

重庆大学化学化工学院 AI 化学与智能创造微专业

《化学信息学与数据基础》

课程教学大纲

一、课程名称： 化学信息学与数据基础

二、学时和学分： 32 学时， 2 学分

三、授课对象：

面向全校理工医类专业本科二年级及以上学生开放，重点支持化学与材料、化工与制药、生物与环境、计算机与自动化、信息与人工智能等方向学生理解化学数据在计算机内存中的物理存储形式与逻辑结构。但建议具备基本编程能力、信息检索与数据处理素养，并具备将问题结构化与工程化表达的学习意愿。

四、先修课程：

建议已完成《大学化学》或《有机化学》，具备基础的化学结构认知。建议同步修读或已修读本微专业《编程基础与数据工具》。

五、开课单位： 化学化工学院

六、授课教师：

序号	教师姓名	工号	职称	所在单位
1	陈书森	33230	副教授	化学化工学院
2	孙耿	30009203	副教授	化学化工学院
3	勾茜	T0091	教授	化学化工学院

七、课程描述：

本课程是数智化学微专业的底层逻辑课，旨在剥开化学软件的“黑盒”，探究计算机如何“存储”、“理解”并“优化”化学物质。课程将化学对象映射为计算机科学中的核心概念：通

过图论 (Graph Theory) 理解分子的拓扑连接, 利用深度优先搜索 (DFS) 与栈 (Stack) 剖析 SMILES 编码。特别地, 在 3D 结构部分, 课程引入分子力场概念, 从数据角度解析“球-弹簧”模型, 将构象优化过程抽象为基于坐标矩阵的梯度下降问题。

此外, 课程还涵盖化学数据库的关系型与非关系型结构设计, 以及谱学数据的数组处理。通过本课程, 学生将建立起“化学-数据-算法”的闭环思维, 为后续的 AI 化学与分子模拟打下坚实基础。

表 1 教学环节对应的教学内容

教学环节	教学内容
分子图论与存储结构	图 (Graph) 与分子映射、邻接矩阵与连接表、树结构与遍历算法、RDKit 图对象操作
线性符号与算法逻辑	SMILES 编码原理、深度优先搜索 (DFS) 与回溯、栈 (Stack) 数据结构应用、InChI 层级逻辑
3D 结构与分子力场	坐标矩阵与线性代数、刚体变换 (旋转/平移)、力场能量函数模型、梯度下降与结构优化算法
数据组织与检索	PubChem、Reaxys、CSD 的检索逻辑与数据导出、关系型与文档型数据库 (SQL vs NoSQL)、子结构匹配算法、API 自动化获取
谱学数据处理	常见谱学数据类型 (IR/Raman/UV-Vis/NMR) 的基础原理、数据格式与预处理、信号离散化与插值、平滑与去噪算法 (卷积核)

八、教学目标

教学环节	知识贡献	能力贡献	素质贡献
分子图论与存储结构	理解“分子即图”的抽象逻辑及计算机存储化学连接表的方式	能够利用邻接矩阵或链表达分子结构, 使用代码操作图对象	培养抽象建模能力, 建立将实体物质映射为抽象数据结构的思维
线性符号与算法逻辑	掌握 SMILES 背后的堆栈操作与 DFS 遍	具备不依赖软件手动解析简单 SMILES 的	强化算法逻辑意识, 理解序列化算法在信

	历顺序，理解 InChI 的唯一性算法	能力，能编写脚本处理字符串逻辑	息压缩与传递中的作用
3D 结构与分子力场	理解 3D 坐标的矩阵本质及力场的数学原理	能够运用矩阵运算进行分子空间变换，利用梯度下降思想解决构象优化问题	建立计算思维，理解物理原理（能量最低）向数学算法（最优化求解）的转化
数据组织与检索	理解不同数据库架构（表 vs 文档）的适用场景及化学数据的层级结构	能够设计合理的数据 Schema，编写程序解析嵌套型数据（JSON/XML）	培养系统工程观，从数据架构与标准化的角度审视化学信息的组织
谱学数据处理	理解模拟信号数字化的过程及数组在科学计算中的核心地位	能够开发自动化脚本对实验数据进行批量读取、清洗与特征提取	提升数据严谨性，树立“Garbage In, Garbage Out”的数据质量批判意识

九、教学方法：

本课程强调“CS + Chem”的跨界融合教学法。

类比教学：将 CS 概念与化学对象强关联（如：结构优化=梯度下降，分子=图）。

底层代码剖析：不仅调用 RDKit 接口，更尝试用 Python 基础代码（Numpy/Dict）模拟实现简单的能量计算函数，让学生理解力场不仅仅是软件里的一个按钮，而是一系列数学公式的集合。

可视化交互：利用 Jupyter Notebook 动态展示优化过程中分子坐标随能量梯度变化的轨迹。

十、考核及成绩评定方式：

平时作业（30%）：基础数据结构练习（如：用邻接矩阵表示给定的分子、计算简单分子的键伸缩能）。

编程实验（30%）：完成核心算法的简单实现（如：坐标矩阵的旋转、JSON 数据的解析提取）。

期末项目（40%）：化学数据转换器与优化器

要求学生编写一个程序，批量读取分子 SMILES，生成 3D 初始坐标（随机或基于规则），调用力场（或手写简单势能函数）对结构进行优化（Minimize），并输出最终的 PDB/SDF 文件。重点考核对“结构生成-能量评估-坐标更新”这一数据流的理解。

十一、教材/参考书目：

教材

Chemoinformatics: A Textbook, Johann Gasteiger.

化学信息学与计算化学参考书

1. Molecular Modelling: Principles and Applications, Andrew Leach. (重点阅读力场与优化算法章节)
2. Python Data Science Handbook, Jake VanderPlas. (侧重 Numpy 矩阵运算)
3. RDKit Documentation (Force Field & Conformer Generation 部分)