# 分子模拟的数据和模型的生成与处理

阿里云云服务协议说明

**研究背景：**

分子模拟是理论化学、计算物理、分子生物学、材料科学等领域的基本方法，也是化学工程、新材料设计、药物设计等应用行业的基础理论工具。分子模拟的效果取决于分子模型，然而长期以来这些领域都面临着一个核心困难：分子模型难以兼顾精度和效率——量子力学模型准而不快 ，经验力场模型快而不准。以新材料研发为例，使用量子力学模型，当体系稍微复杂时计算量即不可承受；使用经验力场方法，则是为了一个材料，一整个团队连续几年的试错，面临着巨大的不确定性，和不可复制性。当前药物发现和其他材料科学领域的科研工作也面临着类似的困境。

为解决这个核心困难，近年来，若干国内外团队探索使用机器学习方法进行分子建模。例如，英国剑桥大学的 G. Csanyi 团队使用 Bispectrum 和光滑交叠位置方法对原子环境进行描述，并以核方法作为拟合量子力学势能面。 德国哥廷根大学的 J. Behler 团队使用 Behler-Parrinello 对称函数对原子环境进行描述，以深度神经网络拟合量子力学势能面。 国内西安交通大学的丁向东课题组使用经验势作为模型，使用机器学习方法优化模型参数，实现了对金属锆的建模，等等。但是，这些研究组的工作并未能突破计算模拟到真实体系的差距。真实情况多尺度、多组分、结构复杂的特点需要非常通用的模型构造和非常高效的处理。

鄂维南团队发展的 Deep Potential 系列方法，成功实现了对量子力学数据的有效生成和充分利用，突破了打通大规模应用的最后一道壁垒 。该方法不仅广泛用于硅、锗等传统材料，而且应用于有机分子、多元合金、电池材料、陶瓷材料、催化剂等分子及材料体系的建模和模拟，在新材料、药物和化学合成等领域中展现了巨大的应用潜力。

在此基础上，本项目拟展开如下研究：

1. 利用 DP 系列方法对合金体系如 MgZn WH 等体系的合金量子力学数据进行有效建模。

2. 利用 DP 系列方法对药物小分子与蛋白质结合过程的自由能数据进行建模。
3. 对化学小分子的 SMILE 表达，以及图表达等非结构数据进行建模，进而探索分子生成新工具。

**项目相关性：**

对分子模拟过程中产生的量子力学数据以及化学分子这一特殊非结构化数据的数学认识，有助于进一步拓展深度学习的应用范围。在此过程中，项目的主要计算开销包括：

1. 批量合金体系的第一性原理数据生成
2. 合金体系数据有的有效势函数模型生成与优化
3. 基于化学小分子 SMILE 式或分子图式等非结构化数据的分子生成模型的训练与优化。

其中预计使用 GPU 资源：25000 卡时；CPU 资源 400 万核时。

**此次委托外包项目主要是为解决，该项目在数据生成和模型处理的计算资源来源问题。**

**项目合理性与经济性：**

计算资源一般有三种解决方案：

1. 通过采买机器进行集群的自行搭建。该方案存在初期基础投入大，可扩展性差以及后期管理运维成本高等问题。因此项目组决定不采取该方案。
2. 向超算中心进行机时采买。超算机时在机器的可扩展性方面，以及 GPU 型号、CPU 型号均符合本项目的要求。但项目组同时向多家超算运营商询价，得到的报价是 GPU：8 元/卡时。CPU:0.1 元/核时。若要达到项目所需的计算资源，总费用将不低于 60 万元。

3. 通过调用云度闲置计算资源。首先云端资源在可扩展性，以及 GPU、CPU 型号符合本项目要求。同时阿里云平台在常规云服务器租赁外，还提供 SPOT 竞价实例的机器使用方式。GPU 单价可低至 4 元/卡时，CPU 单价低至 0.04 元/核时。本项目预计总费用，将不超过 30 万元。

因此经过严格的调研与比价，本项目决定通过从阿里云采买闲置计算资源，为本项目的开展提供计算资源保障。