**材料计算研究基础**

**背景**

**材料计算的背景（现有方法），加速材料计算的方法（各种尝试）**

如何提高新材料研发设计的效率，成为材料科学工作者的首要目标。人们先后采用了一系列手段来筛选具有期望属性值的材料：传统实验合成试错法，基于计算模拟的高通量计算方法和基于历史数据的统计学习预测方法。

传统实验合成试错法得到的实验结果真实可靠，但是存在着成本高、周期长、需要借助多年的人工经验这些显而易见的缺点。在材料理论模型与计算材料科学的不断发展的条件下，人们开始采用基于密度泛函理论（DFT）的第一性原理方法、分子动力学方法、蒙特卡洛方法等计算模拟方法的高通量计算手段。

高通量计算方法作为材料计算科学的主要手段极大地提高了材料的发现与发展的效率，然而，典型的高通量计算任务设计会面临材料空间大、计算耗时长的问题，因此，如何提高高通量计算的效率成为材料研究人员越来越感兴趣的话题。

随着高通量计算技术的广泛应用，出现了大量的可获得计算材料或预测材料属性的数据库（如ICSD、NOMAD、AFLOWLIB、CSD等）.受数据驱动原理启发，机器学习作为一种well-established的统计学技术逐渐被引入到材料数据的挖掘中，获取有价值的信息，来减少材料高通量计算的任务量，并发现了一些新的材料。

**方法**

**在机器学习方面，在原生计算方面**

MIP材料研究团队在材料信息学方面做了深入研究，在机器学习方面，探索了类金刚石的性能、群速度，能带与材料原子属性、化合物属性方面的关联关系。

另一方面，针对具体高通量计算上，我们发现典型的材料高通量计算任务计算周期比较长，而现有的材料机器学习应用仅仅是简单的筛选出具有期望属性的化合物，减少执行完整高通量计算的条目数，无法真正提高单个高通量计算任务的效率。考虑到一般的材料高通量计算任务往往需要经历若干个计算步骤，如结构驰豫优化，静态自洽，能带计算，输运计算等。这些计算步骤明显的具有顺序性、阶段性、有针对性的的特点，且在计算过程中，会产生大量的计算属性信息，如bader、bandgap等材料属性，以及执行计算任务本身带来的信息，如计算时间，收敛次数等。因此可以将机器学习技术引入到高通量计算的整个过程中，对高通量计算任务本身进行优化与指导，则有望可以直接为单个高通量计算任务减负。

**结果：**

**实现了band-gap预测来加速性能判断，实现了基于机器学习的高通量计算系统**

目前，我们已实现了材料大数据查重算法，过滤掉重复的材料体系，并实现了单组元单组分的材料数据掺杂、置换的技术。在机器学习方面，我们实现了能带，类金刚石化合群速度相关属性的机器学习预测工具，以提高材料人员的研究效率。此外，我们设计了一种自评价式高通量计算框架—SEHC，该框架为高通量计算架构引入了self-evaluation机制，可使得在高通量的计算过程中，通过预判断设定的指标信息来提前停止不符合期望的材料计算任务，减少无效的计算量。我们在254个化合物空间中寻找具有较高群速度的热电材料时，设计了具有两个带预测单元的Stage的Pipeline模型，并实现了基于SEHC的高通量计算框架原型。经试验验证，该新型高通量计算框架相对于普通高通量计算系统，加入了自评估方法，如机器学习方法，可以优先筛选出符合期望的体系进行计算，此外，其在使用了Pipeline模型以后，较普通机器学习预测方法，可以提高11%的预测准确率。通过在已有pipeline中替换新设计的high-throughput stage可以快速的实现新的研究目的的高通量计算系统原型，提高研发了高通量计算任务设计效率。