1 导 论

“执古之道，以御今之有。”

关键问题：理解数据、信息、模型、规则、规律等概念。

教学目标：（1）理解数据与信息的区别；

（2）了解化学信息、生物信息数据。

1.1 数据、信息与模型

数据是信息的载体，化学生物信息学的核心任务是从数据中提取信息[1]。信息既可以是静态的、描述性的，也可以是动态的、能够预测新数据的。例如，模型就是一种动态信息的表现形式。在这里，我们可以将模型暂时理解为一个函数，简记为$Y=f(X)$。在化学领域，函数中的X可以代表分子的化学结构数据；在生物学领域，则可以代表序列数据，例如DNA或蛋白质的序列。当X发生变化时，Y也会随之改变。

举例来说，在化学中，当化合物结构x_1发生化学变化，转变成新化合物x_2时，其溶解度也从y_1变为新化合物的溶解度y_2。类似地，在生物学中，当蛋白质序列x_1突变为新序列x_2时，原蛋白质的溶解度y_1也会变为新蛋白质的溶解度y_2。通过这一过程，我们可以记录一组数据，(x_1, y_1)，(x_2, y_2)，…。基于这些数据，我们利用化学信息学、生物信息学等技术，可以构建模型$Y=f(X)$，它代表了对数据的处理与理解。以溶解度为例，模型描述了小分子化合物的结构与溶解度之间的变化关系。这种变化关系本身就是一种信息，它帮助我们理解数据背后的规律，并为预测新数据提供依据。

通过这个案例，我们看到，数据是对客观事物的符号表示，而信息是对数据的解释和组织。假设我们测量了10个化合物的溶解度，获得了一组数据(x_1, y_1)，(x_2, y_2)，…，(x_{10}, y_{10})。此时，我们利用不同的化学信息学技术或者模型架构，使用相同的数据，却能够获得不同的模型。因此，数据是客观存在的（但可能是带有偏见的），而模型并不是客观的（是对数据处理后的“认识”）。假设以某一种模型架构训练获得的模型记作$Y=f^1(X)$。接下来，我们预测了10个新化合物的溶解度，模型的预测值为(x'_{11}, y'_{11})，(x'_{12}, y'_{12})，…，(x'_{20}, y'_{20})。由于不确定预测结果是否正确，我们进一步实验验证了这些化合物的溶解度，实验结果为(x_{11}, y_{11})，(x_{12}, y_{12})，…，(x_{20}, y_{20})。我们发现这些化合物的溶解度与模型预测的不完全一样，差值可以用于评估模型的准确性。现在我们有了20条溶解度的数据，其中第一组数据依然是(x_1, y_1)，(x_2, y_2)，…，(x_{10}, y_{10})，在第二组实验进行的过程中是不变的。但是当我们获

得 20 条数据后，重新利用相同的模型架构获得了第二个模型 $Y=f^2(X)$。我们通常认为，对于同样的技术，新的模型考虑了更多的数据，比旧的模型更加准确，即 $f^2(X)$ 比 $f^1(X)$ 更准确。通过这个案例，我们看到，数据点是固定的，数据集合是可以变化的，信息和模型会随着新数据的加入而不断变化，数据背后的客观规律则是不变的（图 1-1）。

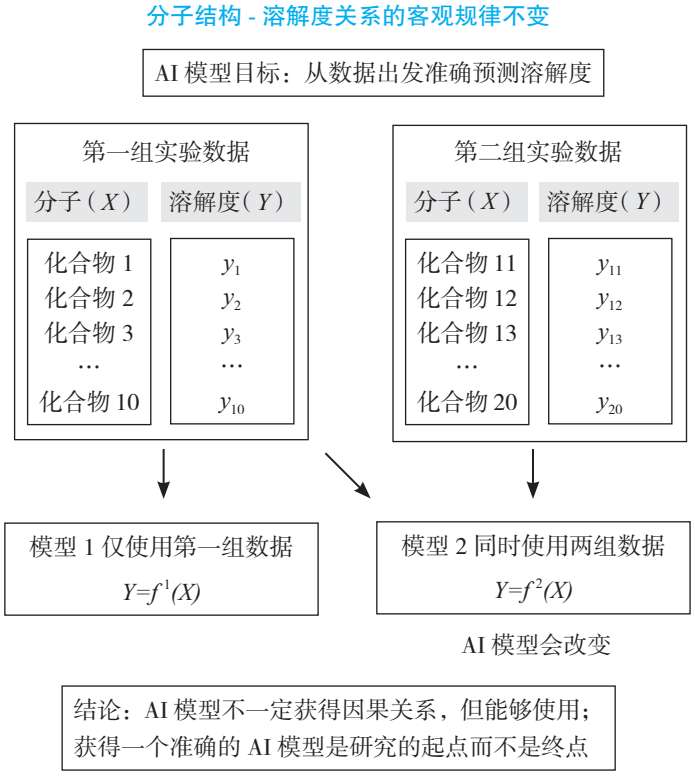

图 1-1 AI 模型未必代表客观规律

1.2 规律、规则与机制

我们总是希望先理解客观规律，再利用客观规律来指导实践，例如使用牛顿方程预测物体运动轨迹。通过溶解度的案例，我们发现无论是化学信息学还是生物信息学，我们获得的模型，都不是客观的。那么为什么我们还要研究？答案是，由于化学和生物的复杂性，我们目前还无法准确预测结果，比如溶解度。我们希望通过数据的累积，以及化学生物信息学技术的发展，逐渐理解这些规律。这个过程是迭代进行的，当数据集更新或者技术升级时，模型需要更新，我们的认识也会发生变化。

假设我们利用人工智能（artificial intelligence，AI）技术，最终能够构建出一个完全符合“客观规律”的模型，也就是预测值 (x', y') 和真实值 (x, y) 在任意情况下都相等。理论上，我们能够利用一个“可解释性”模块，在一定程度上，从模型中总结

出“客观规律”[2]。虽然根据万用近似定理[3]，深度神经网络可以逼近任意函数，但目前无法保证逼近的函数是正确的。在我们还没有完全验证 AI 发现的“规律”时，我暂时称之为“AI 规则”。规律的英文是“law”或者“principle”，而规则的英文是“rule”。规律往往更加普适，具有因果性[4]，而规则很多时候需要满足多个条件才能够使用（有可能只是一种相关性）。目前的 AI 和化学生物信息学，可以看成是技术手段。我们能够利用这些技术手段，构建模型，发现规则，利用规则预测新的数据，再进行实验验证。

以蛋白质结构预测为例，蛋白质折叠的规律就是自由能最小化，理论上我们能够利用这个规律直接根据序列预测蛋白质的 3D 结构。不幸的是，即使我们掌握了牛顿定律和量子力学，依然无法准确预测蛋白质 3D 结构。AlphaFold2 的出现证明了从数据出发的 AI，在这个任务上的表现比根据自由能最小化构建出的物理模型更加准确。在其他多个任务上，比如疾病突变预测、化合物性质预测等，AI 也实现了比其他传统模型更高的准确率。因此，能够准确预测结果的 AI 模型，蕴含了一部分正确的 AI 规则，能够辅助人类理解底层规律。

生物学中的规律与机制本质上是不同层级的解释框架，规律作为底层不可违背的自然法则（如热力学定律），为生命现象提供普适性约束；而机制则是特定系统中各部分的相互作用方式（如 DNA 复制、神经信号传递），解释具体过程“如何发生”。两者之间的鸿沟源于生命系统的复杂性，即生物体并非简单遵循物理规律，而是通过多层次结构的整合与演化，形成独特的适应性策略。例如，蛋白质折叠虽受能量最小化驱动（规律层面），但其折叠路径依赖氨基酸序列的特定排列（机制层面），这种层级跃迁使得生物学既受物理化学规律支配，又表现出不可还原的复杂性。理解这一区别，关键在于认识到生物学规律往往隐含在跨学科框架中，而机制则是生命为适应环境演化出的具体解决方案。

在规律与机制之间，可能存在一个中间研究层，这里简称为“系统原理”，它揭示复杂系统中简单规则如何涌现出高阶功能。例如，单个神经元通过离子通道的开关（机制）遵循电化学规律，但数亿神经元构成的神经网络却能产生意识、记忆等涌现特性；蚁群中个体仅遵循局部互动规则，群体却展现出路径优化、分工协作等智能行为。这些现象无法仅用底层规律或孤立机制解释，而需引入自组织理论等工具，研究多组分交互产生的非线性效应。这一层级的研究不仅填补了微观与宏观之间的解释空白，更揭示了生命系统如何在物理约束下演化出鲁棒性、可塑性等独特性质。

生物学系统原理的核心挑战在于整合能量、信息与熵的交互关系。生命通过消耗能量维持低熵状态（如细胞代谢），同时利用信息编码（DNA）实现跨代稳定性，这种“逆熵”特性在热力学规律与生物机制之间构建了桥梁。例如，光合作用将光能转化为化学能（能量流动），驱动碳固定反应（化学规律），而叶绿体的结构优化（机制）则是数十亿年自然选择对能量转化效率的极致追求。更深层的探索涉及生命起源

问题，例如简单分子如何通过自催化反应形成耗散结构，突破从非生命到生命的相变临界点。这类研究将热力学、信息论与进化理论相结合，试图回答为何生命必然趋向复杂化，以及是否存在普适的“生命设计原则”。未来生物学的突破或依赖于对“系统原理”的普适性建模，这需要跨学科方法的创新，而 AI 很有可能就是构建这个“系统原理”的重要一环。例如，结合复杂系统理论与合成生物学，可构建最小生命模型验证涌现假说。这些方向共同指向一个目标，即建立连接物理规律与生命现象的数学框架，揭示进化如何“探索”物理定律允许的可能性空间。

1.3 人工智能

AI 是一个广泛的概念，涵盖了机器学习、深度学习[5]。简而言之，AI 的目标是让机器具备类似人类的思考、学习和解决问题的能力。早期的 AI 系统主要基于逻辑规则（如 if 语句），被称为专家系统，它们通过预设的逻辑规则来模拟人类决策。然而，随着技术的发展，现代 AI 更注重从数据中自动学习规则。尽管 AI 强调“智能”，但其实现过程中仍然离不开大量的“人工”参与。这些人工工作主要体现在数据的收集与处理、模型代码的编写与部署，以及对模型结果的解读与分析上。本书的方法论是从数据出发，构建模型，最终总结出规则。我们将系统介绍数据的类型及其来源、数据的测量与表示方法［有时也叫表征（representation）］、信息的提取方式，以及如何通过底层规律约束结果。此外，书中还将探讨模型的构建过程、常用算法的选择，重点将放在深度学习技术上，包括近年来备受关注的预训练大模型（图 1-2）。

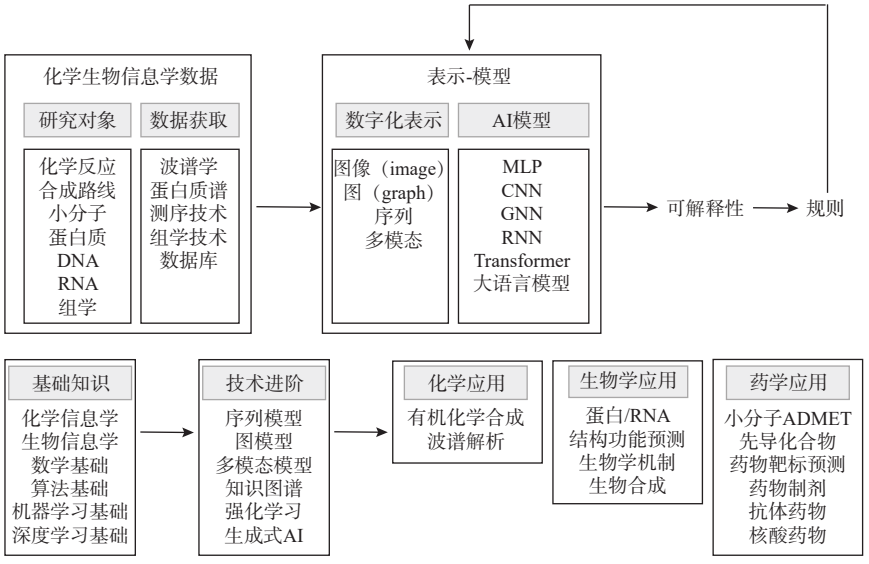

图 1-2 本书框架：从数据到规则，从基础到技术再到应用

在本书中，算法（algorithm）必不可少。算法是计算机程序的基础，指导计算机

一步一步地完成任务，从简单的计算到复杂的决策。例如，在数据库中，排序算法用于将数据按照特定顺序排列；搜索算法（如深度优先搜索、广度优先搜索等）用于查找特定元素。在化学生物信息学中，动态规划算法用于解决最优化问题，如最长公共子序列、编辑距离等；贪心算法用于在每一步选择局部最优解，期望得到全局最优解。在深度学习中，优化算法例如随机梯度下降、Adam 等算法，被认为是关键 AI 技术[6]。

本书将从化学信息和生物信息两个学科的基础与应用研究出发，全面探讨 AI 在化学合成、波谱解析、生物信息学基础研究、合成生物学以及制药领域的应用。AI 的三大核心要素是数据、算法和算力。对于一般研究人员而言，数据和算法是主要的学习内容，但对于团队领导者来说，算力是必须优先考虑和筹备的关键资源。在深度学习时代，算力的不足将严重限制研究的进展。总之，本书旨在为读者提供一个全面的视角，帮助理解 AI 在化学生物信息学中的应用，并为相关领域的研究与实践提供方法论支持。通过结合理论与实践，我们希望读者能够在快速发展的 AI 领域中找到自己的研究方向，并为未来的科学突破贡献力量。

1.4 化学信息学

许多化学信息是描述性的，例如“硫化氢具有臭鸡蛋味”，这类信息通常以人类普遍接受的事实形式存在。化学数据库系统记录了大量事实性数据，如化合物的 CAS 号、结构式、熔点等。通过对这些数据的收集和分析，我们能够利用化学信息学技术构建模型，预测未知化合物的性质。例如，在小分子药物设计中，化合物的合成和活性测试通常耗时且成本高昂，而定量结构 - 活性关系（QSAR）模型可以预测分子的成药性，从而筛选出有潜力的候选分子进行深入研究。

化学信息学的研究涵盖分子表示（molecular representation）、QSAR 模型构建、波谱预测、反应条件预测、反应产物预测以及逆合成路线设计等[7, 8]。在化学应用中，核心规律是“自由能最小化”原理，即系统倾向于达到自由能最低的状态。尽管化学领域存在大量理论和规则，但它们都遵循这一基本原理。AI 技术的引入提升了化学信息学模型的预测准确性，但化学本身的理论框架并未受到根本性挑战。相反，将这些理论和规则融入 AI 模型，往往能进一步提升其预测能力，例如在化学反应预测中加入模板能够提升预测的准确性。然而，AI 时代对化学家的一大挑战在于，机器人可能逐步替代人类完成复杂的合成化学实验[9]。尽管这一领域目前仍存在局限性，但随着 AI 和具身智能技术的发展[10]，自动化合成化学实验将迎来新的突破和升级。

1.5 生物信息学

生物信息学技术的进步推动了生命科学和医学诊断的发展。通过分析生物信息，研究人员能够更全面地了解细胞和组织状态，从而深入理解疾病机制[11]。然而，与化学相比，生物学的理论和技术尚未完全成熟。以转录组学为例，单细胞测序技术目

前仍无法精确测量每个基因的表达量，源头数据的获取仍存在技术瓶颈。生物信息学需要在数据不完整的情况下，借助算法和实验验证，逐步揭示生物学的机制、规则和规律。此外，生物学似乎并未完全遵循化学中的“自由能最小化”原理。正如薛定谔所言，“生命以负熵为生”，从分子和原子层面理解生命的本质将是一个漫长的探索过程。生物信息学正是在这种复杂性和不完备性中，不断推动科学边界的扩展。

中心法则（central dogma）阐明了生物体内遗传信息的传递过程：DNA 作为遗传信息的载体，通过转录生成 RNA，再通过翻译合成蛋白质。在转录过程中，DNA 的一条链作为模板合成 RNA；而在翻译过程中，mRNA 携带遗传信息，指导核糖体将氨基酸按特定顺序连接成蛋白质。蛋白质是细胞功能的主要执行者，直接决定生物体的性状。

中心法则：DNA → RNA → 蛋白质

中心法则为分子生物学研究提供了基础框架，是理解生命现象的核心理论。序列数据（如 DNA、RNA 和蛋白质序列）是生物信息学中最常见的数据类型。随着 AI 的发展，大模型技术为生物信息学注入了新的活力。例如，DNA、RNA 和蛋白质的大语言模型能够挖掘传统序列比对算法无法捕捉的高级语义信息（如三维结构）。本书将回顾生物信息学的经典算法，并探讨大模型在这一领域的应用。

1.6 本书内容框架

本书从基础讲起，包括化学信息学基础（含有机化学基础）、生物信息学基础、数学基础、算法基础、人工智能基础（机器学习、深度学习）等，进阶到高级的 AI 技术，如序列模型、图模型、多模态模型、知识图谱、强化学习、生成式 AI 等。最后进入应用，主要方向是 AI 辅助的化学合成、生物合成和制药。本书虽然不讲编程，但是读者需要掌握 Python 并且通过项目实践 [12]，才能真正掌握书中阐述的理论和技术。

AI 在化学合成领域的应用正在迅速改变传统的研究模式。化学合成的核心目标是设计并合成具有特定功能的分子，这一过程通常需要大量的实验试错和复杂的反应优化。AI 技术通过分析化学数据，能够预测反应路径、优化反应条件，并设计出高效的合成路线，起到降本增效的作用。此外，AI 在催化剂设计、反应机制研究以及副产物预测等方面也展现出巨大潜力。通过结合量子化学计算和机器学习，AI 能够精确预测分子的电子结构和反应活性，从而指导实验设计。

生物合成是利用生物体或其组成部分（如酶）来合成目标分子的过程。AI 在这一领域的应用主要体现在代谢工程、酶工程和合成生物学等方面。通过分析基因组、转录组和代谢组数据，AI 可以预测和优化代谢途径，设计出高效的生物合成路线。例如，AI 可以帮助研究人员识别关键的酶和基因，优化其表达水平，从而提高目标产物的产量。此外，AI 还可以通过模拟和优化酶的结构与功能，设计出具有更高催化效率和特异性的酶。在合成生物学中，AI 能够辅助设计和构建人工生物系统，实

现复杂化合物的生物合成。通过结合 AI 和自动化实验平台，研究人员可以快速筛选和优化生物合成路径，加速新产品的开发。AI 在生物合成中的应用不仅提高了生产效率，还推动了绿色化学和可持续发展。

AI 在制药领域的应用正在革命性地改变药物研发的流程。传统的药物研发周期长、成本高，且成功率低。AI 技术通过整合多源数据（如基因组数据、蛋白质结构数据、临床试验数据等），能够加速药物发现和开发过程。在药物发现阶段，AI 可以通过虚拟筛选和分子对接技术，快速识别潜在的药物候选分子。AI 还可以预测分子的药代动力学性质和毒性，从而优化药物设计。在临床试验阶段，AI 能够分析患者数据，识别潜在的生物标志物，优化试验设计，提高试验成功率。此外，AI 在个性化医疗中也发挥着重要作用，通过分析患者的基因组和临床数据，AI 可以为患者提供个性化的治疗方案。

思考与讨论

1. 数据与信息的区别是什么？
2. AI 模型能够学到规律吗？
3. 化学信息学与生物信息学的交集有哪些？
4. AI 为化学生物信息学提供了哪些新机遇？

课后拓展资料二维码

参考文献

[1] 徐华 . 数据挖掘 : 方法与应用 [M]. 北京 : 清华大学出版社 , 2014.

[2] BOTTOU L. From machine learning to machine reasoning[J]. Machine Learning, 2014, 94(2): 133-149.

[3] HORNIK K, STINCHCOMBE M, WHITE H. Multilayer feedforward networks are universal approximators[J]. Neural Networks, 1989, 2(5): 359-366.

[4] PEARL J, MACKENZIE D. The book of why: The new science of cause and effect[M]. London: Penguin, 2018.

[5] RUSSELL S, NORVIG P. 人工智能 : 一种现代的方法 [M]. 殷建平 , 祝恩 , 刘越 , 等 , 译 . 北京 : 清华大学出版社 , 2013.

[6] 高随祥 , 文新 , 马艳军 , 等 . 深度学习导论与应用实践 [M]. 北京 : 清华大学出版社 , 2019.

[7] 林亚维 , 胡晓松 , 郑铮 . 化学信息学 [M]. 北京 : 化学工业出版社 , 2019.

[8] ENGEL T. Basic overview of chemoinformatics[J]. Journal of Chemical Information and Modeling,

2006, 46(6): 2267-2277.
［9］BÉDARD A C, ADAMO A, AROH K C, et al. Reconfigurable system for automated optimization of diverse chemical reactions[J]. Science, 2018, 361(6408): 1220.
［10］MEHR S H M, CRAVEN M, LEONOV A I, et al. A universal system for digitization and automatic execution of the chemical synthesis literature[J]. Science, 2020, 370(6512): 101.
［11］樊龙江 . 生物信息学 [M]. 北京 : 科学出版社 , 2021.
［12］尹仁诚 . 零基础学 Python 编程 : 从入门到实践 [M]. 崔光善 , 译 . 天津 : 天津科学技术出版社 , 2022.