DOI:10.12113/202409008

人工智能在细胞穿透肽预测中的研究进展

李尧尧[1,2], 李钧翔[3,4], 李旭辉[5], 张 瑾[2*]

(1. 浙江理工大学 生命科学与医药学院,杭州 310018;2. 嘉兴大学 生物与化学工程学院,浙江 嘉兴 314000;
3. 浙江清华长三角研究院 衰老科学创新研发中心,浙江 嘉兴 341001;4. 禾美生物科技(浙江)有限公司,浙江 嘉兴 341001;
5. 浙江省应用酶学重点实验室(浙江清华长三角研究院),浙江 嘉兴 314006)

摘 要:细胞穿透肽(Cell penetrating peptides, CPPs)是指能通过直接转运或内吞作用进入细胞的多肽,一般不超过30个氨基酸。CPPs可携带多种活性物质进入细胞,有望成为新型药物的递送载体。传统实验方法获得CPPs具有工作量大、通量低、周期长等问题;随着计算生物学的发展,基于机器学习算法的人工智能模型提高了候选CPPs的预测效率。本文介绍了基于支持向量机、随机森林、极限学习机、极端随机树和其他深度学习的CPPs预测方法,最后探讨了人工智能预测CPPs存在的挑战和对未来的展望。本文旨在为从事CPPs相关研究的学者利用计算生物学工具预测和初步筛选CPPs提供基础和泛化的理论指导。

关键词:人工智能;细胞穿透肽;CPP预测;机器学习

中图分类号:R978.1+6 **文献标志码**:A **文章编号**:1672-5565(2026)01-014-09

Research progress of artificial intelligence in cell penetrating peptide prediction

LI Yaoyao[1,2], LI Junxiang[3,4], LI Xuhui[5], ZHANG Jin[2*]

(1. *School of Life Sciences and Medicine, Zhejiang Sci-Tech University, Hangzhou* 310018, *China*;
2. *School of Biological and Chemical Engineering, Jiaxing University, Jianxing* 314001, *Zhejiang, China*;
3. *AGECODE R&D CENTER, Yangtze Delta Region Institute of Tsinghua University, Jianxing* 314001, *Zhejiang, China*;
4. *Hemei Biotechnology (Zhejiang) Co., Ltd., Jianxing* 314001, *Zhejiang, China*;5. *Key Laboratory of Applied Enzymology, Zhejiang Province(Yangtze delta region institute of Tsinghua University), Jianxing* 314001, *Zhejiang, China*)

Abstract:Cell penetrating peptides(CPPs) refer to polypeptides that can enter cells through direct transport or endocytosis, and generally do not exceed 30 amino acids. CPPs can carry a variety of active substances into cells and is expected to become a new drug delivery carrier. The traditional experimental method to obtain CPPs has many problems, such as heavy workload, low flux and long cycle. With the development of computational biology, the artificial intelligence model based on machine learning algorithm improves the prediction efficiency of candidate CPPs. This paper introduces the prediction method of CPPs based on support vector machine, random forest, extreme learning machine, extreme random tree and deep learning, and discusses the influence of sequence feature extraction and insufficient training set on the accuracy of artificial intelligence prediction of CPPs. It is believed that with the development of artificial intelligence technology, researchers will be able to develop a CPPs prediction model with higher accuracy and stronger generalization ability.

Keywords:Artificial intelligence; Cell penetrating peptide; CPPs prediction; Machine learning

收稿日期:2024-09-08;**修回日期**:2025-02-06;**网络首发日期**:2025-05-21.

网络首发地址:https://link.cnki.net/urlid/23.1513.q.20250521.1025.002

资金项目:国家自然科学基金项目(No.32172708);浙江省自然基金重点项目(No.LZ23C170002).

*通信作者:张瑾,博士,教授,研究方向:脂代谢调控及生物活性物质的鉴定与应用.E-mail: zhangjin7688@163.com.

引用格式:李尧尧, 李钧翔, 李旭辉, 张 瑾.人工智能在细胞穿透肽预测中的研究进展[J].生物信息学,2026,24(1):14-22.

LI Yaoyao, LI Junxiang, LI Xuhui, ZHANG Jin.Research progress of artificial intelligence in cell penetrating peptide prediction[J].Chinese Journal of Bioinformatics,2026,24(1):14-22.

细胞穿透肽(Cell penetrating peptides, CPPs)是一类以非受体依赖方式,非经典内吞方式直接穿过细胞膜进入细胞的多肽分子,它们的长度一般不超过 30 个氨基酸,可以为线性多肽或环化多肽[1]。1988 年,Green 等[2]首次报道具有跨过多种细胞膜作用的短肽 Tat,并将其命名为细胞穿透肽。

CPPs 的一个重要特点是可以穿过细胞膜,并能携带不同大小和性质的生物活性物质进入细胞[3],其运载能力与其他载体相比,具有稳定、快速、高效、广谱、易吸收的特点,并且具有低毒性和无细胞类型限制等优势,在一定浓度范围内不会受到机体环境的影响而造成细胞损伤[4]。因此,CPPs 可被开发为药物递送载体[5-6]。此外,CPPs 能携带护肤品中的有效成分进入皮肤细胞内发挥作用,可用作外敷类美容成分的载体[7-8]。

根据 CPPs 的氨基酸物理化学特性,CPPs 可以分为阳离子 CPPs、两亲性 CPPs 和疏水性 CPPs[9]。其中,阳离子 CPPs 中的氨基酸残基主要由精氨酸和赖氨酸组成;两亲性 CPPs 则主要由赖氨酸组成,序列中还包含有一定数目的亲水或疏水性的氨基酸残基,其空间构象为 α-螺旋结构;疏水性 CPPs 只包含非极性残基,有一个疏水的基序或者疏水的化学基团,这些基序或基团对穿膜起着非常重要的作用[10]。CPPs 进入细胞的过程见图 1。

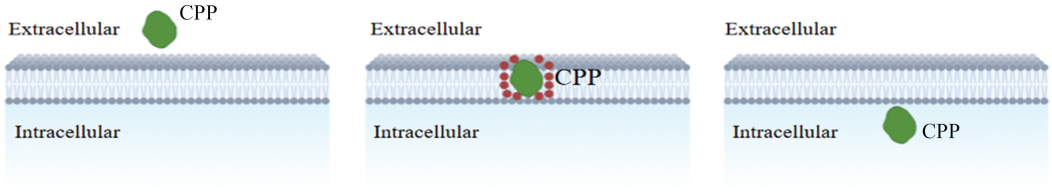

图 1　细胞穿透肽进入细胞的过程

Fig.1　Process of cell penetrating peptides entering cells

注:CPP(绿色)极化细胞膜,形成水孔(红色)。然后 CPP 利用水孔穿过细胞膜。最后极化消散,水孔消失.(彩图见电子版).

传统的 CPPs 鉴定方法通常采用细胞试验,具有直观、准确度高等优点。Tran D P 等[11]在鉴定多肽细胞穿透能力时,通过合成荧光素修饰的多肽,将 HeLa 细胞在荧光素多肽存在下孵育,最后用显微镜观察绿色荧光是否分布在 HeLa 细胞中,从而鉴定多肽是否为 CPPs。Gautam Ankur 等[12]用 FITC 染料标记多种不同的多肽,分别在培养基中处理 HeLa 细胞,采用流式细胞仪对摄取效率进行了分析。用已知的 CPPs(TAT)作为对照,比较这些多肽的细胞穿透能力以及具体的摄取效率。总体而言,细胞试验方法工作量大、成本高、耗时长且难以实现高通量。

利用计算方法辅助预测 CPPs 具有成本低、通量高的特点,前期方法主要包括基于序列比对和基于分子描述符的多元统计等[13-14]。这两种方法都是根据已知的经验规则或者模式对一类多肽的某些生化属性和细胞穿透能力之间的关联进行统计或建模,利用序列对比的方式对其进行识别[15]。其中序列比对是一种将两条或多条序列按照一定规律排列并进行对比的序列分析方法,其计算逻辑是找出待测序列和数据库中目标序列的相似性[13]。该方法的优点是简单直观,相对易于实现,但是如何给出一个合理优化的相似性度量准则是一个挑战,到目前还没有公认的标准。不仅如此,对于分歧较大的序列,预测的准确率以及算法的时间复杂度也都有待提高。还有,如果出现与训练数据匹配度极低的样本,就无法判断[16]。基于分子描述符的多元统计学方法是一种通过量化结构-活性关系的预测方法。这种方法的预测能力取决于数学模型中涉及的分子描述符的质量[14]。研究人员重点从结构方面,研究了 CPPs 基本序列中氨基酸对细胞穿透肽性质所产生的影响。电荷量、基本序列、肽链长度、线性、手性以及多肽骨架的改变都对 CPPs 可能产生影响[17]。这种方法的不足之处在于只计算描述符的总和而不考虑氨基酸的顺序,这就导致某 CPPs 的打乱序列类似物与之具有相同的预测值,但这序列类似物有可能不是 CPPs[18]。因此,学者们致力于开发 CPPs 预测的新算法。

1　人工智能简介和 CPPs 国内外研究现状

人工智能(Artificial intelligence, AI)是通过计算机程序或系统呈现人类思维和智能的技术,它通过捕获数据潜在的概念和关系,独立分析并学习数据从而实现模拟人类思维的目的。AI 一词最早由 John McCarthy 于 1956 年提出[19],随着算法的优化及算力的提升,AI 的发展正在融入到人类生产、生活的方方面面[20-21]。机器学习(Machine learning)是

AI 的一个分支,使用计算方法直接从数据中“学习”信息,而不依赖于预定方程模型,类似于人类从经验中学习。当可用于学习的样本数量增加时,这些算法可自适应提高性能[22]。深度学习(Deep learning)是机器学习的分支,是一种基于对数据进行表征的学习,使用包含复杂结构或由多重非线性变换构成的多个处理层对数据进行高层抽象的算法。

随着基因组、转录组、蛋白组、代谢组等技术的发展,生物数据处于爆炸式增长时代,生物信息学研究的重点从积累数据转移到如何解读数据[23]。大量的数据,在生物信息的存储、获取、处理、浏览及可视化等方面,都对理论算法和软件的发展提出了迫切的需求。另外,由于基因组数据本身的复杂性也对理论算法和软件的发展提出了迫切的需求[24]。而以机器学习与深度学习为代表的人工智能方法正适合于处理这种数据量大、含有噪声并且缺乏统一理论的领域。在 CPPs 的相关研究领域,国内的研究主要集中在 CPPs 的序列设计和改造方面,利用计算生物学方法对 CPPs 进行优化[25-26]。一些团队在药物递送方面进行了探索,尝试将 CPPs 与抗癌药物、基因治疗载体结合,提升治疗效果[27]。国外研究者已建立了多种基于机器学习方法的计算模型,用于预测 CPPs 的生物活性,并优化 CPPs 的设计。在药物递送系统方面,研究者们探索了它们在递送抗体、核酸和小分子药物中的效果[28]。部分研究还关注了 CPPs 的化学空间,探索了 CPPs 的结构和功能多样性。通过统计方法评估 CPPs 的化学空间,并将其与预测细胞渗透性的常规过滤器相关联[29]。

2 传统机器学习与深度学习在 CPPs 预测中的应用

传统机器学习方法通常包括支持向量机(SVM)、随机森林(Random forest)、极限学习机(ELM)等,这些方法通常依赖于特征工程,通过将输入数据转化为具有代表性的特征来进行预测。其优势在于较为高效且易于解释,但需要较强的领域知识和特征工程。深度学习方法包括卷积神经网络(CNN)、循环神经网络(RNN)等,主要利用神经网络模型进行学习和预测,通常包含多个隐藏层,能够从数据中自动学习复杂的模式和特征,尤其适合处理大规模数据和复杂的非线性关系。在 CPPs 预测中,深度学习展现了巨大的潜力,尤其是在处理氨基酸序列或三维结构数据时。然而,它也有较高的数据需求和计算资源要求。

2.1 传统机器学习在 CPPs 预测中的应用

使用传统机器学习进行 CPPs 预测主要分为以下四个阶段(图 2):

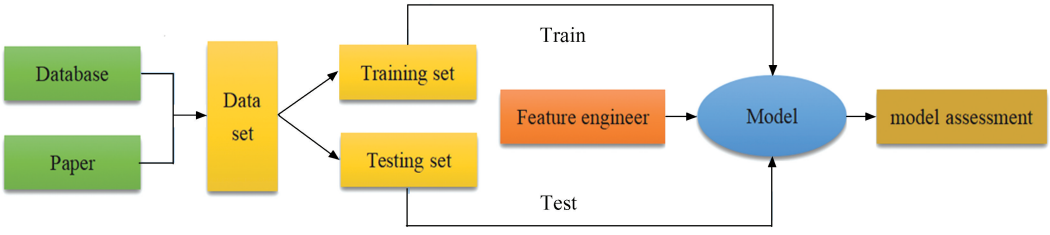

图 2 机器学习预测细胞穿透肽流程图

Fig.2 Flow chart of predicting cell penetrating peptides by machine learning

第一阶段是数据集准备。候选肽序列通常从经过验证的数据库和相关文献中收集。为了构建高质量的预测模型,通常需要训练集和独立的测试集[30]。训练集用于模型训练,测试集用于验证训练模型的可转移性和可靠性。

第二阶段是特征工程。这个阶段由特征表示和特征优化组成。对于特征表示,通常使用各种特征描述符来捕获多肽的特征,包括组成特征[例如氨基酸组成(AAC)和二肽组成(DAC)]、二元谱、基于序列的特征和理化特征等。为了提高特征表示能力,通常通过删除一些不相关的特征来优化特征表示[31]。

第三阶段是模型构建和预测。对上一阶段的最佳特征使用机器学习算法进行训练。对于待预测肽序列,使用特征向量进行编码,然后输入到训练的模型中。最终,开发的预测模型将提供可靠的预测结果。

最后一个阶段是模型的性能评估。通过模型最终在测试集上的表现,对模型现有的预测能力进行性能评估。SE(敏感性)、SP(特异性)、ACC(准确率)、MCC(马修斯相关系数)和 AUC(ROC 曲线下与坐标轴围成的面积)等评价指标已广泛应用于多个生物信息学领域[32]。这些指标计算公式如下:

$$SE = \frac{TP}{TP + FN} \times 100\%$$

$$SP = \frac{TN}{TN + FP} \times 100\%$$

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} \times 100\%$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \times 100\%$$

其中 TP、TN、FP 和 FN 分别表示真阳性、真阴性、假阳性和假阴性的数量。SE 和 SP 分别测量两类的预测能力:阳性和阴性。ACC、MCC 和 AUC 评估预测模型的整体性能[33]。ROC(受试者工作特征曲线)曲线以 SE 为纵坐标、1-SP 为横坐标绘制。AUC 值在 0~1 之间,越接近 1,模型的预测性能越好、越稳健。

2.1.1　基于支持向量机的细胞穿透肽预测

支持向量机(Support vector machine, SVM)(图 3(a))是一类按监督学习(Supervised learning)方式对数据进行二元分类的广义线性分类器(Generalized linear classifier)[34]。

Sanders 等[35]开发了一种基于 SVM 的方法来识别潜在的 CPPs。预测模型以肽的基本生化特性(包括肽的长度、净电荷、疏水性和亲脂性、供体氢键的数量等 23 种特性)为特征进行训练。作者特别关注了训练数据集的平衡问题,使用三个不同的基准数据集证明了平衡数据集对准确预测的重要性。模型在不平衡数据集的准确率为 80.69%,而在平衡数据集的准确率可以达到 91.72%。

Gautam 等[36]提出了一种基于 SVM 的预测模型(CellPPD),并建立了一个公共网络服务器(https://webs.iiitd.edu.in/raghava/cellppd/index.html)来预测 CPPs。在 CellPPD 中,使用的数据集共有 708 个肽,以不同的特征表示算法,如氨基酸组成(AAC)、二肽组成(DAC)、二进制轮廓、基序特征和理化性质(PCP),用于训练不同的预测模型。经测试,基于混合特征的预测模型实现了 97.40%的最大准确率,优于基于单个特征的模型。

Tang 等[37]开发了 C2Pred,一种基于优化 DAC 作为特征的预测器。作者通过方差分析对特征集进行优化。当采用 164 个二肽的最佳特征时,模型准确度最高可达 83.6%。

2.1.2　基于随机森林的细胞穿透肽预测

随机森林(Random forest, RF)(图 3(b))是一个包含多个决策树的分类器,其输出预测结果由所有决策树的分数集合确定[38]。Chen 等[39]发了一个基 RF 的 CPPs 预测模型。该模型在一系列理化性质上进行训练,包括伪氨基酸组成(PseAAC)、分子体积、极性、密码子多样性、静电荷和二级结构。通过最小冗余最大相关性和增量特征选择来选择优化的特征。预测模型的总体准确率为 83.45%。

Wei 等[40]提出了一个名为 SkipCPP-Pred 的预测模型。该模型采用自适应 k-skip-2-gram 算法来提取特征,准确率提高到 90.6%。作者在另一项研究[41]中,提出了一个名为 CPPred-RF 的两层预测器,其中第一层是区分 CPPs 和非 CPPs,而第二层是预测 CPPs 的摄取效率。预测模型在综合特征上进行了训练,这些特征结合了四个基于序列的描述符。与 SkipCPP-Pred 相比,CPPred-RF 在同一基准数据集上的预测准确率提高到 91.6%。

Qiang 等[42]提出了一个名为 CPPred-FL 的预测模型。CPPred-FL 引入了特征表示学习策略,以从使用多个特征描述符构建的机器学习模型中学习类和概率信息,例如理化性质(PCP)、组合信息和位置特定信息等。CPPred-FL 的最佳整体准确度高达 92.1%。与之前的研究相比,准确度虽然没有显著提高,但用于训练预测模型的特征要少得多。该特征表示策略探索了一种提取高表达特征的有效新方法。

2.1.3　基于极限学习机的细胞穿透肽预测

极限学习机(Extreme learning machine, ELM)是一类基于前馈神经网络(Feedforward neuron network, FNN)构建的机器学习系统或方法,适用于监督学习和非监督学习问题[43]。Pandey 等[44]开发了一个基于 ELM 的模型(KELM-CPPpred)。该预测模型利用了六种不同的特征描述符,包括 AAC、DAC、PseAAC 和三种混合特征(Hybrid-AAC、Hybrid-DAC 和 Hybrid-PseAAC)。作者基于不同特征分别开发了模型,其中,KELM-CPPpred 在基于混合 AAC 的 KELM-CPPpred 模型的准确度最高,为86.98%。

2.1.4　基于极端随机树的细胞穿透肽预测

极端随机树(Extremely randomized trees, ERT)属于一类广泛用于开发分类和回归模型的集成方法[45]。Manavalan 等[46]提出了一个两层模型(MLCPP)来预测 CPPs 及其吸收效率。作者从多肽的氨基酸序列计算的信息中提取特征,包括氨基酸组成(AAC)、氨基酸指数(AAI)、二肽组成(DPC)、理化性质(PCP)和组成-过渡-分布(CTD)。该方法采用 ERT 算法训练第一层 CPPs 预测模型,准确率为 89.6%,第二层吸收效率预测模型采用 RF 训练,准确率为 72.5%。

2.2　深度学习在 CPPs 预测中的应用

使用深度学习进行 CPPs 预测流程与机器学习

基本相同,区别在于深度学习直接从数据中学习高级特征。因此,深度学习只需要将数据编码,而不需要进行特征表示和特征优化。

卷积神经网络(Convolutional neural networks, CNN)(图 3(d))是深度学习中非常常见的算法[47]。在生物学领域,CNN 在基因组序列特征预测、医学图像处理和药物开发等方面应用广泛[48-50]。循环神经网络(Recurrent neural network, RNN)(图 3(c))是一类以序列(Sequence)数据为输入,在序列的演进方向进行递归(Recursion)且所有节点(循环单元)按链式连接的递归神经网络(Recursive neural network)[51]。长短期记忆(Long short-term memory, LSTM)网络是一种特殊的 RNN,主要是为了解决长序列训练过程中的梯度消失和梯度爆炸问题[52]。

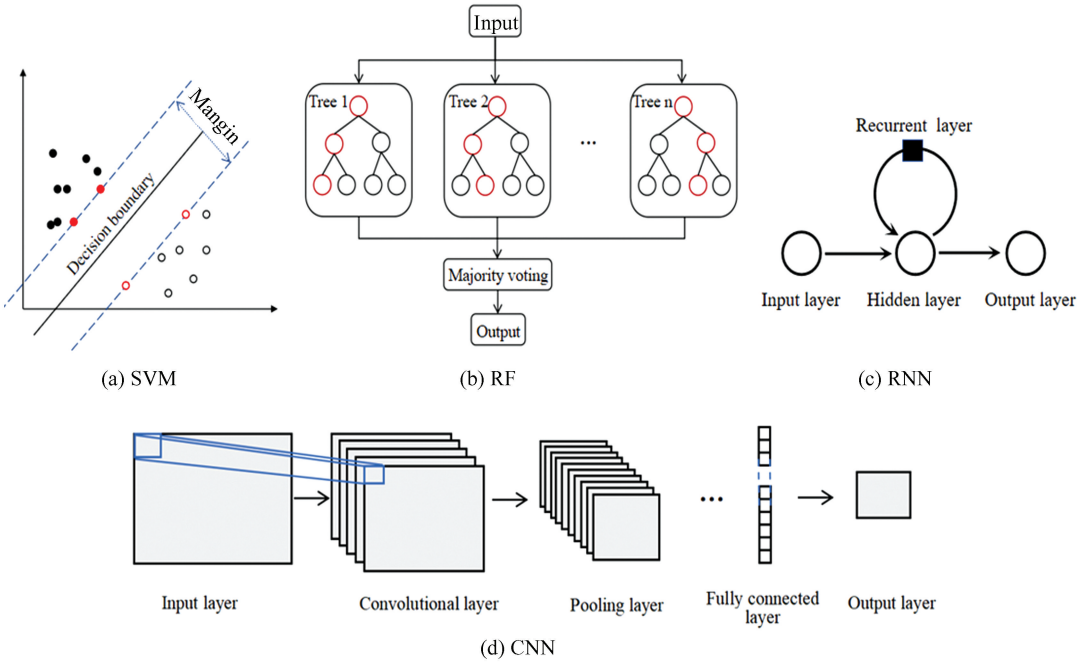

图 3 传统机器学习和深度学习算法示意图

Fig.3 Schematic diagram of machine learning and deep learning

注:(a)表示支持向量机算法,通过寻找最佳的超平面来将数据点分开,以最大化两个类别之间的间隔。这个超平面就是决策边界;(b)表示随机森林算法,构建多棵决策树并将其结果结合起来,最终的预测结果是对所有树的输出进行投票来得到;(c)表示循环神经网络算法,输入层接收输入数据;隐藏层处理序列数据并维护隐藏状态;循环层接受当前的输入和上一时间步的隐藏状态,经过一定的计算后生成新的隐藏状态;输出层将隐藏层的输出转换为最终预测结果;(d)表示卷积神经网络算法,输入层接收输入数据;卷积层通过卷积操作提取特征;池化层减少特征的维度,降低计算复杂度,同时保留重要特征;全连接层将卷积层和池化层提取的特征映射到最终的分类输出;输出层产生最终的预测结果.

在 CPPs 的预测应用中,方春等[53]提出了一种基于字符嵌入和混合深度学习框架的 CPPs 预测方法,名为 CPPDeep。该方法将原始肽序列作为输入,并使用字符嵌入方法来表示序列,采用 CNN-LSTM 组合架构来进行自动的特征提取。该模型包括四个主要层:嵌入层、CNN 层、LSTM 层以及全连接输出层。嵌入层用来学习字符嵌入,为每个字符训练出一个紧凑的向量表示,并将肽序列编码为特征矩阵,作为 CNN 网络的输入;CNN 层用来提取特征矩阵中有用的结构信息;然后,LSTM 层提取对整体预测任务有意义的数据中的长期相关性和上下文信息;最后,全连接层采用激活函数运算来产生模型的输出结果。

在测试集上,CPPDeep 模型取得优异的 AUC 性能(0.97)。准确率(ACC)和真阳性率(即敏感性 SE)分别达到 84.6%和 94.6%。

Park 等[54]提出了一种基于五种不同深度学习架构的预测方法 AiCPP,每种架构都包括嵌入层、长短期记忆(LSTM)层和注意力层。该方法通过纳入大量阴性数据集,减少预测假阳性的偏差来提高模型的特异性或正确识别非 CPPs 的能力。作者从 11 046 种人类蛋白中,使用滑动窗口方法将肽序列切成重叠的 9 个氨基酸片段,去除片段中的重复项。模型的训练集包括 7 165 个阳性(CPPs)肽,14 408 个阴性(非 CPPs)肽和 11 046 343 个来自人类蛋白的阴性肽。

模型的准确率(ACC)性能达到 86.0%,AUC 达

到 0.927。此外,AiCPP 模型对非 CPPs 具有更高的特异性($SP = 0.893$)。特异性的提高可归因于训练集使用了大量来自人类蛋白的阴性数据,显著减少了假阳性的数量(表 1)。

表 1　基于机器学习与深度学习的细胞穿透肽预测模型

Table 1　Models of cell penetration peptide prediction based on machine learning and deep learning

模型	算法	特征表示	年份	数据集	准确率/%
Sanders2011[35]	SVM	肽的基本生化特性	2011	CPP/NCPP:111/111	91.72
CellPPD[36]	SVM	序列信息和二元图谱	2013	CPP/NCPP:708/708	97.40
C2Pred[37]	SVM	二肽组成	2016	CPP/NCPP:411/411	83.60
Chen2015[39]	RF	伪氨基酸组成和 5 种氨基酸性质	2015	CPP/NCPP:111/34	83.45
SkipCPP-Pred[40]	RF	K-skip-2-gram	2017	CPP/NCPP:462/462	90.60
CPPred-RF[41]	RF	四个基于序列的描述符	2017	CPP/NCPP:462/462	91.60
CPPred-FL[42]	RF	PCP、组合信息和位置特定信息	2018	CPP/NCPP:462/462	92.10
KELM-CPPpred[44]	KELM	AAC、DAC、PseAAC 和三种混合特征	2018	CPP/NCPP:408/374	86.98
MLCPP[46]	ERT	AAC、AAI、DPC、PCP 和 CTD	2018	CPP/NCPP:427/427	89.60
CPPDeep[53]	CNN/LSTM	自动提取	2019	CPP/NCPP:582/582	84.60
AiCPP[54]	LSTM	自动提取	2023	CPP/NCPP:7165/11060751	86.00

3　总结与讨论

细胞穿透肽(CPPs)作为一种新型的分子载体,在药物递送、护肤分子渗透皮肤细胞等领域展现了广泛的应用潜力,为效应分子顺利进入细胞提供了创新的途径。尽管 CPPs 在多种应用中表现出良好前景,如何高效筛选新型 CPPs 以及判断候选多肽分子是否具有 CPP 功能,仍然是当前领域中的一大挑战。

本文首先概述了 CPPs 的定义与功能,并简要回顾了传统的实验预测方法与计算方法。进一步,深入探讨了基于人工智能(AI)技术的 CPPs 预测方法,特别是 AI 在解析大量 CPPs 序列与功能关系中的强大优势。通过现有的 CPPs 预测模型,研究人员能够识别肽序列与细胞穿透性之间的关系,并高效预测特定条件下的细胞穿透能力。例如,Sanders 等[35]通过支持向量机(SVM)模型,从鸡蛋白质组中筛选出 4 个预测为 CPP 的肽段,并成功验证其穿透性;Park 等[54]则利用 AiCPP 模型从蛋白序列 APP 中发现新的 CPP 序列,并验证其在 MCF-7 细胞中的穿透性。这些研究表明,人工智能在 CPPs 筛选中的应用,不仅提升了筛选效率,也为新型 CPP 的发现提供了有力支持。

在 CPPs 预测领域,AI 模型的精度是评估其性能的重要标准之一。然而,单一的准确率指标并不能全面反映模型的实际表现[55],尤其是在类别分布不均的情况下。高准确率可能掩盖了模型在少数类别识别上的不足,这在实际应用中可能导致效果不佳[56]。因此,评估模型时,应综合考虑准确率、泛化能力、灵敏度、特异性等多个指标。本文提到的几种 CPPs 预测模型中,CellPPD 模型表现出了最高的预测准确率,但其在未知数据集上的泛化能力仍需进一步验证;而 CPPred-RF 和 MLCPP 模型不仅能够预测多肽是否为 CPP,还能进一步预测 CPPs 的摄取效率。此外,CPPDeep 和 AiCPP 模型在灵敏度和特异性方面表现出了显著优势,展示了人工智能在 CPPs 预测中的多维度优化(表 1)。

尽管近年来 CPPs 预测模型的性能已有显著提升,依然面临一些挑战,这些挑战为未来的研究提供了改进方向。首先,数据准备阶段仍存在高质量样本稀缺的问题。例如,公共数据库 CPPsite 2.0[57]中虽收录了近2 000 个实验验证的 CPPs,但去除高度相似序列后,剩余的数据不足以支持机器学习模型的高质量训练。此外,阴性样本的选择多依赖于未标注为 CPP 的随机肽,无法完全确保这些肽为非 CPP。为此,持续收集并验证 CPPs 数据,扩展数据库规模与质量,显得尤为重要。

其次,在特征表示阶段,短肽通常难以形成稳定的二级结构,导致大多数预测工具依赖一级序列信息提取特征[58]。如何有效利用不同类型的序列信息,仍然是一个亟待解决的难题。借鉴自然语言处理(NLP)中的词嵌入技术,可能为肽序列的高效表示与处理提供新思路,从而提升模型的识别能力。

在算法选择上,将其他领域的先进技术,如图像识别,融合到 CPPs 预测中,可能带来新的突破。同时,集成学习方法的应用,能够结合多种模型的优势,进一步提升预测精度。最后,模型实验验证阶

段,通过加强实验数据与人工智能预测结果之间的反馈循环,有助于不断优化模型,提升预测的准确性与可靠性。

综上所述,人工智能技术的应用为 CPPs 预测提供了高效、精准的工具,并推动了该领域的快速发展。未来的研究不仅需加强高质量数据的积累与验证,还应优化计算算法,提高模型的泛化能力与实际应用效果。这些进展将在很大程度上促进 CPPs 在生物医药等领域的应用,进一步加速其技术转化与实际落地。

参考文献(References)

[1] BÖHMOVÁ E, MACHOVÁ D, PECHAR M, et al. Cell-penetrating peptides: A useful tool for the delivery of various cargoes into cells[J]. Physiological Research, 2018, 67(Suppl 2):S267-S279. DOI: 10.33549/physiolres.933975.

[2] GREEN M, LOEWENSTEIN P M. Autonomous functional domains of chemically synthesized human immunodeficiency virus tat trans-activator protein[J]. Cell, 1988, 55(6): 1179-1188. DOI: 10.1016/0092-8674(88)90262-0.

[3] ZORKO M, LANGEL Ü. Cell-Penetrating Peptides[J]. Methods in Molecular Biology, 2022, 2383:3-32. DOI: 10.1007/978-1-0716-1752-6_1.

[4] KARDANI K, MILANI A, H SHABANI S, et al. Cell penetrating peptides: the potent multi-cargo intracellular carriers[J]. Expert Opinion on Drug Delivery, 2019, 16(11): 1227-1258. DOI: 10.1080/17425247.2019.1676720.

[5] CHABLANI L, SINGH V. Cell-Penetrating peptides as passive permeation enhancers for transdermal drug delivery[J]. AAPS PharmSciTech, 2022, 23:266. DOI: 10.1208/s12249-022-02424-4.

[6] NASROLLAHI S A, TAGHIBIGLOU C, AZIZI E, et al. Cell-penetrating peptides as a novel transdermal drug delivery system[J]. Chemical Biology and Drug Design, 2012, 80(5):639-646. DOI: 10.1111/cbdd.12008.

[7] SHIN H J, LEE B K, KANG H A. Transdermal properties of cell-penetrating peptides: applications and skin penetration mechanisms[J]. American Chemical Society Applied Bio Materials, 2024, 7(1): 1-16. DOI: 10.1021/acsabm.3c00659.

[8] MORTAZAVI S M, MOGHIMI H R. Skin permeability, a dismissed necessity for anti-wrinkle peptide performance[J]. International Journal of Cosmetic Science, 2022, 44(2): 232-248. DOI: 10.1111/ics.12770.

[9] 孙春萌, 沈雁, 涂家生. 细胞穿膜肽的研究进展[J]. 中国药学杂志, 2013, 48(14): 1143-1146. DOI: 10.11669/cpj.2013.14.002.
SUN Chunmeng, SHEN Yan, TU Jiasheng. Advance on the research of cell-penetrating peptides[J]. Chinese Pharmaceutical Journal, 2013, 48(14): 1143-1146. DOI: 10.11669/cpj.2013.14.002.

[10] 王红梅, 王春玲, 邹艳, 等. 细胞穿膜肽的分类和穿膜机制研究进展[J]. 德州学院学报, 2021, 37(6): 12-17. DOI: 10.3969/j.issn.1004-9444.2021.06.003.
WANG Hongmei, WANG Chunling, ZOU Yan, et al. Research progress on classification and membrane-penetrating mechanism of cell membrane-penetrating peptides[J]. Journal of Dezhou University, 2021, 37(6): 12-17. DOI: 10.3969/j.issn.1004-9444.2021.06.003.

[11] TRAN D P, TADA S, YUMOTO A, et al. Using molecular dynamics simulations to prioritize and understand AI-generated cell penetrating peptides[J]. Scientific Reports, 2021, 11:10630. DOI: 10.1038/s41598-021-90245-z.

[12] GAUTAM A, SHARMA M, VIR P, et al. Identification and characterization of novel protein-derived arginine-rich cell-penetrating peptides[J]. European Journal of Pharmaceutics and Biopharmaceutics, 2015, 89:93-106. DOI: 10.1016/j.ejpb.2014.11.020.

[13] EDGAR R C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput[J]. Nucleic Acids Research, 2004, 32(5): 1792-1797. DOI: 10.1093/nar/gkh340.

[14] FERNÁNDEZ-TORRAS A, COMAJUNCOSA-CREUS A, DURAN-FRIGOLA M, et al. Connecting chemistry and biology through molecular descriptors[J]. Current Opinion in Chemical Biology, 2022, 66:102090. DOI: 10.1016/j.cbpa.2021.09.001.

[15] CHATZOU M, MAGIS C, CHANG J M, et al. Multiple sequence alignment modeling: Methods and applications[J]. Briefings in Bioinformatics, 2016, 17(6): 1009-1023. DOI: 10.1093/bib/bbv099.

[16] HANSEN M, KILK K, LANGEL U. Predicting cell-penetrating peptides[J]. Advanced Drug Delivery Reviews, 2008, 60(4-5):572-579. DOI: 10.1016/j.addr.2007.09.003.

[17] RAMAKER K, HENKEL M, KRAUSE T, et al. Cell penetrating peptides: A comparative transport analysis for 474 sequence motifs[J]. Drug Delivery, 2018, 25(1): 928-937. DOI:10.1080/10717544.2018.1458921.

[18] DE OLIVEIRA E C L, DA COSTA K S, TAUBE P S, et al. Biological Membrane-penetrating peptides: computational prediction and applications[J]. Frontiers in Cellular and Infection Microbiology, 2022, 12:838259. DOI: 10.3389/fcimb.2022.838259.

[19] MCCARTHY J, MINSKY M L, ROCHESTER N, et al. A Proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955[J]. AI Magazine, 27(4):12. DOI: 10.1609/aimag.v27i4.1904.

[20] BAHOO S, CUCCULELLI M, GOGA X, et al. Artificial intelligence in Finance: A comprehensive review through

bibliometric and content analysis[J]. Sn Business & Economics, 2024,4:23. DOI: 10.1007/s43546-023-00618-x.

[21]THOMAS L B, MASTORIDES S M, VISWANADHAN N A, et al. Artificial intelligence: Review of current and future applications in medicine[J]. Federal Practitioner, 2021,38(11):527-538. DOI: 10.12788/fp.0174.

[22]GREENER J G, KANDATHIL S M, MOFFAT L, et al. A guide to machine learning for biologists. Nature reviews[J]. Molecular Cell Biology, 2022,23(1):40-55. DOI: 10.1038/s41580-021-00407-0.

[23]HAGEN J B. The origins of bioinformatics[J]. Nature Reviews. Genetics, 2000,1(3):231-236. DOI: 10.1038/35042090.

[24]Genomics and bioinformatics[J]. Nature, 1999,399: 85-88. DOI: 10.1038/20013.

[25]曹赞霞, 刘磊, 王吉华. 分子动力学模拟研究穿透肽的跨膜机制及引导新肽设计[J]. 科学通报, 2014, 59(22): 2160-2168. DOI: 10.1360/N972014-00329.
CAO Zanxia, LIU Lei, WANG Jihua. Molecular mechanism of translocation and design of cell penetrating peptides based on molecular dynamics simulation[J]. Chinese Science Bulletin, 2014, 59(22): 2160-2168. DOI: 10.1360/N972014-00329.

[26]ZHANG Kun, NIE Qichang, CHI-KONG LAU T, et al. Rational design of L-rna aptamer-peptide conjugate for efficient cell uptake and G-quadruplex-mediated gene contro[J]. Angewandte Chemie, 2024, 63(9): e202310798. DOI: 10.1002/anie.202310798.

[27]范博, 金明姬, 黄伟, 等. 细胞穿膜肽在药物递送系统中的研究进展[J]. 药学学报, 2016, 51(2):264-271. DOI: 10.16438/j.0513-4870.2015-0990.
FAN Bo, JIN Mingji, HUANG Wei, et al. The development of cell-penetrating peptides in drug delivery system[J]. Acta Pharmaceutica Sinica, 2016, 51(2):264-271. DOI: 10.16438/j.0513-4870.2015-0990.

[28]ZORKO M, JONES S, LANGEL Ü. Cell-penetrating peptides in protein mimicry and cancer therapeutics[J]. Advanced Drug Delivery Reviews, 2022,180:114044. DOI: 10.1016/j.addr.2021.114044.

[29]FALATO L, GESTIN M, LANGEL Ü. Cell-Penetrating peptides delivering siRNAs: An overview[J]. Methods in Molecular Biology, 2021,2282:329-352. DOI: 10.1007/978-1-0716-1298-9_18.

[30]CHEN V, YANG Muyu, CUI Wenbo, et al. Applying interpretable machine learning in computational biology-pitfalls, recommendations and opportunities for new developments[J]. Nature Methods, 2024,21(8):1454-1461. DOI: 10.1038/s41592-024-02359-7.

[31]ZOU Quan, ZENG Jiancang, CAO Liujuan, et al. A novel features ranking metric with application to scalable visual and bioinformatics data classification[J]. Neurocomputing, 2016,173:346-354. DOI: 10.1016/j.neucom.2014.12.123.

[32]WEI Leyi, DING Yijie, SU Ran, et al. Prediction of human protein subcellular localization using deep learning[J]. Journal of Parallel and Distributed Computing, 2018,117:212-217. DOI: 10.1016/j.jpdc.2017.08.009.

[33]WEI Leyi, ZHOU Chen, CHEN Huangrong, et al. ACPred-FL: A sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides[J]. Bioinformatics, 2018,34(23):4007-4016. DOI: 10.1093/bioinformatics/bty451.

[34]NOBLE W S. What is a support vector machine? [J] Nature Biotechnology, 2006,24:1565-1567. DOI: 10.1038/nbt1206-1565.

[35]SANDERS W S, JOHNSTON C I, BRIDGES S M, et al. Prediction of cell penetrating peptides by support vector machines[J]. PLOS Computational Biology, 2011,7(7): e1002101. DOI: 10.1371/journal.pcbi.1002101.

[36]GAUTAM A, CHAUDHARY K, KUMAR R, et al. *In silico* approaches for designing highly effective cell penetrating peptides[J]. Journal of Translational Medicine, 2013,11:74. DOI: 10.1186/1479-5876-11-74.

[37]TANG Hua, SU Zhendong, WEI Huanhuan, et al. Prediction of cell-penetrating peptides with feature selection techniques[J]. Biochemical and Biophysical Research Communications, 2016,477(1):150-154. DOI: 10.1016/j.bbrc.2016.06.035.

[38]BREIMAN L. Random forests[J]. Machine Learning, 2001,45:5-32. DOI: 10.1023/A:1010933404324.

[39]CHEN Lei, CHU Chen, HUANG Tao, et al. Prediction and analysis of cell-penetrating peptides using pseudo-amino acid composition and random forest models[J]. Amino Acids, 2015,47(7):1485-1493. DOI: 10.1007/s00726-015-1974-5.

[40]WEI Leyi, TANG Jijun, ZOU Quan. SkipCPP-Pred: An improved and promising sequence-based predictor for predicting cell-penetrating peptides[J]. BMC Genomics, 2017,18(Suppl 7):742. DOI: 10.1186/s12864-017-4128-1.

[41]WEI Leyi, XING Pengwei, SU Ran, et al. CPPred-RF: A sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency[J]. Journal of Proteome Research,2017,16(5):2044-2053. DOI: 10.1021/acs.jproteome.7b00019.

[42]QIANG Xiaoli, ZHOU Chen, YE Xiucai, et al. CPPred-FL: A sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning[J]. Briefings in Bioinformatics,2020,21(1):11-23. DOI: 10.1093/bib/bby091.

[43]HUANG Guangbin, ZHU Qinyu, SIEW C K. Extreme learning machine: A new learning scheme of feedforward

neural networks[C]//2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541), July 25-29, 2004, Budapest, Hungary: IEEE, 2004: 985-990. DOI: 10.1109/IJCNN.2004.1380068.

[44] PANDEY P, PATEL V, GEORGE N V, et al. KELM-CPPpred: Kernel extreme learning machine based prediction model for cell-penetrating peptides[J]. Journal of Proteome Research, 2018, 17(9): 3214-3222. DOI: 10.1021/acs.jproteome.8b00322.

[45] GEURTS P, ERNST D, WEHENKEL L. Extremely randomized trees[J]. Machine Learning, 2006, 63, 3-42. DOI: 10.1007/s10994-006-6226-1.

[46] MANAVALAN B, SUBRAMANIYAM S, SHIN T H, et al. Machine-learning-based prediction of Cell-penetrating peptides and their uptake efficiency with improved accuracy [J]. Journal of Proteome Research, 2018, 17(8): 2715-2726. DOI: 10.1021/acs.jproteome.8b00148.

[47] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324. DOI: 10.1109/5.726791.

[48] YAO Zhou, ZHANG Wenjing, SONG Peng, et al. DeepFormer: A hybrid network based on convolutional neural network and flow-attention mechanism for identifying the function of DNA sequences[J]. Briefings in Bioinformatics, 2023, 24(2): bbad095. DOI: 10.1093/bib/bbad095.

[49] FUKUMA R, YANAGISAWA T, KINOSHITA M, et al. Prediction of IDH and TERT promoter mutations in low-grade glioma from magnetic resonance images using a convolutional neural network[J]. Scientific Reports, 2019, 9: 20311. DOI: 10.1038/s41598-019-56767-3.

[50] ZHANG Chengcheng, LU Yao, ZANG Tianyi. CNN-DDI: A learning-based method for predicting drug-drug interactions using convolution neural networks[J]. BMC Bioinformatics, 2022, 23: 88. DOI: 10.1186/s12859-022-04612-2.

[51] GRAVES A, MOHAMED A-R, HINTON G. Speech recognition with deep recurrent neural networks[C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing, May 26-31, 2013, Vancouver, BC, Canada: IEEE, 2013: 6645-6649. DOI: 10.1109/ICASSP.2013.6638947.

[52] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780. DOI: 10.1162/neco.1997.9.8.1735.

[53] 方春, 孙福振, 李彩虹, 等. 基于深度学习和字符嵌入的细胞穿透肽预测[J]. 计算机仿真, 2019, 36(10): 353-358. DOI: 10.3969/j.issn.1006-9348.2019.10.073.
FANG Chun, SUN Fuzhen, LI Caihong, et al. Prediction of cell penetrating peptides based on deep learning and character embedding [J]. Computer Simulation, 2019, 36(10): 353-358. DOI: 10.3969/j.issn.1006-9348.2019.10.073.

[54] PARK H, PARK J H, KIM M S, et al. In Silico screening and optimization of cell-penetrating peptides using deep learning methods [J]. Biomolecules, 2023, 13(3): 522. DOI: 10.3390/biom13030522.

[55] DOMINGOS P. A few useful things to know about machine learning[J]. Communications of the ACM, 2012, 55(10): 78-87. DOI: 10.1145/2347736.2347755.

[56] LEVER J, KRZYWINSKI M, ALTMAN N. Model selection and overfitting[J]. Nature Methods, 2016, 13: 703-704. DOI: 10.1038/nmeth.3968.

[57] AGRAWAL P, BHALLA S, USMANI S S, et al. CPPsite 2.0: A repository of experimentally validated cell-penetrating peptides[J]. Nucleic Acids Research, 2016, 44(D1): D1098-103. DOI: 10.1093/nar/gkv1266.

[58] YAMASHITA H, KATO T, OBA M, et al. Development of a cell-penetrating peptide that exhibits responsive changes in its secondary structure in the cellular environment[J]. Scientific Reports, 2016, 6: 33003. DOI: 10.1038/srep33003.