人工智能在抗癌肽研发中的应用与挑战

张志星, 邓 华, 唐 赟*

（华东理工大学药学院, 上海市新药设计重点实验室, 上海 200237）

摘 要 抗癌肽（anticancer peptides, ACPs）因其高效低毒和高选择性优势成为研究焦点, 而基于人工智能的 ACPs 识别和设计方法较传统实验方法成本低廉、成功率高且能够探索更广阔的序列空间。本文重点介绍了人工智能技术在 ACPs 生成和识别过程中的应用, 包括深度生成模型探索新型 ACPs 设计以及基于机器学习和深度学习的 ACPs 识别方法。此外, 文章还讨论了当前研究中存在的模型可复现性和可解释性不足、缺乏经过实验验证的阴性数据等挑战, 并对未来研究方向提出展望, 以期为 ACPs 的研发提供新思路。

关键词 抗癌肽; 人工智能; 机器学习; 深度学习

中图分类号 TP181; R914.2 **文献标志码** A **文章编号** 1000−5048(2024)03−0347−10

doi: 10.11665/j.issn.1000−5048.2024040201

引用本文 张志星, 邓华, 唐赟. 人工智能在抗癌肽研发中的应用与挑战 [J]. 中国药科大学学报, 2024, 55(3): 347 − 356.

Cite this article as: ZHANG Zhixing, DENG Hua, TANG Yun. Applications and challenges of artificial intelligence in the development of anticancer peptides[J]. *J China Pharm Univ*, 2024, 55(3): 347 − 356.

Applications and challenges of artificial intelligence in the development of anticancer peptides

ZHANG Zhixing, DENG Hua, TANG Yun*

Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China

Abstract Anticancer peptides (ACPs) have become a focal point of research due to their high efficacy, low toxicity, and high selectivity. Methods of ACP identification and design based on artificial intelligence (AI) surpass traditional experimental techniques in cost-efficiency, success rate, and the ability to investigate a broader sequence space. This article highlights the application of AI technology in the generation and identification of ACPs, including the exploration of new ACP design through deep generative models and ACP identification methods based on machine learning and deep learning. Furthermore, it discusses challenges in current research, such as insufficient model reproducibility and interpretability, and a lack of experimentally validated negative data. Future research directions are proposed to provide new insights for the development of anticancer peptides, aiming to enhance the understanding and development of ACPs.

Key words anticancer peptides; artificial intelligence; machine learning; deep learning

This study was supported by the National Natural Science Foundation of China (No. U23A20530)

癌症是仅次于心血管疾病的全球第二大死因[1]。2022 年全球新增癌症病例约 2 000 万、死亡病例约 974 万, 且预计到 2050 年将有超过 3 500 万新增癌症病例, 比 2022 年增加 77%[2]。2022 年中国新增癌症病例约 482 万、死亡病例约 257 万, 中国新增癌症和死亡病例位列全球第一且呈现增长趋势[3]。

目前癌症的治疗方法主要包括手术、化疗、放

收稿日期 2024-04-02 * **通信作者** Tel: 021-64251052 E-mail: ytang234@ecust.edu.cn

基金项目 国家自然科学基金项目（No. U23A20530）

疗、免疫治疗和靶向治疗。手术治疗通过切除肿瘤达到治疗目的，但对于晚期癌症患者效果有限。化疗和放疗虽广泛应用，但常伴随严重的不良反应和耐药问题，影响患者生活质量。免疫治疗和靶向治疗作为新兴治疗方式，虽显示出较好的疗效，但因成本高昂、选择性有限且对特定人群有效，限制了其普遍应用。因此，现有癌症治疗方式虽取得一定成效，仍需开发高疗效、高选择性、低副作用的新型抗癌药物。在此背景下，具有生产成本低、序列选择性高、组织渗透性高、毒性低、免疫原性低等优点的抗癌肽（anticancer peptides, ACPs）逐渐进入人们的视野并得到广泛关注。

ACPs 是一类具有抗肿瘤活性的生物活性肽，其结构特征决定了其独特的理化性质，进而赋予其特殊的抗癌机制和功能。ACPs 通常由 2~50 个氨基酸残基组成，相对分子质量在 50~5 000。在氨基酸组成上，ACPs 富含带正电荷的赖氨酸、精氨酸和组氨酸[4]，以及疏水性较强的亮氨酸，这些残基可赋予 ACPs 阳离子的特性和疏水性。此外，二肽组成（dipeptide composition, DPC）分析表明，ACPs 中排名前 10 位的特征由富含 R 基团性质为正电性、疏水脂肪链以及疏水芳香环的氨基酸残基的二肽组成（前 10 位特征为 KK、AK、KL、AL、KA、KW、LA、LK、FA 和 LF），这进一步佐证了含有带正电荷、疏水性强的氨基酸的比例是鉴别 ACPs 的重要特征[5]。二级结构上，ACPs 主要形成 α-螺旋和 β-折叠构象，其余 β-转角、无规卷曲等构象较少形成。

ACPs 的这些结构特征直接决定了其性质对癌细胞的作用方式（图 1）。由于含有大量正电荷和疏水基团，ACPs 表现出阳离子性和高度疏水性。阳离子性使 ACPs 能与阴离子性的肿瘤细胞膜发生静电作用力；疏水性则使 ACPs 能渗透到癌细胞膜的疏水区域，促进在癌细胞膜上形成孔道[6]，但疏水性过高会降低抗癌活性并增加对正常细胞的毒性[7]。ACPs 的抗肿瘤机制多样，其中，大部分 ACPs 通过破坏肿瘤膜结构来发挥作用[8]。此外，其他常见的机制包括诱导肿瘤细胞凋亡、抑制肿瘤血管生成以及参与免疫调节等。

目前报道的 ACPs 破坏膜结构的机制有孔道形成（pore formation）、地毯模型（carpet model）、膜溶解（membrane dissolution）和脂质-肽结构域形成（lipid-peptide domain formation）等[4]。虽然每种机制的特异性不同，但所有提出的肽诱导的膜破裂机制都会造成细胞质泄漏，最终导致细胞死亡，而这些机制均源于 ACPs 与癌细胞膜成分的亲和结合。因为癌细胞膜表面富含阴离子成分如磷脂、唾液酸化糖蛋白等，所以与表现出阳离子特性的肽易发生静电吸引。同时癌细胞膜流动性增高、微绒毛多而表面积大，也为肽分子渗透提供了条件[9]。相比之下，健康细胞膜呈现中性且流动性较低，因此与 ACPs 的相互作用较弱。癌细胞膜和健康细胞膜的差异使 ACPs 在靶向癌细胞时具有良好选择性。总之，ACPs 的结构决定了其独特的理化性质，而这些性质进一步赋予其特异的抗肿瘤活性和功能。目前已有多个 ACPs 药物被美国食品药品监督管理局（Food and Drug Administration, FDA）批准用于癌症的诊断和治疗，例如亮丙瑞林、曲普瑞林和卡非佐米等。

ACPs 最初来源于天然产物，这些天然产物覆盖了广泛的生物种类，包括细菌、真菌、植物和动物。除了从天然来源中直接分离纯化，利用生物信息学分析毒液动物的转录组和蛋白质组数据，也能为筛选肽类先导化合物提供线索[10]。此外，基于诸如噬菌体显示、酵母显示等体内外展示技术，能在体内或体外系统中快速构建大规模多肽文库（数量级达到 10^{10}~10^{15}），并高通量筛选出潜在的靶向性抗癌先导化合物[10]。通过化学修饰手段能进一步改善

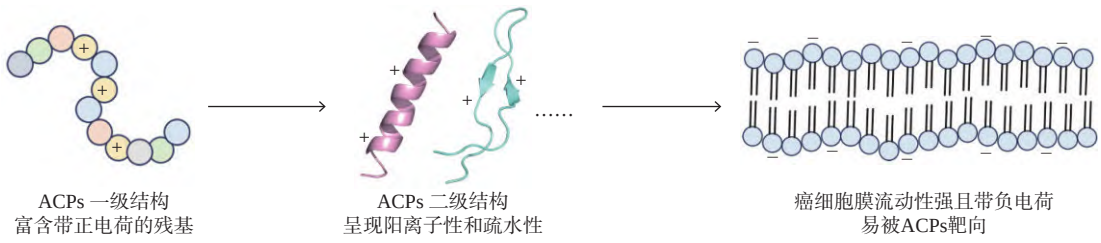

图 1 抗癌肽（anticancer peptides, ACPs）的结构与性质和功能的关系

肽的先导化合物的抗癌特性、稳定性和生物利用度等性质, 常见的策略包括在识别的切割位点进行特定的修饰、N 到 C 末端的环化、肽类似物的设计和肽二级结构元素的模拟等。

天然肽筛选、体内外展示和化学修饰等策略等湿实验方法在当前 ACPs 的发现中发挥了重要作用, 但存在着耗时、昂贵、成功率等缺点。随着人工智能(artificial intelligence, AI)技术的发展, 机器学习(machine learning, ML)和深度学习(deep learning, DL)方法正广泛应用于 ACPs 发现并已取得一定进展, 有望弥补湿实验方法的不足。本文也将着重介绍 ACPs 发现的 AI 方法及其进展。

1　用于模型训练的抗癌肽数据

1.1　数据库

ACPs 数据库及数据集是 ACPs 研究的宝贵资源, 精确、全面的 ACPs 数据库及数据集对于开发和验证 ACPs 生成和识别模型至关重要。

随着对 ACPs 研究的不断深入, 相关数据快速积累, 构建专门的数据库收集和管理这些数据就显得尤为必要。目前, 已经有多个公开可访问的数据库收录了 ACPs 的序列、活性、结构等信息, 为开发设计新型 ACPs 及其计算生物学研究提供了重要的数据支撑, 可获取到 ACPs 的相关数据库见表 1。其中, CancerPPD 是第 1 个专门收录 ACPs 的在线数据库[11], 除了提供肽段序列信息, 还预测并给出了每个收录肽段的二级和三级结构信息。该数据库自问世以来就一直是 ACPs 研究领域的重要资源。ApInAPDB[12] 是首个提供诱导凋亡 ACPs 信息的专业数据库, 该数据库不仅提供肽的功能、结合靶标与亲和力、IC_{50} 等相关信息, 还提供了 ACPs 的其他理化性质和结构信息。

另外, 一部分抗菌肽(antimicrobial peptides, AMPs)具有抗癌活性, 因此一些收集 AMPs 数据的数据库中也涵盖了部分 ACPs 数据。其中, APD(Antimicrobial Peptide Database)是 AMPs 数据库的代表性数据库, 其最新版本为 APD3[13−14]。除 APD 外, DADP(Database of Anuran Defense

表 1　可获得 ACPs 的数据库

生物活性肽类型	数据库	描述	开发年份	更新时间	唯一条目数/总条目数[a]	ACPs数量
ACPs	CancerPPD[11]	经实验验证的ACPs和蛋白质数据库	2015	未获得相关信息	3612	3491
	ApInAPDB[12]	凋亡诱导ACPs数据库	2022	未获得相关信息	818	818
	TumorHoPe[19]	肿瘤归巢肽综合数据库	2012	未获得相关信息	704/744	744
AMPs和ACPs	APD3[14]	AMPs数据库(含ACPs)	2016	2024年1月	3940	290
	DADP[15]	防御肽数据库, 由AMPs和ACPs组成	2012	未获得相关信息	1923/2571	108
	DBAASP v.3[16]	AMPs数据库(含ACPs)	2020	未获得相关信息	21509	3599
	DRAMP 3.0[17]	AMPs数据库(含ACPs)	2022	2023年11月	22528	163
	LAMP2[18]	AMPs和ACPs数据库	2013	2016年12月	23253	未计数
	dbAMP 2.0[20]	用于探索具有转录组和蛋白质组数据的功能活性和理化特性的AMPs的综合数据库	2019	2021年11月	28709	2290
	CAMPR3[21]	AMPs数据库(含ACPs)	2015	未获得相关信息	10247	未计数
	YADAMP[22]	AMPs数据库(含ACPs)	2012	2013年3月	2525	未计数
不限	SATPdb[23]	带注释的肽数据库, 由 20 个肽数据库和2个数据集组成。涵盖ACPs、抗寄生虫肽、细胞穿透肽、毒性肽等10多个类别的肽数据	2015	未获得相关信息	19192	1099
	StraPep[24]	已知生物活性肽的结构数据库	2018	未获得相关信息	1312/3791	未计数
	PlantPepDB[25]	植物肽数据库	2020	未获得相关信息	3848	未计数
	THPdb[26]	FDA 批准的治疗肽和蛋白质数据库	2017	未获得相关信息	852	未计数
	BioPepDB[27]	食品来源的生物活性肽数据库	2018	2018年1月	4807	635
	MBPDB[28]	源自牛奶蛋白质的生物活性肽数据库	2017	2024年1月	691	18

a: 此处统计的是数据库中收录的所有生物活性肽数据, 而不仅限于ACPs数据; 访问数据库及统计数据的日期为2024年3月20日; AMPs: 抗菌肽(antimicrobial peptides)

Peptides)是一个专门收集两栖动物防御肽的手工构建数据库[15]。DBAASP 则是提供 AMPs 分子动力学模拟信息的数据库，最新版本为 v3 版本[16]。此外，DRAMP[17]、LAMP[18] 等 AMPs 数据库也都涵盖了部分 ACPs 数据。最后，一些通用的收集生物活性肽的数据库中也收录了部分 ACPs 数据。

通过以上数据库，研究人员可以方便获取大量 ACPs 的序列、活性、结构和理化性质等数据，这为在分子水平上阐明 ACPs 的作用机制，并开发设计新型抗癌肽类药物奠定了重要的数据基础。与此同时，这些数据库也为借助深度生成模型生成 ACPs、从庞大的肽序列空间中识别 ACPs 等计算生物学研究提供了宝贵的数据资源。可以预见，随着相关研究的持续深入，未来也必将有更多的数据库应运而生，为 ACPs 研究奠定坚实的基础。

1.2 基准数据集

构建高质量的基准数据集是开发计算方法识别 ACPs 的第一步也是关键的一步。目前已有多个研究者从公共数据库和文献中收集并构建了多个基准数据集，用于训练和测试 ACPs 的 ML 和 DL 模型。表 2 中具体列出了现有基准数据集及相关信息。这些数据集大多包含实验验证的 ACPs 作为阳性数据(即正例)，以及从 Uniprot 数据库中随机选取的非抗癌肽(non-ACPs)序列作为阴性数据(即负例)。

在构建基准数据集时，需要注意消除数据中存在的同源偏差和冗余序列。大多数研究采用 CD-HIT 程序[33] 去除序列同一性高于一定阈值(如 90% 或 100%)的肽段。此外，一些研究者还设置了序列长度的限制，以获得更合理的数据分布。例如 TY_MD、TY_AD、TY_BD 和 TY_IND 这四个数据集中，肽段序列长度无限制[29]；而 SA_TRAIN、SA_IND 和 SA_RAND 数据集则只保留长度在 15~100 个氨基酸的肽段序列[31]。

表 2 现有的 ACPs 基准数据集

数据集	序列同一性[a]/%	ACPs数量	non-ACPs数量	总数
TY_MD[29]	100	225	2250	2475
TY_AD[29]	100	225	1372	1597
TY_BD[29]	100	225	225	450
TY_IND[29]	100	50	50	100
ZOH[30]	90	138	206	344
SA_TRAIN[31]	100	217	3979	4196
SA_IND[31]	100	40	40	80
SA_RAND[31]	100	-	2000	2000
Chen_S1[32]	100	138	206	344
Chen_S2[32]	100	150	150	300
H-C[5]	100	126	205	331
LEE[5]	100	422	422	844

a: 在构建数据集时，研究者去除了数据集中两两序列同一性超过90%或100%的肽段。

在正式的基准数据集中，ACPs 和 non-ACPs 的数量存在较大差异。为了避免训练模型时产生偏差，一些研究者会构建平衡数据集，使阴阳性数据数量相等，如 TY_BD 数据集[29]。而另一些研究则保留了原始的不平衡数据分布，如 SA_TRAIN 数据集中阴性数据数量远多于阳性数据[31]。不平衡的数据分布更能反映现实情况，但也可能导致模型对少数类过度拟合。因此，在模型训练时需要采取适当的策略来处理数据的不平衡问题。

除了用于模型训练的数据集外，一些研究还分别构建了 TY_IND[29]、SA_IND[31] 和 Chen_S2[32] 等独立测试数据集。这些独立测试集中的序列与训练集无重叠，可用于评估模型在未见数据上的泛化能力和可靠性。独立测试集的构建有助于全面衡量模型的预测性能，避免过拟合问题。

总的来说，现有的这些基准数据集在序列来源、构建方式、数据分布等方面存在一定差异，为开发 ACPs 生成和识别模型提供了多样化的数据支持。不同的数据集特点也反映了不同研究者在构建数据集时的不同考虑。在实际建模过程中，可以根据具体情况选择合适的基准数据集，或结合多个数据集的优势进行模型训练和测试，以提高模型的泛化能力和稳健性。同时，随着实验技术的发展和数据的不断积累，构建更大规模、更高质量的 ACPs 数据集也是未来的发展趋势。

2 人工智能助力抗癌肽生成

深度生成模型是利用深度神经网络模拟训练样本分布的先进技术，在计算机视觉、自然语言处理、游戏开发、计算生物学和药物研发等领域引起了革命性的变化。常见的深度生成模型包括变分自编码器(variational autoencoder, VAE)、生成对抗网络(generative adversarial network, GAN)、能量基模型、自回归模型和规范化流等[34]。深度生成模型通过学习数据集的分布规律来生成新的数据点，不

仅能够为新数据点分配可能性或概率, 衡量其是否来源于训练数据集的分布, 还能采样或生成与训练数据性质相似的新数据点。此外, 深度生成模型通过指定数据的生成过程, 能够进行特征学习和因果推理, 探索无限或指数大的组合搜索空间, 从而为从头设计生物活性肽等应用提供了可能。

活性肽的深度生成模型基于 DL 的数据驱动方法, 需要先将肽类数据通过 k-mer、独热编码(one-hot encoding)、嵌入(embedding)等特征表示方法转化为机器可读的格式, 再通过深度生成模型生成新的肽类, 最后对筛选得到的候选肽通过湿实验验证检验活性(图 2)。

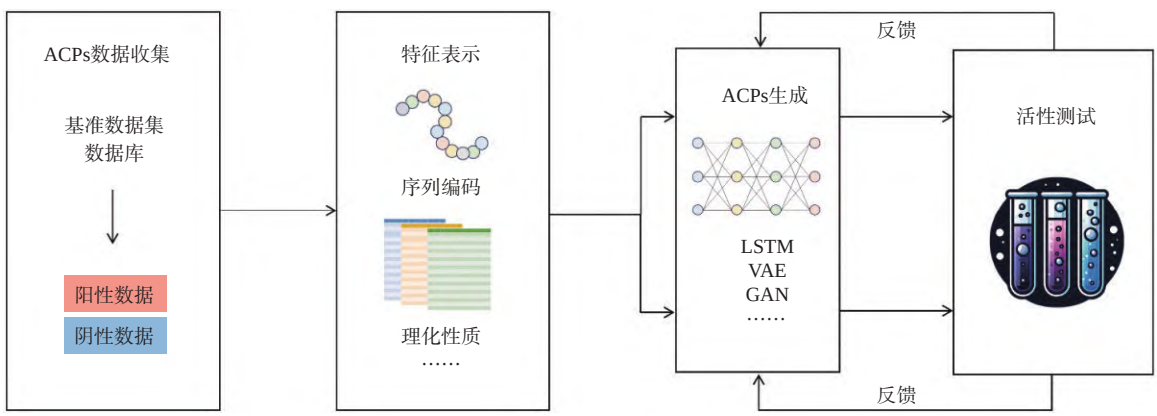

图 2　ACPs 的深度生成模型构建

LSTM: 长短期记忆神经网络;VAE:变分自编码器;GAN:生成对抗网络

最初的活性肽生成模型是一个基于长短时记忆的递归神经网络(recurrent neural network, RNN)生成模型, 其生成的肽序列中 82% 被预测为具有抗菌活性, 优于随机生成的对照组(65% 被预测为具有抗菌活性)[35]。此后, 深度生成模型已被用于生成不同类型的活性肽, 如 AMPs、ACPs、细胞穿透肽、信号肽等[36]。

目前, 在 ACPs 生成方面的研究还较少, 下面具体介绍两个相关研究:

一项研究中训练了一个长短期记忆神经网络(long short term memory, LSTM)单元的 RNN 模型, 首先使用 10 000 个假定为 α-螺旋阳离子两亲性肽序列进行训练, 然后通过迁移学习的方法在已知 26 种具有抗癌活性的 ACPs 上进行了微调[37]。经过优化的模型生成了 1 000 个新颖的氨基酸序列, 其中 98% 是独一无二的。在所合成测试的 12 种肽中, 10 种表现出对乳腺癌 MCF7 细胞的活性, 6 种能够高度选择性地杀伤癌细胞而不影响人红细胞[37]。这项研究显示了深度生成模型在无需显式结构-活性关系信息的情况下, 能从已知的 ACPs 中隐式提取设计规则, 从头设计全新的 ACPs。

另一项研究提出了基于 GAN 的 GANDALF 肽设计系统, 旨在生成针对特定蛋白靶点的肽[38]。与现有 GAN 方法只能生成小分子不同, GANDALF 采用两个网络分别生成肽序列和结构, 并纳入活性原子等数据, 能生成完整的肽结构和预测肽与靶点的结合亲和力。研究中生成了针对 PD-1、PDL-1 和 CTLA-4 3 个靶点的多种肽, 发现生成的最佳肽与 FDA 已批准药物的结合亲和力和三维结合适配度相当, 且序列独特[38]。该工作展示了深度生成模型在设计具有良好生物活性和结构特征的 ACPs 方面的潜力。

深度生成模型为 ACPs 发现提供了新的高效计算工具, 但仍面临一些挑战, 比如怎样更好地评估和筛选生成肽、有效整合肽结构信息、将优化算法与生成模型相结合等, 需要持续的研究和探索[34]。此外, 为所生成的肽分子与实验数据间建立反馈机制, 通过主动学习不断优化生成模型, 也是未来的一个重要方向。

虽然目前大多数研究集中在生成具有单一功能(如抗菌或抗癌)的肽上, 但未来的工作还可探索设计复合活性肽、多肽药物和肽疫苗等更加复杂的肽分子。此外, 大语言模型(large language model, LLM)技术方兴未艾, 已有多个具有生成功能的蛋白质大语言模型[39−42]被报道, 在蛋白质生成方面表现出色。随着未来应用的推广和创新, 大语言模型

技术也将成为 ACPs 生成的一大助力。

总之，随着实验数据的不断积累、算法的不断精进以及计算能力的不断提高，以深度生成模型为代表的 AI 技术在 ACPs 生成方面应用前景广阔。

3 人工智能助力抗癌肽识别

3.1 概 述

ACPs 展现出了结构简单、选择性好、不易引起多重耐药性等多种优势，有望成为临床上理想的抗癌药物。然而，如何快速准确地从广阔的序列空间中识别出有抗癌活性的序列成为 ACPs 研发中的一大难题。实验方法如高通量筛选、生物活性测试等 ACPs 识别方式耗时费力、成本高昂且成功率低。而 AI 技术通过自动化识别模型的构建，革命性地改变了 ACPs 的识别过程。具体应用上，主要通过 ML 和 DL 两种 AI 方法建立 ACPs 识别模型（图 3），其优势如下：（1）AI 通过计算方法预测 ACPs，只需将氨基酸序列输入训练好的模型即可进行识别，绕过了漫长的实验过程，大大加快了发现进程；（2）AI 算法在准确识别 ACPs 方面表现出色，优于传统方法；（3）通过减少对成本高昂的实验方法的依赖，AI 驱动的方法为识别 ACPs 提供了更经济实惠的解决方案。（4）AI 模型可以高效处理和分析大量序列数据，探索更为广阔的序列空间，而这是传统方法所难以胜任的。

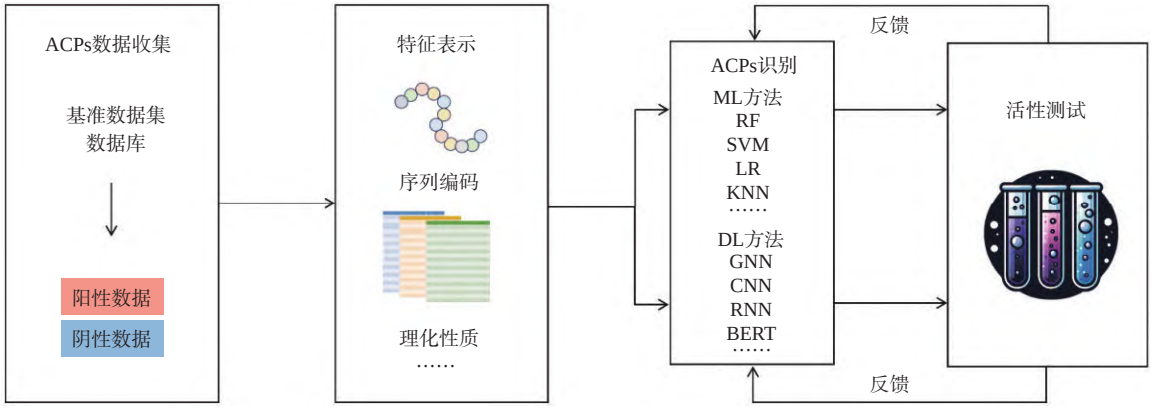

图 3 ACPs 识别模型构建

ML:机器学习;RF:随机森林;SVM:支持向量机;LR:逻辑回归;KNN:K-近邻;DL:深度学习;GNN:图神经网络;CNN:卷积神经网络;BERT: 双向变换器模型

近十年来，识别 ACPs 的 AI 模型如雨后春笋般涌现，已有数十种识别方法被报道，多篇综述对目前已有的 ACPs 识别模型进行了系统介绍[43−45]。以下选取有一定代表性的案例对不同种类的 ACPs 识别模型进行介绍。

3.2 基于机器学习的抗癌肽识别

构建可靠的 ACPs 识别的 ML 模型通常包括 4 个关键步骤[45]：（1）收集数据并进行处理，构建低冗余的训练集和独立测试集；（2）从数据中提取特征以涵盖序列信息的多个方面；（3）使用常用的 ML 分类器进行模型训练和评估；（4）开发网络服务器或独立程序。

以下对 4 个具有代表性的基于 ML 的 ACPs 识别方法进行介绍。

AntiCP 2.0[46] 采用支持向量机（support vector machine, SVM）、随机森林（random forest, RF）、K-近邻（k-nearest neighbors, KNN）、人工神经网络（artificial neural network, ANN）、极端随机树（extra trees）和岭回归（ridge regression）等算法，利用氨基酸组成（amino acid composition, AAC）、二肽组成（dipeptide composition, DPC）、端基组成（terminus composition, TC）和二值编码（binary profiles, BP）等特征，在主要数据集（861 个 ACPs 和 861 个 non-ACPs）和备用数据集（970 个 ACPs 和 970 个 non-ACPs）的验证集上分别取得 75.43% 和 92.01% 的准确率（即 accuracy，下同）。

ACPred-Fuse[47] 采用 RF 算法，生成 114 种特征描述符，涵盖 29 种特征编码，对应训练 RF 模型后获得概率和类别新特征，再结合其他特征通过序列化特征选择获得最优特征集。在十折交叉验证

和独立数据集验证中, ACPred-Fuse 分别达到 88.2% 和 86.8% 的准确率。

mACPpred[48] 研究中构建了高质量的训练集(266 个 ACPs 和 266 个 non-ACPs)和独立测试集(157 个 ACPs 和 157 个 non-ACPs), 对 7 种特征编码进行序列化特征选择训练 SVM 模型, 再将 7 个模型的预测概率作为新特征输入 SVM 进行整合。该方法在交叉验证和独立数据验证中分别取得 91.7% 和 91.4% 的准确率。

ACPred-FL[49] 研究中首先构建了训练集(250 个 ACPs 和 250 个 non-ACPs)和独立测试集(82 个 ACPs 和 82 个 non-ACPs), 然后从 7 种编码生成 40 种特征描述符, 训练相应的 SVM 模型并将模型预测作为新特征。最后利用序列化特征选择获得最优五维特征集, 在十折交叉验证和独立测试集验证中分别取得 91.4% 和 85.7% 的准确率。

近年来, ML 在 ACPs 识别领域取得了长足进展, 现有模型通过有效整合多种序列编码特征以及使用多种分类器, 显著提升了预测性能。这些模型不仅在训练集上普遍取得了 80% 以上的高准确率, 而且通过独立测试集验证证实了良好的泛化能力。尽管如此, 目前这些“黑盒”模型缺乏可解释性, 难以揭示预测的内在决策机制。同时, 评估模型时基准数据集有时并不一致, 也制约了模型间的公平对比和评估。因此, 在未来需要开发更具解释性的 ML 模型, 并建立标准的基准测试集, 以推动该领域的持续发展。

3.3　基于深度学习的抗癌肽识别

基于 DL 的 ACPs 识别模型首先通过构建训练和独立测试数据集, 然后利用 DL 算法结合特定的特征编码方法来训练模型(或者自动从肽序列中学习特征编码), 最终通过交叉验证和独立数据集验证的方式评估模型的预测准确率和性能。

ACP-DL[50] 研究中构建了两个新的基准数据集 ACP740(包含 740 个 ACPs 和 non-ACPs)和 ACP240(包含 240 个 ACPs 和 non-ACPs)。ACP-DL 是一种 LSTM 神经网络模型, 利用 k-mer 稀疏矩阵和二元特征编码对肽序列进行表示, 并输入 LSTM 模型进行训练。在 ACP740 和 ACP240 数据集上的五折交叉验证中, ACP-DL 分别达到 81.48% 和 85.42% 的准确率。

ACPred-LAF[51] 采用的特征编码方法与传统的肽的特征编码方法有着本质的区别。它引入了一种新颖的多感知和多尺度嵌入算法, 通过自动学习和提取 ACPs 序列的上下文序列特征, 来自适应地调整每个氨基酸残基的表示向量。这种方法的核心在于将每个氨基酸残基映射到一个低维稠密向量上, 该向量作为残基的表示, 并利用多感知和多尺度嵌入来捕获 ACPs 序列的不同语义信息和上下文信息。在交叉验证实验中, ACPred-LAF 在 ACP740 和 ACP240 数据集上的准确率分别为 94.40% 和 89.48%, 优于其他方法。

CL-ACP 模型[52] 将 ACPs 的二级结构信息和原始序列作为特征空间输入, 并采用 LSTM 和卷积神经网络(convolutional neural network, CNN)分别提取上下文依赖性和局部相关性, 同时引入多头自注意力机制增强序列表达。在 ACP736(包含 375 个 ACPs 和 361 个 non-ACPs)、ACP240 和 ACP539(包含 189 个 ACPs 和 350 个 non-ACPs)数据集上的五折交叉验证中, CL-ACP 分别取得 83.83%、87.92% 和 84.41% 的准确率, 展现出良好的预测性能。

在 ACPs 识别领域, DL 模型的引入标志着一大进步, 它们通过自动提取序列特征, 优化识别流程, 提高了识别精度和效率。与传统的 ML 模型相比, DL 方法能够处理更复杂的非线性关系和大规模数据, 无需繁琐的特征工程, 这使得 DL 在识别潜在的 ACPs 方面更具有优势。然而, DL 模型也存在着一些问题, 如对大量标记数据的依赖、模型解释性差和计算资源要求高等。尽管 ML 方法在特征工程方面提供了更多的灵活性和可解释性, 但它们在处理复杂模式识别任务时的性能往往不如 DL 模型。总的来说, DL 在 ACPs 识别上的应用正逐步克服这些挑战, 展现出强大的潜力, 预示着在药物发现和生物信息学研究领域的未来发展方向。

3.4　其他抗癌肽识别方法

以上介绍了主流的基于 ML 和 DL 的 ACPs 识别方法, 以下介绍一些其他的 ACPs 识别方法, 包括基于集成学习的 ACPs 识别方法以及 ACPs 两级分类预测器。

集成学习模型不仅可以集成传统的 ML 模型, 也可以集成 DL 模型, 甚至是 ML 和 DL 模型的组合。集成学习的核心目标是结合多个模型的预测, 以提高整体预测的准确性和鲁棒性。已发表的基于集

成学习的 ACPs 识别模型包括 ACPredStackL[44]、PreTP-EL[53] 和 PreTP-Stack[54] 等，这 3 个模型均集成了多种 ML 算法，具有较好的预测性能和鲁棒性。而目前集成 DL 模型的 ACPs 识别模型还鲜有报道。

在 ACPs 的识别场景中，以上介绍的方法都只是识别随机的多肽序列是否具有抗癌活性，而无法识别对候选多肽可能靶向的癌症组织。而 ACPs 识别的两级分类预测器首先通过一个二分类模型识别多肽序列是否为 ACPs，之后可通过一个多标签分类模型预测其可能靶向的癌症组织。目前 ACP-MLC[55] 和 ACPScanner[56] 实现了 ACPs 的两级预测，并取得了较好的结果。但两者的表现均受限于现有的 ACPs 数据规模，只能覆盖一些数据量稍多的癌症组织，未来数据规模扩大或者针对小样本处理的算法显著改善方能进一步提高 ACPs 两级分类预测器的预测准确性，扩展应用到更多癌症组织的分类。

4 总结与展望

AI 技术在 ACPs 的发现领域已展现出巨大潜力。在 ACPs 生成方面，深度生成模型可探索无限或指数大的组合搜索空间，为从头设计 ACPs 开辟了新途径。在 ACPs 识别方面，基于 ML 和 DL 的 ACPs 识别模型可自动化从海量肽序列信息中识别潜在的 ACPs，大幅提升了识别的准确度和效率，降低了依赖成本高昂实验方法的需求。然而，尽管取得了显著进展，当前 AI 在 ACPs 发现中的研究仍面临多个挑战，包括模型的可复现性较差、可解释性不足以及缺乏经过实验验证的阴性数据等。

未来研究应着重于提高模型的可复现性和可解释性，并积极探索获取和利用经过实验验证的阴性数据的新方法。为提升模型的可复现性，需在研究中公开分享数据集、模型参数及训练过程的详细信息，并鼓励采用开源软件和平台促进研究成果的共享。提高模型的可解释性，有助于研究者理解模型的决策过程，进而促进 AI 模型在药物发现和生物机制研究中的应用。通过与实验生物学家的紧密合作，获取并整合更多经过实验验证的 ACPs 和 non-ACPs 数据，将进一步提高模型的预测准确性和泛化能力。

同时，还需解决现有模型对大量标记数据的依赖、模型解释性差以及计算资源要求高等问题。探索如何更好地评估和筛选生成肽、有效整合肽结构信息、将优化算法与生成模型相结合等方面，也是未来的重要研究方向。此外，增加生成肽与实验数据的反馈，通过主动学习不断优化生成模型，以及探索设计复合活性肽、多肽药物和肽疫苗等更加复杂的肽分子，将进一步推动 AI 在 ACPs 发现领域的应用。

随着实验数据的不断积累、算法的不断精进以及计算能力的不断提高，以 AI 技术为代表的新方法将在 ACPs 的发现与识别领域展现出更广阔的前景。通过解决上述挑战，AI 技术在 ACPs 领域的应用将更加深入和广泛，为抗癌药物的发现和开发提供有力支持，最终惠及广大癌症患者。

References

[1] Sung H, Ferlay J, Siegel RL, *et al*. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. *CA Cancer J Clin*, 2021, **71**(3): 209-249.

[2] International Agency for Research on Cancer. Cancer Today[EB/OL]//gco. iarc. who. int. (2024-02-01)[2024-03-05].https://gco.iarc.who.int/today/en/dataviz/tables?mode=population&types=1.

[3] Han BF, Zheng RS, Zeng HM, *et al*. Cancer incidence and mortality in China, 2022[J]. *J Natl Cancer Cent*, 2024, **4**(1): 47-53.

[4] Norouzi P, Mirmohammadi M, Houshdar Tehrani MH. Anticancer peptides mechanisms, simple and complex[J]. *Chem Biol Interact*, 2022, **368**: 110194.

[5] Manavalan B, Basith S, Shin TH, *et al*. MLACP: machine-learning-based prediction of anticancer peptides[J]. *Oncotarget*, 2017, **8**(44): 77121-77136.

[6] Huang YB, Wang XF, Wang HY, *et al*. Studies on mechanism of action of anticancer peptides by modulation of hydrophobicity within a defined structural framework[J]. *Mol Cancer Ther*, 2011, **10**(3): 416-426.

[7] Glukhov E, Burrows LL, Deber CM. Membrane interactions of designed cationic antimicrobial peptides: the two thresholds[J]. *Biopolymers*, 2008, **89**(5): 360-371.

[8] Xie MF, Liu DJ, Yang YF. Anti-cancer peptides: classification, mechanism of action, reconstruction and modification[J]. *Open Biol*, 2020, **10**(7): 200004.

[9] Chiangjong W, Chutipongtanate S, Hongeng S. Anticancer peptide: Physicochemical property, functional aspect and trend in clinical application (Review)[J]. *Int J Oncol*, 2020, **57**(3): 678-696.

[10] Muttenthaler M, King GF, Adams DJ, *et al*. Trends in peptide drug discovery[J]. *Nat Rev Drug Discov*, 2021, **20**(4): 309-325.

[11] Tyagi A, Tuknait A, Anand P, *et al*. CancerPPD: a database of anticancer peptides and proteins[J]. *Nucleic Acids Res*, 2015, **43**(Database issue): D837-D843

[12] Faraji N, Arab SS, Doustmohammadi A, *et al*. ApInAPDB: a database of apoptosis-inducing anticancer peptides[J]. *Sci Rep*, 2022, **12**(1): 21341.

[13] Wang Z, Wang GS. APD: the antimicrobial peptide database[J]. *Nucleic Acids Res*, 2004, **32**(Database issue): D590-D592

[14] Wang GS, Li X, Wang Z. APD3: the antimicrobial peptide database as a tool for research and education[J]. *Nucleic Acids Res*, 2016, **44**(D1): D1087-D1093.

[15] Novković M, Simunić J, Bojović V, *et al*. DADP: the database of anuran defense peptides[J]. *Bioinformatics*, 2012, **28**(10): 1406-1407.

[16] Pirtskhalava M, Amstrong AA, Grigolava M, *et al*. DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics[J]. *Nucleic Acids Res*, 2021, **49**(D1): D288-D297.

[17] Shi GB, Kang XY, Dong FY, *et al*. DRAMP 3.0: an enhanced comprehensive data repository of antimicrobial peptides[J]. *Nucleic Acids Res*, 2022, **50**(D1): D488-D496.

[18] Zhao XW, Wu HY, Lu HR, *et al*. LAMP: a database linking antimicrobial peptides[J]. *PLoS One*, 2013, **8**(6): e66557.

[19] Kapoor P, Singh H, Gautam A, *et al*. TumorHoPe: a database of tumor homing peptides[J]. *PLoS One*, 2012, **7**(4): e35187.

[20] Jhong JH, Yao LT, Pang YX, *et al*. dbAMP 2.0: updated resource for antimicrobial peptides with an enhanced scanning method for genomic and proteomic data[J]. *Nucleic Acids Res*, 2022, **50**(D1): D460-D470.

[21] Waghu FH, Barai RS, Gurung P, *et al*. CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides[J]. *Nucleic Acids Res*, 2016, **44**(D1): D1094-D1097.

[22] Piotto SP, Sessa L, Concilio S, *et al*. YADAMP: yet another database of antimicrobial peptides[J]. *Int J Antimicrob Agents*, 2012, **39**(4): 346-351.

[23] Singh S, Chaudhary K, Dhanda SK, *et al*. SATPdb: a database of structurally annotated therapeutic peptides[J]. *Nucleic Acids Res*, 2016, **44**(D1): D1119-D1126.

[24] Wang J, Yin TL, Xiao XW, *et al*. StraPep: a structure database of bioactive peptides[J]. *Database*, 2018, **2018**: bay038.

[25] Das D, Jaiswal M, Khan FN, *et al*. PlantPepDB: a manually curated plant peptide database[J]. *Sci Rep*, 2020, **10**(1): 2194.

[26] Usmani SS, Bedi G, Samuel JS, *et al*. THPdb: database of FDA-approved peptide and protein therapeutics[J]. *PLoS One*, 2017, **12**(7): e0181748.

[27] Li QL, Zhang C, Chen HJ, *et al*. BioPepDB: an integrated data platform for food-derived bioactive peptides[J]. *Int J Food Sci Nutr*, 2018, **69**(8): 963-968.

[28] Nielsen SD, Beverly RL, Qu YY, *et al*. Milk bioactive peptide database: a comprehensive database of milk protein-derived bioactive peptides and novel visualization[J]. *Food Chem*, 2017, **232**: 673-682.

[29] Tyagi A, Kapoor P, Kumar R, *et al*. In silico models for designing and discovering novel anticancer peptides[J]. *Sci Rep*, 2013, **3**: 2984.

[30] Hajisharifi Z, Piryaiee M, Mohammad Beigi M, *et al*. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test[J]. *J Theor Biol*, 2014, **341**: 34-40.

[31] Vijayakumar S, Ptv L. ACPP: a web server for prediction and design of anti-cancer peptides[J]. *Int J Pept Res Ther*, 2015, **21**(1): 99-106.

[32] Chen W, Ding H, Feng PM, *et al*. iACP: a sequence-based tool for identifying anticancer peptides[J]. *Oncotarget*, 2016, **7**(13): 16895-16909.

[33] Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases[J]. *Bioinformatics*, 2001, **17**(3): 282-283.

[34] Bond-Taylor S, Leach A, Long Y, *et al*. Deep generative modelling: a comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models[EB/OL]. *arXiv*, 2021: 2103.04922. http://arxiv.org/abs/2103.04922.

[35] Müller AT, Hiss JA, Schneider G. Recurrent neural network model for constructive peptide design[J]. *J Chem Inf Model*, 2018, **58**(2): 472-479.

[36] Wan FP, Kontogiorgos-Heintz D, de la Fuente-Nunez C. Deep generative models for peptide design[J]. *Digit Discov*, 2022, **1**(3): 195-208.

[37] Grisoni F, Neuhaus CS, Gabernet G, *et al*. Designing anticancer peptides by constructive machine learning[J]. *ChemMedChem*, 2018, **13**(13): 1300-1302.

[38] Rossetto A, Zhou WJ. GANDALF: peptide generation for drug design using sequential and structural generative adversarial networks[C]//Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. Virtual Event USA. ACM, 2020. doi: 10.1145/3388440.3412487.

[39] Madani A, McCann B, Naik N, *et al*. ProGen: language modeling for protein generation[EB/OL]. *arXiv*, 2020: 2004.03497. http://arxiv.org/abs/2004.03497.

[40] Nijkamp E, Ruffolo JA, Weinstein EN, *et al*. ProGen2: exploring the boundaries of protein language models[J]. *Cell Syst*, 2023, **14**(11): 968-978. e3.

[41] Ferruz N, Schmidt S, Höcker B. ProtGPT2 is a deep unsupervised language model for protein design[J]. *Nat Commun*, 2022, **13**(1): 4348.

[42] Chen B, Cheng XY, Li P, *et al*. xTrimoPGLM: unified 100B-

scale pre-trained transformer for deciphering the language of protein[EB/OL]. *arXiv*, 2024: 2401.06199. http://arxiv.org/abs/2401.06199.

[43] Basith S, Manavalan B, Hwan Shin T, *et al*. Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening[J]. *Med Res Rev*, 2020, **40**(4): 1276-1314.

[44] Liang X, Li FY, Chen JX, *et al*. Large-scale comparative review and assessment of computational methods for anti-cancer peptide identification[J]. *Brief Bioinform*, 2021, **22**(4): bbaa312.

[45] Hwang JS, Kim SG, Shin TH, *et al*. Development of anticancer peptides using artificial intelligence and combinational therapy for cancer therapeutics[J]. *Pharmaceutics*, 2022, **14**(5): 997.

[46] Agrawal P, Bhagat D, Mahalwal M, *et al*. AntiCP 2.0: an updated model for predicting anticancer peptides[J]. *Brief Bioinform*, 2021, **22**(3): bbaa153.

[47] Rao B, Zhou C, Zhang GY, *et al*. ACPred-Fuse: fusing multi-view information improves the prediction of anticancer peptides[J]. *Brief Bioinform*, 2020, **21**(5): 1846-1855.

[48] Boopathi V, Subramaniyam S, Malik A, *et al*. mACPpred: a support vector machine-based meta-predictor for identification of anticancer peptides[J]. *Int J Mol Sci*, 2019, **20**(8): 1964.

[49] Wei LY, Zhou C, Chen HR, *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.

[50] Yi HC, You ZH, Zhou X, *et al*. ACP-DL: a deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation[J]. *Mol Ther Nucleic Acids*, 2019, **17**: 1-9.

[51] He WJ, Wang Y, Cui LZ, *et al*. Learning embedding features based on multisense-scaled attention architecture to improve the predictive performance of anticancer peptides[J]. *Bioinformatics*, 2021, **37**(24): 4684-4693.

[52] Wang HQ, Zhao J, Zhao H, *et al*. CL-ACP: a parallel combination of CNN and LSTM anticancer peptide recognition model[J]. *BMC Bioinformatics*, 2021, **22**(1): 512.

[53] Guo YC, Yan K, Lv HW, *et al*. PreTP-EL: prediction of therapeutic peptides based on ensemble learning[J]. *Brief Bioinform*, 2021, **22**(6): bbab358.

[54] Yan K, Lv HW, Wen J, *et al*. PreTP-stack: prediction of therapeutic peptides based on the stacked ensemble learing[J]. *IEEE/ACM Trans Comput Biol Bioinform*, 2023, **20**(2): 1337-1344.

[55] Deng H, Ding M, Wang YM, *et al*. ACP-MLC: a two-level prediction engine for identification of anticancer peptides and multi-label classification of their functional types[J]. *Comput Biol Med*, 2023, **158**: 106844.

[56] Zhong GL, Deng L. ACPScanner: prediction of anticancer peptides by integrated machine learning methodologies[J]. *J Chem Inf Model*, 2024, **64**(3): 1092-1104.

［专家介绍］唐赟，博士，华东理工大学药学院教授、博士生导师。1996 年博士毕业于中国科学院上海药物研究所，随后在瑞典、美国、加拿大等地留学工作 8 年，2004 年回国任复旦大学教授，同年参与华东理工大学药学院创建，曾任副院长 10 年，目前为药学学科负责人。2005 年入选上海市首批“浦江人才计划”，2008 年入选教育部“新世纪优秀人才支持计划”。有超过 30 年的计算机辅助药物设计经验，承担过国家自然科学基金等 20 余项科研项目，已发表 SCI 论文 280 余篇，获得计算机软件著作权 14 项，授权专利 6 项，主编、参编教材、专著和译著 10 余本，已为制药行业培养硕博人才 100 多名。曾获得上海市育才奖、宝钢优秀教师奖、上海市教学成果一等奖、药明康德生命化学研究奖，编著的《药物设计学》获得 2022 年中国石油和化学工业优秀图书奖-优秀教材一等奖。