**论文阅读：Benchmark of Machine Learning Force Fields for**

|  |
| --- |
| 阅读论文：  **《 Benchmark of Machine Learning Force Fields for Semiconductor Simulations: Datasets, Metrics, and Comparative Analysis》** |

* 作者单位:

SAIT(三星高等技术研究院)

* 论文发表期刊

neurips

* 代码链接

https://github.com/SAITPublic/MLFF-Framework

**论文主要内容**

1. 引入了两个新的机器学习力场(MLFF)基准数据集SAMD23,反映了在各种情况下对SiN和HfO进行的半导体模拟。这为研究人员提供了标准的数据集来开发和评估MLFF模型在半导体材料模拟中的性能。
2. 提供了一个框架来促进MLFF模型的开发。该框架可能包含数据预处理、模型训练、性能评估等组件,方便研究人员快速开发和测试模型。
3. 为SiN和HfO提供了基准测试,并提出了五个模拟指标来评估MLFF模型在模拟中的预测性能和外推能力。这些指标为全面评价模型在半导体材料模拟中的表现提供了标准。
4. 通过对10个MLFF模型进行比较分析,提出了一个基线训练方案和模型选择策略,以在实际模拟中使用性能最优的模型。这为研究人员在众多MLFF模型中选择最适合特定半导体模拟任务的模型提供了指导。

**数据集生成方法**

论文中主要生成了SiN和HfO这两个数据集：

* SiN数据集:

1. 通过110个独立的DFT模拟生成,包括各种结构,如非晶态、晶体(α,β和γ)、一维或二维缺陷、表面和孤立的纳米团簇。
2. 模拟在不同温度、应变和系综下进行,每个单元中的原子数量范围从16到510。
3. SiN模拟包括各种SiN化学计量比以及仅含Si和仅含N的结构。
4. 数据集由这些模拟中以9 fs的采样间隔获得的样本组成。
5. 为了生成OOD数据集,使用非晶态SiN作为初始结构,并进行熔融-淬火-弛豫模拟。

* HfO数据集:

1. 使用改进的熔融-淬火-退火(m-MQA)方法创建,其中温度变化。
2. 为了获得高熵结构,在m-MQA方法中加入了三个5000 K极高温度下的预熔阶段。
3. 在五种HfO2晶体(即单斜、四方、立方和两种正交结构)以及七个随机生成的结构上,使用m-MQA方法进行了12个MD场景(产生60个模拟条件)。
4. 每个结构有96个原子(32个Hf和64个O),以保持HfO2化学计量比。
5. 在HfO2晶体的单元格内随机分布32个Hf和64个O原子,生成随机结构。
6. 在进行的12次模拟中,10个结构(五个晶体和五个随机结构)的模拟和其余两个随机结构分别用作ID和OOD数据集的来源。
7. 在三个预熔和熔化阶段,以9 fs的采样间隔从每个源采样快照,在淬火和退火阶段,采样间隔为12 fs。

**评测指标**

* 受力和能量

|  |
| --- |
| 原子能量和力的均方根误差（RMSE）之和，用 EF 指标表示 |

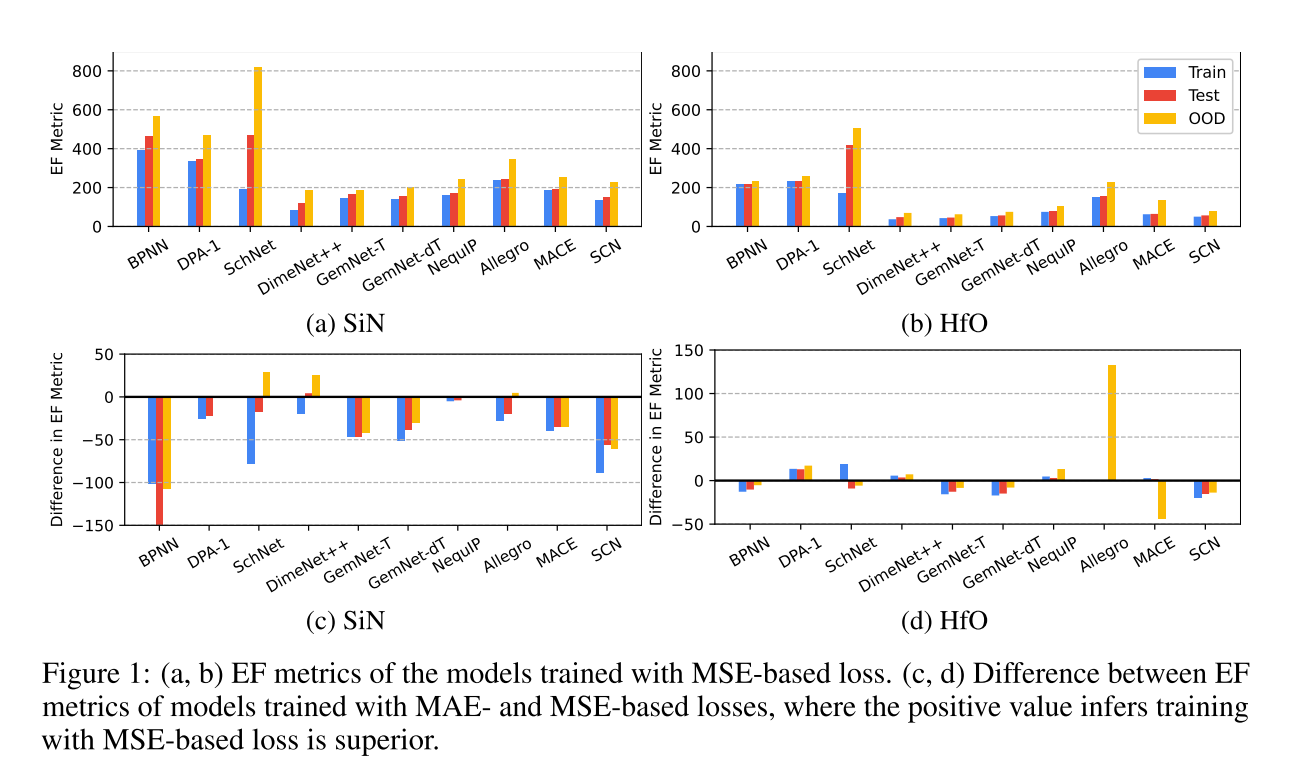
* 损失函数：

**Benchmark方法**

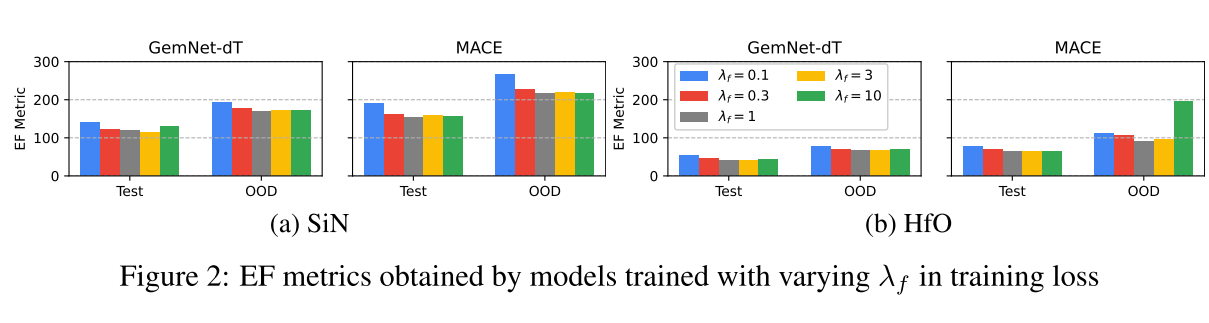
benchmark使用了以下10个模型进行训练测试（包括基于分子描述符的模型和基于图神经网络的模型）:

1. BPNN (Back-Propagation Neural Network)
2. DPA-1 ( Potential - Atomic One)
3. SchNet
4. DimeNet++
5. GemNet-T/-dT
6. NequIP
7. Allegro
8. MACE
9. SCN (Spectral Chemical Network)

在这个基准测试中,研究者探索了不同的归一化略对机器学习力场(MLFF)模型训练结果的影响。他们使用了三种不同的策略:每原子归一化、每原子中心化和不归一化。 通过比较使用这三种归一化策略训练的模型的错误,他们发现即使在OOD(out-of-distribution)错误方面,差异也很小。这表明,即使不采用归一化,MLFF模型中的内部表示也可以与每原子值很好地对齐,并用于模拟与训练数据集中不同大小的结构。



**两个训练因素对机器学习力场(MLFF)模型能的影响**



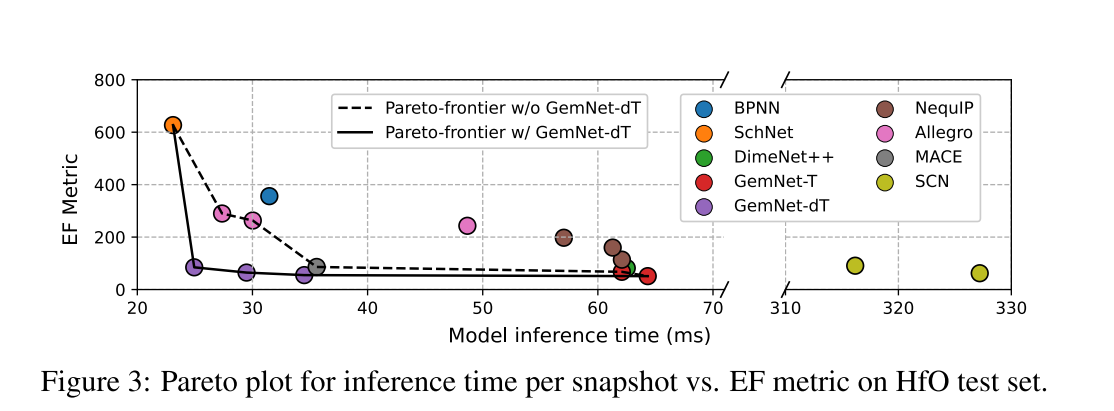
探讨了两个训练因素对机器学习力场(MLFF)模型能的影响:基于MSE和MAE的损失函数,以及力损失系数(λf)。

1. 基于MSE vs. 基于MAE的损失函数: 研究者使用基于MAE的损失函数训练模型,并将其与使用基于MSE损失函数训练的模型进行了比较。结果表明,对于SiN数据集,基于MAE的损失函数明显优于基于MSE的损失函数。然而,对于HfO数据集,优劣难以确定,大多数情况下,EF指标的差异较小。值得注意的是,Allegro在HfO的OOD(out-of-distribution)集上性能显著下降,而MACE的精度得到提高。
2. 力损失系数(λf)的影响: 对于所有模型,力误差在EF指标方面占主导地位。因此,如果通过改变力系数λf设计训练损失以减少力损失,测试力误差也会相应减少。为了研究λf的影响,研究者使用基于MAE的损失函数训练了GemNet-dT和MACE,其中λf的范围从0.1到10。结果表明,与λf=1相比,使用λf<1训练的模型误差增加。然而,MACE在HfO上的结果表明,增加λf并不总是导致力误差的减少,λf>10的结果可能会更差。因此,在超参数调整中探索合适的λf时,建议在1≤λf<10的范围内训练MLFF模型。

**推理时间和精度之间的权衡**

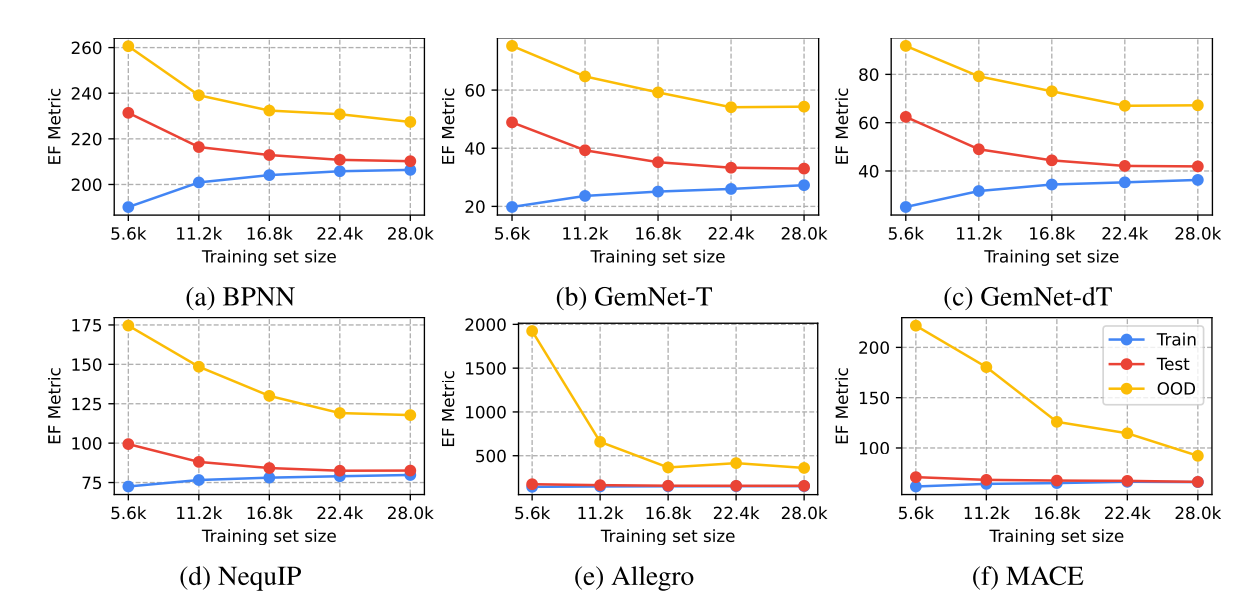
这一部分主要探讨了模型的推理时间和精度之间的权衡,以及训练数据量对模型性能的影响。

* 推理时与精度的权衡（模型大小的影响）: 研究者绘制了每个模型在HfO测试集上的推理时间和EF指标之间的帕累托图(Pareto plot),这两个因素呈现出权衡关系。他们在V100 GPU上使用批量大小为1的情况下对HfO测试集进行了五次推理,模拟了模拟过程中的逐步过程。为了进一步探索这种权衡,他们训练了特征维度不同的模型,相对变化范围从{0.25x, 0.5x, 2x}中选择。



结果表明,所有GemNet-dT模型都位于帕累托前沿曲线上。与能量类似,GemNet-dT仅通过GNN中的前向过程预测力,而不是在MLFF中的常规计算,即力是通过能量对原子位置的导数计算得出,以满足能量守恒定律。GemNet-dT的直接力预测不满足能量守恒定律,导致其在性能基准测试中的结果明显不如GemNet-T。然而,GemNet-T的推理时间是GemNet-dT的2.2倍,且需要更大的GPU内存,这意味着GemNet-dT更适合通过在更短的时间内处理更多原子来进行大规模模拟。因此,考虑到上述权衡,基于直接力预测的模型开发方法是一个很有前途的方向。尽管SchNet速度最快,但它产生的误差最高,因此可能会产生不可靠的模拟结果。同样采用直接力预测的SCN实现了与GemNet-dT相当的EF指标;然而,其计算需求高出10倍以上,使其不适合模拟。除了GemNet-dT,Allegro和MACE显示了由虚线表示的帕累托前沿结果。如果解决了测试集和OOD集之间EF指标的显著差异,它们可以用于需要满足能量守恒定律的模拟中。

* 数据量对模型性能的影响（数据集大小的影响）: 研究者还展示了数据量的影响,这可能有助于MLFF研究人员采取培训策略。他们从HfO训练集中随机抽取20%、40%、60%和80%的快照,并使用这些子集训练模型。结果表明,对于所有模型,训练集、测试集和OOD集的EF指标随着训练集大小的增加而降低。值得注意的是,当使用20%的训练数据时,大多数模型在训练集上的EF指标优于测试集,这表明过拟合。然而,即使只使用20%的训练数据,GemNet-dT和MACE在测试集上的性能也优于其他模型。

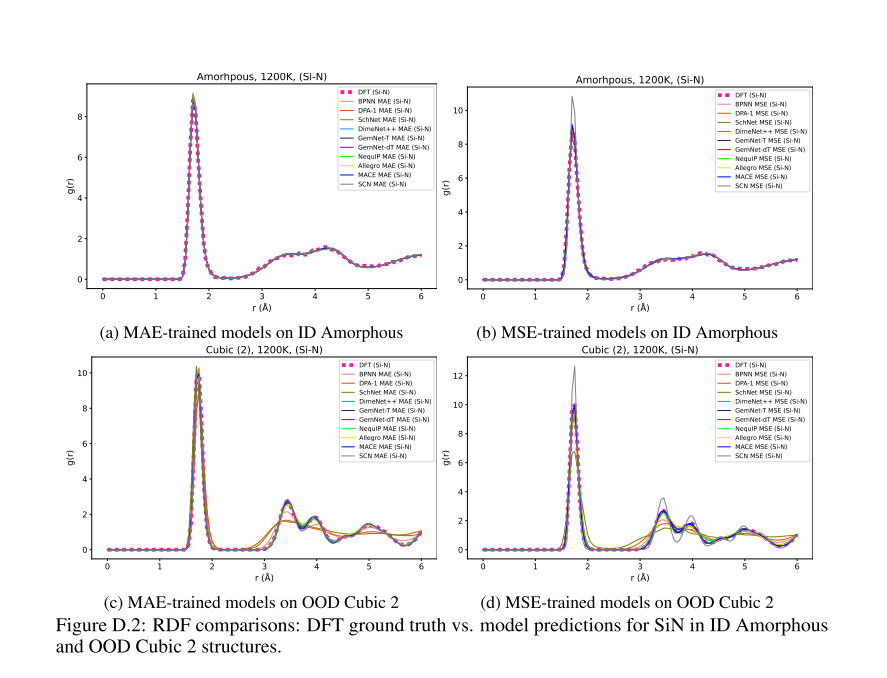


**预测材料性质方面的准确性和可靠性**

这一部分主要介绍了用于评估机器学习力场(MLFF)模型在预测材料性质方面的准确性和可靠性的一组指标,并讨论了基于这些指标的模型比较结果。

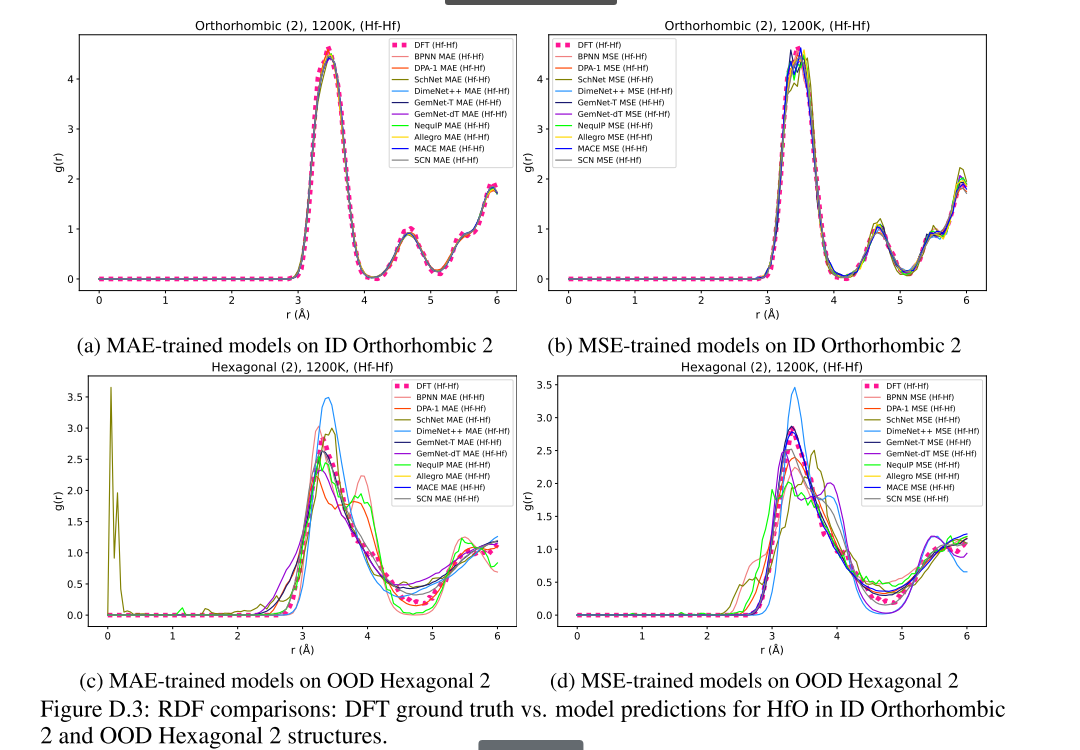
**动态指标:**

* 径向分布函数(RDF):捕捉相对于所选参考粒子距离的密度变化。
* 角度分布函数(ADF):表征围绕参考粒子的粒子的角度分布。 RDF和ADF是从高温分子动力学(MD)轨迹中得出的,用于评估模型在受高热能引起的湍流或活跃运动影响的不同原子环境中的稳定性。



**静态指标:**

* 体积模量(B0):表示薄膜对压力和外部应力的响应,表明其灵活性、硬度和抵抗体积变化的能力。
* 平衡体积(V0):有助于确定薄膜的厚度-残余应力关系,使其状态(应变或松弛)得到评估,并优化生长过程。
* 势能曲线(PEC):评估在稀疏原子环境中具有少量原子的系统,这与固态结构数据集中遇到的密集周期性边界条件不同。



* 结果与讨论:
* 总体而言,对于SiN和HfO,GemNet-T、NequIP、MACE和SCN取得了突出的结果。
* 基于描述符的模型BPNN和DPA-1在两个数据集上表现中等。
* SchNet在两个数据集的B0方面表现出显著的不匹配,并且泛化性能最差。
* DimeNet++和GemNet-dT在HfO的EF指标方面表现出色,但在OOD的模拟指标方面表现不一致,表明现有评估方法的局限性,强调需要与材料模拟和性质相关的评估指标。
* 与SiN相比,HfO只使用一个成分比(1:2)生成,因此评估OOD性能相对具有挑战性。
* OOD的EF指标与模拟指标的皮尔逊相关性高于ID评估的EF指标,表明模拟指标可以作为评估MLFF模型外推性能的指南。
* Allegro在HfOOOD评估中得分为0,原因是高温模拟过程中不真实的原子碰撞导致模拟中断。