“十四五”普通高等教育本科部委级规划教材

普通高等学校生物制药产教融合系列教材

人工智能与智慧制药

Rengong Zhineng Yu Zhihui Zhiyao

配套多媒体资源

孙先宇 何卫刚 王颖◎主编

中国纺织出版社有限公司
国家一级出版社
全国百佳图书出版单位

“十四五”普通高等教育本科部委级规划教材

普通高等学校生物制药产教融合系列教材

人工智能与智慧制药

Rengong Zhineng Yu Zhihui Zhiyao

孙先宇 何卫刚 王颖◎主编

中国纺织出版社有限公司

图书在版编目（CIP）数据

人工智能与智慧制药 / 孙先宇，何卫刚，王颖主编
. --北京：中国纺织出版社有限公司，2023. 11
“十四五”普通高等教育本科部委级规划教材
ISBN 978-7-5229-0961-5

Ⅰ. ①人… Ⅱ. ①孙… ②何… ③王… Ⅲ. ①人工智能－应用－制药工业－高等学校－教材 Ⅳ. ①TQ46－39

中国国家版本馆 CIP 数据核字（2023）第 241134 号

责任编辑：金 鑫 闫 婷 责任校对：江思飞
责任印制：王艳丽

中国纺织出版社有限公司出版发行
地址：北京市朝阳区百子湾东里 A407 号楼 邮政编码：100124
销售电话：010—67004422 传真：010—87155801
http：//www. c-textilep. com
中国纺织出版社天猫旗舰店
官方微博 http：//weibo. com/2119887771
三河市宏盛印务有限公司印刷 各地新华书店经销
2023 年 11 月第 1 版第 1 次印刷
开本：787×1092 1/16 印张：10. 5
字数：195 千字 定价：49. 80 元

普通高等学校生物制药产教融合系列教材
编委会成员

《人工智能与智慧制药》编委会

主　编　孙先宇　常熟理工学院

何卫刚　常熟理工学院

王　颖　常熟理工学院

编　委　（按姓氏笔画排序）

王　颖　常熟理工学院

邓先清　井岗山大学

孙先宇　常熟理工学院

何卫刚　常熟理工学院

张立秋　嘉兴学院

张　扬　常熟理工学院

陈梦玲　常熟理工学院

郭利军　上海药坦药物研究开发有限公司

冀　宏　常熟理工学院

前　言

药物的发现和设计包括漫长而复杂的步骤，如靶点选择和验证、先导化合物优化、临床前试验和临床试验以及生产实践。在过去的几十年里，药物化学家和生物科学家致力于以最大的效率开发靶向治疗药物。创新药物研究的成本和时间消耗是药物设计和开发过程中的巨大障碍。为了最大限度地克服这些挑战和障碍，全球研究人员越来越多地借助虚拟筛选和分子对接等计算方法，大大提高先导物的筛选效率，但这些技术也伴随诸如不准确性和效率仍不理想等问题。

人工智能的兴起和不断完善，包括深度学习和机器学习算法，尤其是其在药物研究领域的应用已经成为一种可能的解决方案。相较传统方式，它能以简单科学的方式解决现实问题，克服药物设计和发现过程中的问题和障碍。新技术的实现，足以消除传统计算方法中遇到的挑战，在一定程度上减少了药物研究过程的时间消耗和成本。生物和化学科学家将人工智能算法广泛应用于药物设计和发现过程中。基于人工智能和机器学习原理的计算建模为化合物的识别和验证、靶标识别、肽合成、药物毒性和理化性质评估、药物监测、药物疗效和有效性以及药物重新定位提供了一条很好的途径。随着人工智能原理以及机器学习和深度学习算法的出现，来自化学文库的化合物虚拟筛选（包括超过 1 亿种化合物）变得简单且高效。此外，人工智能模型消除了由于靶标相互作用而产生的毒性问题。

为了更好地介绍人工智能在药物研究各领域的应用及相关技术，我们组织编写了此书。本书积极响应党的二十大精神，简要讨论了人工智能从机器学习到深度学习的演变，以及大数据参与药物发现过程的革命性变化。全书分为 7 章。第 1 章概述了人工智能的起源和发展，重点论述了经典机器学习方法和最新机器学习方法；第 2 章论述了人工智能在蛋白质结构方向中的应用，包括蛋白质活性及毒性预测，以及蛋白质与蛋白质间的相互作用；第 3 章主要论述了人工智能在药物筛选过程中，对化合物成药性相关的物理、化学性质的预测，以及结构与活性关系预测；第 4 章主要论述了人工智能在改进传统药物发现过程方面的应用，如一次和二次筛选、药物毒性、药物释放和监测、药物剂量的有效性、药物重新定位和多药效以及药物–靶标相互作用；第 5 章从联合治疗、评估药物反应及药物–药物相互作用等方面，介绍了人工智能在药物治疗方面的应用。第 6 章论述了多组学数据和生物网络，介绍了药物代谢和药效相关基因，并给出人工智能的药物基因识别方法和应用实践。第 7 章主要介绍了制药工业智能制造的

系统构架，人工智能在药物生产过程中的应用场景，以及我国制药行业智能制造的发展现状。

本书由常熟理工学院生物与食品工程学院孙先宇、何卫刚，常熟理工学院计算机学院王颖共同编写。其中第1章、第2章由孙先宇编写，第3章由何卫刚编写，第4章、第5章、第6章由王颖编写，第7章由上海药坦研究开发有限公司郭利军编写。

由于人工智能技术涉及广泛且发展迅速，加之编者视野及水平限制，书中难免存在疏漏和不足，敬请广大读者批评指正。

编者

2023年9月

目　录

本书 PPT

第 1 章
人工智能

人工智能作为新一代数字技术的典型代表，逐渐从专业领域走向实际应用。在日常生活领域，人类围棋冠军被电脑击败，智能手机利用了面部识别算法，自动驾驶汽车在街道上行驶；在医学领域，食品和药物管理局已经允许临床医生在不同的医疗领域使用人工智能，如人工智能现在可以常规检测糖尿病视网膜病变，而不需要眼科医生来确认该算法的诊断结果。

作为新一轮产业变革的核心驱动力，人工智能将催生新的技术、产品、产业、业态、模式，从而引发经济结构的重大变革，实现社会生产力的整体提升。当前，人工智能发展进入了新阶段，涉及数学、神经生理学、计算机科学、信息控制论、生物学、语言学、心理学等多门学科，是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的交叉性、边缘性学科。人工智能的研究内容包括知识表示和知识图谱、自动推理、专家系统、群智能算法等，目标是使机器能完成一些原来只有人类才能完成的复杂性工作（图 1-1）。

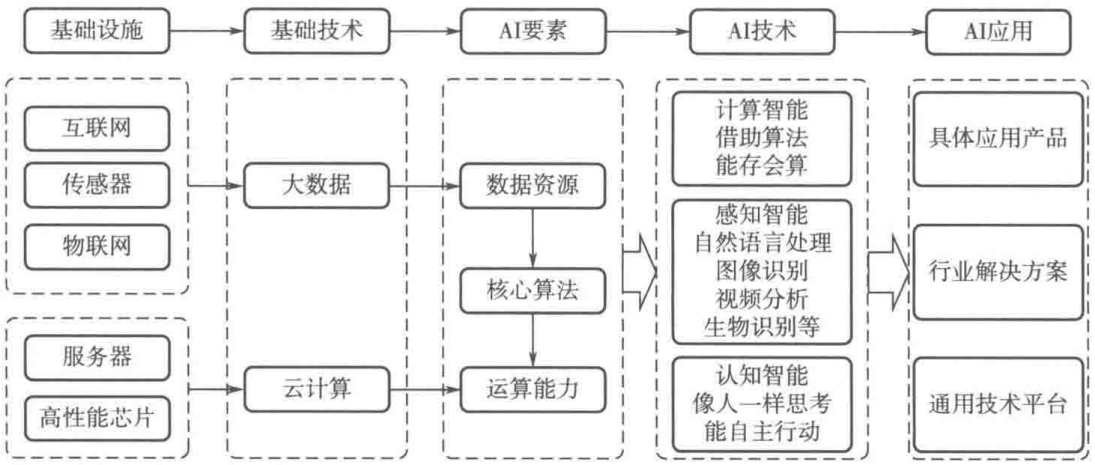

图 1-1　人工智能产业链

人工智能是一个总括术语，指计算机程序能够像人类一样思考和执行行为，而机器学习则超越了将数据与决策树、隐马尔可夫模型等算法一起输入机器的范畴，这有助于机器在不被明确编程的情况下进行学习。一些人将机器学习描述为主要的人工智能应用程序，而另一些人则描述为人工智能的一个子集。后来，随着神经网络的发展，机器可以像人脑一样对输入的数据进行分类和组织，这进一步促进了人工智能的进步。20 世纪前后，出

现了“深度学习”一词。深度学习是机器学习的子集，机器学习本身就是人工智能的子集，因此，它们之间的关系类似于人工智能>机器学习>深度学习。

机器学习要么使用监督学习，其中模型被训练为使用标记数据，这意味着输入已经被相应的优选输出标签标记，要么使用无监督学习，在其中模型被培训为使用未标记数据，但从输入数据中寻找重复模式。另外还有使用监督和非监督学习相结合的半监督学习；自监督学习是一种特殊情况，它采用两步过程，即无监督学习为未标记的数据生成标签，其最终目标是建立监督学习模型；强化学习是一种机器学习，它在恒定反馈回路的帮助下随着时间的推移改进了算法。最后是深度学习，其中有许多层机器学习算法，被称为模仿人脑的受大脑启发的算法家族，但需要高计算能力才能进行训练和大数据分析。

机器学习的起源可以追溯到 1943 年，当时 McCulloch 和 Pitts 发表了一篇题为《神经活动中固有思想的逻辑演算》的文章，他们在文章中给出了有史以来第一个神经网络的数学模型。Alan M. Turing 在 1950 年发表的开创性论文中对机器学习的概念进行了理论化。1952 年，Arthur L. Samuel 为 IBM 编写了一个棋盘游戏程序，推广了“机器学习”一词。1957 年，Frank Rosenblatt 开发了用于图像识别的感知器。Henry J. Kelley 于 1960 年开发了连续反向传播模型，Stuart Dreyfus 于 1962 年开发了一个仅基于链式规则的更简单版本。1965 年，Ivakhnenko 和 Lapa 开发了第一个可工作的深度学习网络。1980 年左右，福岛邦彦开发了一种名为新认知机的人工神经网络，该网络具有多层设计，可以帮助计算机学习如何识别视觉模式，他还开发了基于动物视觉皮层组织的卷积神经网络。

1.1 人工智能概述

人工智能，最早由约翰·麦卡锡在 1956 年的达特茅斯会议上提出，用来描述“制造智能机器的科学和工程”。麦卡锡最初的描述今天仍然成立，但细节上有了丰富的扩展。简单地说，人工智能可以被看作是对计算的研究，它使感知、推理和行为预测成为可能。人工智能本身可以被用来执行各种任务，但另一种用途是利用人工智能算法来增强人类智能，而不是取代它，这个概念被称为增强智能。

斯坦福大学的尼尔斯教授认为：“人工智能致力于使机器智能化，智能化是衡量实体在一定环境中反应和判断的定量值。”麻省理工学院的温斯顿教授认为：“人工智能是研究如何使计算机去做只有人类才能做的智能工作。”这些说法反映了人工智能学科的基本思想和基本内容，即人工智能是研究人类智能活动的规律，构造具有一定智能的机器系统，最终让计算机完成人类大脑才能胜任的工作。总地说来，人工智能是研究、开发用于模拟、延伸和扩展人的智能理论、方法、技术及应用系统的一门科学。

人工智能的探索道路充满曲折和起伏，发展并非一凡风顺，经历了 20 世纪中期的人工智能浪潮期，也经历了 20 世纪七八十年代的沉寂期，最终在 21 世纪初迎来了发展黄金期。随着大数据、云计算、物联网等信息技术的发展，泛在感知数据和图形处理器等计算平台推动人工智能技术飞速发展，大幅跨越了科学与应用之间的鸿沟。人工智能自 1956 年以来的发展历程，大至可以分为以下 6 个阶段（图 1-2）。

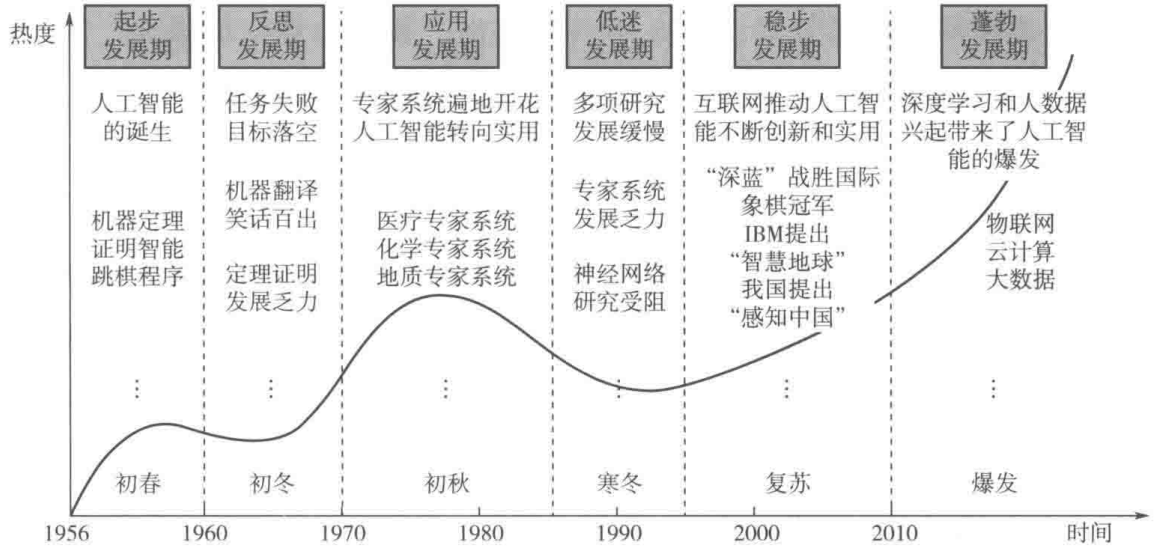

图 1-2　人工智能的发展历程

（1）起步发展期：1956—20 世纪 60 年代初。

达特茅斯会议确立了人工智能这一术语，之后又陆续出现了如跳棋程序、感知神经网络软件和聊天软件等，并且人们用机器证明的办法去证明和推理一些定理，相继取得了一批令人瞩目的研究成果，掀起人工智能发展的第一个高潮。

（2）反思发展期：20 世纪 60—70 年代初。

人工智能发展初期的突破性进展大大提升了人们对人工智能的期望，人们开始尝试更具挑战性的任务，并提出了一些不切实际的研发目标。然而，接二连三的失败和预期目标的落空（例如，无法用机器证明两个连续函数之和还是连续函数，机器翻译闹出笑话等），使人工智能的发展走入低谷，人工智能进入第一次寒冬。

（3）应用发展期：20 世纪 70 年代初—80 年代中期。

20 世纪 70 年代出现的系统模拟人类专家的知识和经验解决特定领域的问题，实现了人工智能从理论研究走向实际应用、从一般推理策略探讨转向专门知识运用的重大突破。人工智能在医疗、化学、地质等领域取得的成功，推动人工智能进入应用发展的新高潮。

（4）低迷发展期：20 世纪 80 年代中—90 年代中期。

随着人工智能的应用规模不断扩大，专家系统存在的应用领域狭窄、缺乏常识性知

识、知识获取困难、推理方法单一、缺乏分布式功能、难以与现有数据库兼容等问题逐渐暴露出来，人工智能进入第二次寒冬。

（5）稳步发展期：20 世纪 90 年代—2010 年。

网络技术特别是互联网技术的发展，加速了人工智能的创新研究，促使人工智能技术进一步走向实用化。1997 年，国际商业机器公司（IBM）“深蓝”超级计算机战胜了国际象棋世界冠军卡斯帕罗夫；2002 年，iRobot 公司打造出全球首款家用自动扫地机器人；2006 年出现深度学习技术；2008 年 IBM 提出“智慧地球”的概念，与此同时，Siri、Alexa、Cortana 等语音识别应用在智能手机上得到应用，以上都是这一时期的标志性事件。

（6）蓬勃发展期：2010 年以来。

随着大数据、云计算、互联网、物联网等信息技术的发展，泛在感知数据和图形处理器等计算平台推动以深度神经网络为代表的人工智能技术的飞速发展，大幅跨越了科学与应用之间的“技术鸿沟”，诸如图像分类、语音识别、知识问答、人机对弈、无人驾驶等人工智能技术实现了从“不能用、不好用”到“可以用”的技术突破。同时，这一轮人工智能发展的影响已经不局限于学界，政府、企业、非营利机构都开始拥抱人工智能技术。人们身处的第三次人工智能浪潮仅仅是一个开始，人工智能的高速发展将揭开一个新时代的帷幕，迎来爆发式增长的蓬勃发展期。

作为一个多学科领域，人工智能涉及来自不同学科的知识，如计算机科学、数学、心理学、语言学、哲学、神经科学、人工心理学和许多其他领域。这些领域的知识和工程进步，帮助人工智能从纯理论研究发展到解决我们生活各个方面问题的智能系统。

1.2 机器学习

从 20 世纪 50 年代提出机器学习，到今天多媒体、图形学、网络通信、软件工程乃至体系结构芯片设计，都能找到机器学习技术的身影。机器学习已经成为最重要的技术进步源泉之一。为了使读者对机器学习有一个初步的了解，本节将对机器学习的发展历程和相关基础概念、范畴进行说明，使读者对机器学习有一个基本的认识。

人工智能的一个主要分支被称为机器学习（图 1-3）。机器学习可以被定义为一组能够从经验中学习和改进的算法，而无须为特定的任务进行显式的编程。这一特性使机器学习本质上不同与经典的计算方法。机器学习与其他类型的计算机编程的不同之处在于，它使用统计、数据驱动的规则将算法的输入转换为输出，这些规则是从大量示例中自动派生的，而不是由人类明确指定的（图 1-4、图 1-5）。

图 1-3　机器学习与相关学科

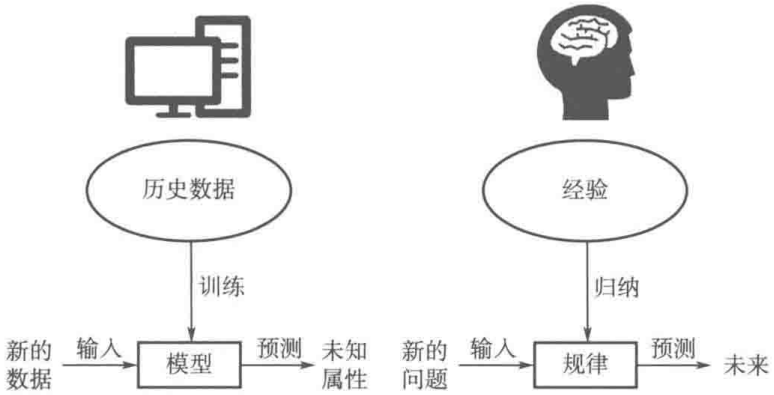

图 1-4　机器学习与人类学习对比

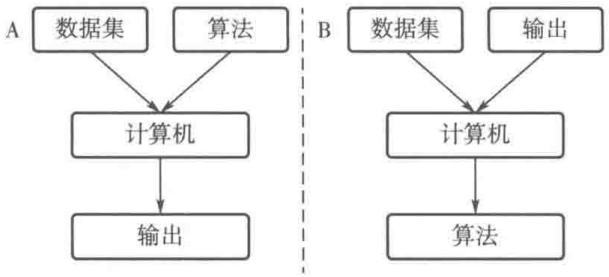

图 1-5　经典编程与机器学习模式

机器学习算法的流行是随机森林、支持向量机、人工神经网络和深度学习的发展而来的。由于人工神经网络和深度学习在医药人工智能中占主导地位，本节将更详细地讨论它们。

机器学习是人工智能发展到一定时期的必然产物。在 20 世纪 50—70 年代初，人工智能的研究处于“推理期”，那时的人们以为只要赋予机器逻辑推理能力，机器就能具有智能。然而随着研究的发展，人们逐渐认识到仅仅具有逻辑推理能力是无法实现人工智能的。E. A. Feigenbaum 等人认为，要使机器学习具有人工智能，就必须设法使机器具有知

识。在他们的倡导下，从20世纪70年代中期开始，人工智能的研究进入“知识期”。在这一时期，大量专家系统问世，在很多应用领域取得了大量的成果。但人们也意识到专家系统面临“知识工程瓶颈”，也就是说，人们把知识总结出来交给计算机是相当困难的。于是，一些学者想到让机器自己去学习知识。

20世纪80年代，从样例中学习的一大主流是符号主义学习，其代表包括决策树和基于逻辑的学习。典型的决策树学习以信息论为基础，以信息熵的最小化为目标，直接模拟了人类对概念进行判定的树形流程。而基于逻辑的学习，其著名代表是归纳逻辑程序设计，可以看作机器学习与逻辑程序设计的交叉。这两种方法各有特点，决策树简单易用，直至今天仍然是机器学习常用技术之一。基于逻辑的学习具有很强的知识表达能力，可以较容易地表达出复杂数据关系。在20世纪90年代中期之前，“从样例中学习”的另一主流技术是基于神经网络的连接主义学习。连接主义学习在20世纪50年代取得了很大发展，但因为早期人工智能的研究者对符号表示特别偏爱，而且连接主义自身也遇到了很大的障碍，当时的神经网络只能处理线性问题，甚至都处理不了“异或”这种简单问题。直到J. J. Hopfield利用神经网络求解“流动推销员问题”这个著名的NP难题取得重大进展，才使连接主义重新受到人们的关注。1986年，D. E. Rumelhart等人发明了著名的BP算法，产生了深远的影响，使连接主义的发展突飞猛进。而且BP算法一直是应用最为广泛的机器学习算法之一。

20世纪90年代中期，“统计学习”闪亮登场并迅速占领“从样例中学习”的主流舞台，其代表性技术是支持向量机。其实这方面的研究早在20世纪60—70年代就已经开始，统计学习理论在那时就已经打下了基础，但直至20世纪90年代才成为机器学习的主流技术。一方面是因为有效的支持向量机算法在20世纪90年代初才被提出，并且其优越的性能在20世纪90年代中期文本分类的应用中才得以显现；另一方面，正是在连接主义学习技术的局限性凸显之后，人们才把目光转向以统计学习理论为直接支撑的统计学习技术。在支持向量机被普遍接受后，该技巧被人们利用到机器学习的每个角落，该方法也成为机器学习的基本内容之一。

21世纪初期，随着社会进入大数据时代，数据量和计算设备的发展使连接主义技术焕发出新的生机，掀起了以“深度学习”为名的热潮。所谓的深度学习，狭义地说就是很多层神经网络。在涉及语音、图像等复杂对象的应用中，深度学习技术具有优越的性能。虽然深度学习模型复杂度高，参数较多，但如果下功夫“调参”，把参数调节好，其性能往往较好。因此，深度学习虽然缺乏严格的理论基础，但是显著降低了机器学习应用者的门槛，为机器学习走向工程实践带来诸多便利。

机器学习是现阶段实现人工智能应用的主要手段和方法，在人工智能体系中处于基础

与核心的地位，它被广泛地应用于计算机视觉、语音识别、自然语言处理、数据挖掘等领域（图 1-6）。

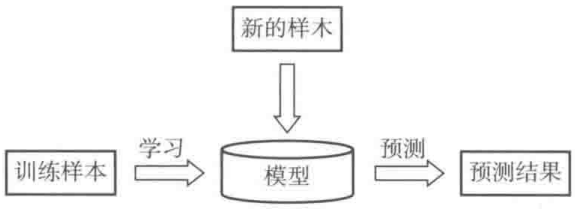

图 1-6　机器学习任务的一般流程

机器学习策略可以广泛地分为无监督学习和监督学习。无监督学习的重点是发现数据集中变量之间的潜在结构或关系，而监督学习通常涉及将观察结果分类为 1 个或多个类别或结果。监督学习需要一个具有预测变量和标记结果的数据集。特征选择对于预测建模至关重要，而机器学习对它特别有用。

例如，在医学上，当观察结果通常有“病例”或“对照”等标签时，就可以进行预测建模，并且这些观察结果与相关特征如年龄、性别或临床变量配对。假设考虑一个医生希望预测某患者是否会在入院后 30 天内再次入院。对于这个困难的问题，机器学习技术已被证明可以改进传统的统计方法。我们假设的临床医生拥有一个大但“混乱”的电子健康记录数据集。通常，电子病历包括一些变量，如病历号、药物处方、检测指标、影像学数据等。利用人口统计学、实验室值和生命体征，各种不同的算法在临床显著程度上优于逻辑回归。

事实上，人们很难决定在预测模型中应该包含哪些变量。当自变量大于观测变量时，拟合逻辑回归模型在代数上是困难的。人们通常使用单变量方差筛选或正向逐步回归等技术。不幸的是，这些方法导致的模型不倾向于在其他数据集中进行验证，并且不适用于患者使用。

对于机器学习，变量之间经常存在复杂的相互作用。这种交互作用的数量和质量很难用传统的方法来描述。通过机器学习，我们可以捕获和使用这些复杂的关系。由无监督学习设计的特征也经常被纳入监督学习模型中，在比较机器学习特征选择的方法中证明了机器学习特征选择的实用性。

缺乏可解释性可能是大多数机器学习算法的主要缺点，包括人工神经网络。换句话说，不可能准确地理解一个网络是如何近似于一个特定的函数。这种“暗箱”行为的一个直接结果是，不可能预测输入的微小变化将如何影响网络的预测能力。作为人工神经网络输入的图片，一些难以查觉的微小差别，可能会导致网络性能的急剧下降。相反，对同一张图片进行重大修改，使人眼无法再识别它们，并没有改变人工神经网络对图像进行分类

的方式。在实践中，当使用人工神经网络代替线性模型时，预测性能的提高应平衡可解释性的损失。

在药物发现项目中使用的大多数学习任务和技术可分为七大类：监督学习、无监督学习、半监督学习、主动学习、强化学习、迁移学习和多任务学习。每一类都有自己的特征性的优点和局限性。

（1）监督学习。

监督学习是指在存在标记的样本数据中进行模型训练的过程，是机器学习中应用最为成熟的学习方法。其中数据存在标记的主要功能是提供误差的精确度量，也就是当数据输入到模型中得到模型预测值，能够与真实值进行比较得到误差的精确度量。在监督学习的过程（即建立预测模型的过程）中，可以根据误差的精确度量对预测模型进行不断调整，直到预测模型的结果达到一个预期的准确率，这样模型的准确性可以得到一定的保证。监督学习常见的应用场景有分类问题和回归问题。两者的区别主要在于对待预测的结果是否为离散值，若待预测的数据是离散的，此类学习任务称为分类；若待预测的数据为连续的，则此类任务称为“回归”。在分类问题中只涉及两个类别的分类问题，人们一般称其中一个为正类（positive class），一个为反类（negative class）。当涉及多个类别时，则称为多分类任务。常见的监督学习应用包括基于回归或分类的预测性分析、垃圾邮件检测、模式检测、自然语言处理、情感分析、自动图像分类等。

监督学习用于描述预测任务，因为其目标是预测或分类感兴趣的特定结果。监督学习已被应用于包括人口统计学、临床和社会预测因素在内的大型数据结构。

（2）无监督学习。

与监督学习相对应，在不存在标记的样本数据中建立机器学习模型的过程称为无监督学习。由于不存在标记数据，所以有绝对误差的衡量。无监督学习中得到的模型大多是为了推断此数据的内在结构，其中应用最广、研究最多的就是“聚类”，其可以根据训练数据中数据之间的相似度，对数据进行聚类（分组）。经过聚类得到的簇也就是形成的分组可能对应一些潜在的概念划分，进而厘清数据的内在结构。如一批图形数据通过聚类算法可以将三角图形确定一个集合，圆点图形确定一个集合。经过这样的过程可以为下一步具体的数据分析建立基础，但需要注意，聚类过程仅能自动形成簇结构，但是簇对应的具体语义要使用者来进行命名和把握。其实从过程也可以看出无监督学习方法在于寻找数据集的规律性，这种规律性不一定要达到划分数据集的目的，也就是说不一定要对数据进行“分类”，而且无监督学习方法所需训练数据是不存在标记的数据集，这就使无监督学习比监督学习用途更广，如分析一堆数据的主分量或者分析数据集有什么特点都可以归为无监督学习。常见的无监督应用包括对象分割、相似性检测、自动标记、推荐引擎等。