赛道 B：人工智能范式的物理化学家

物理和化学研究的对象日益复杂化、高维化，传统的研究范式主要是依赖于“穷举”、“试错”、“重复”等手段。面对庞大的化学空间，配方和工艺等各种参数的搜索常常止步于局部最优，无法进行配方和工艺参数的全局最优搜索。

图 1. 全球首个数据智能驱动的全流程机器化学家。

中国科技大学机器化学家平台实现了大数据与智能模型双驱动下的化学合成-表征-测试全流程开发，在软硬件方面已全面超过欧美同类装置，作为唯一装载了计算大脑、理论模型和开放式操作系统的智能平台，它具有更强的化学智能和广泛的化学品开发能力，目前已涵盖光催化与电催化材料、发光分子、导电分子、磁学、光学薄膜材料等，且适用范围将随平台升级和拓展继续扩大。

该平台可采用机器智能去查找和阅读文献，从海量研究数据中汲取专家经验，在前人知识与数据的基础上提出科学假说并制定实验方案；调度 2 台移动机器人和 15 个自主开发的智能化学工作站，完成高通量合成、表征、测试的化学实验全流程；通过配套的后台操作系统，实现了数据的自动采集、处理、分析和可视化，并装载了云端数据库，可实时调用和更新数据库信息；独有的计算大脑通过大数据挖掘和分析，调用物理模型、理论计算物理和化学、量子物理、量子化学、

机器学习、深度学习和贝叶斯优化等方法，让智能模型融入底层的理论规律与复杂的化学实验演化，使得机器科学家更加理解化学，更加擅长化学创造。其工作基本原理如图 1 所示。

机器人系统、工作站和智能化学大脑都是最先进的，将对化学科学产生巨大影响。该成果脱离了传统试错研究范式的限制，展现了“最强化学大脑”指导的智能新范式的巨大优势，引领化学研究朝着知识理解数字化、操作指令化、创制模板化的未来趋势前进，确立了我国在智能化学创新领域的全球领先地位。

基于数据的预测研究是机器化学家平台从事的重要工作之一。附件 data.csv 中提供了 20 万个化学分子的物理化学性质的数据，其中，第一列“id”用于区分不同的分子，第二列“class”是它们的类别，$y_1 \sim y_3, x_1 \sim x_{100}$是它们的 103 个物理化学性质。predict.csv 中有 2580 个分子的$x_1 \sim x_{100}$性质数据。submit.csv 中需要预测填写这 2580 个分子的$y_1 \sim y_3$和“class”。

基于以上背景，请你们的团队根据附件给出的数据，通过数据分析与建模的方法，帮助实验室里机器化学家解决以下问题：

问题 1 请对题目所给数据进行预处理，明确你们处理数据必要性和所采用的处理方法。研究y_2与分子 id 之间是否有一定的函数关系，尝试直接通过 id 预测y_2；将 predict.csv 预测结果填入在附件 submit.csv 文件中。

问题 2 对附件 data.csv 中的$y_2 \sim y_3, x_1 \sim x_{100}$进行数据分析，选择不超过 10 个特征指标，建立$y_1$的预测模型，将 predict.csv 预测结果填入在附件 submit.csv 文件中。

问题 3 请分析y_3与$y_1 \sim y_2, x_1 \sim x_{100}$之间的函数关系，建立数学模型预测$y_3$，研究$y_1 \sim y_2, x_1 \sim x_{100}$中，哪些特征指标对$y_3$预测结果的影响较大？并对所选择的指标进行灵敏度分析，将 predict.csv 预测结果填入在附件 submit.csv 文件中。

问题 4 请分析 class 与$y_1 \sim y_3, x_1 \sim x_{100}$指标之间的关系，基于物理化学性质，建立分子的类别预测模型，分析$y_1 \sim y_3, x_1 \sim x_{100}$中哪些特征指标对分类的结果影响较大？将 predict.csv 预测结果填入在附件 submit.csv 文件中。

问题 5 在不局限于特征选择的情况下，你们是否有更好的方法，提高模型的

预测精度，请详细描述你们的方法，并重新对y_1, y_3以及类别 class 进行预测，论证你们预测方法的优越性。

附件：

（1）原始数据集：data.csv；

（2）预测数据集：predict.csv；

（3）提交数据集：submit.csv.

注：提交论文时，同时将预测结果以文件 submit.csv 提交到参赛平台，不要改变文件的格式。