT和DPA3-MFT在介电数据集、弹性数据集和钙钛矿数据集上的逐折平均绝对误差图分别见图6、图7 和图8。50个折的平均绝对误差分布表明，对于DPA3-FT和DPA3-MFT而言，只有少数具有挑战性的折导致了整体平均绝对误差，这与参考文献 [48]中先前报道的基线模型观察结果一致。尽管这些模型具有相似的分布特征，如表4所示，DPA3-MFT的表现始终优于DPA3-FT和基线模型。

D. 实现细节

D.1. 表征分析细节

图 2(d-f) 展示了预训练模型和微调模型的原子（节点）表征。不同化学物种的原子以不同颜色表示，而同种原子则根据其共价

邻居的数量进一步区分。这里，我们将距离小于两原子共价半径之和的 1.15 倍的两个原子视为彼此的邻居。我们从 OOD 测试集中随机抽取每种原子种类在各类局部共价环境下的 2000 个原子表示，用于后续分析。随后，我们使用 t 分布随机邻域嵌入（t-SNE）[71] 通过 Scikit-learn 软件包进行降维，将 128 维原子表示转换为 2 维，以便可视化 [51]。我们计算二维表征空间中的 Davies–Bouldin（DB）[18] 指数，通过 Scikit-learn 软件包定量评估表示聚类的清晰度和分离度。DB 指数的定义为

$$\text{DB index} = \frac{1}{N}\sum_{i=1}^{N}\max_{i\neq j}\left(\frac{S_i + S_j}{M_{ij}}\right) \tag{13}$$

其中， M_{ij} 是被选作簇 i 和 j特征的向量之间的距离，而 S_i 和 S_j 分别是簇 i 和 j的离散度。具体而言， S_i 和 M_{ij} 定义为

$$S_i = \left\{\frac{1}{T_i}\sum_{j=1}^{T_i}|X_j - A_i|^q\right\}^{\frac{1}{q}} \quad M_{ij} = \left\{\sum_{k=1}^{n}|a_{ki} - a_{kj}|^p\right\}^{\frac{1}{p}} \tag{14}$$

其中， T_i 是簇 i中的向量数量， A_i 是簇 i的质心， a_{ki} 是 n维向量 a_i的第 k个分量，该向量是簇 i的质心。这里我们定义 $p = 2$ 和 $q = 1$，其中 M_{ij} 表示质心之间的欧几里得距离， S_i 则表示簇 i 中的向量到簇 i质心的平均欧几里得距离。DB 指数值越低，表明簇之间的区分度越高、越紧密。

DPA-3.1-3M 和 JMP-L 均采用边表示，对预设截断半径内的原子对关系进行编码。我们对 DPA-3.1-3M 下游任务中两种常见边类型——碳-氢（C-H）对和碳-氧（C-O）对的边表示进行可视化，并对 JMP-L 的 C-O 边表示进行可视化。随后，根据成对欧氏距离，将这些边进一步划分为近、中、远三类。我们将原子对之间的距离小于两个原子共价半径之和的 1.15 倍的边定义为近边，即这些原子对形成共价键。我们将原子对之间距离在 3 埃至 4 埃之间的边定义为中边。中边中的原子对不成键，但其特征会根据角度信息进行更新；在 DPA-3.1-3M 中，角度截断为 4 埃。我们将原子对之间距离在 5 埃至 6 埃之间的边定义为远边，因为 DPA-3.1-3M 的截断半径为 6 埃。与中边相比，远边同样不成键，但原子对的特征更新不包含角度信息。与原子表征分析类似，我们从 OOD 测试集中随机抽取每种边类型的 2000 个边表示，再使用 t-SNE 将高维边表示转换到二维空间进行可视化，并使用 DB 指数评估边表示聚类的清晰度和分离度。

我们使用 1-瓦瑟斯坦距离（$\mathcal{W}_1$）正式量化预训练模型和微调模型的原子表示分布之间的相似性，或等价地说，漂移；本文中简记为 $\mathcal{W}$。实践中，我们分别对预训练模型和微调模型采样 5000 个原子表示。离散最优传输问题定义如下：

$$\mathcal{W} = \min_{\gamma}\sum_{i,j}\gamma_{i,j}C_{i,j} \tag{15}$$

其中， $\gamma \in \mathbb{R}^{5000\times5000}$ 是最优传输矩阵， $C \in \mathbb{R}^{5000\times5000}$ 是代价矩阵，$C_{i,\ j}$ 定义为原始 128 维空间中第 i 个预训练表征与第 j 个微调表征之间的欧几里得距离。结果通过 POT 包 [23]计算得出。

D.2. 模型架构细节

沿用正文中的记号，我们将用于学习原子表示的图神经网络骨干网络记为 f_θ，其可训练参数为 θ；将力场任务头和下游属性预测头分别记为 g_ϕ 和 h_ψ，其可训练参数分别为 ϕ 和 ψ。 f_θ 和 g_ϕ 的架构与 DPA-3.1-3M [83] 完全相同，本文不再详述。

对于下游性质预测任务，我们考虑一个带标签的训练数据集 $\mathcal{D} = \{(x_k,\ y_k)\}_{k=1}^{K}$，其中每个数据点由一个三维原子结构 x_k 组成，而 $y_k \in \mathbb{R}^d$ 表示感兴趣的分子或材料性质（例如，

带隙或介电常数）。形式上，每个结构表示为 $x_k = (R_k,\ Z_k)$，其中 $R_k \in \mathbb{R}^{N_k \times 3}$ 表示 N_k 原子的笛卡尔坐标，$Z_k \in \mathbb{N}^{N_k}$ 表示它们的原子序数。目标性质的维度记为 $d \geq 1$，其中 $d > 1$ 支持同时预测多个性质。在本研究中，我们将研究范围限定为 $d = 1$，即每次只预测单个性质。一般而言，h_ψ 的形式为

$$\hat{y}_k = h_\psi(f_\theta(x_k)) = \begin{cases} \sum_{i=1}^{N_k}[\text{MLP}(f_\theta(x_k)_i) + p(Z_{ki})], & \text{if } y_k \text{ is extensive.} \\ \frac{1}{N_k}\sum_{i=1}^{N_k}[\text{MLP}(f_\theta(x_k)_i) + p(Z_{ki})], & \text{if } y_k \text{ is intensive.} \end{cases} \tag{16}$$

其中，MLP 指多层感知机，本研究采用的具体配置为 3 层，每层 240 个神经元。此外，$p(Z_{ki})$ 是根据原子序数 Z_{ki}确定的原子性质贡献常数偏置。该性质偏置通过对训练数据中的性质进行最小二乘拟合确定。更准确地说，我们有K 个训练数据帧，在第 k 帧中有 N_{kz} 个原子，其原子序数为 z，该帧的标注性质记为 y_k。随后以最小二乘意义求解一个线性系统。

$$\sum_z N_{kz} p(z) = \begin{cases} y_k, & \text{if } y_k \text{ is extensive} \\ N_k y_k, & \text{if } y_k \text{ is intensive} \end{cases} \quad k = 1, ..., K \tag{17}$$

我们将 QM9 中的最高占据分子轨道、LUMO、GAP，以及 MatBench 中的介电常数、弹性、钙钛矿视为强度性质；而将 QM9 α、零点振动能和 C_v 视为广延性质。存在三个特殊情况：介电常数、$\langle R^2 \rangle$ 和 μ。对于介电常数，我们不使用性质偏置项 $p(Z_{ki})$（公式（16）），因为测试集中存在训练集中未出现过的元素种类。对于 QM 9 $\langle R^2 \rangle$ 和 μ，我们采用定制的任务头架构，以更好地利用先验信息。具体而言，对于 [70, 67, 4] 预测，我们采用与先前工作相同的预测头形式 $\langle R^2 \rangle$；对于偶极矩标量的预测，我们采用偶极矩向量的范数 [84] μ。

D.3. 模型超参数

表中展示了 DPA3 模型从头训练、微调（FT）和多任务微调（MFT）所用的超参数。8。DPA3 根据 N 每个特定系统中的原子数采用动态批大小。具体而言，Auto：$X = \lceil \frac{X}{N} \rceil$ 其中 N 表示每个系统中的系统数，$\lceil \cdot \rceil$ 表示向上取整函数，即将数值向上取整至最近的整数。能量、力和维里损失前因子中的记号“→”表示这些前因子会随学习率同步调整。此外，DPA3 采用了一种名为 SiLUT（带双曲正切函数的 SiLU 阈值函数）的定制激活函数，以提高训练过程中的数值稳定性，其定义如下：

$$\text{SiLUT:t}(x) = \begin{cases} \text{SiLU}(x), & \text{if } x \leq t. \\ \text{Tanh}(a \cdot (x - t)) + b, & \text{if } x > t. \end{cases} \tag{18}$$

其中，t 表示阈值，指示何时从 SiLU 切换到双曲正切函数。常数 a 和 b 根据t 确定，以确保激活函数在阈值处的一阶和二阶连续性。

关于 JMP 模型的超参数，我们采用了 28 的批大小，并进行了 24 轮训练。其余所有超参数均按照参考文献 [67] 设置，适用于从头训练和微调。

D.4. 数据效率实验详情

如图 3(a) 和图 3(b) 所示，我们通过在 QM9 带隙（GAP）和 MatBench 折射率（介电常数）预测任务上开展一系列实验，评估 MFT 的数据效率。在这些实验中，通过对原始数据进行随机下采样，使训练集大小跨越三个数量级。该方法用于检验模型在极低数据区域中的性能，其中 QM9 GAP 任务的数据点最少仅为 114 个，MatBench 介电常数任务的数据点最少仅为 14 个。

对于 QM9 GAP 预测任务，我们使用不同的随机种子，为每种数据规模生成 3 个独立的训练子集，并进行 3 次独立运行以确保统计可靠性。结果以分布内和分布外测试集的均方根误差（RMSE）报告，并以训练数据量为横轴绘制。

表 8：在 QM9 和 MatBench 上训练的 DPA3 模型的超参数。

	超参数	QM9	MatBench
常规	优化器	Adam	Adam
	学习率调度	Exp	Exp
	最大学习率	1e-3	1e-3
	最小学习率	1e-5	1e-5
从头训练/微调	训练步数	0.1M	0.1M
	损失函数	MSE	平滑平均绝对误差
	每个 GPU 的批次大小	自动：512	自动：512
	GPU数量	1	1
	激活函数	Tanh	Tanh
MFT	训练步数	0.1M	0.2M
	力场任务头	SPICE2	OMat24
	能量损失前因子	$0.2 \rightarrow 20$	$0.2 \rightarrow 20$
	力损失前因子	$100 \rightarrow 60$	$100 \rightarrow 60$
	维里损失前因子	-	$0.02 \rightarrow 1$
	性质损失前因子	1	1
	每个 GPU 的批量大小（性质任务）	自动：512	自动：512
	每个 GPU 的批量大小（力场任务）	自动：128	自动：128
	激活函数（性质输出头）	Tanh	Tanh
	激活函数（力场头）	SiLUT:3.0	SiLUT:3.0
	GPU数量	2	1

在 MatBench 介电性质预测任务中，我们使用随机划分评估分布内性能，并使用留一簇划分（LOCO）进行分布外评估，涵盖不同的训练数据规模。与只有一个折的 QM9 GAP 不同，MatBench 介电任务采用多个折，具体而言，随机划分使用 5 个折，LOCO 使用 50个折。实际操作中，我们选择了全数据集微调期间测试集平均绝对误差（MAE）最接近所有折的平均 MAE 的那个折。后续的数据效率实验均使用这一代表性折。

对于介电常数任务的随机划分实验，我们沿用 QM9 GAP 所采用的采样策略，即每种数据规模使用 3 个独立训练子集，并分别运行 3 次。不过，为了减轻介电常数任务中 LOCO 划分固有的较高方差，我们采用了更严格的采样策略。具体而言，对于从 1 到 $1/16$ 的数据规模，我们使用 6 个不同的随机子集进行了 6 次独立运行。对于较小的数据规模，包括 $1/32$、$1/64$、$1/128$ 和 $1/256$，则将试验次数增加到 10 次独立运行，并使用 10 个不同的随机子集。我们将所得平均绝对误差绘制为训练数据量的函数，以展示模型的数据效率。

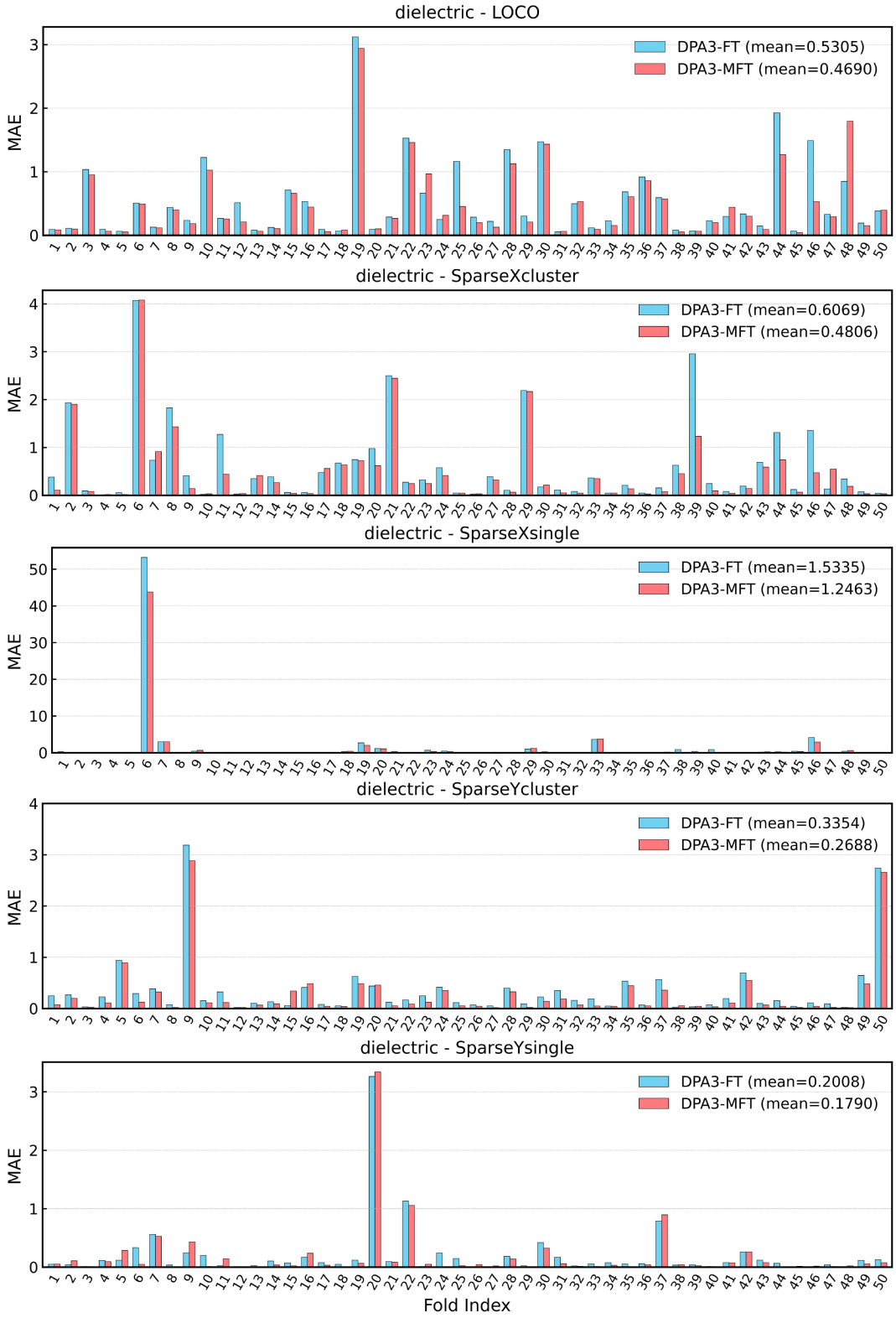


图 6：在五个不同的分布外划分中，DPA3-FT 和 DPA3-MFT 在介电数据集上各折的平均绝对误差分布。

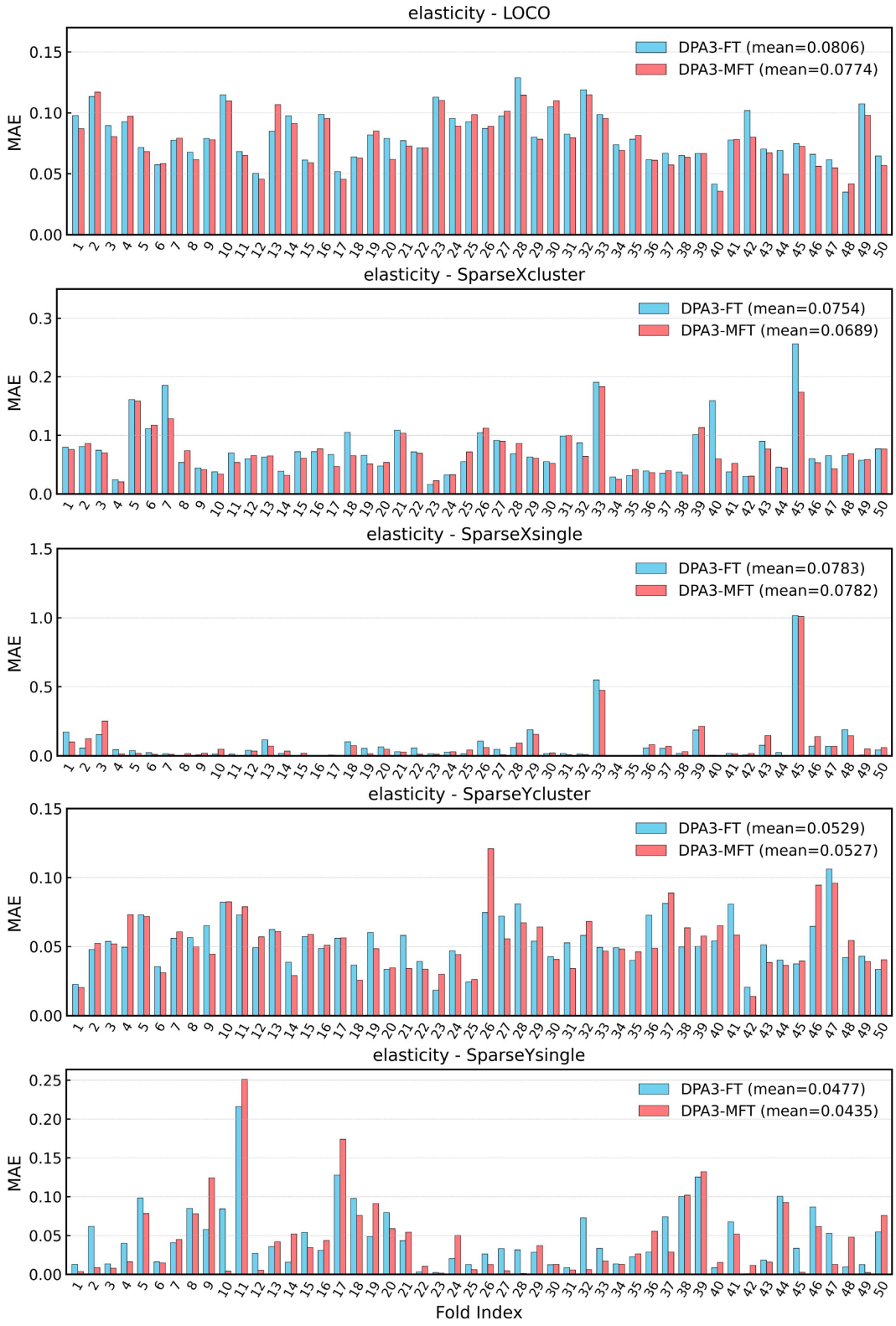


图 7：DPA3-FT 和 DPA3-MFT 在弹性数据集的五种不同分布外划分上各折的平均绝对误差分布。

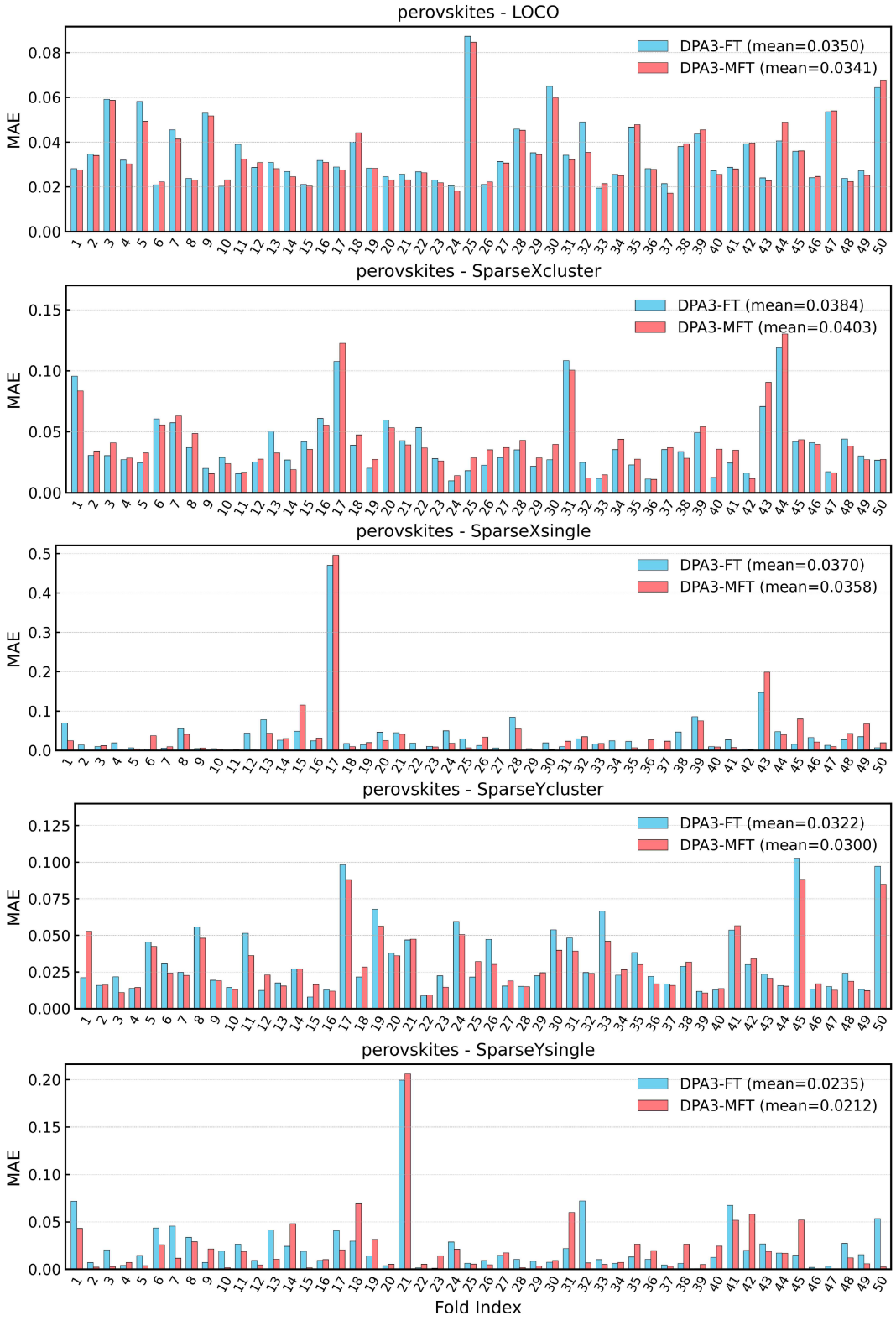


图 8：DPA3-FT 和 DPA3-MFT 在钙钛矿数据集的五种不同分布外划分中各折平均绝对误差的分布。