首先，我们将带你走进 DPA-2 —— 一个大规模、多任务预训练的原子模型。它不仅代表了当前机器学习在分子模拟领域的前沿探索，也为构建通用势能面模型提供了全新思路。”

势能面（PES）是分子模拟的地基。它决定了原子如何运动、如何相互作用。但目前的做法总是在精度和效率之间做取舍。量子力学方法，像 DFT，非常精确，但计算成本高得惊人。而传统的经验力场虽然快，但又太不准。我们需要的是，既快又准的‘新势力’。

这时，机器学习登场了。但别高兴太早，想要训练一个靠谱的机器学习势能函数，首先得有大量高质量的标注数据——也就是 DFT 算出来的能量和力。可这些数据的获取，堪比烧钱机器，动辄百万 CPU 小时。而且，模型常常只能‘学会特定领域’，面对新材料、新结构，就彻底蒙了。

那我们该怎么训练模型呢？有两种思路。第一是‘单任务训练’：数据一致，模型专注，效果稳定，但拓展性差。第二种是‘多任务预训练’：用来自多个来源的数据共同训练，让模型学会更多领域的知识，之后再根据任务微调。这种方式更灵活，也更像人类的通识学习方式。

如果我们想造出一个真正通用的‘大型原子模型’，我们希望它具备四个特质：一是广泛适用，泛化能力强；二是遵循物理对称性，比如旋转、平移不改变物理本质；三是能量守恒，也就是说它预测的力必须是能量梯度推导出来的；四是光滑连续，特别是在进行动力学模拟时，这一点至关重要。

DPA-2 就是这样一个符合所有标准的模型。它不仅在多种任务上展现出优异的泛化能力，还能广泛适用于合金、半导体、电池材料甚至是药物分子。它的成功，证明了多任务训练在构建通用原子模型中的巨大潜力。

可以说，DPA-2 不只是一个模型，更像是一座桥梁，把高精度的量子模拟和高效率的大规模计算连接起来。它为下一代机器学习势能函数的崛起，铺平了道路。

现在，让我们来认识一下 LAM，也就是大原子模型的整体工作流程。整个流程图分为三个主要阶段：多任务预训练、下游任务微调，以及知识蒸馏。每一部分都相互衔接，共同构建出一个既通用又高效的原子建模框架。

在预训练阶段，所有的 DFT 标注数据集共享同一个描述子网络，用来捕捉原子系统的结构与物理对称性。每个数据集还配备了自己专属的能量拟合网络。描述子的参数在所有数据上共同优化，而拟合网络只对各自的数据进行训练。这样的多任务机制既保留了个性，又训练出一个统一的、高泛化能力的原子表示。

接下来是微调阶段。在处理下游任务时，我们保留预训练得到的统一描述子，拟合网络可以选择重新初始化，或使用来自某个上游任务的结构。由于描述子已囊括了大部分关键信息，拟合网络的初始值影响较小。这一步的目标，是快速适应新的势能面建模任务，实现精准预测。

为了提升实际应用中的效率，我们引入了知识蒸馏。预训练+微调后得到的大模型作为老师，使用分子动力学模拟采样新的结构，并打上标签。然后，训练一个简化版的学生模型，比如没有注意力机制的 DPA-1。通过多轮“老师教学生”的过程，学生模型不断进化，最终兼具准确性与运算效率，非常适合部署到下游任务中。

整个 LAM 工作流，从多任务预训练、到灵活微调、再到高效蒸馏，形成了一个既能全面学习、又能轻装上阵的智能系统。它为原子建模、材料发现和分子模拟等任务，提供了统一、高效、可靠的新范式。

接下来，我给你介绍 LAM 的预训练数据。在深入细节之前，我们先明确数据需求。第一，覆盖范围要广，包括多种化学成分和原子构型，确保模型在后续任务中遇到的任何场景都能在预训练阶段有所体现。第二，依赖多种计算条件下生成的高质量 DFT 标注，保证多任务训练的鲁棒性，并使模型具备对全新体系的零样本泛化能力。

这里是表格 1，对本研究所用的数据集进行了汇总。我们自制了多种数据集，包括金属合金、电池阴极材料、金属纳米簇和类药分子；同时还使用了社区贡献的数据，涵盖钙钛矿、固态电解质、半导体、水簇等。每行列出了训练样本数、测试样本数、总体规模，以及预训练时用于多任务采样的权重。

多任务预训练完成后，我们评估 DPA-2 对全新任务的预测能力，也就是所谓的零样本泛化。表格 2 显示，无论是合金形成能，还是分子反应势垒，DPA-2 在各项下游基准测试中都保持了稳定低误差，充分证明了多样化加权预训练对迁移性能的大幅提升。

综上，通过精心挑选并根据多样性与相关性赋权的丰富数据集，多任务预训练让 DPA-2 内化了大量原子级知识。结果是：模型可以快速微调，并且在全新任务上无需额外训练就能交出高精度的答案，真正向‘通用机器学习势能函数’迈出了重要一步。

现在我们来介绍 DPA-2 模型的核心：原子表征描述子（Descriptor）。

图中展示了 DPA-2 描述子的完整结构，它主要由两大部分组成：RepInit 初始化模块 和 RepFormer 表征更新模块。

这套架构的设计目标，是从原始的原子结构中提取出包含丰富物理语义的特征，作为下游任务的输入。

在 DPA-2 中，原子之间的交互关系被分为三种不同的通道：

第一种是单原子通道，用于捕捉每个原子的本地属性；第二种是旋转不变的双原子通道，用来编码与几何相关的标量关系；

第三种是旋转协变的双原子通道，用于捕捉与方向相关的矢量性质。

这三者结合起来，共同构成了丰富而稳定的结构表示。

在 DPA-2 中，双原子特征 的初始值由 环境编码器 env 模块生成，

它会在一定的截断半径范围内，考虑每个原子的邻近结构，并采用平滑衰减函数对远处的原子权重降低。

而单原子特征则是由 repinit 层进行初始化，学习每个原子本身的结构语义。

初始化完成后，所有原子特征会通过 12 层 RepFormer 网络进行更新。

在每一层中：单原子特征会经历卷积、对称化、多层感知机（MLP）、以及局部自注意力等模块；而双原子通道则会经过 MLP、点积操作以及门控注意力机制。所有这些操作以消息传递的方式进行，使得原子之间的依赖能够跨越距离传播。

DPA-2 模型在架构设计上严格遵循物理规律，具备以下重要属性：对于平移、旋转与原子置换操作都保持不变；力的计算方式为能量对原子位置的梯度，压力则通过应力张量计算，所有模块均具备二阶连续性，确保能量守恒，在分子动力学模拟中不会出现数值漂移。

多任务预训练为 DPA-2 带来了极强的泛化能力。

与单任务训练版本或传统模型（如 MACE）相比，DPA-2 在能量与力的加权平均误差（WARMSE）方面表现出明显优势，

特别是在涉及不同分子系统的下游测试任务上，其零样本泛化能力尤为突出。

接下来，我们讲解多任务预训练的 DPA-2 模型在下游任务中的泛化能力。如图所示，展示了多个模型在零样本测试中的性能对比结果。

在分析多任务方案的泛化性之前，我们首先对 DPA-2 进行了单任务基准测试。

在 ANI-1x 数据集上，DPA-2 的测试精度明显优于原始的 ANI-1x 模型。

在 OC20 数据集上，它的准确率与 GemNet-OC84 相当，并且优于 Equiformer V2、NequIP、Allegro 和 MACE 等多个先进模型。

这说明，即使不使用多任务训练，DPA-2 本身已经具备非常强的建模能力。

接下来，我们将 DPA-2 应用于所有预训练数据集，采用多任务训练策略。

结果显示，这种训练方式对原始数据集上的性能影响极小。

举例来说，力的 WARMSE 仅从 111.1 增加到 116.3 meV/Å，能量误差也只是从 14.9 增加到 18.6 meV/atom。

这表明多任务学习的引入并未带来明显的性能损耗，却为后续泛化能力打下了坚实基础。

为测试其泛化能力，我们将预训练完成的 DPA-2 模型，直接应用于多个下游任务中，不进行任何微调，这种方式称为“零样本泛化”。实验结果表明，性能的提升主要源自多任务预训练策略本身，而不是模型结构的差异。

虽然零样本泛化已经表现优秀，但我们还可以进一步通过微调提升模型表现。

与从零开始训练相比，经过预训练的 DPA-2 在下游任务中表现出更高的样本效率。

举例来说，在 H2O-PBE0TS-MD 任务中，为达到相同的能量精度，微调模型所需的训练样本数量是从头训练模型的百分之一，甚至更少。

这表明，预训练为模型带来了极强的迁移能力，也显著降低了下游任务的标注数据成本。

总的来说，多任务预训练不仅保持了 DPA-2 在原始数据集上的高精度，还显著增强了其在下游任务中的泛化与迁移能力。

无论是在零样本场景还是微调场景中，DPA-2 都展示出领先的性能。

接下来我将为讲解我们在模型蒸馏和评估方面的工作。

我们微调过的DPA-2模型在准确度方面表现非常出色，但它的参数量非常庞大，导致计算开销非常大。这意味着在实际应用中，虽然模型精度高，但运行起来非常耗时且资源消耗大，限制了它的推广和应用。

为了克服这个瓶颈，我们采用了知识蒸馏的方法。简单来说，就是让一个大而复杂的“老师模型”——也就是微调后的DPA-2，把它学到的知识“教给”一个更小、更轻量的“学生模型”，这里是无注意力机制的压缩版DPA-1。

我们并不是直接用全部数据训练学生模型，而是先让老师模型在下游任务中少量数据上产生高质量的训练样本，然后用这些样本训练学生模型。这样既保证了学生模型能覆盖关键的物理配置空间，又大大节省了训练资源。

这里这张图展示了不同模型在下游任务中的样本效率。

横轴表示训练所需的下游数据量，纵轴则是模型预测能量或力的均方根误差。

从图中可以看出，蒸馏后的模型在使用更少数据的情况下，仍然能达到和复杂模型相当的预测精度，体现了非常高的样本利用效率。

经过蒸馏，学生模型不仅大幅提升了推理速度，甚至单个GPU卡能模拟的最大体系规模提升了近100倍

更令人欣喜的是，蒸馏模型的精度几乎和微调的DPA-2持平，说明我们在保证模型轻量化的同时，没有牺牲其预测能力。

为了验证蒸馏模型的实际表现，我们在三个真实的物理系统上做了深入测试：

首先，在水的分子动力学模拟中，模型计算的径向分布函数和角度分布函数与昂贵的第一性原理AIMD模拟几乎完全一致，说明模型能精准反映水分子间的结构关系。

其次，在锂离子扩散模拟中，模型预测的扩散常数与之前的高水平DP-PBE和密度泛函理论结果相吻合，验证了模型对动力学性质的捕捉能力。

最后，在铁电钙钛矿固溶体中，蒸馏模型成功重现了温度驱动的相变过程，且相变温度随着组成比例变化的趋势也与实验观察高度一致，展示了模型在复杂多组分体系中的可靠性。

综上所述，通过知识蒸馏，我们得到了轻量且高效的模型，极大提升了计算速度和模拟规模，同时保持了与复杂模型相媲美的高精度。

这使得我们的模型非常适合用于大规模、长时间的分子模拟，推动了多尺度、多物理过程模拟的前沿研究。

最后，我们总结 DPA-2 模型的研究成果，并展望其未来发展方向

本文提出了 DPA-2，一种面向大型原子模拟（LAM）的新型模型架构。

它搭配了完整的流程：包括多任务预训练、微调、知识蒸馏以及最终部署，构成了一个从训练到应用的完整闭环系统。

研究得出三点关键结论：

第一，DPA-2 通过多任务预训练显著增强了泛化能力，能够处理多达 73 种元素。

第二，与 MPtrj 预训练的 MACE 模型相比，DPA-2 在零样本评估中的能量误差降低了 52%，力误差降低了 59%；相较于单任务预训练的 DPA-2 模型，也有超过 50% 的改进。

第三，在下游任务中，仅用原始数据集的百分之一甚至千分之一，DPA-2 也能达到接近的精度，大幅减少了数据需求。

尽管取得显著进展，目前的预训练数据集中仍存在不足，特别是对二维材料的覆盖严重缺失，限制了模型的泛化性。

因此，DPA-2 的发展应被视为一个长期过程。我们呼吁社区共同参与，持续收集多样化的训练数据，引入具体应用场景的测试，并建立自动化流程来完成数据预处理、模型训练与版本迭代等任务。

为实现这一目标，我们发起了 OpenLAM 计划，致力于推动开放式的 LAM 模型发展生态。欢迎大家在 AIS Square 平台关注并参与 OpenLAM 项目。

感谢您的聆听！