

# 前言

生命科学作为一门探索生命奥秘的学科，始终与人类的命运息息相关。从最初的博物学到现代的基因组学，科学家们的每一次突破，都深刻影响着我们对生命的理解与对疾病的防治。随着技术的飞速发展，尤其是信息技术和计算科学的飞跃，生命科学的研究方式和范式发生了革命性的变化。特别是近几十年来，生物学大数据的积累和智能化分析手段的应用，使得我们能以全新的视角审视生命的复杂性。人工智能技术，尤其是深度学习的兴起，正在极大地改变我们对生物世界的理解。它让我们不仅仅停留在表象之上，而是深入到数据背后的模式与规律，去发现那些肉眼无法捕捉的深层联系。

近年来，随着基因组学、蛋白质组学、代谢组学等数据的积累，我们能够在海量数据中挖掘出潜在的规律和信息，而这一切都离不开智能化的分析手段。本书不仅介绍了人工智能在生命科学中的最新应用，还通过具体的案例，展示了智能算法如何帮助我们解决传统方法无法触及的复杂问题，以及如何在蛋白质结构预测、基因变异解析、疾病机制探索、药物研发等方面取得突破。在对深度学习关键模型和组学数据进行简要介绍的基础上，本书重点介绍了我们在生物多组学、动物学、植物学和微生物领域的最新研究工作。

本书是对笔者多年从事的生命科学大数据智能研究的系统总结。全书共分为7章：第1章概述了人工智能的基本概念、发展历程及其在生命科学中应用的现状与前景；第2章介绍了人工智能的主要学习范式和深度学习的关键模型，包括卷积神经网络（CNN）、循环神经网络（RNN）、注意力机制、自监督学习模型及生成式模型；第3章阐述了组学大数据的诞生、组学数据资源、组学数据的结构、组学数据的特征，以及传统统计方法在组学数据中的应用，重点探讨了数据的高维性和异质性问题；第4章探讨了人工智能在多组学研究中的应用，包括蛋白质组学和代谢组学中的智能分析方法，如蛋白质定位预测和小分子性质预测；第5章讨论了人工智能在动物研究中的应用，涵盖动物分类、行为、遗传、疾病模型等领域的应用，并分析了面临的挑战；第6章阐述了人工智能在植物研究中的应用，涵盖植物表型分析、智慧农业等；第7章介绍了人工智能在微生物研究中的应用，包含微生物生态学与环境监测、病原微生物、食品微生物、农业微生物、微生物与合成生物学等领域。

本书的另一个重点是对“跨学科合作”这一理念的深入阐述。在生命科学大

数据的研究中，我们不仅需要生物学家的专业知识，更需要计算机科学家、数学家的智慧和技术。数据的复杂性与多样性，要求我们能够跨越学科的界限，整合各类技术和方法，才能够真正破解生命的密码。正是这种跨学科的协作，让我们能够更好地理解生命现象的规律，并将这些理解转化为实际的应用，服务于人类的健康和福祉。

人工智能在生命科学中的应用，不仅仅是技术的进步，更是一种哲学性的思维革命，它促使我们从微观世界到宏观世界、从实验室到临床，再到社会与伦理，全面反思人类对自然界、对生命的理解。本书的每一章都是对这种深刻变革的呈现，也记录着我们在技术与哲学的交汇处探索生命的奥秘，思考生命未来的足迹。

希望读者能够通过本书学习、理解、体会在人工智能、大数据、生命科学等交叉领域进行研究的基本方法和思路。

本书的完成，离不开诸多师生在科研与实践中的参与和贡献。特别感谢合作者陈宗青、潘建波、冯金旺、张文锋、杨宪、冉瑞生老师，以及课题组的罗泽宇、吴锦洲、刘俊豪、孙亚雯、王瑞、仇鑫、肖续、朱梦黎、施韵、孟晨、朱星源、李鑫、何东林、李睿盟、伍习松、林语、孙涛、余炳秋、杨秀苓、李颖、蒋楠、刘桂荣、邹歆等同学。

由于作者水平和精力有限，书中难免有错误和疏漏，恳请读者不吝指正。

# 目 录

|                          |    |
|--------------------------|----|
| 第 1 章 人工智能发展历史           | 1  |
| 1.1 人工智能的思想雏形            | 1  |
| 1.2 人工智能的发展历程            | 2  |
| 1.2.1 起源与发展初期            | 2  |
| 1.2.2 寒冬期                | 4  |
| 1.2.3 复苏与繁荣              | 4  |
| 1.3 人工智能在生命科学中应用的现状与前景   | 8  |
| 1.3.1 现状                 | 8  |
| 1.3.2 面临的问题              | 11 |
| 1.3.3 发展前景               | 12 |
| 参考文献                     | 13 |
| 第 2 章 人工智能算法介绍           | 14 |
| 2.1 人工智能的主要任务和学习范式       | 14 |
| 2.1.1 监督学习               | 14 |
| 2.1.2 无监督学习              | 15 |
| 2.1.3 半监督学习              | 16 |
| 2.1.4 强化学习               | 17 |
| 2.1.5 计算机视觉、自然语言处理和多模态学习 | 18 |
| 2.2 深度学习的关键模型            | 20 |
| 2.2.1 卷积神经网络             | 22 |
| 2.2.2 循环神经网络             | 24 |
| 2.2.3 注意力机制              | 24 |
| 2.2.4 自监督学习模型            | 25 |
| 2.2.5 生成模型               | 26 |
| 2.2.6 深度学习的高速化技术         | 27 |
| 2.3 总结                   | 28 |
| 参考文献                     | 30 |

|                            |     |
|----------------------------|-----|
| <b>第3章 生命科学中的组学大数据</b>     | 34  |
| 3.1 组学大数据的诞生               | 34  |
| 3.1.1 组学数据介绍               | 35  |
| 3.1.2 从跨物种到泛组学的扩展          | 52  |
| 3.2 组学数据资源                 | 56  |
| 3.2.1 国家基因组科学数据中心          | 56  |
| 3.2.2 美国国立生物技术信息中心         | 56  |
| 3.2.3 RNA 相关数据库            | 60  |
| 3.2.4 蛋白质相关数据库             | 60  |
| 3.2.5 代谢物数据库               | 61  |
| 3.3 组学数据的结构与特征             | 62  |
| 3.3.1 组学数据的结构类型            | 63  |
| 3.3.2 组学数据的主要特征            | 69  |
| 3.4 传统统计分析方法在组学数据中的应用      | 72  |
| 3.4.1 差异表达分析               | 74  |
| 3.4.2 功能富集分析               | 76  |
| 3.4.3 以化学计量基因组学为例          | 77  |
| 参考文献                       | 81  |
| <b>第4章 人工智能在多组学研究中的应用</b>  | 86  |
| 4.1 人工智能在组学数据中的应用          | 86  |
| 4.1.1 机器学习与深度学习在组学研究中的应用概述 | 86  |
| 4.1.2 智能算法赋能组学数据整合         | 92  |
| 4.2 蛋白质组学中的人工智能            | 96  |
| 4.2.1 蛋白质定位预测              | 96  |
| 4.2.2 蛋白质修饰与相互作用预测         | 119 |
| 4.3 代谢组学中的人工智能             | 135 |
| 4.3.1 基于多模态信息表示的小分子性质预测    | 135 |
| 4.3.2 基于图对比学习的小分子性质预测      | 151 |
| 参考文献                       | 161 |
| <b>第5章 人工智能在动物研究中的应用</b>   | 169 |
| 5.1 动物研究数据集                | 170 |
| 5.2 动物科学不同领域的应用            | 172 |
| 5.2.1 动物分类与保护              | 172 |

|                                  |            |
|----------------------------------|------------|
| 5.2.2 动物行为与发育 .....              | 176        |
| 5.2.3 动物遗传、进化和疾病模型 .....         | 182        |
| 5.2.4 动物育种与健康管理 .....            | 192        |
| 5.2.5 古生物学 .....                 | 198        |
| 5.3 在动物研究中使用 AI 的挑战 .....        | 200        |
| 5.3.1 数据相关的挑战 .....              | 201        |
| 5.3.2 算法相关的挑战 .....              | 202        |
| 5.3.3 伦理相关的挑战 .....              | 203        |
| 5.3.4 技术相关的挑战 .....              | 204        |
| 5.3.5 应对挑战的策略 .....              | 205        |
| 5.3.6 未来展望 .....                 | 206        |
| 参考文献 .....                       | 206        |
| <b>第 6 章 人工智能在植物研究中的应用 .....</b> | <b>217</b> |
| 6.1 植物研究数据集 .....                | 218        |
| 6.2 人工智能在植物学不同领域的应用 .....        | 220        |
| 6.2.1 植物表型识别 .....               | 220        |
| 6.2.2 植物基因组与多组学分析 .....          | 225        |
| 6.2.3 作物性状预测与精准育种 .....          | 228        |
| 6.2.4 植物病害预测与诊断 .....            | 231        |
| 6.2.5 植物营养与产量预测 .....            | 234        |
| 6.2.6 植物与环境互作建模 .....            | 238        |
| 6.2.7 植物适应性研究 .....              | 241        |
| 6.2.8 植物基因编辑 .....               | 246        |
| 6.2.9 药用植物与活性成分预测 .....          | 251        |
| 6.3 植物学研究中人工智能面临的挑战与展望 .....     | 254        |
| 6.3.1 技术融合驱动的植物学研究创新 .....       | 254        |
| 6.3.2 植物学研究中人工智能面临的挑战 .....      | 255        |
| 6.3.3 人工智能促进植物学研究跨学科发展 .....     | 256        |
| 6.3.4 人工智能赋能植物可持续发展 .....        | 256        |
| 6.3.5 植物研究智能化的未来展望 .....         | 257        |
| 参考文献 .....                       | 258        |

第 7 章 人工智能在微生物研究中的应用 ..... 269

7.1 微生物研究数据集 ..... 270

7.2 人工智能在微生物不同领域的应用 ..... 274

7.2.1 16S 核糖体 RNA ..... 274

7.2.2 微生物组学 ..... 277

7.2.3 微生物与生态 ..... 284

7.2.4 微生物与疾病 ..... 287

7.2.5 微生物与食品 ..... 293

7.2.6 微生物与农业 ..... 295

7.2.7 微生物与合成生物学 ..... 298

7.3 人工智能在微生物领域研究中的挑战与展望 ..... 301

7.3.1 人工智能在微生物多领域研究结合中的挑战 ..... 301

7.3.2 人工智能在微生物多领域研究结合中的展望 ..... 304

7.3.3 结论 ..... 308

参考文献 ..... 309

# 第 1 章 人工智能发展历史

## 1.1 人工智能的思想雏形

人工智能（artificial intelligence, AI）是研究如何利用计算机系统模拟、延伸和增强人类智能的一门交叉学科。长期以来，人类始终希望通过技术手段构建能够感知环境、处理信息并完成特定任务的智能系统。随着计算机科学、数学和工程技术的发展，这一设想逐渐从理论走向现实，并形成了现代人工智能研究领域。

根据智能系统能力范围的不同，人工智能通常可分为弱人工智能（weak AI）和强人工智能（strong AI）。弱人工智能主要针对特定任务设计，具备有限的感知与学习能力，但不具有自主意识。当前的图像识别、语音识别与推荐系统等均属此类。相比之下，强人工智能则设想构建具备通用推理能力、自我调整机制，甚至自主意识的系统，其行为不再完全依赖外部设定，可能具有独立目标。虽然其目前仍停留在理论构想阶段，但已引发对控制力、安全性与伦理规范的广泛讨论。

在当前的技术体系中，机器学习（machine learning, ML）是实现人工智能的核心方法。它通过从大规模数据中学习输入输出模式并优化模型参数，使系统具备基于经验的自我改进能力。深度学习（deep learning, DL）利用多层神经网络结构实现对复杂特征的自动抽象，在图像处理、自然语言理解等领域已取得显著成果。强化学习（reinforcement learning, RL）则强调智能体在环境中通过反馈信号进行试错优化，逐步形成最优策略。这类学习方式已在智能机器人、自动驾驶、资源调度等场景中取得初步成果。

图 1-1 展示了当前人工智能在生命科学领域中各类机器学习算法的应用占比情况。其中，逻辑回归以 33.0% 的占比位居首位，反映出其在处理分类问题、建模简单关系等场景中的广泛应用；主成分分析（12%）和随机森林（10.3%）分别在数据降维与复杂非线性问题建模中占据重要地位；支持向量机（4.9%）在小样本学习任务中具有优势；卷积神经网络（6.8%）作为深度学习的代表性算法，在图像识别等领域的应用占比凸显了其对复杂特征提取的能力。整体来看，传统机器学习算法与深度学习算法的占比分布，既体现了经典方法在实际场景中的持续价值，也反映了深度学习在复杂任务中的崛起趋势，为理解人工

智能技术的应用生态提供了直观参考。

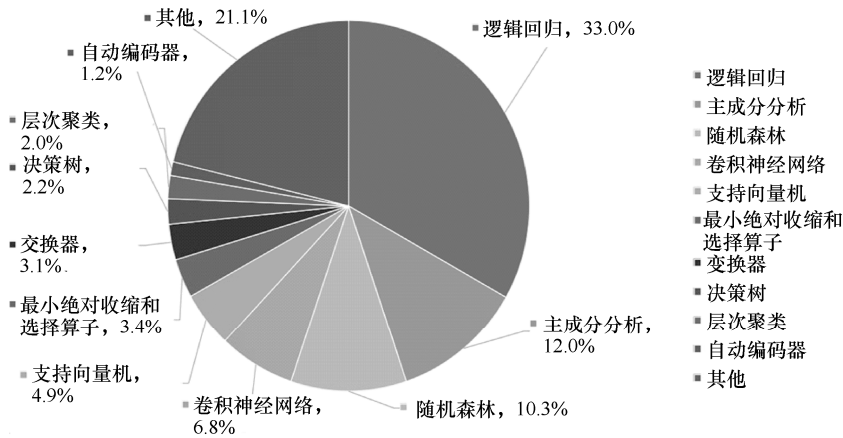


图 1-1 2010~2024 年人工智能在生命科学领域常用算法占比饼图（彩图请扫封底二维码）  
根据“机器学习算法”（如 Random Forest, Principal Component Analysis, Logistic Regression, Support Vector Machine 等）与“生物学相关领域”（如 Biological Sciences, Medical Sciences, Genetics, Pharmaceutical Sciences, Immunology 等）关键词组合，在 PubMed 平台执行检索，筛选并统计 2010~2024 年已发表的研究文献

从理论构想到现实应用，人工智能的发展经历了从规则驱动到数据驱动、从专用系统到通用能力不断拓展的过程。近年来，随着算法创新、算力提升以及海量数据的积累，人工智能已逐渐成为推动科技进步和产业变革的重要力量。然而，系统能力快速提升的同时，也带来了安全性、可靠性、可解释性、隐私保护以及伦理治理等方面的新挑战。如何在促进技术创新的同时建立完善的监管与治理体系，确保人工智能朝着安全、可信、可控的方向发展，已成为学术界、产业界和社会各界共同关注的重要议题。

## 1.2 人工智能的发展历程

### 1.2.1 起源与发展初期

人工智能的思想根源可追溯至古代哲学家和数学家对“形式化”推理及机械思维的早期探索。早在古希腊时期，哲学家们就开始思考思维的本质和逻辑推理的规律，这些早期的哲学探索为后来人工智能的理论发展埋下了种子。而逻辑学的发展为人工智能奠定了重要的理论基础，从亚里士多德的“三段论”到现代数理逻辑，逻辑推理的形式化和规范化为人工智能实现机器推理提供了关键的理论支撑。

在人类文明经历的三次工业革命中，技术革命始终推动着社会变革。第一次

工业革命（1760~1840 年）以蒸汽机为标志实现机械化生产；第二次工业革命（1870~1914 年）通过电力与流水线实现规模化制造；第三次工业革命（1960~2000 年）以计算机和信息技术催生数字化时代。20 世纪中叶，随着计算机技术的逐渐成熟，人工智能迎来了诞生的契机。艾伦·图灵提出的计算理论指出，通过对简单符号“0”和“1”进行操作的机器，能够模拟任意复杂的数学推理过程，这一理论为人工智能的诞生提供了关键的理论支撑，被视为现代计算机科学和人工智能的基石。由此可见，第三次工业革命所形成的技术积累，为第四次工业革命核心技术——人工智能的产生提供了重要基础。与此同时，控制论、信息论及神经生物学等领域的同期发现，促使研究人员开始思索构建“电子大脑”的可能性，为人工智能的发展开辟了新的方向。

1943 年，沃伦·麦卡洛克与沃尔特·皮茨设计的“人工神经元”，为模拟人类大脑的神经网络结构提供了雏形。他们提出的模型基于简单的数学规则，模拟了神经元的兴奋和抑制过程，证明了通过大量简单单元的相互连接可以实现复杂的计算功能，这一成果激发了后续对神经网络的研究兴趣。1950 年，图灵发表的“计算机与智能”一文，提出了“图灵测试”，即通过观察机器与人类之间的对话，判断机器是否具备智能，展示了“机器智能”的可行性，引发了人们对人工智能的广泛关注和深入思考，也为人工智能的发展指明了一个重要的研究方向。

1956 年，在达特茅斯学院举行的“达特茅斯会议”，标志着人工智能研究领域的正式确立。此次会议汇聚了约翰·麦卡锡、马文·明斯基、克劳德·香农等众多顶尖学者，他们后来成为人工智能领域的先驱者。在会议上，研究者探讨了用机器模拟人类智能的各种可能性，并首次使用了“人工智能”这一术语。在这一时期，研发出的程序取得了令人瞩目的成果，计算机不仅能够学会国际跳棋策略，还能解决代数与文字的难题、证明逻辑定理，甚至实现英语对话。这些成就使得人工智能迅速成为热门研究领域。20 世纪 50 年代末至 60 年代初，英国和美国的许多大学纷纷建立人工智能实验室，加大对该领域的研究投入，人工智能迎来了发展的初期繁荣阶段。

在这一阶段，人工智能的研究焦点集中于符号主义这一研究路径。符号主义认为人工智能可以通过对符号的逻辑运算和推理来实现，致力于开发基于规则的系统，将人类的知识和推理过程用符号表示并输入计算机。例如，在定理证明领域，研究人员开发出了能够自动证明数学定理的程序，通过对逻辑规则的应用和搜索，找到从前提到结论的证明路径。这种方法在一些特定领域取得了一定的成功，但也暴露出了局限性，如难以处理不确定和不完整的信息，以及对复杂现实世界问题的适应性较差。

### 1.2.2 寒冬期

20 世纪 60~70 年代，研究人员对创造通用智能机器充满信心，赫伯特·西蒙曾预言“在未来二十年内，机器能够胜任所有人类能够完成的工作”，马文·明斯基也认为“在一代人的时间内，创造‘人工智能’的问题将会得到实质性的解决”。然而，他们低估了人工智能问题的复杂性，实际研究进展远未达到预期目标。

在这一时期，人工智能研究面临诸多技术难题。一方面，早期的人工智能系统依赖于手工编写大量的规则和知识，随着问题领域的扩大和复杂性的增加，这种方法变得难以维护和扩展，知识获取的瓶颈逐渐显现；另一方面，计算机硬件性能有限，计算能力无法满足处理大规模复杂问题的需求，使得许多人工智能算法在实际应用中效率低下。

1974 年，詹姆斯·莱特希尔爵士在其评估报告中指出，当时的人工智能研究在解决实际问题方面进展缓慢，未能达到预期的效果。同时，美国国会倾向于资助更具成效的项目，导致美国和英国政府中断了对人工智能探索性研究的资金支持。此外，马文·明斯基和西摩·佩珀特在 *Perceptrons: An Introduction to Computational Geometry*（《感知器》）一书中提出的观点，被误读为证明人工神经网络在解决现实世界问题上毫无用处，这使得人工神经网络的研究受到极大抑制。在多重因素的影响下，人工智能的发展进入了“寒冬期”，在此期间，人工智能项目难以获得资金支持，研究进展缓慢，许多研究人员纷纷转向其他领域，大量的人工智能实验室关闭，相关研究陷入停滞。

尽管如此，仍有部分研究人员在坚持探索。他们开始反思传统人工智能研究方法的局限性，并尝试寻找新的途径。例如，一些研究人员开始关注启发式搜索算法，通过引入启发式信息来引导搜索过程，提高问题解决的效率。这些探索虽然在当时没有引起广泛关注，但为后来人工智能的复苏奠定了基础。

### 1.2.3 复苏与繁荣

20 世纪 80 年代初期，专家系统的商业化为人工智能研究带来了新的转机。专家系统是一种旨在模拟人类专家经验和分析技巧的人工智能程序，它将领域专家的知识和经验以规则的形式存储在计算机中，通过推理引擎对输入的问题进行分析和解答。例如，在医疗诊断领域，MYCIN 系统能够根据患者的症状、体征和检查结果，诊断感染性疾病并推荐治疗方案；在工业生产领域，XCon 系统可以根据用户的需求配置计算机系统，提高生产效率和准确性。这些专家系统在实际应用中取得了显著的经济效益，到 1985 年，人工智能市场估值超过 10 亿美元。

同时，日本的第五代计算机项目也推动了英国和美国恢复对人工智能学术研究的资助，人工智能研究再度活跃起来。

然而，从1987年Lisp机器（Lisp machine）市场的衰退开始，人工智能研究再次遭遇挫折，进入了第二个较长的低迷期。Lisp机器是专门为人工智能研究和开发的计算机，因其高昂的成本和有限的应用范围，市场需求逐渐萎缩，导致相关企业纷纷倒闭。这一事件对人工智能产业造成了沉重打击，资金投入减少，研究进展受阻。

在此期间，一些研究人员开始探索新的研究方向，如“次符号”方法。罗德尼·布鲁克斯专注于设计能够移动和自持生存的机器，提出了“包容架构”，强调智能体通过与环境的交互逐步构建复杂的行为，而不是依赖于预先构建的复杂知识体系和推理系统。朱迪亚·珀尔、卢特菲·泽德等学者发展出基于合理推断处理不完备或不确定信息的方法，如贝叶斯网络（Bayesian network, BN）和模糊逻辑（fuzzy logic），这些方法能够更好地处理现实世界中的不确定性。杰弗里·辛顿等研究者在“联结主义”（connectionism）及其涉及的神经网络研究方面取得重要突破，1986年，他们提出了反向传播算法（backpropagation algorithm, BP），解决了多层神经网络的训练问题，使得神经网络能够学习复杂的非线性关系。1990年，杨立昆展示的卷积神经网络（convolutional neural network, CNN）能够识别手写数字，为神经网络的广泛应用奠定了基础。

20世纪末至21世纪初期，人工智能利用形式化的数学方法，结合针对特定问题制定的策略，在学术界逐步重建声誉。这种“聚焦”与“规范化”的研究方法促进了人工智能与统计学、经济学及数学等其他学科的交叉合作，使研究成果更具可验证性。在数据挖掘领域，人工智能算法能够从大量数据中发现有价值的模式和规律，为商业决策提供支持；在优化问题中，遗传算法、模拟退火等人工智能方法能够在复杂的搜索空间中找到最优解。至2000年，人工智能领域的研究解决方案得到广泛应用，尽管在20世纪90年代这些方案往往未被明确标识为“人工智能”。

图1-2梳理了人工智能发展历程中具有标志性意义的技术突破与事件节点，涵盖从基础理论到重要产品的演进轨迹。1943年，人工神经元模型的提出奠定了神经网络基础；1950年，图灵测试为机器智能设定评判标准；1956年，“人工智能”术语正式确立；1972年，MYCIN等专家系统得到应用；1990年，CNN实用化；2012年，AlexNet推动深度学习爆发；2015年，阿尔法狗展现了强化学习的非凡能力；2020年，GPT-3开启大语言模型时代；2024年，GPT-4o实现了更自然的多模态交互；2025年，更多中小型模型崛起。以上这些事件完整呈现了人工智能从理论构想走向不同技术分支的关键脉络。

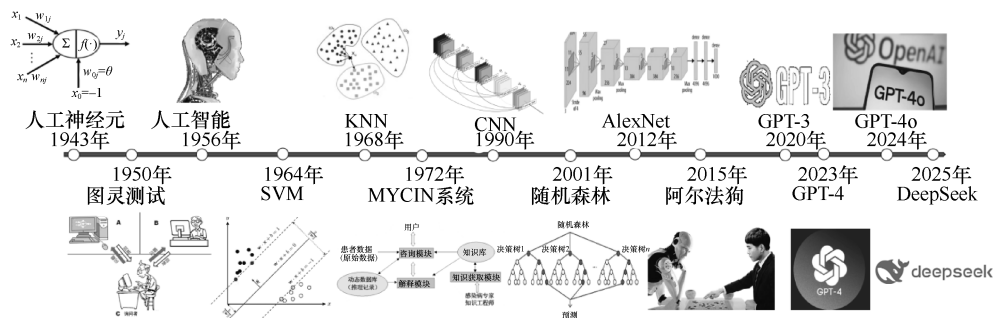


图 1-2 人工智能发展时间轴（彩图请扫封底二维码）

自 2012 年以来，深度学习迅速成为人工智能领域广泛采用的方法。深度学习的突破性成果得益于硬件的显著进步，如运算速度更快的计算机、图形处理单元（graphics processing unit, GPU）和云计算技术，同时也离不开广泛可用的数据，包括精心策划的数据集（如 ImageNet）。2012 年，AlexNet 在 ImageNet 图像识别竞赛中以巨大优势夺冠，展示了深度学习在图像识别领域的强大能力。随后，深度学习在语音识别、自然语言处理、自动驾驶等领域取得了一系列重大突破。例如，在语音识别方面，深度学习模型大幅提高了语音识别的准确率，使得语音助手如 Siri、小爱同学等得到广泛应用；在自然语言处理领域，预训练语言模型如生成式预训练变换模型（generative pre-trained transformer, GPT）系列、双向编码器表示的变换器（bidirectional encoder representations from transformer, BERT）等的出现，实现了对自然语言理解和生成能力的显著提升。

深度学习引发了公众对人工智能的浓厚兴趣，吸引了大量资金投入，2015～2019 年，机器学习领域的出版物数目增加了 50%。2016 年，在机器学习会议上，“公平性”与“技术滥用”成为突出话题，相关研究受到更多关注，“对齐问题”也逐渐成为学术探讨的重要议题。随着人工智能技术的快速发展，人们开始意识到其可能带来的社会和伦理问题，如算法偏见、数据隐私泄露等，因此对人工智能的监管和规范也日益受到重视。

2010 年代末至 2020 年代初，人工智能领域取得一系列突破性进展。2015 年，Google DeepMind 研发的“阿尔法狗”战胜世界围棋冠军，展示了人工智能在复杂策略游戏领域的强大能力，其采用的强化学习和深度学习相结合的方法，为人工智能研究提供了新的思路。2020 年，OpenAI 推出的 GPT-3 通过大规模的无监督学习，在语言理解和生成任务方面表现出色，能够生成高质量、类人的文本，引发了人们对自然语言处理领域的广泛关注。2023 年，OpenAI 推出的 GPT-4，其多模态能力和推理性能显著提升，能处理文本和图像输入，并在复杂问题解决方面展现出更强大的能力；2024 年，OpenAI 发布的 GPT-4o 以更自然的对话能力

获得广泛应用。这些成果促使大型企业大幅增加人工智能研究的投入。

图 1-3 呈现了 2010~2024 年，中国、美国、意大利、德国和英国在人工智能领域的发展趋势。从整体来看，五国的相关指标均呈现持续增长态势，反映出全球人工智能产业在这一时期的快速扩张。其中，美国和中国的增长曲线尤为显著，不仅起点较高，且增速领先于其他三国，体现了两国在人工智能核心技术研发、市场应用及政策支持方面的突出地位；德国、英国作为欧洲科技强国，保持稳步增长，在特定领域展现出特色优势；意大利则在整体增长趋势中紧随其后，反映了人工智能技术在全球范围内的普及与渗透。该图直观展示了人工智能发展的地域格局，凸显了国际上在技术竞争与合作中的动态平衡，为理解全球人工智能产业的演化提供了重要视角。

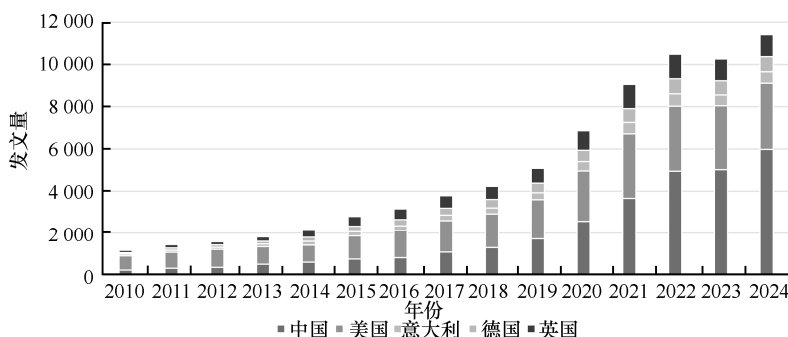


图 1-3 2010~2024 年人工智能在生命科学领域国家发文量统计柱状图（彩图请扫封底二维码）  
通过在 PubMed 平台执行高级检索，使用关键词组合 artificial intelligence (AI), machine learning, deep learning, neural network, 限定文献类型为 journal article 和 review, 时间范围为 2010 年 1 月 1 日至 2024 年 12 月 31 日。检索结果按通讯作者所属国家/地区进行分组统计，选取发文量排名前五的国家（中国、美国、意大利、德国、英国），提取其逐年发表量数据作为分析基础。对检索结果中符合上述算法关键词的文献进行人工审阅和归类（部分文献可能涉及多种算法，统计时计入其提及的主要方法或核心方法）。统计每种算法在上述生命科学相关文献中被明确提及或作为主要方法应用的总条目数，然后将每种算法对应的文献条目数除以所有统计的目标算法文献条目总数，可得出其在统计样本中的相对占比

截至 2024 年，人工智能技术的应用范围进一步扩大，全球人工智能市场规模已达到约 1640 亿美元，并预计以年复合增长率（compound annual growth rate, CAGR）为 36.6% 的速度持续增长。在各个行业，人工智能都展现出了巨大的应用潜力。在医疗行业，人工智能辅助诊断系统能够帮助医生更准确地诊断疾病，提高治疗效果；在金融行业，人工智能用于风险评估、欺诈检测和投资决策，提升金融服务的效率 and 安全性；在交通行业，自动驾驶技术正在逐步改变人们的出行方式。2025 年，中小型人工智能模型开始崛起，例如，DeepSeek 推出的 DeepSeekV3 和 DeepSeekR1，以及 AlibabaGroup 发布的 QwQ-32B，以更低的计算需求提供高效性能，受到市场的广泛关注。这些模型的出现为人工智能的应用

带来了更多的灵活性和可扩展性，降低了应用门槛，使得更多的企业和开发者能够利用人工智能技术解决实际问题。

### 1.3 人工智能在生命科学中应用的现状与前景

自 2010 年以来，人工智能技术在设备更新的推动下逐渐成为一种重要的工具，尤其是机器学习和深度学习方面，在提升计算机视觉、提升自然语言处理和语音识别的准确率，以及变革生命科学研究方面发挥了重要作用。人工智能技术始终引领着科学创新的潮流<sup>[1]</sup>。从医学诊断<sup>[2]</sup>到气候跟踪<sup>[3]</sup>，其研究范畴广泛且不断演进。其中，生命科学研究领域在人工智能所取得的重大进展中占据关键地位<sup>[4]</sup>。

生命科学研究专注于动植物和微生物的分类、行为、生理、发育、遗传、进化、疾病模型构建及古生命科学等方面。但是却长期面临着管理和解释庞大复杂数据集的挑战。随着人工智能技术，尤其是机器学习<sup>[5]</sup>及其子领域深度学习<sup>[6]</sup>的发展和大数据的兴起，人工智能与大数据在生物研究中开启了智能研究时代。

然而，尽管人工智能在生命科学研究中的应用已有较长时间，但相较于生命科学的其他研究领域，其在研究中的应用仍显滞后。究其原因，一方面可能是计算资源效率低下且缺乏足够的生物数据集，阻碍了其应用；另一方面，生命科学研究者可能缺乏掌握和应用这些方法所需的基础知识，不确定如何为研究目标选择最合适的模型。同时，不断发展的、复杂的人工智能模型架构，如 BERT<sup>[7]</sup>、T5 (text-to-text transfer transformer 5)<sup>[8]</sup>、ESM2 (evolutionary scale modeling-2, ESM2)<sup>[9]</sup>等，使生命科学研究人员难以跟上最新进展。尽管强大的计算资源和广泛的生物数据集日益可得，可能促使这些算法在未来成为主流，但许多生物学家对这些技术仍然知之甚少。为了更高效地开展生命科学研究工作，研究者需掌握在何时、为何以及如何应用这些方法，并对相应的数据类型有基本了解。

本节将介绍人工智能及其主要任务，阐明关键模型、数据集以及目前存在的困难和挑战，进而深入探究生物与人工智能相结合的领域，研究人工智能应用如何彻底改变生命科学研究的各个方面；通过分析现实世界的案例研究以及预测未来方向，全面概述人工智能在加深我们对生物王国的理解以及未来可能开启的潜在领域方面的作用。

#### 1.3.1 现状

人工智能的发展已渗透至科技与产业的各个维度，其技术突破与多元应用正持续重构社会发展图景。在通用技术领域，人工智能通过算法、算力与数据

的协同进化，构建起从基础层到应用层的完整生态体系。基础层中，芯片技术成为核心竞争力——英伟达公司的 GPU 凭借强大并行计算能力，支撑起深度学习训练的半壁江山；谷歌公司的张量处理单元（tensor processing unit, TPU）作为专用集成电路，以定制化架构优化推理效率，使阿尔法狗等系统的实时决策成为可能；苹果公司的神经引擎则聚焦终端设备，将人脸识别、语音交互等人工智能任务本地化处理，兼顾隐私保护与响应速度。传感器技术如同人工智能的“感知器官”，如视觉传感器通过 3D 结构光技术实现毫米级精度的环境建模、声音传感器借助降噪算法将语音识别准确率提升至 97% 以上，它们为智能设备赋予了类人的感知能力。

与此同时，中国在人工智能核心技术领域也取得了显著进展。在芯片技术方面，华为公司的昇腾系列（如昇腾 910）凭借达芬奇架构，在深度学习训练和推理方面提供强大算力，并与百度公司的飞桨等框架深度协同，构建“端、边、云”全场景人工智能基础设施。阿里巴巴公司的含光 800 芯片则专注于云端人工智能推理，尤其在视觉识别方面表现卓越，其性能和能效比均达到全球领先水平，已在“拍立淘”和“城市大脑”等内部业务中大规模应用。百度公司的昆仑芯 2 代人工智能芯片采用 7nm 工艺，搭载 XPU-R 架构，集成了编解码子系统，实现了“编解码+AI 一站式处理”，在互联网搜索、智慧城市和工业质检等领域广泛部署。此外，寒武纪公司、燧原科技、壁仞科技等企业也在各自的人工智能芯片领域展现出强大的竞争力。

在传感器技术方面，中国在 3D 视觉和语音 AI 领域亦有突破。视觉传感器方面，3D 结构光和飞行时间（time of flight, ToF）技术发展迅速，广泛应用于人脸识别、VR/AR 以及汽车驾驶辅助等场景，奥比中光、舜宇智能光学等企业是行业中的佼佼者。声音传感器及语音 AI 方面，语音识别技术准确率已超 95%，中国科学院声学研究所、腾讯天籁实验室、一知智能等机构和企业 在降噪算法上取得显著进展，有效提升了嘈杂环境下的语音通信和识别体验。科大讯飞、百度、阿里巴巴、腾讯等公司在语音识别市场占据领先地位。

在人工智能生态系统构建上，国产深度学习框架的崛起为算法、算力与数据的协同进化提供了坚实支撑。百度飞桨（PaddlePaddle）作为中国首个成熟的深度学习开源框架，在中国市场所占份额处于领先地位，并与华为昇腾 AI 联合打造软硬一体的人工智能大模型平台。华为的昇思（MindSpore）框架也与昇腾 AI 处理器紧密协同，致力于实现易开发、高效执行和全场景覆盖。这种软硬件协同被视作中国人工智能产业发展的核心竞争力，正加速推动人工智能应用的广泛落地和产业链的高质量发展。

技术层的突破则以机器学习为核心引擎，推动计算机视觉、自然语言处理

(natural language processing, NLP) 等领域跨越式发展。计算机视觉在安防场景中展现出惊人精度,例如,百度大脑的 1:N 人脸识别系统准确率达 99.7%,可在百万级人脸库中秒级定位目标;自然语言处理领域,GPT-4 等大模型可实现上下文理解与多模态生成,其编写代码、分析财报等能力已接近人类专业水平。机器学习算法更成为各行业效率革新的“杠杆”,例如,在金融领域,智能风控系统通过千万级特征变量分析,将信贷违约预测的准确率提升 35%;在工业领域,AI 视觉检测系统将产品缺陷率从 0.3%降至 0.02%,颠覆了传统质检模式。

在生物医学影像分析领域,深度学习算法宛如一位不知疲倦的“影像侦探”,为疾病诊断带来革命性变革。通过对海量 X 射线(X-ray)、计算机断层扫描(computed tomography, CT)、磁共振成像(magnetic resonance imaging, MRI)等医学影像数据的深度“研习”,计算机视觉模型已具备精准识别影像中细微病变的超凡能力。以肺癌早期筛查的关键环节——肺结节检测为例,人工智能系统能够以极快的速度扫描肺部 CT 影像,敏锐捕捉毫米级的肺结节,并借助复杂算法精准判断其良恶性概率,为医生的诊断决策提供有力支持。在乳腺癌诊断方面,基于深度学习的图像分析技术更是大显身手,它能够敏锐捕捉乳腺 X 光片中的微小钙化点与肿块特征,显著提升乳腺癌的早期检出率,让更多患者能够在疾病萌芽阶段就得到及时诊治。部分前沿的人工智能影像诊断系统,对特定疾病诊断的准确率已与经验丰富的放射科医生旗鼓相当,甚至在细微特征识别等方面更胜一筹,展现出人工智能在医学影像分析领域的巨大潜力。

药物研发流程因人工智能的强势介入,正经历着脱胎换骨的变革。传统药物研发宛如一场耗时耗力的马拉松,漫长的周期与高昂的成本令人望而却步。如今,机器学习算法成为科研人员的得力助手,能够对海量化学分子结构数据、生物活性数据进行高效分析,精准预测潜在药物靶点与化合物的结合能力。例如,生成式对抗网络(generative adversarial network, GAN)和变分自编码器(variational autoencoder, VAE)等先进人工智能技术,宛如分子结构设计的“神奇画笔”,能够创造出全新的化学分子结构,极大地拓展药物研发的分子库,为新药研发开辟广阔空间。AlphaFold 的问世更是在蛋白质结构预测领域掀起一场“风暴”,这个基于深度学习模型的“结构破解大师”,成功攻克了困扰生物学界数十年的蛋白质三维结构预测难题,为基于蛋白质结构的药物设计按下“加速键”,为靶向药物研发奠定了不可或缺的关键基础。

在动物学研究中,人工智能深度融入多个研究方向。借助古生物学数据库,科研人员利用人工智能分析史前动物数据,探索生物演化奥秘;AP-10K 数据集助力动物姿势估计研究,通过计算机视觉技术精准捕捉动物姿态变化。在野生动物监测方面,研究者可以运用深度神经网络对农场建筑中的野生动物和家畜图像进