

译者序

AI for Science，即智能辅助科研，从 2021 年前后开始越来越频繁地出现在科研工作者的视野中。国际顶级学术期刊之上，人工智能（AI）辅助计算分子、设计药物、挖掘数据逐渐成为常态，国内的科技管理部门也在有识之士的推动下启动了多个重大专项，支撑 AI 在科研领域的应用。在日常工业、科技和生活中，AI 的渗透面更是广得多、深得多，可谓无处不在，在此不再一一列举。

智能辅助科研思想的首次出现，可追溯到 2016 年，就在人们为 AlphaGo 首次战胜人类围棋世界冠军而或欢呼或惊叹之时，科学家已经在尝试将机器学习等 AI 工具用于科学问题的求解。芝加哥大学丰田中心的许锦波教授尝试用深度神经网络预测蛋白质的三维结构，普林斯顿大学的鄂维南和 Roberto Car 团队也开始着手运用机器学习构建原子间相互作用的势函数。

2020 年前后，智能辅助科研正式走入人们的视野。DeepMind 推出的 AlphaFold2 在 CASP14 大赛中轰动世界，深度势能团队的 DeePMD 以“AI+ 物理模型 +HPC”新范式获得高性能计算界的“诺贝尔奖”——戈登·贝尔奖。此后 FermiNet、DeePKS、DM21 等优秀开源工具也相继诞生。这些工具与成果的出现，意味着智能辅助科研已经得到真正意义的重视，科学家在很多领域不用再“造轮子”，而是可以直接使用新方法、新工具进行科学研究。

2016~2021 年是智能辅助科研的“概念导入期”：智能辅助科研的概念被初步证明，智能辅助科研思想的潜力被核心圈层认可，但更多处在萌芽阶段。该阶段的主要目标是 AI 在各个科学场景应用的概念落地，主要工作是定义出迫切需要也十分适合使用 AI 来求解的关键科学问题，并实现算法领域从 0 到 1 的突破。2021 年后，智能辅助科研进入大规模基础建设期，我国在 2022 年和 2023 年分别在国家自然科学基金委员会与科学技术部新列了该领域的重点专项。另外，AI 的头部公司开始打造平台和面向工业及大众应用的工作流，吸引科研工作者和社会普通用户使用和贡献数据，进一步推进各自模型的进步，并在材料、生物、环境、制造等领域起到显著的加速研发的作用。这些工作都进一步加速了已经精进的智能辅助科研领域的发展速度。

智能辅助科研：深度学习革命

Artificial Intelligence for Science:
A Deep Learning Revolution

尽管智能辅助科研领域的国内外科研论文众多，但系统总结进展的著作较少。2023年4月，该领域的三位计算机和数据科学家出版了 *Artificial Intelligence for Science: A Deep Learning Revolution*。

智能辅助科研是一个快速发展的新兴领域，新概念与新方法不断涌现，部分术语尚无统一或确切的中文译法。此外，本书作为一部涵盖数学、物理、化学、材料、农学、医学等多个学科成果的综合性文集，各章节由不同领域的专家撰写，写作风格与表述深度存在一定差异。

针对新兴概念与创新性表述，译者力求忠实原文，尽可能还原作者的原意与思想。书中若有翻译疏漏或内容不足之处，恳请广大读者不吝指正，以利后续改进。

译者

2025年7月于北京

目 录

译者序

第一部分 智能辅助科研简介

1

第 1 章 人工智能辅助科研 / 3

- 1.1 科学发现简史 / 3
- 1.2 人工智能、机器学习和深度神经网络 / 4
- 1.3 人工智能与现代科学发现 / 5
- 1.4 本书介绍 / 6
- 1.5 总结 / 7
- 参考文献 / 8

第 2 章 本书概要 / 9

- 2.1 概述 / 9
- 2.2 基本方法与内涵：第 3~10 章 / 10
- 2.3 探索应用领域：第 11~26 章 / 11
- 2.4 智能辅助科研生态圈：第 27~33 章 / 14
- 2.5 智能辅助科研前景：第 34~38 章 / 16
- 2.6 结语：AI 工具与概念：第 39 章和第 40 章 / 17

第二部分 基本方法与内涵

2

第 3 章 人工智能时代的数据驱动科学：从模式到实践 / 21

- 3.1 引言 / 21
- 3.2 发现收敛的模式 / 22
- 3.3 将模式转化为正式过程 / 25

3.4 大型数据集的可扩展性和可计算性的设计原则 / 26

3.5 规模经济——从 SkyServer 到 SciServer / 27

3.6 开放科学和大型开放数据集 / 29

3.7 长期高价值数据集 / 30

3.8 未来发展趋势与展望 / 34

致谢 / 35

参考文献 / 35

第 4 章 数据科学大背景下的人工智能 / 37

4.1 引言 / 37

4.2 美国视角下的数据科学简史 / 38

4.3 数据科学的“4 + 1”模型 / 39

4.4 “4 + 1”模型视角下的人工智能 / 41

4.5 可理解的 AI / 43

4.6 结论 / 43

参考文献 / 44

第 5 章 AlphaFold——蛋白质折叠问题的终结还是更伟大事物的开端 / 46

5.1 引言 / 46

5.2 背景 / 47

5.3 结果 / 48

5.4 DeepMind 是如何赢得 CASP 的 / 50

5.5 AlphaFold 蛋白结构数据库 / 51

5.6 AlphaFold 真的解决了蛋白质折叠问题吗 / 51

5.7 对实验结构生物学家的影响 / 53

参考文献 / 54

第 6 章 人工智能在天文学中的应用 / 55

6.1 引言 / 55

6.2 早期应用：数字巡天 / 56

6.3 挑战日益增加：时域天文学 / 56

6.4 应用日益广泛 / 58

6.5 小结和展望 / 59

参考文献 / 59

第 7 章 基于机器学习的复杂仪器设计与优化 / 63

7.1 引言 / 63

7.2 机器学习的挑战和要求 / 65

7.3 优化仪器操作性能 / 66

7.4 仪器设计优化 / 72

7.5 结论 / 74

参考文献 / 74

第 8 章 实验设备中的人工智能和机器学习 / 77

8.1 实验设备的快速发展 / 77

8.2 什么是“人工智能” / 79

8.3 将 AI/ML 应用于实验科学所面临的挑战 / 80

8.4 案例研究 / 81

8.5 AI/ML 在实验和计算中的未来机遇 / 90

8.6 AI/ML 用于发现和理解科学全局 / 92

参考文献 / 93

第 9 章 第一台艾次超级计算机：加速人工智能辅助科研及在更多领域的应用 / 95

9.1 Fugaku：驱动智能辅助科研的尖端 CPU 型超级计算机 / 95

9.2 智能辅助科研的应用——案例研究 / 98

9.3 智能辅助科研的架构——过去、现在和未来 / 102

参考文献 / 103

第 10 章 人工智能辅助科研的基准测试 / 106

10.1 引言 / 106

10.2 人工智能在科学研究中的作用 / 108

10.3 方法和优点 / 112

10.4 未来发展方向 / 114

致谢 / 114

参考文献 / 115

第三部分 探索应用领域

3

天文学与宇宙学

第 11 章 射电天文学与平方千米阵列 / 119

- 11.1 现代射电天文学与平方千米阵列 / 119
- 11.2 SKA 的大数据挑战：为什么需要人工智能 / 121
- 11.3 用于射电天文学图像分析的人工智能 / 122
- 11.4 射电天文学时间域分析中的人工智能 / 124
- 11.5 SKA 的 AI 挑战 / 126
- 参考文献 / 127

第 12 章 天文学中的人工智能：机器学习的崛起 / 131

- 12.1 引言 / 131
- 12.2 机器学习在天文学中的崛起 / 132
- 12.3 开放数据与开放科学 / 133
- 12.4 开放框架的出现与机器学习的民主化 / 134
- 12.5 流数据环境下机器学习面临的挑战 / 134
- 12.6 物理驱动的机器学习方法 / 136
- 12.7 机器学习的可扩展性 / 137
- 12.8 科学中的培训与软件工程 / 138
- 12.9 未来的机遇与挑战 / 139
- 参考文献 / 139

气候变化

第 13 章 人工智能辅助净零排放 / 141

- 13.1 引言 / 141
- 13.2 人工智能在净零排放中的应用 / 142
- 13.3 面向人工智能的净零排放 / 150
- 13.4 结论与未来展望 / 151
- 参考文献 / 152

第 14 章 人工智能在气候科学中的应用 / 155

- 14.1 引言 / 155
- 14.2 气候建模 / 155
- 14.3 气候模型数据分析 / 159

- 14.4 人工智能在气候科学中的地球观测 / 161
- 14.5 人工智能/机器学习将如何改变气候科学 / 164
- 参考文献 / 166

第 15 章 应用人工智能加速聚变能源研究 / 171

- 15.1 引言 / 171
- 15.2 核聚变反应堆在运行和控制中的挑战 / 172
- 15.3 仿真的加速和验证 / 174
- 15.4 设计优化 / 177
- 15.5 AI 与聚变的未来 / 177
- 参考文献 / 177

第 16 章 人工智能为灵活弹性电网赋能 / 180

- 16.1 引言 / 180
- 16.2 电网运行虚拟化 / 181
- 16.3 人机协作以实现高效的电网运营 / 183
- 16.4 分布式能源资源促进高效的电网服务 / 185
- 16.5 本章小结 / 187
- 参考文献 / 187

第 17 章 人工智能和机器学习在从太空观测地球中的应用 / 190

- 17.1 引言 / 190
- 17.2 遥感中的人工智能和机器学习 / 190
- 17.3 观测目标 / 191
- 17.4 辐射转移、人工智能和机器学习 / 192
- 17.5 机器学习在解释遥感图像中的应用实例 / 195
- 17.6 结束语 / 196
- 参考文献 / 196

第 18 章 植物和农业研究中的人工智能 / 200

- 18.1 引言 / 200
- 18.2 前景和预期 / 200
- 18.3 生物学挑战 / 201
- 18.4 治理挑战 / 202

- 18.5 数据挑战 / 203
- 18.6 结论：AI 成为农业发展的促进平台 / 205
- 18.7 资助说明 / 206
- 参考文献 / 206

健康

第 19 章 人工智能和病理学：指导治疗和预测结果 / 209

- 19.1 引言 / 209
- 19.2 数字病理学：简要背景 / 209
- 19.3 组织病理学数字化中人工智能的应用 / 210
- 19.4 总结 / 216
- 参考文献 / 218

第 20 章 人工智能在流行病学建模中的作用 / 221

- 20.1 引言 / 221
- 20.2 建模方法 / 222
- 20.3 应用 / 224
- 20.4 决策支持工具 / 230
- 20.5 讨论和总结 / 232
- 参考文献 / 233

生命科学

第 21 章 大人工智能：大数据与大理论融合下的虚拟人构建 / 237

- 21.1 引言 / 237
- 21.2 狭义的人工智能：传统人工智能 / 238
- 21.3 人工智能能走多远 / 239
- 21.4 传统人工智能的问题 / 240
- 21.5 大人工智能的崛起 / 240
- 21.6 应对混乱 / 242
- 21.7 药物发现中的大人工智能 / 242
- 21.8 大人工智能应用于虚拟人类 / 243
- 21.9 大人工智能的未来 / 245
- 参考文献 / 246

生命科学

第 22 章 定义生命科学领域中机器学习标准的路线图 / 249

- 22.1 引言 / 249
- 22.2 挑战 / 250
- 22.3 新兴解决方案 / 251
- 22.4 机会 / 254
- 22.5 结论 / 255
- 参考文献 / 255

材料科学与工程

第 23 章 人工智能辅助材料科学研究 / 257

- 23.1 引言 / 257
- 23.2 人工智能使构建材料模型成为可能 / 258
- 23.3 应用程序模式 / 260
- 23.4 自主科学 / 261
- 23.5 挑战 / 262
- 23.6 结论与展望 / 264
- 参考文献 / 265

第 24 章 人工智能加速材料发现 / 268

- 24.1 引言 / 268
- 24.2 人工智能如何推进材料科学和工程领域的发展 / 269
- 24.3 人工智能辅助材料研发：挑战与机遇 / 271
- 24.4 人工智能辅助材料研发：解释性应用 / 272
- 24.5 人工智能辅助材料研发：前景和更加广泛的影响 / 274
- 24.6 总结 / 274
- 参考文献 / 274

粒子物理学

第 25 章 人工智能与实验粒子物理学 / 277

- 25.1 引言 / 277
- 25.2 粒子水平推断 / 278
- 25.3 事件水平推断 / 280
- 25.4 实验水平推断 / 282
- 25.5 模拟 / 285
- 25.6 结论 / 286
- 参考文献 / 286

第 26 章 人工智能与理论粒子物理学 / 290

- 26.1 引言 / 290
- 26.2 代理模型：大海捞针 / 290
- 26.3 ML 模型作为拟合函数 / 293
- 26.4 使用 ML 进行多维积分 / 293
- 26.5 LQCD / 294
- 26.6 用于高效生成规范构型的 ML 方法：应用于格点规范理论 / 295
- 26.7 对比两种黑盒子：LQCD 与 ML / 296
- 26.8 ML 预测平均值的偏差纠正的通用方法 / 298
- 26.9 物理感知 ML 模型的例子 / 299
- 26.10 无监督 ML 方法 / 300
- 26.11 探索可能理论的空间 / 300
- 26.12 从深度神经网络解读额外维度 / 302
- 26.13 结论 / 302
- 参考文献 / 303

第四部分 智能辅助科研生态圈

4

第 27 章 Schema.org 在科学研究数据中的应用 / 309

- 27.1 引言 / 309
- 27.2 使用 Schema.org 增强 Web 搜索 / 310
- 27.3 数据集搜索：专用搜索门户 / 312
- 27.4 Bioschemas（生物数据库）：生命科学的 Schema.org / 313
- 27.5 使用生物化学方案的应用程序 / 315
- 27.6 结论与未来研究方向 / 318
- 参考文献 / 319

第 28 章 人工智能耦合高性能计算 workflow / 321

- 28.1 引言 / 321
- 28.2 无处不在的学习范式 / 321
- 28.3 “处处学习” 范式示例 / 323
- 28.4 机器学习和科学计算流应用 / 327

28.5 讨论 / 329

参考文献 / 330

第 29 章 人工智能用于科学可视化 / 334

29.1 引言 / 334

29.2 具体的挑战 / 335

29.3 科学可视化与人工智能 / 336

29.4 DNN 应用案例：可视化替代模型 / 339

参考文献 / 341

第 30 章 科学领域人工智能中的不确定性量化 / 346

30.1 引言 / 346

30.2 深度学习中的不确定性量化综述 / 347

30.3 去伪存真 / 353

30.4 结语 / 355

参考文献 / 356

第 31 章 面向下一代全球网络集成系统与测试平台的人工智能 / 358

31.1 引言 / 358

31.2 AI 带来的创新 / 359

31.3 开放研究面临的挑战 / 361

31.4 真实网络的统计数据：美国能源科学网络 (ESnet) / 363

31.5 背景：当前和新兴的网络技术 / 363

31.6 广域网流量预测案例研究 / 366

31.7 迈向构建自动驾驶网络 / 376

31.8 展望与下一步计划 / 378

31.9 结论 / 378

参考文献 / 379

第 32 章 人工智能在最优实验设计与决策中的应用 / 382

32.1 引言 / 382

32.2 贝叶斯定律 / 383

32.3 先验构建 / 384

32.4 在不确定性下的决策制定 / 386

32.5 最优实验设计 / 387

32.6 结论：从精炼概率到扩大可能性 / 389

参考文献 / 390

第 33 章 FAIR：让数据为人工智能使用做好准备 / 393

33.1 FAIR 原则：数据科学的先决条件 / 393

33.2 机器可读的元数据的重要性 / 394

33.3 FAIR 评估体系：进展如何 / 397

33.4 工业中的 FAIR 数据策略 / 398

33.5 前瞻 / 400

参考文献 / 400

第五部分 智能辅助科研前景

5

第 34 章 科学大型语言模型 / 405

34.1 引言 / 405

34.2 从自然语言到科学语言 / 412

34.3 机遇和未解决的问题 / 414

34.4 结论 / 417

参考文献 / 417

第 35 章 自动驾驶汽车领域的人工智能 / 424

35.1 引言 / 424

35.2 传感器 / 424

35.3 视听系统架构 / 425

35.4 感知 / 426

35.5 总结 / 427

参考文献 / 428

第 36 章 自动化人工智能驱动科学发现未来 / 430

36.1 引言 / 430

36.2 人工智能主导的科研自动化 / 431

36.3 关键里程碑 / 432

36.4 人工智能引领科学发现的未来 / 434

36.5 人工智能辅助科研的评级 / 435

36.6 结论 / 436

参考文献 / 437

第 37 章 科学智能系统中的反思能力 / 439

37.1 引言：80 年来的人工智能辅助科研 / 439

37.2 人工智能对科学的重要性 / 440

37.3 科研导向的通用智能能力体系：深思型人工智能 / 441

37.4 人工智能科学系统的具体能力：核心能力 / 442

37.5 科学智能系统中的反思能力 / 443

37.6 结论 / 446

参考文献 / 446

第 38 章 机器学习与因果推理的接口 / 448

38.1 引言 / 448

38.2 解释深度神经网络的预测 / 448

38.3 利用连续处理推断非线性因果关系 / 453

38.4 结论 / 456

参考文献 / 457

第六部分 结语：AI 工具与概念

6

第 39 章 深度学习与机器学习概述 / 463

39.1 引言 / 463

39.2 经典人工智能/机器学习 / 464

39.3 深度学习组件 / 465

39.4 深度学习网络 / 468

参考文献 / 473

第 40 章 各章讨论的主题、概念和人工智能方法 / 475

第一部分

智能辅助科研简介

Alok Choudhary, 西北大学电子和计算机工程系, 美国伊利诺伊州埃文斯顿市。E-mail: a-choudhary@northwestern.edu

Geoffrey Fox, 弗吉尼亚大学生物复杂性研究所和计算机科学系, 美国弗吉尼亚州夏洛茨维尔市。E-mail: vxj6mb@virginia.edu

Tony Hey, 英国科技设备委员会卢瑟福·阿普尔顿实验室, 英国迪德科特镇。E-mail: tony.hey@stfc.ac.uk

1.1 科学发现简史

“第四范式”由图灵奖得主吉姆·格雷（Jim Gray）首次提出，用来描述面向未来的数据密集型科学探索^[1]。1000多年来，科学仅仅是基于观察得来的经验。1687年，艾萨克·牛顿（Isaac Newton）的《自然哲学的数学原理》一书出版，其理论建立在开普勒（Kepler）和伽利略（Galileo）的科学发现之上^[2]。该书中提出了三大运动定律。这三大定律既定义了经典力学，又为他的引力理论奠定了基础。这些工作奠定了通过理论来解释科学现象的范式，即科学探索的第二范式。约200年后，牛顿力学定律被麦克斯韦（Maxwell）用于统一电磁学方程^[3]，后又在20世纪初，被用于发展量子力学领域中的薛定谔（Schrödinger）方程^[4]。在过去的几百年里，实验观察和理论计算这两种范式一直是科学探索和发现的基础。

吉姆·格雷是美国国家科学院计算机科学与电信委员会的成员，该委员会发布了一项关于未来计算机发展的研究报告。该委员会提出，计算科学是科学探索的第三范式。吉姆·格雷说：

“随着理论模型在解决众多复杂问题时难以直接分析求解，人们不得不逐渐转向对这些模型进行仿真模拟。这一转变在过去的数千年间引领了科学探索的后半程。当前，模拟实验已经积累了海量数据。与此同时，实验科学的数据规模也大幅增加。实际上，现今的研究人员已不再单纯依赖远程观测，而是通过传输至数据中心的各类先进仪器设备得到的研究数据来‘观察’，并通过计算机进一步解析和研究这些信息。”

吉姆·格雷得出结论，科学研究的情况已经发生了变化^[1]：

“新的研究模式的运作机制在于，数据在被软件进行处理之前，先由精密仪器测试得到或通过模拟生成，并将产生的信息和知识存储在计算机中。在这个过程中，科学家往往只能在相当后期的阶段对这些数据进行审查分析。这种高度依赖于数据的新科学范式，其技术方法与传统计算科学截然不同，

因此，值得将数据密集型科学从计算机科学中区分出来，作为科学探索的全新‘第四范式’。”

1.2 人工智能、机器学习和深度神经网络

计算机诞生后不久，这一领域的早期先驱之一艾伦·图灵（Alan Turing）就提出了关于“机器智能”概念的想法。1950年，在其题为《计算机与智能》的论文中，他以独到且闻名遐迩的方式探讨了“机器是否能够思考”这一问题^[5]。然而，另一位计算机领域的开拓者约翰·麦卡锡（John McCarthy）在1956年的研讨会上首次提出了“人工智能”这一术语。约翰·麦卡锡不仅率先在麻省理工学院创立了人工智能研究实验室，随后又在斯坦福大学建立了类似的研究机构。他进一步将人工智能定义为：

“一种制造智能机器尤其是编写智能程序的科学和工程。”

拉塞尔（Russell）和诺威格（Norvig）在他们最近的人工智能教科书^[6]中介绍了“弱人工智能”和“强人工智能”这两个概念：

“哲学家将形式上表现得具备智能特征的机器称为弱人工智能，而将真实能够思考的机器（不仅是模拟思考）称为强人工智能。”

在将人工智能应用于科学的过程中，我们不会关心两者之间的区别，并默认为通常应用的都是弱人工智能。机器学习（ML）可以被视为人工智能的一个子领域。在乔丹（Jordan）和米歇尔（Mitchell）最近的综述^[7]中，他们将机器学习定义为：

“机器学习解决了如何通过经验让计算机自动改进算法的问题。它是当今发展最快的技术领域之一，是计算机科学和统计学的交叉前沿，也是人工智能和数据科学的核心与前沿。机器学习的进步主要是由两个因素推动，一个是新学习算法和理论的发展，另一个是可在线数据的增多和计算成本的降低。”

作为人工智能的最新领域，深度神经网络（图 1.1）是人工神经网络的机器学习方法的分支^[8]：

“人工神经网络是基于一组被称为人工神经元的连接单元或节点的集合，这些单元或节点对生物大脑中的神经元的功能进行模拟。每一个连接，就像大脑的突触一样，可以将信号传递给其他神经元。接收信号的人工神经元会处理信号，并向连接到信号的神经元发出信号。连接处的‘信号’是一个实数，每个神经元的输出是通过其输入之和的一些非线性函数来计算的，这些连接称为边缘。而神经元和边缘的权重通常会随着学习的进行而调整。”

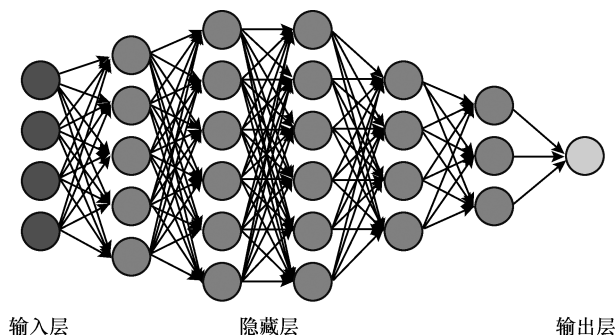


图 1.1 深度神经网络示意图

输入层使用来自大型数据集的数据实例。而隐藏层则用来共同学习函数和模型，这些函数和模型反过来用于预测感兴趣的输出变量（或多个变量）。输出变量可以是一个分类变量（如图像内容或股票价值等数字）。这只是神经网络中的一种，而一些深度神经网络可能有超过 100 个隐藏层。

这些网络中的人工神经元被排列在从输入层到输出层中的每一层中，不同层中的神经元之间存在连接。深度神经网络只是这种网络的一种，而且在输入层和输出层之间还含有大量隐藏的中间层。当人工智能领域尤其是人工神经网络的倡导者杰弗里·欣顿（Geoffrey Hinton）领导的团队在 2012 年赢得 ImageNet 图像识别挑战赛时，深度学习网络开始崭露头角^[9]。他们最初的竞争对手开发的 AlexNet 是一个由 8 个隐藏层组成的深度学习网络，其训练过程主要依赖图形处理器（GPU）进行计算。2015 年，微软研究团队使用了一个深度神经网络，该网络使用 GPU 集群训练了 150 多个层，从而实现了与人类错误率相当的机器识图技术^[10]。

深度学习网络现在是 IT 行业的一项关键技术，并广泛被用于各种重要的商业应用，如图像分类、面部识别、手写转录、机器翻译、语音识别、文本到语音转换、自动驾驶和定向广告。最近，谷歌的英国子公司 DeepMind 利用深度神经网络开发了优秀的围棋软件 AlphaGo 及其衍生产品。然而，人工智能辅助科研更典型的成果是 DeepMind 的 AlphaFold 蛋白质折叠预测系统^[11]。他们最新版本的 AlphaFold 令人吃惊地实现了计算机辅助合成规划（CASP）蛋白质折叠^[12]。诺贝尔奖获得者 Venki Ramakrishnan 曾说过：

“这项计算工作展现了人工智能在计算方面惊人的进步，代表了解决蛋白质折叠问题取得了重大进展。蛋白质折叠问题是生物学中一个 50 年来的重大挑战，许多人在几十年前就已经预测和关注到这个问题。令人兴奋的是，这个问题的解决将从根本上改变生物学研究的许多方面。”^[13]

1.3 人工智能与现代科学发现

如前所述，现代常用的四种科研范式是：①经验科学；②理论科学模型；③计算科学（模拟仿真）；④（大）数据驱动科学。图 1.2 展示了这四种范式。

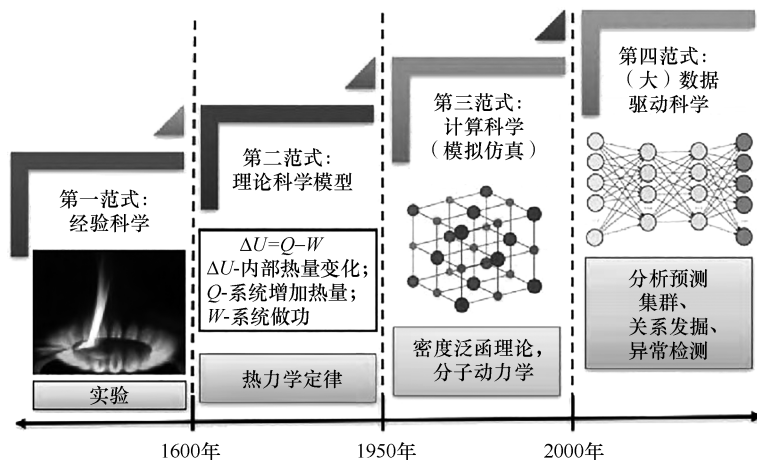


图 1.2 科学发现的四种现有范式

每种范式都有其局限性。实验的进展速度非常缓慢，很难大规模进行，甚至有些条件根本无法完成。此外，一些大型仪器，如大型强子对撞机（LHC）或平方千米阵列（SKA）射电望远镜，其建造和维护成本非常高。每个实验的数据结果通常在各自实验装置中独立分析，因此获得新知识的潜在途径仅限于分析单个实验的输出和输入参数。然而，人工智能技术不仅可以应用于单种实验的单个实例，还可以应用于分析许多此类实验的综合信息，并有助于产生新的科学发现和见解。

原理驱动的模式也有其局限性：首先，需要简化假设以创建模型；其次，模型中的复杂方程组时常无法得到分析解。在实践中，计算机模拟可以在一定程度上解决以上两个难题。事实证明，模拟气候科学、分子动力学、材料科学和天体物理学等各种不同研究领域的计算机模拟都非常成功。使用超级计算机可有效增强建模计算能力。超级计算机的计算能力现已接近艾级水平（相当于每秒 10^{18} 次浮点计算）^①。然而，即使是这样的超级计算机模拟也会在几个方面受到限制。首先，超级计算机价格昂贵，因此全世界目前只有少数此类超算，模拟本身还会受到所模拟数学模型和数据格式的限制。其次，一次模拟仅代表具有一组特定初始条件和约束问题的一个实例。最后，一些模拟可能需要几天或几周才能完成，这就限制了可计算的模拟数量，进而限制了对大参数空间的探索。

此类基于超级计算机模拟受限的一个例子是气候模拟。对气候的模拟是人类面临的一个紧迫问题。美国国家大气研究中心（NCAR）是美国国家科学基金会（NSF）新成立的人工智能与物理学习地球中心（LEAP）的成员之一。NCAR 科学家大卫·劳伦斯（David Lawrence）这样说道^[14]：

“LEAP 对于一个多学科团队来说是一个巨大的机会，可以去探索通过使用机器学习来改进复杂地球模型的可操作性，而这一切都是为了社会的长远利益。”

LEAP 将聚焦使用机器学习技术来改进 NCAR 的地球系统模型（CESM）。CESM 由一系列复杂的基础模型组成，可以模拟大气、海洋、陆地、海冰和冰盖过程的相互作用。然而，CESM 的能力有限，无法整合一些难以模拟的重要物理过程。例如，云的形成和演化就发生在一个十分精细的尺度上，模型无法对此进行表达。另外，地面上的生态过程也过于复杂，无法在模拟中体现。为了将这些重要物理过程的近似值纳入 CESM，气候科学家创建了参数化的描述方法。LEAP 的一个主要目标是通过使用机器学习技术整合大量地球系统观测数据和目前可用的高分辨率模型模拟数据中学到的知识来改进这些近似值。

1.4 本书介绍

本书讲述了如何将人工智能、机器学习和深度学习技术应用于由超级计算机模拟或者现代实验设施产生的海量科学数据集。如今，海量的科学数据来自卫星、基因测序仪、大功率望远镜、X 射线同步加速器、中子源、电子显微镜以及许多大型国际设施 [如位于日内瓦的欧洲核子研究中心（CERN）的大型强子对撞机以及位于汉堡的德国电子加速器装置研究机构（DESY）的

① 译者注：截至目前，中国已有多台艾级超级计算机投入运行。

欧洲 X 射线自由电子激光器 (European XFEL)]。这些数据源自每年产生数百亿字节的数据,并且这些设施的改进预计至少产生超过现在一个数量级的数据。对科学家来说,如何从这些不断增加的数据中提取有意义的科学见解将是一项重大挑战。因此,科学领域中的“大数据”是人工智能辅助科研的前提,也为应用新型人工智能技术提供了一个令人兴奋的机会,这可能将会彻底改变许多科学领域。

目前,在国际上有许多关于应用人工智能技术来管理和分析日益庞大且复杂的数据集的倡议。在美国,美国国家科学基金会已经建立了 18 个国家级人工智能研究所,其合作伙伴覆盖了 40 个州^[15]。此外,美国能源部(DOE)也为美国国家实验室的大型设施和超级计算机提供资金。2019 年,美国能源部实验室组织了一系列大型会议,吸引了数百名科学家、计算机科学家,以及来自工业界、学术界和政府的参与者。该会议旨在研究利用人工智能技术加速或转变其职责范围内相关科学领域研究的机会^[16]。这次会议的目标是

“研究未来十年人工智能、大数据和高性能计算领域的科学机遇,且为抓住这些机遇构思相关重大发展思路、研讨可能面临的重大挑战以及具体实现措施”。

这次会议使用了“人工智能辅助科研”这一术语来广泛代表计算和数据分析领域的下一代方法和科学机遇。人工智能辅助科研包括开发和应用人工智能方法,如机器学习、深度学习、统计方法、数据分析和自动控制。利用数据建立模型,并单独使用这些模型或者结合模拟数据使用这些模型来推进科学研究。会议得出结论,在未来十年,人工智能方法在科学领域的应用有可能会给许多科学研究领域带来重大变革。总之,会议认为人工智能技术可以:

- (1) 加速新材料的设计、发现和评估;
- (2) 推进新硬件和软件系统、仪器和仿真数据流的发展;
- (3) 通过高带宽仪器数据流揭示新科学和理论;
- (4) 通过在控制和分析循环中加入推理能力来改进实验;
- (5) 实现从光源、加速器到仪器探测器和高性能计算数据中心等复杂系统的设计、评估、自主运行和优化;
- (6) 推动自动驾驶实验室和科学工作流程的发展;
- (7) 利用人工智能代理模型,大幅提高 E 级超级计算机和未来超级计算机的能力;
- (8) 自动化大规模创建“FAIR”数据——可发现的、可访问的、可互操作的和可重用的数据。

在英国,位于伦敦的艾伦·图灵研究所是英国国家数据科学和人工智能研究所^[17],该所的“人工智能辅助科研与政府”项目与位于哈威尔地区的英国国家实验室合作,目的是对人工智能辅助科研领域的相关重大工作进行研究^[18]。

1.5 总结

本书充分讨论了人工智能技术在多个科学领域实现重大推进作用的潜能。各章节均由各领域顶尖科学家撰写,他们权威地论述了人工智能技术在其专业领域所面临的挑战和发展前景。

参考文献

- [1] J. Gray (2009). Jim Gray on eScience: A transformed scientific method. In T. Hey, S. Tansley, and K. Tolle (eds.), *The Fourth Paradigm: Data Intensive Scientific Discovery*. Redmond: Microsoft Research.
- [2] I. Newton (1687). *Mathematical principles of natural philosophy*. https://en.wikipedia.org/wiki/Philosophi%C3%A6_Naturalis_Principia_Mathematica (accessed June 2022).
- [3] J. C. Maxwell (1865). A dynamical theory of the electromagnetic field. *Philosophical Transactions of the Royal Society of London*, 155, 459–512.
- [4] E. Schrödinger (1926). An undulatory theory of the mechanics of atoms and molecules. *Physical Review*, 28(6), 1049–1070.
- [5] A. Turing (1950). Computing machinery and intelligence. *Mind*, 49, 433–460.
- [6] S. Russell and P. Norvig (2020). *Artificial Intelligence: A Modern Approach*, 4th edn. Pearson Education Limited, New York.
- [7] M. I. Jordan and T. M. Mitchell (2015). Machine learning: Trends, perspectives and prospects. *Science*, 349, 255–260.
- [8] Artificial Neural Network. https://en.wikipedia.org/wiki/Artificial_neural_network (accessed June 2022).
- [9] A. Krizhevsky, I. Sutskever, and G. E. Hinton (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105.
- [10] K. He, X. Zhang, S. Ren et al. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 27–30 June 2016, Las Vegas, NV, USA, pp. 770–778.
- [11] A. W. Senior, R. Evans, J. Jumper et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577, 706–710.
- [12] J. Jumper, R. Evans, A. Pritzel et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596, 583–589.
- [13] D. Rice (2020). Stunning advance on ‘protein folding’: A 50-year-old science problem solved and that could mean big things. Protein folding discovery a major breakthrough from DeepMind ([usatoday.com](https://www.usatoday.com/story/news/nation/2020/12/03/protein-folding-discovery-major-breakthrough-deepmind/3809693001/)) (accessed June 2022). 3 December 2020, <https://www.usatoday.com/story/news/nation/2020/12/03/protein-folding-discovery-major-breakthrough-deepmind/3809693001/>.
- [14] L. Snyder (2021). NCAR will collaborate on new initiative to integrate AI with climate modeling | NCAR & UCAR News ([news.ucar.edu](https://news.ucar.edu/132809/ncar-willcollaborate-new-initiative-integrate-ai-climate-modeling)) (accessed June 2022). 10 September 2021, <https://news.ucar.edu/132809/ncar-willcollaborate-new-initiative-integrate-ai-climate-modeling>.
- [15] NSF AI Map (accessed June 2022).
- [16] AI for science report | Argonne National Laboratory (anl.gov) (accessed June 2022).
- [17] Home | The Alan Turing Institute (accessed June 2022).
- [18] SCD Scientific Machine Learning (stfc.ac.uk) (accessed June 2022).

第2章

本书概要

Alok Choudhary, 西北大学电子和计算机工程系, 美国伊利诺伊州埃文斯顿市。E-mail: a-choudhary@northwestern.edu

Geoffrey Fox, 弗吉尼亚大学生物复杂性研究所和计算机科学系, 美国弗吉尼亚州夏洛茨维尔市。E-mail: vxj6mb@virginia.edu

Tony Hey, 英国国家科研与创新署卢瑟福实验室, 英国迪德科特镇。E-mail: tony.hey@stfc.ac.uk

2.1 概述

本书分为 6 个部分。

第一部分, “智能辅助科研简介” 包含两章 (第 1 章和第 2 章), 内容涉及科学领域的人工智能基本理念和现状。

第二部分, “基本方法与内涵” 中包含 8 章 (第 3~10 章), 提供了各个领域的典型示例, 为展开讨论奠定了背景。这些章节涵盖了大数据与数据科学的推动作用, 以及 AlphaFold 在人工智能辅助预测蛋白质结构方面的成功案例。随后是观测科学的相关内容, 分别涉及天文学、引力波探测和加速器。第 9 章描述了多个在首台每秒 1 艾次 (10^{18} 次) 超级计算机上进行的智能辅助科研的开创性研究。最后, 讲述了基准测试在推进跨学科科学发现与人工智能发展中的作用。

第三部分, “探索应用领域”, 是本书的核心内容, 包含第 11~26 章, 每两章内容组成一个领域, 共 8 个领域, 讨论了人工智能在科学与工程领域的研究应用。这八个领域按照英文首字母顺序排列, 分别是天文学与宇宙学; 气候科学与气候变化; 涵盖聚变能和电力网格的能源领域; 环境科学与遥感农业; 健康学、病理学和流行病学; 生命科学与利用虚拟人体研究药物, 以及人工智能标准的作用; 材料科学与工程, 包括代理模型仿真的应用; 粒子物理学, 描述人工智能在实验分析和理论理解两方面的应用。

第四部分, “智能辅助科研生态圈”, 包含第 27~33 章, 阐述了促进科学领域中人工智能应用的基础架构。首先讨论了构建大数据库及其分析所需的各种本体论和模式的发展。第 33 章描述了 FAIR (findable, accessible, interoperable, reusable, 即可查找、可访问、可互相操作、可重复利用) 数据的重要性及其实现方式。这部分介绍了支撑现代科学领域复杂工作流程的网络基

基础设施和中间组件，展示了人工智能是如何提升科学可视化的，以及人工智能的影响。多个章节强调了不确定性量化的必要性，因为来自人工智能的结果可能非常不透明，也阐述了当前能够解决此问题的方法。许多科学进步需要国家间数据的共享，在这方面，人工智能能够帮助控制连接各类数据的网络。科学建立在实验之上，第 32 章介绍了设计最优实验和决策的人工智能方法。

第五部分，“智能辅助科研前景”，包含第 34~38 章，描述了一些面向未来具有广泛影响力的理念。第 34 章讲述了自然语言处理（NLP）的进步如何引领科学推理领域的革命。第 35 章则突出了工业领域的进步，这些进步是基于图像处理、自主决策、自动控制等众多科学领域的进步而实现的。第 36~38 章讨论了未来通用人工智能的发展趋势，以及可能带来的深层次问题，这些问题的解答有望突破智能辅助科研目前只能协助寻找规律的状况，实现发现定理、自主思考设计。

第六部分“结语：AI 工具与概念”包含第 39 章和第 40 章。第 39 章是 AI 方法的简明手册，第 40 章则回顾了书中各章节讨论的各种主题、概念和方法。

下面介绍第 3~40 章的具体信息。

2.2 基本方法与内涵：第 3~10 章

第 3 章和第 4 章描述了对人工智能有效辅助至关重要的大数据环境。在第 3 章中，Szalay 讲述了大规模科学数据集创造与发展所需的技能和协作方法。他以模拟结果数据库（如湍流、宇宙和海洋的模拟）以及由 SkyServer 访问的斯隆数字巡天 SDSS 天文数据所示的观测科学为例，并运用了云和 Kubernetes 基础设施。在第 4 章中，Alvarado 和 Bourne 着重指出，人工智能需要数据来学习。这意味着，必须将人工智能放在更广阔的数据科学背景下进行考量，其中包括考虑所学内容的伦理含义、如何理解和解释这些学习成果，以及如何最好地将结果呈现给人类。作者回顾数据科学的历史，提出了一个 4+1 模型，其中数据科学的四大支柱是系统、设计、分析和伦理，而 +1 则代表具体应用领域。

第 5 章“AlphaFold——蛋白质折叠问题的终结还是更伟大事物的开端”中，Jones 和 Thornton 描述了人工智能近期的一个显著成就：AlphaFold2 在 2020 年的 CASP 竞赛中脱颖而出，仅基于氨基酸序列就成功预测出了蛋白质结构的原子坐标。本章详细介绍了蛋白质结构的科学问题，并概述了 AlphaFold2 采用深度学习解决方案，预测出了 100 万种新的蛋白质结构。这令人鼓舞，尽管这只是开端，还有配体和碳水化合物等领域有待探索。AlphaFold2 建立在谷歌早期的创新技术之上，如 AlphaGo 和 Transformers。一家公司通过利用其人力资源、人工智能特长及庞大的计算资源赢得了比赛，这非常具有启发性。我们期待科学领域能更广泛地受益于学术界、政府与工业界的协作。新方法无所不在，并且需要多学科、多机构的合作。

在第 6 章“人工智能在天文学中的应用”中，Djorgovski 及其合作者描述了人工智能在天文学、天体物理学和宇宙学中的应用。这包括综合研究多种波长的天文学。全景天空调查和时域天文学产生了千万亿级的数据流，对于有趣事件的分类与识别需要实时处理。一个典型的问题是，如何从包含数十亿个特征向量、数百维数据中分离并分类出恒星和星系。在过去十年里，