# 人工智能在药物发现中的应用与挑战

梁礼，邓成龙，张艳敏，滑艺，刘海春，陆涛，陈亚东 *

（中国药科大学理学院，江苏 南京 211198）

[摘要] 新药研发存在周期长、费用高和成功率低等特点。人工智能技术是近些年来的热点技术之一，在很多领域都有非常广泛的应用，多种人工智能方法已经成功应用于药物的发现过程。综述总结了常用机器学习方法和深度学习方法在药物研发领域中的应用，同时也提出了人工智能存在的问题和面临的挑战。整体而言，人工智能技术在药物研发领域发展潜力巨大，将为医药发展带来新的机遇和希望。

[关键词] 人工智能；机器学习；深度学习；药物研发

[中图分类号] R9-39；R918 [文献标志码] A [文章编号] 1001-5094（2020）01-0018-10

# Application and Challenges of Artificial Intelligence in Drug Discovery

LIANG Li, DENG Chenglong, ZHANG Yanmin, HUA Yi, LIU Haichun, LU Tao, CHEN Yadong

*( School of Science, China Pharmaceutical University, Nanjing 211198, China )*

[Abstract] The development of new drugs usually requires a long time and huge cost with low rate of success. Artificial intelligence (AI), which is one of the hottest technologies in recent years, has been widely used in many fields. A variety of AI-based methods have been successfully applied to the discovery of drugs. This paper summarizes the applications of common machine learning and deep learning methods in the field of drug research and development, and raises the problems and challenges for artificial intelligence. With its great potential in the field of drug research and development, artificial intelligence will bring new opportunities and promising future for the development of medical and pharmaceutical sciences.

[Key words] artificial intelligence; machine learning; deep learning; drug discovery

随着疾病多样性和药物耐药问题频出，药物需求日益增加，但新药研发存在研发周期长、成本高和成功率低等风险。一般而言，一个创新药从研发到最后上市，需要花费数十亿美元和10～15年的时间[1]。尽管投入高，耗时长，小分子药物最终上市的成功率仅为13%，失败风险较高[2]。计算机辅助药物设计能极大地缩短药物研发时间，提高药物研发成功率。传统的药物筛选方法有分子对接、药效团匹配和相似性搜索等。近年来随着计算机计算能力的高速发展和大数据时代的到来，人工智能助力药物研发迎来了极大的发展机遇。

近年来计算机辅助药物设计在药物发现领域也不乏一些成功的案例。中国药科大学陆涛教授课题组[3]的Flt3（Fms-like tyrosine kinase）小分子抑制剂正在进行I期临床试验，该抑制剂从先导化合物的发现到后续的优化评价均是在计算机辅助药物设计的指导下完成。英属哥伦比亚大学Li等[4]利用计算机辅助药物设计方法，从苗头化合物发现到候选化合物性质评价，完成雄激素受体抑制剂的临床前研究，并已将成果转让。加州大学Manglik等[5]利用基于结构的药物设计方法发现了一类新型的具有止痛作用的阿片受体激动剂。来自Insilico Medicine和药明康德等机构的研究人员[6]开发了一种人工智能算法（GENTRL模型），在21天内就设计出了DDR1（discoidin domain receptor 1）激酶抑制剂的潜在分子结构，并在46天内完成初步生物学验证。GENTRL模型只用了46天的时间，就完成了传统方法用数月到数年的时间所完成的工作，大大节省了药物的研发时间和高昂的研发费用。

人工智能与药物研发相结合应用的主要场景包括药物靶点预测、高通量筛选、药物设计和药物的吸收、分配、代谢、排泄和毒性（absorption，distribution，metabolism，excretion and toxicity，ADMET）等性质预测。人工智能涵盖了机器学习和深度学习，而深度学习又属于机器学习的子领域。机器学习算法在药物研发领域被广泛用于分类和回归预测等方面。与机器学习相比，深度学习适合处理大数据，模型也相对复杂。随着大数据时代的到来和计算机性能的不断增

接受日期：2020-01-01

*通讯作者：陈亚东，教授；

研究方向：药物分子设计与合成；

Tel：025-86185163；E-mail：ydchen@cpu.edu.cn

强，近年来越来越多的人工智能算法模型被提出、如图 1 所示，最早应用于药物发现领域的有决策树，随机森林和支持向量机等机器学习模型，随着计算机性能的不断提高和大数据时代的到来，深度神经网络、卷积深度网络和循环神经网络等深度学习算法逐渐发展，其在药物发现领域的应用也越来越广泛。本文将主要介绍机器学习和深度学习方法在药物发现领域的应用。

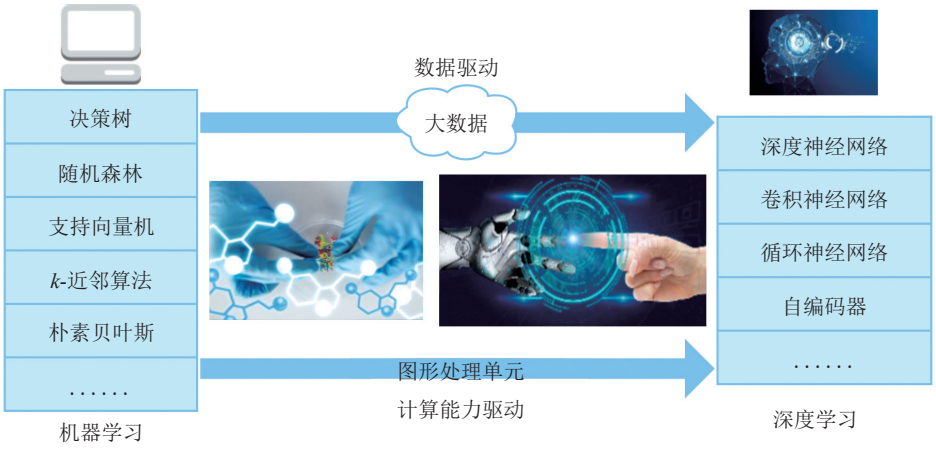



**图 1 人工智能算法模型**

**Figure 1 Model of artificial intelligence algorithm**

## 1 人工智能算法模型简介

在过去的 10 年间，人工智能在很多领域都有广泛的应用。继机器学习后，深度学习模型被提出并应用于药物发现领域。常见的机器学习算法包括决策树（decision tree）、随机森林（random forest）、支持向量机（support vector machine，SVM），*k*-最近邻算法（*k*-nearest neighbor model）和朴素贝叶斯（Naïve Bayes）算法。深度学习和机器学习的主要区别是数据量的大小及模型的复杂度，深度学习模型更复杂，需要的数据量也更大。深度学习属于机器学习的子领域，近年来随着计算性能的高速发展及图形处理单元（graphics processing unit，GPU）的应用，深度学习模型的应用越来越广泛，主要有深度神经网络、卷积神经网络、循环神经网络和自编码器。

### 1.1 决策树和随机森林

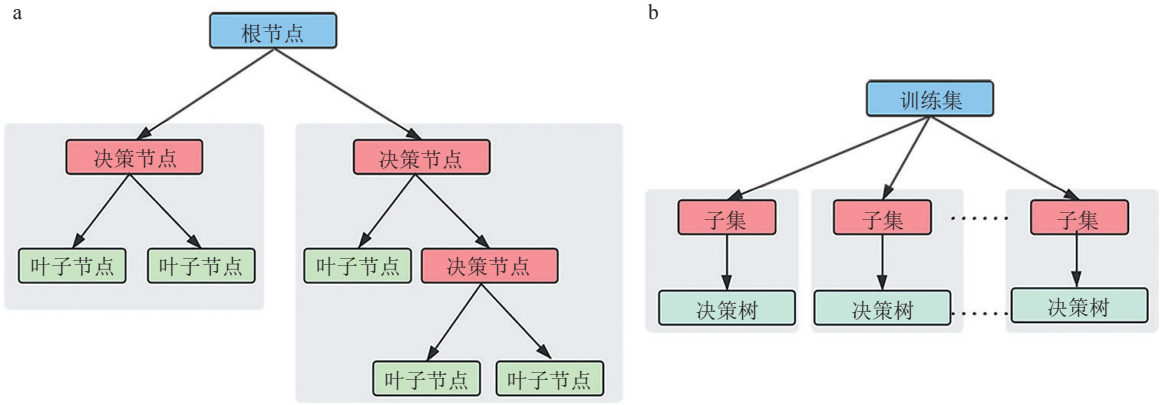
决策树是一种将决策流程以树状结构清晰表示的机器学习方法，本质上是通过一系列规则对数据进行分类的过程。如图 2a 所示，在决策树模型中，每个决策树的非叶节点表示一个特征属性上的测试，每个分支代表这个特征属性在某个值域上的输出，而每个叶子节点存放一个类别。选择属性和剪枝是构建决策树的 2 个基本步骤。首先，选择根节点属性对输入分子进行测试，依据是否符合根节点属性将分子划分到下一个决策节点，再根据决策节点的属性向下划分子节点，重复该过程直到最终划分到叶子节点。其次，决策树分支过多容易导致模型过拟合，需要使用修剪算法对生成的树进行剪枝，降低树结构的复杂性。

随机森林是通过构建多个决策树对样本进行训练并预测的一种分类器，其最终输出的类别是由每个决策树输出的类别的众数而决定，如图 2b 所示是一个随机森林模型。每棵树根据如下算法来建造 [7]：用 $N$ 来表示训练样本的个数，从 $N$ 个训练样本中以有放回抽样的方式，取样 $N$ 次，用来训练一个决策树；随机从每个样本的 $M$ 个属性中选取 $m$ 个属性，然后从 $m$ 个属性中通过信息增益选择一个属性作为该节点的分裂属性，直到该节点不能分裂为止；重复以上步骤构建大量的决策树，从而形成随机森林。随机森林在训练过程中会对数据进行有放回的随机抽样，因此与决策树相比随机森林不太可能过拟合数据，而且对数据分类的准确度也较高。

### 1.2 支持向量机

SVM 由 Vidyasagar 等 [8] 在 1998 年提出，它能够处理小数据集中的高维变量，可以用于分类和回归问题，但更多用在分类问题上。如图 3 所示，对于线性可分数据集，SVM 模型通过映射空间中的点来分离不同的类别，这样能使不同类别的点之间的边界最大化。对于线性不可分数据集，SVM 使用核映射将非线性数据集放入高维特征空间用于线性分类。SVM 在数据分

类领域应用广泛，在某些方面其分类效果要强于其他机器学习方法。




a：决策树模型；b：随机森林模型

**图 2 人工智能算法模型**

**Figure 2 Model of artificial intelligence algorithm**

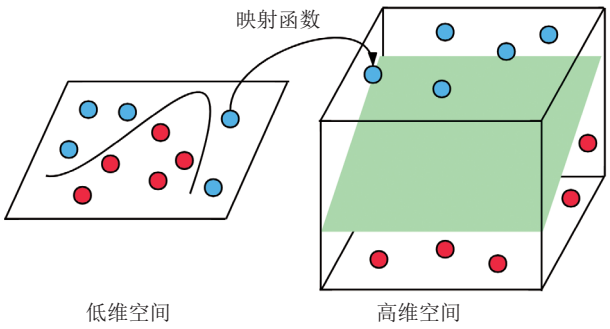



**图 3 支持向量机模型**

**Figure 3 Support vector machine model**

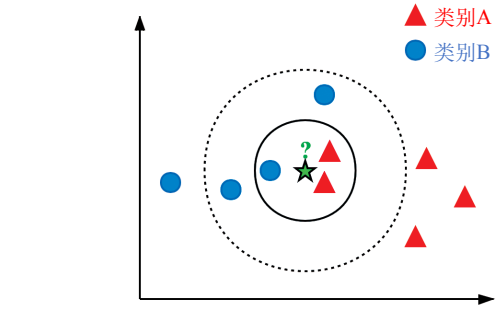
### 1.3 *k*-最近邻算法

*k*-最近邻算法是一种用于分类和回归的无监督学习算法，由 Cover 和 Hart 在 1968 年提出 [9]。如图 4 所示，*k*-最近邻算法基于某种距离度量找出训练集中与测试样本最靠近的 *k* 个训练样本，然后基于这 *k* 个“邻居”的信息来进行预测，其核心思想是如果一个样本在 *k* 个最邻近的大多数样本属于某一个类别，则该样本也属于这一个类别。*k*-最近邻算法是所有机器学习算法中最简单而且容易操作的一种算法，常用于化合物分类。在 *k*-最近邻模型中，每一个化合物代表一个样本，分子描述符代表化学特征空间，如果一个化合物在化学特征空间中的 *k* 个最相邻的大多数化合物属于活性化合物，则该化合物理论上有较大的可能性也为活性化合物。

### 1.4 朴素贝叶斯算法

朴素贝叶斯分类器是应用最为广泛的分类算法之一，如图 5 所示是贝叶斯公式，对于事件 A 和 B，P( B|A ) 就是指在事件 A 发生的条件下，事件 B 发生的概率，又称条件概率，P（B）和 P（A）是没有前提条件时事件 B 和事件 A 发生的概率，又称先验概率。朴素贝叶斯算法最早由 Duda 和 Hart 在 1973 年提出 [10]，根据贝叶斯原理来处理分类和回归问题 [11]。贝叶斯分类器只需要少量的训练数据即可估计出一些必要的参数，能够在许多复杂的条件中取得较好的效果。

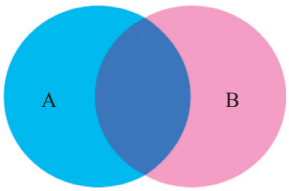
### 1.5 深度神经网络




▲：A类化合物；●：B类化合物；★：待分类化合物

**图 4 *k*-最近邻算法模型**

**Figure 4 *k*-nearest neighbor model**

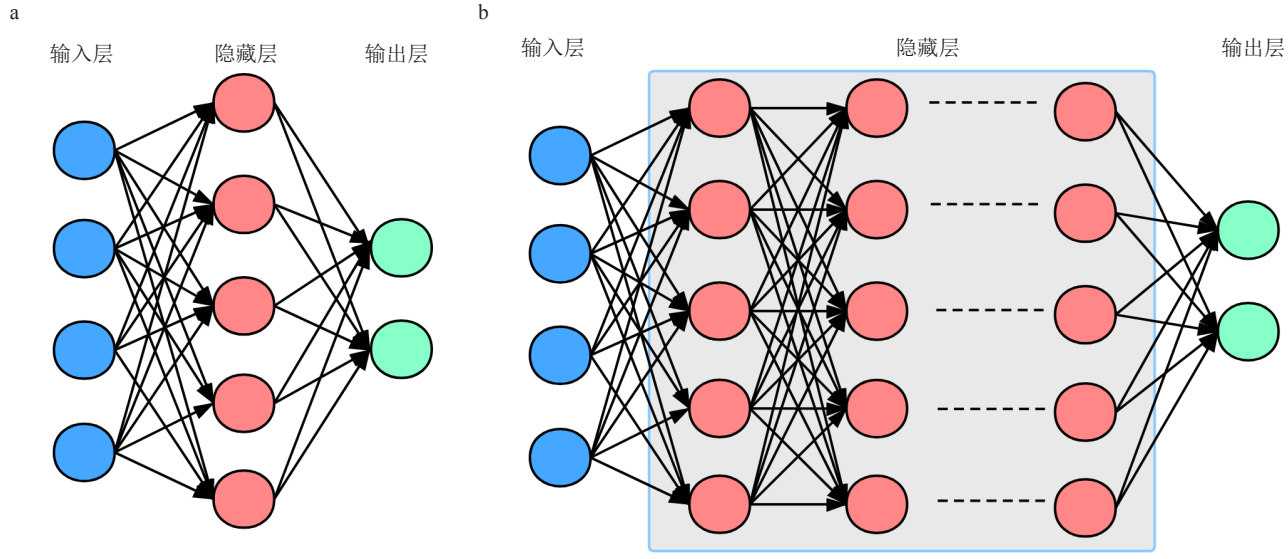



$$P(B|A) = P(B) \times \frac{P(B|A)}{P(A)}$$

**图 5 朴素贝叶斯模型**

**Figure 5 Naïve Bayes model**

人工神经网络（artificial neural network，ANN），如图 6a 所示，由输入层（蓝色）、一个隐藏层（红色）和输出层（绿色）3 部分组成，每层都包含若干个神经元，ANN 最早来源于 1943 年 McCulloch 等 [12] 的计算模型，19 世纪 60 到 80 年代现代人工神经网络开始发展并应用于不同领域，但 ANN 对训练数据容易出现过拟合问题，其很快被其他机器学习算法如支持向量机代替。随着计算机性能的发展，新的深度学习算法开始涌现，其中包括深度神经网络（deep neural network，DNN）。如图 6b 所示，DNN 本质上是具有多个隐藏层的 ANN，它是最早应用于药物发现的深度学习算法之一。




a：人工神经网络结构；b：深度神经网络结构

**图 6 人工神经网络与深度神经网络模型**

**Figure 6 Models of artificial neural network and deep neural network**

### 1.6 卷积神经网络

卷积神经网络（convolutional neural network，CNN）是一种前馈神经网络，它在图像识别领域的表现优异。如图 7 所示，CNN 的核心一般由卷积层（绿色方块）、池化层（蓝色方块）和全连接层（蓝色圆圈）3 个部分组成，最后一列为输出层，其中卷积层是最重要的一个部分，该层的参数由一系列过滤器又称卷积核组成，使用不同的卷积核对输入数据进行卷积可以提取不同的特征，随着原始特征的不断提取压缩，最终能提取到高层次的特征。卷积层的优点在于其通过权值共享策略极大地缩小了参数的规模并逐渐建立空间和结构的不变性 [13]。池化层也称为下采样层，它用来压缩特征空间，池化层可以降低噪声的影响和参数的规模，提高模型的鲁棒性。每个卷积层连接池化层构成卷积模块，一个 CNN 通常有多个卷积模块，用以提取特征。最后模型中会有一个或多个的全连接层，接受卷积模块提取的特征并输出结果。

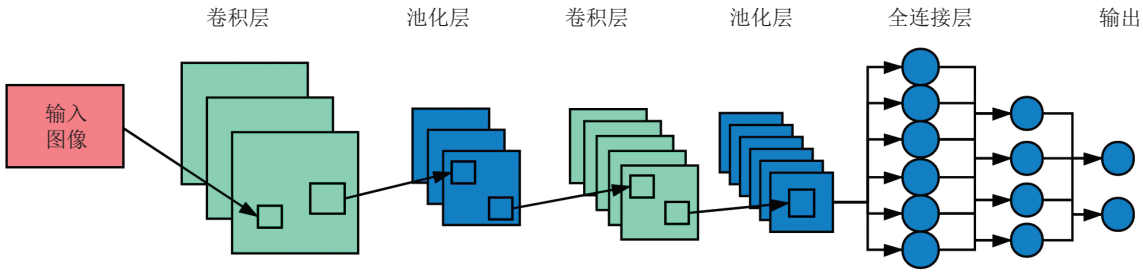
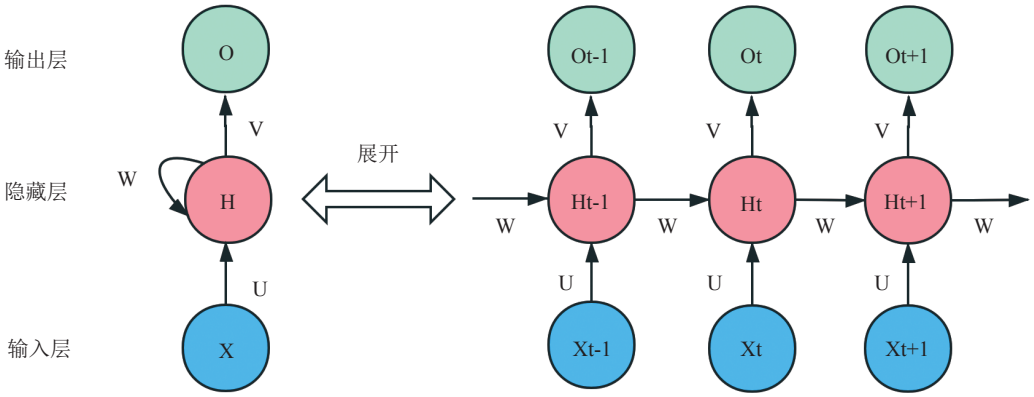



**图 7 卷积神经网络结构**

**Figure 7 Structure of convolutional neural network**

### 1.7 循环神经网络

循环神经网络（recurrent neural network，RNN），如图 8 所示，同样由输入层（蓝色）、隐藏层（红色）和输出层（绿色）3 部分组成，RNN 是一类用于处理序列数据的神经网络，比如时间序列数据，基因和蛋白序列数据或分子线性输入字符串（SMILES）等[14]，与普通的前馈神经网络不同，RNN 在其隐藏层的各节点之间建立了连接，使一个节点的输入不仅包括输入层的输出，还包括上一时刻隐藏层节点的输出，这是 RNN 可用于处理序列数据的重要原因，同时 RNN 也是唯一一个具有记忆能力的神经网络[15]，但却受到短期记忆的影响，因此产生了一些 RNN 的改进算法如长短期记忆网络（long short-term memory，LSTM）和 GRU（gated recurrent unit）算法，RNN 在自然语言处理方面得到了广泛的应用，同时基于 LSTM 和 GRU 算法的 RNN 在从头药物设计中也占据很重要的地位。




注：X、H、O分别为输入层，隐藏层和输出层；U、W、V分别为X到H、H到H和H到O的权值

**图 8 基础循环神经网络结构**

**Figure 8 Structure of recurrent neural network**

### 1.8 自编码器

自编码器（autoencoder，AE），是一种用于非监督学习的神经网络，如图 9 所示，它具有输入层（蓝色）、隐藏层（红色）和输出层（绿色）3 层结构，包含编码部分和解码部分，编码部分是一个将输入层接受到的数据转化为有限数量的隐藏层的神经网络，然后通过解码部分与输出层连接，自编码器的目的在于重构输入数据，典型的就是用于数据降维[16]。自编码器的概念已经广泛应用于生成学习模型，并且经过改进，产生了变分自编码器和条件变分自编码器等，它们在药物分子生成方面具有广泛的应用。

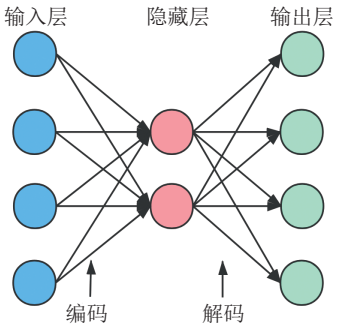



**图 9 自编码器结构**

**Figure 9 Autoencoder model structure**

## 2 人工智能在药物发现中的应用

在当今大数据时代背景下，人工智能已经渗透到各个领域。在药物发现领域，人工智能在药物靶点识别、化合物虚拟筛选和药物性质预测等方面得到越来越广泛的应用，如图 10 所示。

### 2.1 药物靶点识别

靶点是新药研发的基础，因此药物靶点的识别在药物发现过程中尤为重要。近年来也有越来越多的靶点被发现，然而相对于未知的靶点，已发现的靶点只是冰山一角。若能在早期通过计算机预测药物靶点，缩短靶点发现周期，对药物研发具有重要意义。

决策树可用于预测药物靶点，Costa 等[17]基于决策树分类器来预测与疾病相关的基因，最后他们发现了多种转录因子在代谢通路和细胞外定位中的调控作用。基于蛋白靶点的化学结构和几何特征，Nayal 等[18]选取了 99 个蛋白的 99 个药物结合位点和 1 187 个非药物结合位点，然后构建了一个随机森林分类器来预测成药靶点。Kumari 等[19]结合自助法（bootstrap）采样提升了随机森林算法，并成功从非药物靶点中区分出了药物靶点。针对乳腺癌、胰腺癌和卵巢癌等疾病，Jeon

等[20]利用一系列基因数据集构建了一个 SVM 分类器，可将蛋白分为药物靶点和非药物靶点 2 个类别。

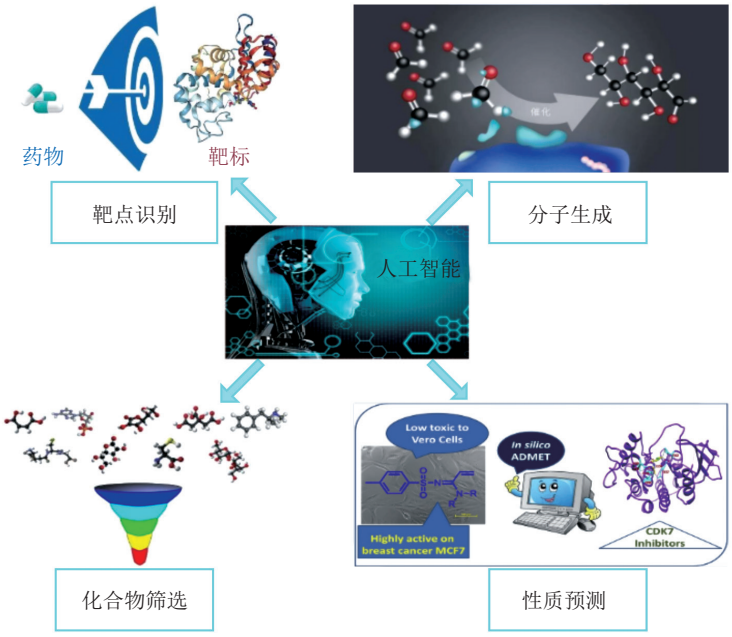



**图 10 人工智能在药物发现中的应用**

**Figure 10 Application of artificial intelligence in drug discovery**

### 2.2 活性化合物筛选

药物在人体内可以同时作用多个靶点，但如果作用于非靶向受体就会引起副作用。人工智能可以对候选化合物进行筛选，更快筛选出作用于特定靶点且具有较高活性的化合物，为后期临床试验做准备。

决策树模型可用于拓扑异构酶 I 抑制剂的分类和预测[21]。Neugebauer 等[22]利用低维定量构效关系描述符建立决策树来预测与蛋白相互作用的抑制剂，并通过建模技术进一步修剪决策树得到真阳率更高的蛋白相互作用抑制剂。王洁雪等[23]采用决策树与随机森林 2 种机器学习方法分别对脾酪氨酸激酶（spleen tyrosine kinase，Syk）抑制剂与非抑制剂建立模型，经过对比，随机森林具有更好的预测精度，采用随机森林模型对 Syk 抑制剂进行虚拟筛选，从 ZINC 分子数据库筛选得到潜在的 Syk 抑制剂分子。Warmuth 等[24]利用 SVM 方法生成最大间隔超平面来从一系列化合物中分离出活性化合物，结果表明 SVM 的分类效果强于其他模型。Poorinmohammad 等[25]建立 SVM 分类模型对抗人类免疫缺陷病毒（human immunodeficiency virus，HIV）肽进行分类，预测准确率达到了 96.76%。SVM 也可以和其他方法结合用于化合物库的虚拟筛选，有研究显示组合 SVM 和分子对接方法筛选化合物库可大大提高活性化合物的命中率和富集因子[26]。贝叶斯模型能够快速有效地识别大型化合物数据库，从化合物库中筛选出活性化合物[27]。贝叶斯分类模型已成功用于许多抑制剂的虚拟筛选，如雷帕霉素蛋白酶抑制剂的虚拟筛选等[28]。*k*-最近邻算法也可与其他特征选择算法相结合。Weidlich 等[29]应用 *k*-最近邻算法，同时结合模拟退火方法与随机森林算法，从 679 个药物分子中筛选抗病毒药物，他们的结果表明改进的 *k*-最近邻算法模型优于随机森林算法模型。

### 2.3 化合物性质预测

药代动力学性质不理想是药物在临床研究阶段研发失败的主要原因。因此在药物研发早期阶段对化合物成药性和安全性进行评估，对于提高药物研发成功率、降低研发成本具有十分重要的意义。

Newby 等[30]构建决策树模型用来预测化合物渗透性和溶解性在药物口服吸收过程中的作用，结果表明低渗透性高溶解性的化合物的肠道吸收率低，然而低溶解性高渗透性的化合物的肠道吸收率高。王昊等[31]利用朴素贝叶斯模型来进行药物不良反应的预测，结果发现贝叶斯网络预测模型对导致呼吸困难发生频率在 1%以上的药物的预测准确率可以达到 86.76%。毒性是新药开发的一项重要指标，在早期就排除一些毒性大的化合物对于新药研发来说非常有利。在 2014 年的 Tox21 数据挑战赛中，Mayr 等[32]用多任务 DNN

建立了 DeepTox 毒性评估模型从而赢得胜利，该模型在 15 项挑战中获得 9 项胜利，并且没有任何一项低于前 5 名。在他们的模型中使用了 Dropout 方法和 ReLu 激活函数，并且通过 GPU 并行计算进行模型训练。CNN 在性质预测方面也有所应用，例如 Wallach 等 [33] 使用蛋白配体复合物结合位点的三维格点作为输入，设计了第一个基于结构的深度 CNN，称为 AtomNet，该网络被用于预测小分子的生物活性。AtomNet 可以在没有活性化合物对照的情况下预测新的活性分子，在 DUDE 基准库测试中，其受试者工作特征（receiver operating characteristics，ROC）曲线下面积（area under the curves，AUC）达到了 0.9，远超先前的对接方法。ROC 曲线对于评价二分类模型非常有用，而且 ROC 曲线可以通过其曲线下面积 AUC 来解读，理想的分类模型 AUC 为 1，随机分类 AUC 为 0.5 [34]，因此 AUC 越接近 1 代表模型能力越强。同样地，Goh 等 [35] 设计了一种通用的深度 CNN，称为 Chemception，该网络被用于预测分子的各种性质如毒性、活性和溶解性等，重要的是该网络接受的输入数据仅为分子的二维图像而不需要其他任何化学信息。他们将该网络与多层感知机深度神经网络（multilayer perceptron DNN，MLPDNN）相比，发现 Chemception 在活性与溶解度的预测方面表现更优异。

### 2.4 分子生成

有效地构建拥有一定规模且高质量的小分子库是药物研发人员一直关注的问题，组合化合物库和枚举化合物库等技术能够迅速地构建大规模的分子库，这类化合物库的重要不足在于分子结构缺乏一定的新颖性，为了扩充化学空间且产生高成药性的分子，研究者们利用深度学习技术设计了不同的分子生成模型。

Segler 等 [36] 利用 RNN 设计了分子生成模型，他们首先用大量的有效的 SMILES 字符串训练了 RNN 模型，在他们的模型中使用了 3 个叠加的 LSTM 层，最终他们生成了 847 995 个新分子，并且这些分子具有一定的多样性，通过计算生成分子的各种性质包括分子量、氢键供体和受体数、脂水分配系数、可旋转键及极性表面积并进行数据降维，发现生成分子的性质与训练集分子表现出良好的相关性，同时证明这些分子适合于虚拟筛选。为了产生对特定靶点具有潜在活性的分子，Segler 等 [36] 使用对不同靶点有活性的小分子分别作为测试集对模型进行了微调，占测试集 14% 的抗金黄色葡萄球菌分子和占测试集 28% 的抗恶性疟原虫分子出现在微调后模型生成的分子中。同样地，Yuan 等 [37] 介绍了一种新的分子生成方法 MIMICS（machine-based identification of molecules inside characterized space），在该方法中，以给定化学子集的 SMILES 字符串作为输入，他们首先使用 RNN 学习这些字符串中字符的概率分布，然后删除无效的结构，最终在 MIMICS 中生成了性质相似但骨架不同的新分子，重要的是在随后的细胞实验中发现新生成的分子中有能够作为血管内皮生长因子抑制剂，证明该方法能够生成结构新颖并且具有类药性的分子。这 2 个案例都说明基于 RNN 生成的分子与模板分子性质相似但骨架新颖，为从头药物设计提供了强大的支持。

Gomez-Bombarelli 等 [38] 提出了一种使用变分自编码器生成分子结构的新方法。与自编码器不同的是，变分自编码器将输入数据编码到隐含空间是不连续的，该方法的编码器将输入分子的离散表示转换成隐含空间的连续向量，随后解码器可将这些连续向量还原成分子离散表示。重要之处在于隐含空间中的分子表示为连续的，因此可以通过随机解码、扰乱或插入等方法产生新的分子，并且通过一些优化算法可以产生期望性质的分子。Lim 等 [39] 使用条件变分自编码器设计了一种分子生成方法，与变分自编码器不同之处在于，其可以在编码和解码过程中施加条件。该方法被证实可以在 10% 误差范围内生成特定属性（如特定的分子量、脂水分配系数、氢键受体和供体、拓扑极性表面积等）的类药分子，并可以在保持其他性质的情况下控制某一种性质。Skalic 等 [40] 提出了借助变分自编码器使用分子三维表现和药理特性来产生新型分子的方法，该方法同时结合了 RNN 和 CNN 方法，最终该方法被证实可以产生具有类药性的分子。

### 2.5 蛋白结构及蛋白配体相互作用预测

了解蛋白质的结构与性质在药物研发初级阶段极为重要，在计算机辅助药物设计中，基于受体结构的药物设计也具有很重要的地位，其中模拟蛋白受体相互作用的分子对接技术应用广泛，不同的对接打分函数也会一定程度影响结果。DNN 在蛋白结构预测方面也有应用，例如 Qi 等 [41] 使用多任务 DNN 构建了一个用于预测蛋白质各种局部性质的预测器，该预测器可以应用于多种目的，例如糖基化位点、扭转角等的预测。由于 CNN 在图像识别领域比较成功，因此开始有人研

究利用 CNN 来评价蛋白配体相互作用，例如 Ragoza 等 [42] 将蛋白配体复合物表示为三维格点作为输入，使用多层 CNN 构建了一个打分函数，该打分函数在结合模式预测和虚拟筛选中的打分表现比 AutoDock Vina 的打分函数更好，但是多层 CNN 构建的打分函数也存在与一般打分函数相似的问题，因此 CNN 在该方面的应用还有一定的改进空间。

## 3 人工智能在药物发现中的机遇与挑战

新药研发具有成本高、研发周期长、成功率低的 3 大高风险性质。近年来随着计算性能的持续提高和先进算法的开发，人工智能快速发展，已应用于药物研发的各个领域。计算机辅助药物设计在药物研发领域早有应用，传统的计算机辅助药物设计更偏向于以靶点和结构信息为核心的计算机辅助药物设计，如基于结构的虚拟筛选和定量构效关系模型构建等，而人工智能是以数据为核心的药物研发模式，因此其在靶点未知和机制未明的复杂疾病药物研发中占有优势。新药研发成本约为 26 亿美元，耗时约 10 年，成功率仅有 6.2%[43]，而人工智能应用于药物研发可大大节省研发成本和时间。报告显示人工智能在化合物合成和筛选方面比传统手段可节约 40% 的时间，每年可为医药企业节约 260 亿美元的化合物筛选成本。

虽然机器学习和深度学习已被用于药物研发的各个领域，但是人工智能在新药研发中的应用才刚刚起步，也面临着诸多挑战。在药物研发领域，数据是人工智能的关键。因此作为一种数据挖掘技术，人工智能模型依赖于大数据的积累，并不能无中生有。用来学习的数据很大程度上会影响模型的性能，因此模型是否有效往往取决于数据的质量。若是数据质量不高，即使使用可靠的算法，也不会获得良好的结果，反而会浪费大量的资源和时间。目前大多数预测模型来源于参差不齐的数据，因此如何获得高质量的数据是人工智能面临的一个主要问题。此外，如何学习训练数据得到泛化能力强的模型也是人工智能的难点及热点。

## 4 总结与展望

计算机辅助药物设计在药物研发领域的应用已经历数十年，随着医药数据的不断积累和计算机性能的不断增强，人工智能在药物设计上的应用也越来越广泛，特别是深度学习技术，为计算机辅助药物设计注入了新的活力，极大地推进药物研发的进程。未来随着数据进一步积累和新的算法出现，人工智能辅助药物设计有望在药物发现领域得到更广泛的应用，更多地覆盖药物设计与发现各个阶段，更大程度地降低药物研发的成本和周期，更好地助力我国创新药物的研发。

## [ 参考文献 ]


[1] DiMasi J A, Grabowski H G, Hansen R W. Innovation in the pharmaceutical industry: new estimates of R&D costs[J/OL]. *J Health Econ*, 2016, 47: 20-33[2020-01-01]. https://www.sciencedirect.com/science/article/abs/pii/S0167629616000291?via%3Dihub. Doi: 10.1016/j.jhealeco.2016.01.012.

[2] Zhong F S, Xing J, Li X T, *et al*. Artificial intelligence in drug design[J]. *Sci China Life Sci*, 2018, 61(10): 1191-1204.

[3] Wang Y, Zhi Y L, Jin Q M, *et al*. Discovery of 4-((7*H*-pyrrolo[2,3-*d*]pyrimidin-4-yl)amino)-*N*-(4-((4-methylpiperazin-1-yl)methyl)phenyl)-1*H*-pyrazole-3-carboxamide(FN1501), an FLT3- and CDK-Kinase inhibitor with potentially high efficiency against acute myelocytic leukemia[J]. *J Med Chem*, 2018, 61(4): 1499-1518.

[4] Li H F, Ban F Q, Dalal K, *et al*. Discovery of small-molecule inhibitors selectively targeting the DNA-binding domain of the human androgen receptor[J]. *J Med Chem*, 2014, 57(15): 6458-6467.

[5] Manglik A, Lin H, Aryal D K, *et al*. Structure-based discovery of opioid analgesics with reduced side effects[J]. *Nature*, 2016, 537(7619): 185-190.

[6] Zhavoronkov A, Ivanenkov Y A, Aliper A, *et al*. Deep learning enables rapid identification of potent DDR1 kinase inhibitors[J]. *Nat Biotechnol*, 2019, 37(9): 1038-1040.

[7] 董师师 , 黄哲学 . 随机森林理论浅析 [J]. 集成技术 , 2013, 2(1): 1-7.

[8] Vidyasagar M. Statistical learning theory and randomized algorithms for control[J]. *Ieee Contr Syst Mag*, 1998, 18(6): 69-85.

[9] Cover T, Hart P. Nearest neighbor pattern classification[J]. *Ieee T Inform Theory*, 1967, 13(1): 21-27.

[10] Duda R O, Hart P E. *Pattern classification and scene analysis*[M]. New York: Wiley-Interscience, 1973: 16.

[11] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers[J/OL]. *Mach Learn*, 1997, 29: 131-163[2020-01-01]. https://link.springer.com/article/10.1023/A:1007465528199.

[12] McCulloch W S, Pitts W. A logical calculus of the ideas immanent in

nervous activity[J/OL]. *B Math Biol*, 1943, 5: 115-133[2020-01-01]. https://link.springer.com/article/10.1007/BF02478259.

[13] Lu H M, Li Y J, Uemura T, *et al*. FDCNet: filtering deep convolutional network for marine organism classification[J]. *Multimed Tools Appl*, 2017, 77(17): 21847-21860.

[14] Yang X, Wang Y F, Byrne R, *et al*. Concepts of artificial intelligence for computer-assisted drug discovery[J]. *Chem Rev*, 2019, 119(18): 10520-10594.

[15] Lavecchia A. Deep learning in drug discovery: opportunities, challenges and future prospects[J]. *Drug Discov Today*, 2019, 24(10): 2017-2032.

[16] Chen H M, Engkvist O, Wang Y H, *et al*. The rise of deep learning in drug discovery[J]. *Drug Discov Today*, 2018, 23(6): 1241-1250.

[17] Costa P R, Acencio M L, Lemke N. A machine learning approach for genome-wide prediction of morbid and druggable human genes based on systems-level data[J/OL]. *BMC Genomics*, 2010, 11 Suppl 5 (Suppl 5): S9[2020-01-01]. https://link.springer.com/article/10.1186/1471-2164-11-S5-S9.

[18] Nayal M, Honig B. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites[J]. *Proteins*, 2006, 63(4): 892-906.

[19] Kumari P, Nath A, Chaube R. Identification of human drug targets using machine-learning algorithms[J/OL]. *Comput Biol Med*, 2015, 56: 175-181[2020-01-01]. https://www.sciencedirect.com/science/article/abs/pii/S0010482514003254. Doi: 10.1016/j.compbiomed.2014.11.008.

[20] Jeon J, Nim S, Teyra J, *et al*. A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening[J/OL]. *Genome Med*, 2014, 6(7): 57[2020-01-01]. https://link.springer.com/article/10.1186/s13073-014-0057-7.

[21] Li B K, Kang X K, Zhao D, *et al*. Machine learning models combined with virtual screening and molecular docking to predict human topoisomerase I inhibitors[J/OL]. *Molecules*, 2019, 24(11): 2107[2020-01-01]. https://www.mdpi.com/1420-3049/24/11/2107. Doi: 10.3390/molecules24112107.

[22] Neugebauer A, Hartmann R W, Klein C D. Prediction of protein-protein interaction inhibitors by chemoinformatics and machine learning methods[J]. *J Med Chem*, 2007, 50(19): 4665-4668.

[23] 王洁雪 , 李瑶 , 杨敏 , 等 . 基于机器学习方法虚拟筛选 Syk 的抑制剂 [J/OL]. *化学研究与应用* , 2019, 7: 1313-1320[2020-01-01]. http://www.cnki.com.cn/Article/CJFDTotal-HXYJ201907013.htm. Doi: 10.3969/j.issn.1004-1656.2019.07.013.

[24] Warmuth M K, Liao J, Ratsch G, *et al*. Active learning with support vector machines in the drug discovery process[J]. *J Chem Inf Comput Sci*, 2003, 43(2): 667-673.

[25] Poorinmohammad N, Mohabatkar H, Behbahani M, *et al*. Computational prediction of anti HIV-1 peptides and *in vitro* evaluation of anti HIV-1 activity of HIV-1 P24-derived peptides[J]. *J Pept Sci*, 2015, 21(1): 10-16.

[26] Xie Q Q, Zhong L, Pan Y L, *et al*. Combined SVM-based and docking-based virtual screening for retrieving novel inhibitors of c-Met[J]. *Eur J Med Chem*, 2011, 46(9): 3675-3680.

[27] Xiong X, Yuan H L, Zhang Y M, *et al*. Protein flexibility oriented virtual screening strategy for JAK2 inhibitors[J/OL]. *J Mol Struct*, 2015, 1097: 136-144[2020-01-01]. https://www.sciencedirect.com/science/article/abs/pii/S0022286015004081. Doi: 10.1016/j.molstruc.2015.05.007.

[28] Wang L, Chen L, Liu Z H, *et al*. Predicting mTOR inhibitors with a classifier using recursive partitioning and naive bayesian approaches[J/OL]. *PLoS One*, 2014, 9(5): e95221[2020-01-01]. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0095221.

[29] Weidlich I E, Filippov I V, Brown J, *et al*. Inhibitors for the hepatitis C virus RNA polymerase explored by SAR with advanced machine learning methods[J]. *Bioorgan Med Chem*, 2013, 21(11): 3127-3137.

[30] Newby D, Freitas A A, Ghafourian T. Decision trees to characterise the roles of permeability and solubility on the prediction of oral absorption[J/OL]. *Eur J Med Chem*, 2015, 90: 751-765[2020-01-01]. https://www.sciencedirect.com/science/article/abs/pii/S0223523414011155?via%3Dihub. Doi: 10.1016/j.ejmech.2014.12.006.

[31] 王昊. 基于机器学习方法的药物不良反应预测 [D]. 厦门: 厦门大学, 2012: 38.

[32] Mayr A, Klambauer G, Unterthiner T, *et al*. DeepTox: toxicity prediction using deep learning[J/OL]. *Front Environ Sci*, 2016, 3: 80[2020-01-01]. https://www.frontiersin.org/articles/10.3389/fenvs.2015.00080/full.

[33] Wallach I, Dzamba M, Heifets A. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery[EB/OL]. (2015-10-10)[2020-01-01]. https://arxiv.org/abs/1510.02855.

[34] Bradley A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms[J]. *Pattern Recogn*, 1997, 30(7): 1145-1159.

[35] Goh G B, Siegel C, Vishnu A, *et al*. Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models[EB/OL]. (2017-06-20)[2020-01-01]. https://arxiv.org/abs/1706.06689.

[36] Segler M H S, Kogej T, Tyrchan C, *et al*. Generating focused molecule

libraries for drug discovery with recurrent neural networks[EB/OL]. (2017-01-05)[2020-01-01]. https://arxiv.org/abs/1701.01329.

[37] Yuan W, Jiang D, Nambiar D K, *et al*. Chemical space mimicry for drug discovery[J]. *J Chem Inf Model*, 2017, 57(4): 875-882.

[38] Gomez-Bombarelli R, Wei J N, Duvenaud D, *et al*. Automatic chemical design using a data-driven continuous representation of molecules[J]. *ACS Cent Sci*, 2018, 4(2): 268-276.

[39] Lim J, Ryu S, Kim J W, *et al*. Molecular generative model based on conditional variational autoencoder for de novo molecular design[J/OL]. *J Cheminformatics*, 2018, 10(1): 31[2020-01-01]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6041224/. Doi: 10.1186/s13321-018-0286-7.

[40] Skalic M, Jiménez J, Sabbadin D, *et al*. Shape-based generative modeling for de novo drug design[J]. *J Chem Inf Model*, 2019, 59(3): 1205-1214.

[41] Qi Y J, Oja M, Weston J, *et al*. A unified multitask architecture for predicting local protein properties[J/OL]. *PLoS One*, 2012, 7(3): e32235[2020-01-01]. https://journals.PLoS.org/plosone/article?id=10.1371/journal.pone.0032235.

[42] Ragoza M, Hochuli J, Idrobo E, *et al*. Protein-ligand scoring with convolutional neural networks[J]. *J Chem Inf Model*, 2017, 57(4): 942-957.

[43] Wong C H, Siah K W, Lo A W. Estimation of clinical trial success rates and related parameters[J]. *Biostatistics*, 2019, 20(2): 273-286.

**[专家介绍]陈亚东**：博士，教授，药物化学专业、药学信息学专业博士生导师。江苏省“青蓝工程”优秀青年骨干教师（2008），江苏省“青蓝工程”中青年学术带头人（2014）。美国密歇根大学（University of Michigan, Ann Arbor）医学院综合癌症中心访问学者。主持和参与了国家自然科学基金、国家重大科技专项“重大新药创制”等多项科研项目。申请国内专利 14 项、国际专利 1 项、授权 3 项；主编或参编学术著作和教材 3 本；在 *J Med Chem*、*Eur J Med Chem.*、*J Chem Inf Model J* 等国际学术期刊发表 SCI 论文 80 多篇。2015 年研究团队发现的 1.1 类抗肿瘤新药临床前候选化合物以 1.5 亿人民币转让给上海复星医药，目前在美国、澳大利亚及中国进行 I 期临床。