项目

材料计算基础术语

VASP：是用于原子尺度材料模拟的计算机程序，其基于第一性原理，可以执行电子结构计算和量子力学分子动力学。

DFT：密度泛函理论

ASE：用Python编程语言编写的原子模拟环境，旨在设置，引导和分析原子模拟。

AENEAT：原子能网络，基于人工神经网络构建和应用原子相互作用势的工具集合。

MD：molecular dynamic，分子动力学。模拟用牛顿力学描述一群粒子的运动。粒子的受力通过系统总势能对其坐标变量求微分得到。总势能由外场势能和粒子间势能（主要）组成。实践中通常把这些作用势表达为一个解析函数，函数形式受物理原理启发，其参数则通过拟合实验结果和第一性原理计算结果来得到，因此称作**经验作用势 (empirical potential)** 或者直接简称为**势** (potential)。

lattices：晶格

phonon：声子

quantum-mechanical：量子力学

eV：电子伏特，一个电子经过1伏特的电位差加速后获得的动能。

Hartree：1H=27.21138eV

Deep potentials(DP)：深度势能

计算的属性

- **自由能**：某一个热力学过程中，系统减少的内能中**可以转化为对外做功的部分**，在一个特定的热力学过程中，系统可对外输出的"有用能量"。当且仅当系统平衡时自由能变化为零。
    - 赫姆霍兹自由能 和 吉布斯自由能G
- **结合能：**是指晶体中各个原子或分子**通过化学键相互结合而形成晶体时释放或吸收的能量。**它表示了晶体的稳定性和结构强度。结合能可以看作是晶体内部相互作用的总和，包括**原子间的键合能和相互作用能。**通过计算晶体的总能量和分离态的原子或分子能量之差，可以得到结合能。
    - eV/atom
- **形成能：**是指形成晶体中一个特定缺陷（如点缺陷、薄膜、界面等）时，相对于无缺陷晶体而言，所需要的能量变化。形成能反映了晶体中缺陷的稳定性和形成难度。形成能可以通过计算晶体中带有缺陷的体系的总能量和无缺陷体系的总能量之差来确定。
- **ZPVE：**零点振动能量
- **polarizability** α：极化率
- **electronic spatial extent** $\langle R^2 \rangle$：电子空间范围
- U0：0K时的内能
- 内能U：298K时的内能。
- **焓**H：
- 熵S：
- 热容量：物质存储热量，温度就会上升。大热容存储一定的热量，升温会比小热容少。
    - **定容热容**C_v：物质体积不变时，定容

    - **定压热容**C_p：物质的压强不变时，定压热容量
- **最高被占轨道** (HOMO)：被占轨道中能级最高的轨道
- **最低空轨道** (LUMO)：空轨道中能级最低的轨道
- **能隙**gap：电子在两个特定的不同状态下的能量差。E(LUMO) - E(HOMO)。
- **带隙**：周期性体系才有能带的概念。分子系统没有能带。
- **费米能**E_F：温度为0K时，一群电子从最低能级E1开始排布，布满所有的量子态后，再继续排E2能级的量子态。把所有电子排完后，所能达到最高能级，对应的能量叫做费米能 E_F。
- 体积模量 Bulk moduli：均质各向同性固体的弹性，可表示为单位面积的力。
- 剪切模量 Shear moduli： 剪切应力与剪切应变之比。
- 泊松比：材料在外力作用下，沿某个方向拉伸时，其在垂直于该方向的横向收缩程度的比例。是横向应变和纵向应变
- w：最高振动频率

ALKEMIE软件

- **高通量计算总流程**：
    - 首先选择合适控件进行建模，通过高通量处理器控件（HT Preprocessor）统一处理。高通量处理器控件（HT Preprocessor）是连接材料结构和第一性原理计算的关键控件。
    - 高通量处理器输出接口可以连接第一性原理VASP实际计算流程相关的控件，即连接多个计算工作流。选择完不同的工作流之后，同样需要将所有不同的计算控件统一连接至Manager管理控件，该控件用来管理所需计算的任务并将任务发送至服务器上的数据库中。
    - 添加完工作流任务后，将管理控件连接至Jobs任务提交控件，即可开始计算。

机器学习势

- 材料建模中的所有原子模拟都要求某种形式势能面PES的输入来描述原子如何相互作用，由此确定原子力。在玻恩-奥本海默近似里，**获得PES**最准确的方法是通过求解 基于固定原子核坐标电子结构的量子力学处理的 薛定谔方程。
    - 最广泛的电子结构求解方法是**DFT**，但时间复杂度是$O(N^3)$的。这使DFT在大材料系统和长模拟时间（MD）中，很耗时。
    - 解决这个问题的普遍方法是 **基于经验的原子间势**。它基于物理或化学观察，来假设原子位置和系统能量间的analytical functional relationship。
    - 量子力学方法对于这种原子模拟来说精度很高，但效率极低，而经验原子间势虽然有效，但通常精度有限。
- ML势：基于 ML灵活分析 形式的 参数化PES的通用描述。与经验原子间势相比，ML 势的灵活性提高了可表示性（准确性），并且分析形式相对于DFT 计算显着提高了效率。

深度势

- 系统能量E是原子坐标R的函数，E=E(R)。原子i的力是势能对坐标的负梯度。
- virial tensor

- loss就是三种的带权均方差和的均值。

知识科普

第一性原理

- 第一性原理（First Principle），是从量子力学理论出发的计算方法，它“号称”自己仅需要原子精细结构常数、电子质量及带电量、原子核质量及电量、普朗克常量和光速这几个已知的参数，便可根据原子核和电子相互作用的原理及其基本运动规律，经过多个近似处理后直接求解薛定谔方程，进而得到材料（几乎）所有的基态性质。**狭义的第一性原理计算，是指基于Hatree-Fock自洽场计算方法的“从头算”（ab initio），广义的第一性原理计算在此基础上还包含了密度泛函理论（DFT）计算。**
- 在实际操作中，除了第一性原理要求的已知参数外，我们还常常加入一些“经验参数”，它们通常来自于第一性原理计算中前人得出的、已得到大量实例验证的规律，或者来自于实验领域最直接的结果，这样的做法通常称之为**“半经验的”**，它能有效地减少计算资源的损耗，进而保证计算工作在最优化的条件下进行。

分子动力学

- 分子动力学（Molecular Dynamics，简称MD），是从经典物理的统计力学出发的计算方法，它通过对分子间相互作用势函数及运动方程的求解，分析其分子运动的行为规律，模拟体系的动力学演化过程，给出微观量（如：分子的坐标与速度等）与宏观可观测量（如：体系的温度、压强、热容等）之间的关系，从而研究复合体系的平衡态性质和力学性质，是研究材料内部流体行为、通道运输等现象有效的研究手段。
- 通过给定势函数，赋予体系初始的坐标和速度，我们会得到一系列包含了整个分子动力学过程的坐标与速度，再通过对坐标与速度的统计，我们可以得到需要的体系相应的热力学与动力学性质。

波函数

- 薛定谔方程中的 Ψ 是一个关于时间和位置的函数，也就是 Ψ(r,t) ，被称为波函数。
- 波函数的模的平方表征粒子在这个位置这个时间出现的概率。

基态和激发态

- 电子的最低能量基态决定了结构和原子核的低能运动。物质的各种形式在很大程度上都是电子基态的表现。
- 对于给定的原子核形成的结构，**电子的激发态就是那些“电子性质”的本质**-包括电导，光学性质，热激发，半导体中的非本征电子的现象等等。
- **电子的激发类型主要有两种：电子的增加或减少，以及电子数恒定的激发。**因为**激发经常可以被粗略地看作是对基态的微扰**，因此**微扰理论**的方法经常是理论上理解和计算这些性质的关键。
- **电子的激发态同样可以和原子核的运动相耦合**，这也就导致了其他的影响例如**电子-声子相互作用（electron-phonon interaction）。**

DFT

- 只保留泡利不相容原理和外部势场（V(r)）约束，根据基态粒子密度n(r)来算电子的动能。

径向基函数

- RBF（Radial Basis Function，径向基函数）是一个函数空间中的基函数，而这些基函数都是径向函数。
- 所谓径向函数（Radial Function） $\varphi(x)$满足这样一种条件：对于某一个固定点c，满足 $\varphi(x) = \varphi(||x - c||)$, 即对于围绕着某固定点 c 的等距的 x，函数值相同。
- 常见的径向函数有很多，高斯函数是其中之一。

反演变换（parity）

- 矢量在反演变换下，方向保持不变，叫偶（even）。方向反向，叫奇（odd）。
  - 在反演变换下，位置矢量会反向，磁场强度矢量则保持不变。

E(3)不变性

- 3D空间上，分子表示应该具有平移、旋转和反演(reflection)不变性。
  - 反演就是(x,y,z) -> (-x, -y, -z)

构象和构型

- 构型：分子中由于各原子或基团间特有的固定的空间排列方式不同而使它呈现出不同的立体结构。一般情况下，构型都比较稳定，一种构型转变为另一种构型则要求共价键的断裂、原子/基团间的重排和新共价键的重新生成。
- 构象：由于分子中的某个原子（基团）绕C-C单键自由旋转而形成的不同的暂时性的易变的空间结构形式。不同的构象之间可以相互转变。

声子

- 传统的晶体表示是静态晶格模型。但晶体中的原子是振动的，对原子的简谐振动进行量子化，可以发现其简谐振动的能量是分裂的。称每一个分裂的能量是简谐振动能量子，即声子。
- 每个声子都对应晶体格点的一种集体振动模式。每个声子代表晶格振动的激发态的一个量子化能级。
- 态密度：每单位能量范围内可用的声子态的数量。
- 固体中的粒子不断运动，将粒子间相互作用近似为谐振子势后可以求解出晶格的独立振动模式，其具有波动形式，称为格波，总数和晶体的自由度数目相等。**振动模式的频率和波矢的对应关系称为色散关系。**倘若原胞中有多于一种粒子，将会有多组独立的色散关系，声学支反应了晶胞的整体运动，光学支反映了晶胞内不同粒子的相对运动。**经正则变换可以将每种振动模式转化为一个独立谐**

振子，其每一份量子化的能量称为一个声子。声子具有能量和准动量，可以与其它粒子发生散射，散射过程满足能量动量守恒。

声子是玻色子，满足玻色统计，**根据分布可计算晶格振动的能量，能量对温度求偏导得到热容。能量的计算依赖于态密度，而态密度依赖色散关系。**德拜模型和爱因斯坦模型对态密度进行了两种不同的近似，分别用于处理色散关系中的声学支和光学支。

- 声子色散关系（PDR）：频率-波矢（波长）关系（能量-动量函数关系）
    - 确定粒子的色散关系就完全确定了粒子的性质和宏观热力学量。
- 声子群速度是色散关系曲线的斜率。

论文们

SAIBench

- 科学人工智能的基准测试不仅关注人工智能模型的准确性（accuracy），还关注计算成本。计算成本可以进一步分为两个阶段：（1）模型达到一定精度的成本，（2）模型经过适当训练后，使用模型进行推理任务的成本。对于第一阶段，标准做法是**根据最佳/平均/最差精度来测量训练时间（挂钟或总 CPU/GPU 时间）**，第二阶段则**测量完成推理任务的吞吐量/延迟**等。
-

Scientific Benchmarking of Parallel Computing Systems

1 模型需报告的指标

- 硬件信息：计算、存储、网络
    - 节点数、处理器、加速器
    - 内存大小和带宽、总线信息、非本地存储信息
    - 网络拓扑和带宽
- 软件信息：
    - 编译器版本、库版本、文件系统、存储系统、内核版本
- 配置：
    - 软件和输入、评估设置和代码是否在线可用

2 常见最佳实践

- 谨慎使用 Speedup
    - 可以被执行时间的下限取代
    - 发布并行加速时，报告base case是单个并行进程还是最佳串行执行，以及base case的绝对执行性能。
- 明确报告单位
    - FLOP、flop/s、B
    - 使用整个 基准、节点集、应用的结果，而不是子集。
        - 如果使用子集应该指明原因

3 分析实验结果

- 对于确定性数据：算术平均值仅用于汇总cost、使用调和平均值来汇总rate、仅当这些不可用时才汇总ratio。
- 对于非确定性数据：报告测量的置信区间

MLPerf HPC

- 其他Benchmark的性能评估：Flops、Flops/Watt、Valid Flops、 Valid Flops/Watt、Throughput、Time to solution
    - MLPerf HPC的：Time to train
- 数据分级、算法收敛和计算性能方面对它们进行了比较。因此，我们对不同子系统的优化有了定量的了解，例如数据的分段和节点加载、计算单元利用率和通信调度，从而通过系统扩展实现整体 > 10 倍（端到端）的性能改进。
- 值得注意的是，我们的分析显示了数据集大小、系统内存层次结构和训练收敛性之间依赖于规模的相互作用，这强调了近计算存储的重要性。
- 为了克服大批量数据并行可扩展性挑战，我们讨论了对大型系统有效的特定学习技术和混合数据和模型并行性。
- 最后，我们对低级内存、I/O 和网络行为方面的每个基准进行了表征，以参数化未来几轮中扩展的roofline性能模型。
- intersection-over-union (IOU)
- 将训练时间报告为主要指标。这种选择创建了一个与用户和客户直接相关的指标，用于捕获端到端性能，包括系统速度和准确性。对规则进行了一些更改，以提高 HPC 设置中科学工作负载基准的相关性。
    - 耗时 = 数据阶段耗时 + （训练耗时 + 评估耗时） + 额外耗时

- 下述指标参数化了扩展的roofline模型 [19]，从而可以描述未来 MLPerf HPC 提交的系统功能特征，并确定提交中硬件和软件级别的剩余优化潜力。
    - 内存带宽：测量加速器DRAM的读取和写入吞吐量。
    - 网络带宽：集体通信的平均网络带宽。（或实际传输的数据量/通信时间）
    - I/O带宽：每个工作线程的平均I/O带宽。

- Epochs、Batch Size、训练吞吐、评估吞吐、time to solution、 计算预算

SCIML-Bench

- 指标：训练耗时、推理吞吐、MAE、MSE、RMSE、准确率、f1_score
- 力的角度误差和幅度误差

知乎-论文们

材料表示学习

- **M3GNet：**MAE、MSE、AUC
  - GNN
- **Ewald-MP：**MAE
  - GNN 消息传递 物理交互

材料生成

> Cond-DFC-VAE 有综述

- VAE、GAN、GNN、GCN
- **Cond-DFC-VAE：**AE、MAE、Average earth-mover distance（EMD）、symmetric mean absolute percentage error (SMAPE) 、sample sizes used to train and eval、键长的绝对百分比变化、CPU时间
  - Average earth-mover distance between predicted atomic sites and the ground truth
  - VAE+GNN
- **CDVAE**：RMSE、Match rate（满足一定特性的重建材料的百分比）
  - 开发了几个具有物理意义的指标来评估生成材料的有效性、属性统计数据和多样性。
    - Validity。根据 Court 等人 (2020) 的说法，只要任何一对原子之间的最短距离大于 0.5°A，结构就是有效的，这是一个相对较弱的标准。如果根据 SMACT 计算总电荷呈中性，则该成分有效。
    - Coverage (COV)。受到 Xu 等人 (2021a) 的启发； Ganea 等人（2021），我们定义了两个覆盖率指标，COV-R（召回率）和 COV-P（精度），来衡量测试集中生成的材料和真实材料的集合之间的相似性。直观上，COV-R 衡量的是正确预测的真实材料的百分比，COV-P 衡量的是高质量预测材料的百分比（详细信息参见附录 G）
    - Property statistics。我们计算生成材料和测试材料的属性分布之间的EMD。属性统计数据是通过从通过有效性测试的材料中随机抽取的 1,000 种有效材料来计算的。
  - VAE+GNN
- FTCP：Validity Rate（满足DFT结构松弛的百分比）、Success Rate（满足用户指定目标属性在容差范围内的百分比）、
  - 广义晶体学表示 + VAE
  - 半监督VAE

材料表征

- Learning to Predict Material Structure from Neutron Scattering Data：
  - 分类：accuracy
  - 回归：MSE、标准差
  - 分类器CNN + 回归器random forest regression (RFR) 和 non-negative least squares (NNLS)
- DeepStruc：MAE
  - 图条件变分自动编码器 graph CVAE

声子计算

- MLIPs：机器学习原子间势

AI4Sci综述 Nature

- 在本次调查中，我们在不同层面进行分析，包括量子物理学、密度泛函理论（DFT）、分子动力学（MD）和连续介质动力学。在近似水平和所处理的尺度方面存在显着差异。具体来说，量子物理学通过求解多体相互作用系统的薛定谔方程来研究电子、质子和中子等粒子的行为和相互作用，以及它们的量子力学特性。量子物理学中的空间尺度通常为原子和亚原子级别，范围从皮米（10−12 米）到纳米（10−9 米）尺度，具体取决于具体问题。 DFT 使用另一种方法将多体相互作用系统映射到多体非相互作用系统来求解电子和离子的薛定谔方程，因此可以深入了解原子、分子和固体等现实材料的电子结构范围从埃（10−10 米）到数百埃。 MD 模拟在更大的尺度上进行，通常使用经验/半经验力场以及不断增长的机器学习力场，从纳米（10−9 米）到微米（10−6 米）尺度进行模拟。 MD 专注于原子和分子在各种热力学系综下随时间的运动和相互作用，从而可以研究动态行为、结构变化、动力学和热力学性质。相比之下，量子物理学旨在解决多体相互作用系统的多体波函数和哈密顿量； DFT 采用了分子和材料实际应用的替代方法； MD 模拟在更大的空间尺度和更长的时间尺度上运行，无需明确处理电子波函数的空间和旋量分量。为了解决更大的尺度并消除粒子的离散特征，偏微分方程（PDE）被用来研究从流体动力学中的微米（10−6 米，例如柯尔莫哥洛夫微尺度）到公里（ 103 米）气候动态。我们在图 3 中比较了不同系统的空间和时间尺度。因此，这项工作的重点领域分为量子系统、原子系统和连续系统。理论水平的选择取决于感兴趣的现象和研究所需的计算复杂性。不同的分析可以互相受益并导致综合分析。

5.小分子

5.5 分子动力学模拟

- ML根据原子坐标预测能量和力 => 通过将学习到的力场与适当的恒温器/恒压器配对来模拟MD => 根据模拟轨迹，可以计算出感兴趣的属性。
-

Flowing match：粗采样MD

- 一个生成神经网络，可以通过将任意概率分布转换为简单、易于采样的先验分布来近似它们
- 经过训练后，归一化流可以生成不相关的样本并计算归一化概率密度、能量和力
-

Machine Learning of Coarse-Grained Molecular Dynamics Force Fields

- CGnet 预测自由能（平均力的势），然后使用该自由能相对于输入坐标的梯度来计算 CG 坐标上的平均力。
- 首先定义粗粒度的含义以及粗粒度模型中应保留哪些物理属性：
    - 粗坐标的选择通常是通过用一个有效粒子替换一组原子来进行的。把高维原子特征表达通过粗粒度矩阵映射到低维粗粒度粒子空间。
    - 学习粗粒度能量函数 U(x; θ)，该函数将与动力学模型结合使用来模拟 CG 分子。

    - 保持平衡分布：即在映射到CG坐标时，粗粒度模型的平衡分布应尽可能接近原子模型的平衡分布。
        - 使用动力学模拟算法，使系统的平衡分布与所用势 U 的玻尔兹曼分布相同；因此，这个目标可以通过加强热力学一致性来实现。
- 定义U(x)：多尺度粗粒度（**力匹配**）48,59 和相对熵方法
-

Coarse-Grained Molecular Dynamics with GNN

- 粗粒度方法是为了延长MD的时间，以获得某些长时模拟才能观察到的属性。
- 粗粒度方法分Top Down和Bottom Up。自顶向下详细制定模型来再现宏观属性。自底向上比起模型细节更注重特定特征的再现。
- 本文用自底向上，需要学习 **自下而上粗粒度的映射** 和 **物理模型组件**：
    - 从细粒度（例如原子）表示到交互位点集（通常称为“珠子”bead）的映射
    - 包含这些珠子的系统的粗粒度物理模型（即哈密顿函数）
- 关注粗粒度模型，并采用自下而上的“力匹配”方案，该方案被制定为监督机器学习问题，以重现小生物分子系统的热力学。
- SchNet 方案来学习特征。
    - 就是CGnet + SchNet
- 寻求粗粒度表示，使得根据给定配置的粗粒度能量函数计算出的力与相应原子表示上的平均力相匹配。

- 力匹配方法就是 最小化 精细计算方法（如从头算）得到的力 和 预测得到的力 间的均方差。
- 在粗粒度模型力匹配方法中，是最小化 平均原子力 和 粗粒度模型得到的力 间的均方差。
- 像比如CGnet手动选择特征（如：距离、平面角和扭转角），该文献仅需bead和距离，来学习特征。

Learning local equivariant representations for large-scale atomistic dynamics

- Allegro 严格局部等边的 DNN 分子间势 架构。
- Allegro 用学习到的等变表示的迭代张量积 来表示 多体势。
- 并行、泛化、性能、准确
- MLIP：基于MPNN的、严格局部描述符的。
    - 等变、以原子为中心的消息传递（MPNN）原子间势准确率极高。但迭代传播会导致每个原子具有许多有效邻居的大感受野，这阻碍了并行计算 并 限制了以原子为中心的消息传递 MLIP 可访问的长度尺度
    - 严格局部描述符原子间势 因严格局部性，不会遇到并行和尺度阻碍，但准确率低。
- 原子系统的物理性质在许多几何对称性（旋转、反转和平移）的作用下保持不变，这些对称性共同构成欧几里得群 E(3)（单独的旋转为 SO(3)，旋转和反转共同构成 O (3)）。标量（例如势能）对于这些对称群运算是不变的。矢量（例如原子力）是等变的，原子几何形状改变则相应改变。

相关工作

- 系统能量可以分解为每个原子能量的组合。
- 每个原子的能量分解为该原子和各邻居原子的成对能量和。
-

NequIP

- 物理属性在原子集的平移/反射/旋转下具有明确定义的变换属性。例如，如果分子在空间中旋转，那么它的原子偶极子或力的矢量也会通过等变变换相应地旋转。
- 等变神经网络能更普遍地表示物理系统的张量属性和张量运算（如向量加法/点积/叉积）。等变神经网络保证在坐标变化下保留物理系统的已知变换属性，因为它们是由等变运算显式构造的。
- E(3)指3D空间中的旋转，反射和平移。
- 原子间势能：给定原子位置和化学种类，得到原子系统的总势能和作用在原子上的力。之前的工作认为原子系统的总势能就是原子势能的和，力就是总势能对原子位置的负梯度。
- NequIP的内部特征是几何张量，具有旋转和反射等变性。
- 以张量场网络(Tenosr-Field Networks)中的层为基础，构建平移不变/旋转等变/parity等变的NN。
- 原子特征是不同阶张量组成的。特征张量的索引 l, p, a, c, m 分别是
  - l：旋转阶数，非负整数
  - p：取(1, -1)，和l共同标记 O(3)不可约表示(irreducible representations)
  - a：原子
  - c：通道（特征向量元素）
  - m：[-l, l]的表示索引
- 对这些由几何对象表示的特征向量 进行的卷积是等变函数（即满足旋转等变/奇偶等变，l层进行等变trans，l->l+1层的结果会对应等变）。且卷积本身满足平移不变（因为滤波器作用于**相对**原子间距离向量）。
- 此外，它们是排列不变的，因为来自不同原子的贡献之和对于这些原子的排列来说，是不变的。

SphereNet

- 原子在3D图中的位置由距离、角度、二面角唯一确定。

等变Transformer

- 嵌入层把 原子类型Z和邻域编码成特征向量x_i。
- 更新层用 修改过的多头注意力机制 计算原子对间的相互作用 以更新隐原子特征表达。
- 输出层用 门控等变block 计算标量原子预测 这些预测被聚合成单个分子预测。

CGCNN

- 直接在晶体结构生成的晶体图之上构建卷积神经网络。
- 在DFT计算数据上进行训练，CNN会自动获取适合预测特定属性的表示。
- 第一个 卷积+隐层 用来更新晶体图的每个节点。
- 第二个 池化+隐层 将整个晶体的向量 映射到 预测输出。

Matformer

- 周期图由3D空间上的常规晶格的最小细胞单元重复组成。
- 除了E(3) 还要满足周期不变性，学习到的表示不能因细胞边界改变而改变，还要能捕获重复模式。
    - 周期性不变性有两个方面：学习到的表示对于最小可重复单位单元的放大和周期性边界的移动应该是不变的。
- 用多边图来表示两个原子在可能出现位置的全部连接。以捕获cell边界相互作用力。

PotNet

- 对所有原子之间的完整势能集进行建模，将对完整原子间势的计算纳入消息传递神经网络中以进行表示学习。
- 关键挑战是捕获因单位晶胞重复而产生的无限范围原子间相互作用。
- 建模原子间势并直接作为边特征！建模所有原子的势集而不是邻居原子间的！
- Matformer还是距离建模，PotNet直接用原子间势建模。

Wasserstein GAN

- 目标是在笛卡尔坐标空间生成分子结构。通过构造体系结构，使生成的欧几里得距离矩阵天然具备旋转和平移不变性。
- Wasserstein GAN，排列不变 critic网络。
- 一次性生成分子结构

Cond-DFC-VAE

- 基于自动编码器的生成深度表示学习管道，用于几何优化的 3D 晶体结构，可同时预测八个目标属性的值。该系统具有高度通用性。
- 我们训练了条件深度特征一致的变分自动编码器和 UNet 分割网络，以学习立方二元合金、三元钙钛矿和 Heusler 化合物的表示。
- 通过使用条件自动编码器，对电子密度图以及相关晶体的每个原子的形成能进行编码，VAE 学会同时对结构和属性进行编码。
- 因此，根据用户定义的形成能条件，从编码空间中随机采样，会产生晶体结构的新示例。
- 此外，对于每个生成的晶体，我们使用 GNN 预测八个相关属性。我们通过将 VAE 生成的结构和 GNN 生成的预测与通过电子结构计算计算出的预测进行比较来验证它们。

欧几里得神经网络直接预测声子态密度

- 仅使用原子种类和位置作为输入来演示声子态密度（DOS）的直接预测。应用欧几里得神经网络，其构造与 3D 旋转、平移和反转等变，从而捕获完整的晶体对称性。通过预测大量高声子比热容材料证明了该网络的潜力。该工作提出了一种探索材料声子结构的有效方法。
- 我们采用欧几里得神经网络 (E(3)NN)，它自然地对 3D 几何进行操作，并且与 3D 平移、旋转和反转等价。E(3)NN 保留输入的所有几何信息，并消除了昂贵的（大约 500 倍）数据增强的需要。此外，网络保留了输入数据的所有晶体学对称性。
- E(3)NN，又叫**e3nn**。
- 欧几里得神经网络可以应用于预测晶体固体中更广泛的特性。
- 仅包含 1200 个示例的小型训练集足以生成有意义的预测，即使在数据增强（支持信息）的情况下，其性能也优于训练有素的卷积神经网络。

VGNN

- 是**e3nn**的加强。
- VVN是最简单的 VGNN，它采用m个原子的晶体结构并输出3m个分支声子能量
  - 最终转换为声子能带结果的向量虚拟节点。3m个虚拟节点。
  - 对于每晶胞有 m 个原子的材料，有 3m 个声子能带，添加虚拟原子的一种明智选择是添加 3m 个虚拟原子，每个虚拟原子输出其中一个能带的预测。
- MVN 是一种更复杂的 VGNN，它对复杂材料显示出更高的精度，但计算成本稍高。（动力矩阵）
  - 对m个原子的晶胞，创建m^2个虚拟节点，V_{ij}捕获真实原子i和j的连接本质。消息传递后V_{ij}成为3*3特征。总共3m*3m。
  - 3是指两个真实原子间三维谐波相互作用。
- k-MVN 是一种 VGNN，可以预测布里渊区任意 k 点处的完整声子能带结构。（动量矩阵）
  - 考虑到晶胞平移，对每一个邻居晶胞都创建一组MVN虚拟节点，t个邻居晶胞就是tm^2个虚拟节点。

MTP

- 在这项工作中，我们提出了一种计算高效且准确的方法来获取 PDR 并在被动拟合矩张量势（**MTP**）的基础上探索其他关键声子特性 [24]。
- 我们在被动训练的 MTP 的基础上描述了我们的方法，该方法可以方便地用于探索低对称性和纳米多孔二维材料的声子特性，并具有增强的稳定性和计算效率。
- 最小化能、力、应力的 AIMD和MTP MAE。能是各个中心原子能的和，采用邻域阶段。
- 输入是原子类型和距离。用MLIP的一种方法MTP构造中心原子的邻域公式，聚合成能。

MLIPs

- PES被描述为局部环境描述符的函数，这些描述符对于同核原子的平移、旋转和排列具有不变性。

- 这些势包括高维神经网络势（NNP），高斯近似势（GAP），谱邻域分析势（SNAP），和矩张量势（MTP）等。

GAN for Crystal Structure Prediction

- 在这项工作中，我们采用基于晶胞和分数原子坐标的免反演晶体表示，并为晶体结构构建了生成对抗网络。
- 该模型用于生成 Mg−Mn−O 三元材料，并对其光电阳极性能进行高通量虚拟筛选（HTVS）性能的理论评估。所提出的生成 HTVS 框架预测了 23 种具有合理计算稳定性和带隙的新晶体结构。
- 开发晶体结构生成模型的一项重要任务就是构建可逆表示，是从材料表示到材料真实结构的可逆性。要能从embedding生成材料结构。
- 3D体素表示要克服的挑战：(1) 将表示反转为材料结构需要用户定义的后处理。 (2)晶体材料的晶胞尺寸受到立方尺度三维网格的限制。 (3)表示是内存密集型的，导致训练时间长。最后，（4）图像本质上不是平移、旋转和超单元不变的。
- 将晶体结构表示为一组原子坐标和细胞参数
- 采用的生成式 HTVS 预测了 23 种新颖的 Mg−Mn−O 结构作为潜在的光电阳极
- 排列不变性是由网络编码中使用的对称操作强加的。
- 生成的GAN由生成器、分类器和批评器组成。
    - 生成器将随机高斯噪声向量 (Z) 和 one-hot 编码合成向量 (Cgen) 作为输入来生成新的 2D 表示。
    - 批评器计算了 Wasserstein 距离，它代表了真实数据分布和训练数据分布之间的差异，并且通过减少这个距离，生成器将生成更真实的材料。
    - 分类器网络从输入的二维表示输出合成向量，用于确保生成的新材料满足给定的合成条件。仅当生成的 2D 表示 ($\tilde{x}$) 作为输入时，分类器的损失才会反向传播到生成器。

FTCP

- 一个能进行一般逆设计的框架（不限于给定的一组元素或晶体结构）。具有在实空间和倒易空间中编码晶体的广义可逆表示，以及来自变分自动编码器的属性结构化潜在空间。
- 我们对“一般逆向设计”的定义是能够根据用户指定的目标属性（或多个属性）产生特定材料（化学和结构）的预测，即解决属性预测的逆问题。
- 假设材料表示必须包含两个关键元素：（1）将结构和化学成分纳入描述符的表示（结构和成分都不同）），以及（2）可逆的表示，使其适合解决反演问题（性质→结构+化学）。
- 允许组合和结构发生变化，从而实现通用和属性驱动的逆向设计
- FTCP = 广义可逆表示 + 倒易空间特征化器 + VAE + 目标学习分支 = 傅里叶变换晶体特性框架
    - 广义可逆表示 = 实空间特征（晶体结构信息） + 倒易空间特征（结构因子/傅里叶变换特征/附加特征器）
        - concatenate horizontally
    - VAE with 额外的连接到潜在空间的目标学习分支。编码器将训练集中的晶体编码到连续的概率潜在空间中，解码器将潜在空间中的任何向量解码到其对应的晶体（对潜在空间进行采样以设计新的晶体）。
- 目标学习分支将（与VAE）协同组织 潜在空间 以反映属性的连续变化。潜在空间又叫“属性结构化”

- 实空间包含类似CIF的特征，例如元素矩阵（描述组成元素）、晶格矩阵（描述晶格参数）、位点坐标矩阵（描述位点的分数坐标）、位点占用矩阵（描述每个元素的占用率）站点），以及元素属性矩阵（元素描述符）。
- 倒易空间特征沿着不同的空间频率投影晶胞中所有 N 个位点的元素描述符，米勒指数通过空间离散傅里叶变换形成 FTCP 矩阵。每个 k 点距 (000) 的距离也被记录并添加到 FTCP 矩阵中。
- 在普通 VAE 的编码器 + 解码器架构之上，潜在空间还连接到用于属性映射的目标学习分支，反映属性梯度（属性结构的潜在空间）。
- VAE的整体损失是三个损失的加权和，即重构损失、KL损失和属性映射损失。P8
    - 前两者是标准的VAE损失，保证了潜在空间的连续性
- 为了从经过训练的属性结构潜在空间中进行采样，我们采用了局部扰动（Lp）策略。我们在训练集中识别满足目标的晶体的潜在点，并以缩放（0.3-3）单位高斯噪声在潜在点附近采样。
- 然后把采样的属性结构样本经过后处理的到重构的FTCP表示。
- 工作流的四个阶段：（1）我们定义预期材料的目标属性，（2）我们根据训练模型进行设计（执行采样和后处理）并获得一些 FTCP 设计的候选，（3）我们通过结构传递这些候选弛豫（在我们的研究中由 DFT 执行）。（4）我们执行第一性原理计算来验证设计候选的属性，并在用户指定的误差范围（容差）内保留满足设计目标的属性。
- 定义三个指标：有效率、成功率和相对随机的改进。
- 形成能目标基于三元晶体、带隙目标基于三元和四元晶体、TE功率因子基于三元和四元目标。
- ICSD分数

CDVAE

- 晶体扩散变分自动编码器（CDVAE），它可以捕获材料**稳定性的物理感应偏差**。通过学习稳定材料的数据分布，解码器在扩散过程中生成材料，将原子坐标移向较低的能态并更新原子类型以**满足邻居之间的键合偏好**。
- 我们的模型还**显式编码跨周期边界的交互**，并尊重**排列、平移、旋转和周期不变性**。我们在三个任务中显着优于过去的方法：1）重建输入结构，2）生成有效、多样化和现实的材料，3）生成优化特定属性的材料。我们还为更广泛的机器学习社区提供了**几个标准数据集和评估指标**。
- **稳定材料**仅存在于所有可能的原子周期排列的低维子空间中：1）原子坐标必须位于量子力学（QM）定义的局部能量最小值内； 2) 整体稳定性还要求结构遵循不同原子类型之间复杂但特定的键合偏好
- 数据分布中所有材料都是稳定的，如果将噪声添加到地面实况结构中，将其降噪回其原始结构可能会增加稳定性。通过设计噪声条件评分网络（NCSN）作为我们的解码器来捕捉这一见解：1）解码器输出将原子坐标驱动到能量局部最小值的梯度； 2) 它还根据邻居更新原子类型，以捕获特定的局部键合偏好
- 为了捕获必要的不变性并对跨越周期性边界的相互作用进行编码，我们使用具有周期性的 SE(3) 等变图神经网络 (PGNN) 作为 VAE 的编码器和解码器。
- 将材料视为3D体素图像，但将图像解码回原子类型和坐标的过程通常会导致有效性较低，并且模型不是旋转不变的（Hoffmann 等人，2019；Noh 等人，2019；Court 等人，2020；Long 等人，2021）。第二个直接编码原子坐标，并将晶格作为向量（Ren et al, 2020; Kim et al, 2020; Chao et al, 2021），但模型通常不具有不变性任何欧几里得变换。另一种相关方法是从 QM 力训练力场，然后应用学习的力场通过最小化能量来生成稳定的材料（Deringer 等人，2018；Chen & Ong，2022）。这种方法在概念上类似于我们的解码器，但它需要额外的力数据，而获取这些数据的成本很高。

- 通过两步过程生成新材料：1）我们从潜在空间中采样 z 并用它来预测材料的 3 个聚合属性：成分 (c)、晶格 (L) 和原子数量 (N)，然后用于随机初始化材料结构 $\backsim$M = ($\backsim$A, $\backsim$X, L)。 2）我们执行 Langevin动力学，以z为条件同时对x和A进行去噪，以提高M的局部和全局稳定性并生成新材料的最终结构。
- 并行优化三个网络： 1）周期性 GNN 编码器 $PGNN_{ENC}$(M)，将 M 编码为潜在表示 z。2）属性预测器MLP_{AGG}预测c/L/N。3）周期性 GNN 解码器 $PGNN_{DEC}$(M|z)，对X和A进行去噪。最小化三个loss的组合。
- 为了捕获跨周期边界的相互作用，我们对 M 和 ~M 采用多图表示（第 3.1 节）。我们还使用适应周期性的 SE(3) 等变 GNN 作为编码器和解码器，以确保模型的排列、平移、旋转和周期性不变性

SyMat

- 可以捕获周期性材料结构的**物理对称性**。 SyMat 通过使用变分自动编码器模型生成原子类型集、晶格长度和晶格角度来生成材料的原子类型和晶格。此外，SyMat采用**基于分数的扩散模型**来生成材料的原子坐标，其中在坐标扩散过程中使用了一种新颖的对称感知概率模型。
- SyMat 在随机生成和属性优化任务上取得了良好的性能。
- SyMat 将材料的原子类型和晶格转换为对称感知生成目标，这些目标由变分自动编码器模型生成 [22]。此外，SyMat 采用新颖的对称感知概率扩散过程，通过基于分数的扩散模型生成原子坐标
- SyMat 方法首先用VAE模型生成原子类型向量 A 和晶格矩阵 L，然后通过基于分数的扩散模型来生成以 A 和 L 为条件的坐标矩阵 P
- 将A和L变换为对对称变换不变的项，并使这些项成为VAE模型的直接生成目标。
- 模型的主体是SphereNet
- 训练得分模型后，我们可以使用它通过退火 Langevin 动态采样算法 [49] 根据给定的 A 和 L 生成坐标矩阵 P。

DeepStruc

- 分布函数是一个统计物理学的概念，表示在一个原子附近的径向原子分布状态，一般有三种形式：径向分布函数（Radial Distribution Function，RDF）、双体分布函数g(r)（Pair Distribution Function，PDF又称对分布函数）以及约化径向分布函数G(r)。
- 使用Cond VAE 从（总散射数据获得的）对分布函数中求解简单的单金属纳米颗粒结构。
- deepstrucc是生成的，这意味着它可以用来构造不在训练集中的结构，即从PDF中求解结构。
- 数据集：ASE构建的3742个独特单金属纳米颗粒结构，$Au_{144}(p\text{-}MBA)_{60}$纳米颗粒/1.8nm的Pt纳米颗粒/2.2/3.4纳米nm的颗粒/$Au_{144}(PET)_{60}$纳米颗粒。
- 预测结构和实际结构间的MAE。

NMR

- 将给定内壳层电子激发产生的光电子峰位移称为化学位移，凭该位移可以断定元素组成/含量/化学状态/分子结构/化学键等关键信息
- 用等变GNN预测全化学位移张量 各向同性化学位移（标量）。
    - TFN和e3nn
- Si(硅) NMR数据集

利用中子散射数据预测材料结构

- 利用中子散射剖面数据进行材料结构预测
- 晶体学对称群 分类器
- 晶体材料的特征在于称为晶胞的基本单元的平移周期性
- 输入包括布拉格轮廓和收集衍射数据的仪器的描述。分类器使用此信息来预测其晶体学类别（七个晶体类别之一）。然后，回归器使用预测的类别以及布拉格衍射图案来预测材料的晶格参数和键角。
- 训练集是由GSAS-II软件生成的每种晶格参数下的衍射图案集，约67GB。
- 橡树岭国家实验室散裂中子源的纳米级有序材料衍射仪 NOMAD 获得的实验注释衍射数据进行验证。
- 分类器是CNN。回归器选择了两种：非负最小二乘回归和随机森林回归。

用于大型 X 射线衍射数据集中可视化和新颖性检测的深度学习

- 用VAE在X射线衍射(XRD)数据集上做VAE潜在空间和XRD图谱的可视化。
- VAE通过X射线衍射数据 学习晶体结构表示。
- 首先，我们在先验中识别学习到的相似性和潜在退化解的区域。接下来，我们开发了一个用于实验XRD图谱和VAE潜在空间的动态可视化工具。可视化潜在空间可以在实验期间进行新颖性检测
- 15000个一维XRD图谱的合成数据集
- 潜在空间是根据晶体结构的主反射轴组织的

CrystalMELA

- 晶体系统分类。该平台提供了两种不同的ML模型，**随机森林和卷积神经网络，以及文献中提供的极度随机树算法。**ML模型是从POW_COD数据库中收集的有机、无机和金属-有机化合物和矿物的28万多个已发表晶体结构的模拟粉末X射线衍射图中学习的。
- 分类 accuray、Top-2 accuracy、Balance accuracy、F1 score、confusion matrix、Precision、Recall

数据集们

- OC20 dataset、OE62
- The Materials Project
- Open Quantum Materials Database
- Novel Materials Database (NoMaD),
- Inorganic Crystal Structure Database (ICSD),
- Cambridge Crystal Structure Data(CCSD),
- and the Crystallography Open Database(COD)
- Perov-5
- Carbon-24
- QMDB
- MP-20

- QM9
- MD17
- GDB
- 3BPA（Learning local equivariant representations for large-scale atomistic dynamics）
- Li3P04
- Ag

Materials Project-2018.6.1 [Chen et al. 2019b]、JARVIS [Choudhary 和 DeCost 2021] 和 MatBench

Perov-5、Carbon-24和MP-20

Materials Project [Jain et al. 2013]、无机晶体结构数据库（ICSD）[D et al. 2019]和剑桥晶体数据库（CSD）

PDF数据集、Venetos等人[2023]使用了由Sun等人[2020]计算的放松结构的一部分从头计算的NMR化学位移张量子集

预测属性

- 回归任务：MAE，MSE，RMSE，MAD
- 分类任务：ROC，AUC

CGCNN

- **晶体**

> MP

- 形成能、绝对能量、带隙、费米能、体积模量、剪切模量、泊松比
- 在钙钛矿数据集上做的消融实验

Schnet

- 晶体和分子

> QM9

- E_{HOMO}、 E_{LUMO}、 能隙gap、ZPVE零点振动能量 、 polarizability极化率α、electronic spatial extent $<R^2>$电子空间范围、0K内能U0、内能U、焓H、定容热熔C_v、 吉布斯自由能G、 μ电偶极矩。

> MP

- 形成能、金属/非金属分类

> MD17、C20富勒烯的分子动力学模拟

- 势能面（总能量）、力场（原子力）

MEGNet

- 晶体和分子

> QM9 Faber等人处理的

- Schnet的QM9指标 + w_1最高振动频率。

> MP

- 形成能、带隙、体积模量、剪切模量、金属/非金属分类

GATGNN

- 晶体

> MP

- 形成能、绝对能量、带隙、费米能、体积模量、剪切模量、泊松比

ALIGNN

- 分子和晶体
- 认为高 MAD:MAE ratio 是良好的预测模型。

> MP

- 形成能、带隙

> JARVIS-DFT

- 形成能、带隙、总能量、凸包上方能量 (ehull)、Vogit体积(Kv)、剪切模量 (Gv)、太阳能电池效率（光谱有限最大效率，SLME）、拓扑自旋轨道溢出(Spillage)、介电常数(ϵx(DFPT)、ϵx(OPT, MBJ))、二维剥离能、电场梯度 (EFG)、最大压电应力 (e_{ij})、最大应变 (d_{ij}) 张量、n型和p型Seebeck系数和功率因数 (PF)、电子 (m_e) 和空穴 (m_h) 有效质量的晶体学平均值、平面波截止、K点长度、磁矩(magnetic moments)。

> QM9

- E_{HOMO}、 E_{LUMO}、 能隙gap、ZPVE零点振动能量 、 polarizability极化率α、electronic spatial extent $<R^2>$电子空间范围、0K内能U0、内能U、焓H、 吉布斯自由能G、 μ电偶极矩

M3GNet

> MP && single-element dataset from Zuo et al &&

- 能、力、应力

> DFT计算

- 声子态密度中心数据、声子色散曲线、德拜温度

> ICSD

- 松弛结构的能(E_{hull})

CHGNet

> Materials Project Trajectory Dataset

- 能、力、磁矩、应力

零碎知识

> 学术干货 | 机器学习对材料数据库的发展与优化 - 知乎 (zhihu.com)

对于材料数据库中的性质，主要存在两个问题，第一，性质算得不够准，现有的大型的计算材料数据库主由一些比较低价的计算方法来构建，比如PBE泛函或者经典动力学，其代表性质是带隙，PBE用来计算带隙时会产生很大的系统误差。所以我们要让材料数据库里面的性质更准。第二，人们想要的一些性质尚未被收录到材料数据库中，这些性质通常需要昂贵的计算或实验来获得，比如材料和其他物质的相互作用以及近期热门的输运性质。对于这类问题，我们需要思考：怎样更高效地构建全新的数据集。

我们一方面希望通过机器学习加速的实验和计算来获取更多数据，另一方面我们希望通过主动学习或者贝叶斯优化的方法来获取最有价值的数据。

主动学习指的是在学习的过程中，先有一个较小的有标签的训练集和一个比较大的没有标签的数据库，我们通过学习已有的训练集，用模型从大的数据库中找到对于提高学习模型最有用的数据，然后拿出来通过实验或者计算得到标签，再放回训练集里，学习这个新的训练集，以此形成一个循环。

除获取新数据之外，我们还可以利用已有的低质量数据集实现所谓的信息传递，比如迁移学习、多精度机器学习的方法，来降低我们对高质量数据集的需求。迁移学习可以简单理解为先用大数据集训练一个模型，这样这个模型里就会包含一些大数据集里学到的知识，然后我们把这些知识转移到针对小数据集的学习模型中。

另一个方法是多精度机器学习，这里有两种思路：第一种是把低精度的数据直接当作输入特征来训练模型，比如在Chen et al.等人的工作中，直接将不同精度的带隙放到卷积神经网络的初始化当中7。第二种思路是调整我们学习的目标，把学习的目标从高精度的数据换成低精度和高精度数据的差值, 通过学差值来提高机器学习的表现，例如The Δ-Machine Learning Approach8。

我们希望通过一些物理上的知识来帮助对材料的机器学习。

> (17 封私信 / 85 条消息) 计算化学领域有哪些著名的数据库，可用于机器学习的模型训练？ - 知乎 (zhihu.com)

对于训练势能面的数据集来说，上古时期主要有两类，一类是GDB数据集的子集，图论工具枚举得到结构，算DFT，仅有稳定的有机小分子。一类是同一分子的不同构象，ANI-1。active learning和transfer learning = ANI-1x和ANI-1ccx

局限性在于，数据集仅包含稳定的有机小分子，但无数研究方向都涉及晶体、表面、离子、过渡态、自由基、大分子和各种强弱相互作用。于是，从真实的MD模拟中采样的方法迅速崛起。这里dataset大多是即时的。

> DP能干这么多？DP在材料科学中的应用综述 - 知乎 (zhihu.com)

量子力学方法准确度高，但效率很低；经验原子间势函数效率高，但是通常准确度有限。

> 电子结构理论（二）计算方法的出现、发展和现状 - 知乎 (zhihu.com)

目前大多数物质电子结构理论和计算方法研究的基础：

- 电子基态的**密度泛函理论（density functional theory）**以及它向激发态的扩展

- **量子蒙特卡洛方法（quantum Monte Carlo methods）**，可以直接处理电子和原子核的相互作用的多体系统
- 电子系统的激发谱的**多体微扰方法（many-body perturbation methods）**
- **计算机技术**，使得实际的计算变得可行，并反过来促进了计算机技术的发展

材料性质

- 力学性能、热学性能、光学性能、电学性能、磁学性能、声学性能、放射性性能、表面性能等。
    1. 力学性能：指材料在外力作用下所表现出来的性质。它包括材料的弹性模量、屈服强度、拉伸强度、压缩强度、弯曲强度、冲击韧性等。
    2. 热学性能：指材料在热作用下的性质。它包括材料的热容、热膨胀系数、热导率、热扩散系数等。
    3. 电学性能：指材料在电场作用下的性质。它包括材料的电阻率、电导率、介电常数等。
    4. 磁学性能：指材料在磁场作用下的性质。它包括材料的磁导率、磁感应强度等。
    5. 表面性能：指材料表面在各种不同环境条件下，所表现出来的性质。它包括材料的表面张力、表面粗糙度等。
    6. 光学性能：指材料在光的作用下的性质。它包括材料的折射率、反射率、透光率等。

材料表示学习

- 材料表示学习的目标是学习一个函数f来预测任何给定材料M的属y，y可以是实数（回归问题）或分类数（分类问题）。晶体的材料表示学习旨在根据晶格结构预测晶体材料的物理和化学性质。
- 主要挑战在于捕获晶体材料的对称性，特别是周期性变换的不变性。
    - 包括：晶胞的E(3)不变性、周期不变、周期性模式捕获。

- 晶体图表示：CGCNN提出通过基于晶胞结构构建的多边晶体图来表示无限晶体结构，该结构被证实为周期不变的。此外，使用成对欧几里得距离可以实现对E(3) 变换不变的多边晶体图。ALIGNN提议进一步包括键角而不破坏周期不变性，M3GNet和CHGNet以类似的方式包含了三体（three-body）交互。为了明确地捕获周期性模式，Matforme提出使用六个自连接边对三个周期向量进行编码，其组合可以完全捕获周期向量的长度和它们之间的角度。最近的两项工作利用Ewald求和来捕获周期性信息。 PotNet提出通过使用多种类型势的无限求和来考虑晶体结构中任意两个节点之间的无限相互作用，从而通过这些求和捕获周期性模式。同样，Ewald-MP通过应用Ewald求和的分解来解决周期性结构中的长程相互作用问题。
- 晶体图神经网络：现代图神经网络在预测材料特性方面的有效性依赖于它们所包含的对称性和高阶信息以及所使用的消息传递方式。现有网络，如 CGCNN、MEGNET、GATGNN和SchNet采用半径晶体图，并在消息传递过程中仅考虑两体距离作为边缘特征。此外，Matformer、PotNet和Ewald-MP在消息传递过程中引入了附加功能，以解决先前方法缺乏周期性信息的限制。

材料生成

- 材料生成领域的目标是通过生成模型学习材料空间上的概率分布p，并从p中采样新颖的晶体材料。
- 关键的挑战是在生成模型的概率建模框架中实现所有对称变换的不变性，即将晶体材料的所有对称性纳入生成模型中。
    - 换句话说，如果两种晶体材料 M 和 M' 可以通过对称变换相互转移，则生成模型应该将它们分配给相同的概率
    - 对称变换包括：E(3)变换和周期变换（晶体材料）。

- 使用两种策略来确保周期性变换的不变性。
- 首先，我们可以通过生成内部对周期性变换不变的3D特征或表示（例如3D体素网格）来隐式确定材料中的原子位置。使用该策略的两种代表性方法是Cond-DFC-VAE和Hoffmann等人提出的方法。它们都将晶体材料转换为3D体素网格，将平滑3D体素网格转换为3D密度图，并使用3D密度图作为生成目标。3D体素网格或3D密度图对于周期性变换具有不变性，但对于E(3) 变换不是不变的，因此两种方法都无法捕获E(3)对称性。
- 另一种策略是直接生成晶格矩阵L和坐标矩阵C，但生成模型需要定制概率建模，使得对于任何整数矩阵K，p(C) = p(C+LK)始终成立。两种早期方法：FTCP和Kim等人提出的生成晶格参数和分数坐标矩阵。很容易地证明晶格参数和分数坐标矩阵对于旋转和反射变换是不变的。然而，分数坐标矩阵假定原子之间有一定的顺序，因此它们不是排列不变的，并且它们对于平移和周期变换也不是不变的。
- 在第二种策略之上，最近的两种方法（CDVAE和SyMat）不直接生成分数坐标矩阵，而是建议以3D图的形式生成晶体材料。他们使用VAE模型生成上述晶格参数，随机初始化原子坐标，并通过分数匹配模型迭代细化原子坐标。在这两种方法中，原子坐标细化都是通过E3NN完成的，这确保了对排列和周期性变换的不变性。不同的是，CDVAE直接将分数匹配应用于原子坐标，无法实现平移不变。而SyMat 将分数匹配应用于原子之间的成对距离，以实现所有的E(3)不变性。

声子计算

- 声子计算的目标是：给定晶体结构（原子位置、原子种类和晶胞的晶格向量），以计算声子性质。性质包括诸如声子色散关系（声子能带结构），声子态密度等。在预测声子态密度、声子色散关系和其他声子计算时，图神经网络模型通常没有原子间作用力的先验知识，并且除了原子位置、质量和原子种类之外还使用其他特征。而在机器学习原子间势（MLIP）中，模型是在从头开始分子动力学轨迹上进行训练的。
- 关键挑战：传统的DFT方法面临精度或计算成本方面的限制。材料科学机器学习的一些最新进展为材料研究提出了一种新范式，然而，原子系统的3D性质以及晶体材料的周期性和对称性使常规神经网络的PDOS学习变得非常复杂，因为这需要昂贵的数据增强来学习3D坐标系的不同旋转和平移。此外，真实空间和倒易空间中的属性之间的差异，以及输出属性的长度要求可变等问题，导致材料表示和神经网络设计仍存在挑战。

- 局部环境描述符：利用机器学习原子间势 (MLIP) 来训练计算高效的从头算分子动力学轨迹，为 DFT 模拟提供了一种替代且有效的方法。在MLIP中，势能面被描述为局部环境描述符的函数，该描述符对于同核原子的旋转、平移和反转具有不变性。进一步的，矩张量势 （MTP）可以近似任何原子间相互作用，因此能够评估与密度泛函微扰理论方法相当的声子特性，且计算量不高。

- 几何图神经网络：基于图神经网络从3D原子位置、原子种类和原子间距离中学习材料属性。例如，用虚拟节点图神经网络（VGNN）来增强 GNN，该网络管理具有可变甚至任意维度的输出属性。Chen等人用E3NN捕获了PDOS的主要特征，用E3NN成功地再现了实验数据中的关键特征，并直接从原子结构预测了大量的高声子比热容材料，而无需昂贵的计算量。

材料表征

- 晶体结构的准确和有效的实验测定是材料领域更根本的挑战。为了确定材料的精确晶体结构，通常会结合使用仪器和表征技术，例如X射线和中子散射以及光谱学。
- 材料表征的目标分两类：晶体结构预测及其逆问题。
    - 晶体结构预测：使用实验表征技术的输出预测晶体结构参数。例如，用X射线衍射的一维光谱，来预测三维晶体结构（可以通过原子位置以及三个晶格向量或其他晶体结构参数来描述）。
    - 逆问题：用晶体结构预测表征方法的输出。
- 关键挑战：从散射图案重建3D晶体结构或相关参数的过程并非易事。如，相位问题：如果没有相位信息，衍射图案可能无法唯一地映射到晶体结构，反之亦然。此外，散射实验的数据可能会根据材料、样本质量以及实验人员提出的科学问题而有所不同。因此，开发可推广到实验数据和模拟数据的机器学习模型至关重要。对于从晶体结构重建散射数据的逆问题，等变性应该被保留，这就要求设计保持对称性和等变的机器学习方法。

- 光谱数据作为输入：大多数现有工作都使用光谱数据来对晶体对称性进行分类（晶类、布拉维晶格或空间群）或在回归问题中查找晶格参数。
    - Garcia-Cardona等人开发了一种CNN分类器，用于根据中子散射数据预测钙钛矿晶体系统和晶格参数。然后，他们为所研究的每个晶体学对称性训练随机森林回归模型，以预测晶格参数。Chituri等人假设晶体系统已知，并使用一维CNN来预测每个晶体系统的晶格参数。一维CNN能够预测晶胞长度，但无法预测单斜晶系或三斜晶系（较低对称类别）的晶胞角度。Corriero等人还使用CNN和随机森林来预测晶体系统和空间群。
    - Greasley和Hosein证明了更传统的监督学习算法，例如支持向量机和补朴素贝叶斯分类器，在多阶段识别方面的性能与神经网络一样好。其他小组也采用了变分自动编码器（VAE）架构。VAE的潜在空间是训练样本的“压缩”表示，因此这可以成为从散射数据中学习材料属性的有意义连续表示的有效策略。
    - 然而，这些方法都以一维光谱作为输入。例如，CNN假设输入数据具有平移不变性，这一特征可能不存在于这些光谱中。因此，在将机器学习模型应用于粉末光谱之前，应考虑其对称性和假设。此外，开发以更有意义的方式利用等变性的方法也是有价值的。
- 光谱数据作为输出：此类问题尚未使用机器学习技术进行广泛的研究。 Cheng等人开发了一种基于ML的框架，可以根据结构（原子坐标和元素种类）预测一维和二维非弹性中子散射 (INS) 谱。这项研究扩展自Chen等人2021年所做的工作（使用等变神经网络来预测态的声子密度）。