Artificial Intelligence Review (2025) 58:156
https://doi.org/10.1007/s10462-025-11148-3

Bridging machine learning and peptide design for cancer treatment: a comprehensive review

Khosro Rezaee[1] · Hossein Eslami[1]

Accepted: 11 February 2025 / Published online: 5 March 2025

Abstract
Anticancer peptides (ACPs) offer a promising alternative to traditional cancer therapies due to their specificity and reduced side effects. The development of ACPs using machine learning (ML) and deep learning (DL) follows a structured process, beginning with sequence collection from in vitro and in vivo experiments. Key features such as hydrophobicity and secondary structure are extracted, and classification models categorize peptides based on their properties. ML models predict anticancer effectiveness, followed by toxicity checks and Structure-Activity Relationship (SAR) analysis to ensure safety and efficacy, with validation tests confirming their activity. This review explores how the automated design of ACPs can be enhanced by leveraging advanced ML and DL techniques. These methods, with their ability to automate feature selection and activity prediction, have significantly improved the efficiency and accuracy of peptide discovery. This structured approach holds high potential to guide researchers in the automated design of ACPs, accelerating the discovery of effective peptides while ensuring safety. Special attention is given to new approaches such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Generative Adversarial Networks (GANs), which show promise in addressing key challenges like data imbalance and computational complexity. Moreover, we examine the latest published research to compare the performance of various ML models in ACP prediction. By considering these advancements and challenges, this review outlines future opportunities for improving the scalability and reliability of ACP discovery using AI-driven techniques. This structured approach underscores the transformative impact of automation in peptide design, pushing the boundaries of modern cancer therapy development.

Keywords Anticancer peptides · Machine learning · Deep learning · Feature selection · Automated peptide design

✉ Khosro Rezaee
Kh.rezaee@meybod.ac.ir

✉ Hossein Eslami
heslami@meybod.ac.ir

[1] Department of Biomedical Engineering, Meybod University, Meybod, Iran

1 Introduction

Cancer remains one of the most significant threats to human life, with high mortality rates in both developing and developed countries (Ortega-García et al. 2020). According to the World Health Organization (WHO) and the International Agency for Research on Cancer (IARC), 18.1 million new cancer cases were reported in 2018, with 9.6 million deaths attributed to the disease (Bray et al. 2018). Cancer is characterized by molecular and genetic alterations that lead to uncontrolled cell growth and proliferation, eventually forming tissue masses in affected areas of the body (Jafari et al. 2022). Under normal conditions, cells undergo growth followed by programmed cell death, known as apoptosis. However, in cancer, cells evade these processes, leading to unchecked growth and division (Saxena et al. 2021). These abnormal cells rapidly multiply and invade surrounding tissues and organs, a process known as metastasis, which further complicates the progression of the disease. In 2020 alone, cancer accounted for nearly 10 million deaths (Anand et al. 2023; Kim et al. 2023).

Common treatment methods include surgery, chemotherapy, radiotherapy, and immunotherapy (Shewach and Kuchta 2009). Surgical interventions can often remove solid tumors quickly, especially in the early to mid-stages of cancer. However, surgery often causes significant damage, including bleeding, infections, and reduced immunity, making it unsuitable for most patients with advanced tumors (Cai et al. 2018). Radiotherapy is commonly used for patients who are not candidates for surgery, while chemotherapy uses chemical agents to target cancer cells throughout the body. However, prolonged chemotherapy can lead to drug resistance and accelerate disease progression, as these agents often affect both cancerous and healthy cells, resulting in severe side effects (Aljabery et al. 2018). Traditional treatments face several challenges that limit their effectiveness, such as inherent or acquired drug resistance (Chiangjong et al. 2020) and metabolic heterogeneity within tumor cells, which can enhance resistance and promote cancer progression (Payandeh et al. 2018a; Payandeh et al. 2018b). Additionally, chemotherapy often lacks precision, as it targets both cancerous and healthy cells, leading to severe side effects.

Given the risks of traditional chemotherapy and radiotherapy, there has been a noticeable shift toward molecular targeting of cancer cells to reduce harm to healthy tissues (Zhang et al. 2019; Zhong et al. 2021). Peptide-based therapies have garnered significant attention due to their high selectivity, effectiveness, and lower toxicity compared to small-molecule drugs. These therapies hold considerable promise in overcoming many challenges faced by conventional treatments, such as drug resistance and side effects (Anand et al. 2023; Henninot et al. 2018; Kaspar and Reichert 2013; Uhlig et al. 2014). In recent years, peptide-based therapies have gained significant attention due to their selectivity and effectiveness compared to traditional treatments.

ACPs are preferred over small-molecule drugs due to their distinct pharmacokinetic (PK), harmacodynamic (PD), structural, and safety advantages. Pharmacokinetically, ACPs often have shorter half-lives, reducing drug accumulation and associated side effects. Their high specificity allows targeted binding to cancer cells, minimizing damage to healthy tissues (Bidwell and Raucher 2009; Rusiecka et al. 2022). Pharmacodynamically, ACPs are designed to interact with cancer cell surface receptors, ensuring precise biological effects and significantly reducing off-target interactions compared to small molecules (Rusiecka et al. 2022; Raucher et al. 2009). Structurally, their amphipathic nature enables them to disrupt cancer cell membranes and induce apoptosis efficiently. Additionally, ACPs are highly modifiable, allowing rapid optimization for enhanced efficacy. From a safety perspective, their similarity to endogenous molecules reduces the risk of severe immune responses, and their distinct metabolic pathways result in better tolerability compared to the extensive hepatic metabolism of small-molecule drugs (Bidwell and Raucher 2009; Rusiecka et al. 2022). These features make ACPs a promising alternative in cancer therapy.

Peptides, thanks to their ability to specifically target cancer cells while minimizing harm to healthy tissues, represent a promising area of research. Furthermore, peptides can be used in combination with other treatment methods, including vaccines, chemotherapy, radionuclide carriers, and hormone therapy, to enhance therapeutic outcomes. Figure 1 illustrates various potential cancer treatment options explored through peptide sequences, demonstrating the multifaceted approach of combining traditional methods like chemotherapy and hormone therapy with newer strategies such as peptide therapy and radionuclide carriers.

Recent research has shown that ML techniques can play a crucial role in the identification and design of ACPs (Chiangjong et al. 2020; Luo et al. 2019). By analyzing the physicochemical properties and functional aspects of peptides, ML models can predict which peptides are most likely to exhibit anticancer activity, streamlining the drug discovery process (Chiangjong et al. 2020). This approach allows for faster and more accurate identification of potential therapeutic peptides, reducing the need for extensive experimental testing (Yang et al. 2019). Integrating ML, particularly deep learning (DL), into ACP identification is a major advancement, offering faster and more accurate predictions than traditional methods (Lv et al. 2021). DL is a subset of machine learning that uses neural networks with multiple layers (often referred to as deep neural networks) to automatically learn and extract

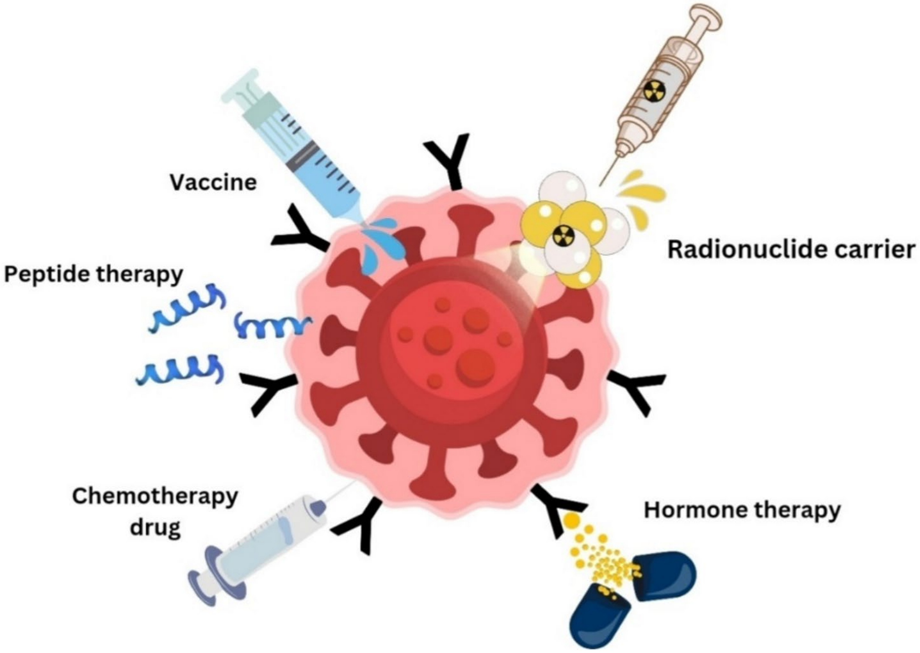

Fig. 1 Exploring potential cancer treatment options through peptide sequences (Anand et al. 2023; Henninot et al. 2018)

complex, nonlinear patterns from large datasets. Unlike traditional machine learning, deep learning models do not require extensive manual feature engineering, as they can learn hierarchical representations directly from raw data. Therefore, DL-based algorithms have the capability to extract complex features from biological data, significantly outperforming traditional methods in the identification of effective peptides (Kaleem et al. 2022). One of the key advantages of DL is its ability to process large datasets, precisely analyzing peptide features and predicting ACPs with high accuracy. These models can automatically recognize critical features using complex algorithms, such as recurrent neural networks (RNNs) and convolutional neural networks (CNN), offering superior accuracy compared to simpler ML approaches (Yao et al. 2023). CNNs are a class of DL models particularly effective in analyzing structured data such as images and peptide sequences. CNNs work by identifying spatial patterns through convolutional layers, making them ideal for feature extraction in anticancer peptide prediction. Moreover, RNNs are a type of DL model specifically designed to handle sequential data. Unlike traditional neural networks (Schneider and Wrede 1994), RNNs use loops to retain information from previous inputs, making them particularly effective for tasks involving time-series data, natural language processing, and sequence-based anticancer peptide prediction.

The purpose of this review is to explore the structures and characteristics of ACPs and assess the efficiency of their identification using ML and DL methods. By employing computational methods, the identification of ACPs can be significantly expedited, reducing the time and costs associated with traditional experimental methods (Yu et al. 2020). Given the limited research on the comprehensive examination of DL and ML techniques for ACP identification, this review seeks to fill these gaps and provide new insights into the field (Yi et al. 2019a, b).

Following this introduction, we delve into the structural analysis of ACPs and examine the ML and DL algorithms applied in their identification. All key terminologies and frequently used abbreviations in this review are included in Table 1 as a glossary for clarity. The primary objective is to provide a comprehensive understanding of the efficiency of these methods, along with their strengths and limitations in cancer therapy. The review is structured as follows: Sect. 2 introduces the characteristics and structures of ACPs. Section 3 explores how ML and DL are applied to the design of ACPs. Section 4, Recent Advances and Comparison, provides a detailed comparison of the latest methodologies used in ACP prediction. Section 5 highlights the challenges and limitations of current approaches. Section 6 discusses future directions and opportunities for improving the design and prediction of ACPs using AI-driven techniques. Finally, Sect. 7 summarizes the key findings and presents future research avenues.

2 Peptides and their characteristics

Antimicrobial peptides (AMPs) are bioactive molecules that offer protection against a range of pathogens, including bacteria, protozoa, fungi, and viruses (Raffatellu 2018). A specific subset of these peptides, known as ACPs, shows significant promise in combating cancer. These peptides are found across a wide variety of organisms, including mammals, plants, birds, amphibians, fish, insects, and microorganisms (Eghtedari et al. 2021). They typically consist of 2 to 50 amino acids (Xie et al. 2020), though this classification can sometimes be imprecise, as longer peptides are occasionally classified as protein fragments.

ACPs possess both hydrophobic and positively charged regions, enabling them to interact with the negatively charged membranes of cancer cells. This selective interaction allows ACPs to target and destroy cancer cells without damaging healthy cells (Shoombuatong et al. 2018), a major advantage over many conventional cancer treatments (Schweizer 2009; Soon et al. 2020). Additionally, ACPs are naturally occurring biological inhibitors that can be easily synthesized, making them an ideal therapeutic candidate for cancer treatment (Soon et al. 2020).

To date, more than 600 peptides have been used in clinical and preclinical studies, with 60 of them approved as drugs (Lau and Dunn 2018; Fuchs et al. 2018). The therapeutic applications of these peptides extend beyond cancer treatment to include drug delivery systems, hormone regulation, inflammation modulation, vaccines, antibiotics, and quorum-sensing molecules (Li et al. 2014; Verbeke et al. 2017; Basith et al. 2018; Manavalan et al. 2018; Tesauro et al. 2019). The first ACP approved by the FDA was Bortezomib (marketed as Velcade$^{®}$), which received approval in 2003 for the treatment of multiple myeloma (plasma cell cancer) and in 2006 for the treatment of mantle cell lymphoma (Micale et al. 2014).

Peptides exhibit a wide variety of structures. Some, like Tritrpticin and Indolicidin, are linear, while others, such as LL-37, BMAP-27, BMAP-28, and Cercopin A, adopt an alpha-helix structure. Additionally, some peptides fold into beta-sheets, as seen in defensins and lactoferrin (Deslouches and Di 2017; Gaspar et al. 2013). This structural diversity, along with their varied functions, plays a crucial role in biological processes, making peptides valuable in the development of drugs, antibodies, enzymes, and sensors (Langan et al. 2019; Quijano-Rubio et al. 2021).

2.1 Classification of anticancer peptides

Since the discovery of "cecropins" as bioactive peptides by Swedish scientist Boman and colleagues in 1980 (Boman et al. 1972), numerous other bioactive peptides have been identified. Many of these peptides demonstrate a wide range of biological functions, including antitumor effects and immune system regulation (Li et al. 2014). Various types of ACPs have since been identified across different organisms, and these peptides are classified in multiple ways depending on their structure, origin, and mode of action (Wang et al. 2016). Typically, ACPs are divided into four categories: cationic α-helical peptides, β-sheet peptides, cationic peptides rich in specific amino acids, and anionic peptides. These categories will be introduced and discussed in the following sections.

2.1.1 Primary structure

The primary structure of a polypeptide or protein refers to the exact sequence of amino acids that make up the chain. Each polypeptide chain has two distinct ends:

- **Amino-terminus (N-terminus)**: This end contains an amino group.
- **Carboxyl-terminus (C-terminus)**: This end contains a carboxyl group.

Table 1 Glossary of key terms and abbreviations used in the study

No.	Term	Explanation	No.	Term	Explanation	No.	Term	Explanation
1	Anticancer Peptides (ACP)	Bioactive peptides that target cancer cells while minimizing harm to healthy cells.	13	Antimicrobial Peptides (AMP)	Peptides with antimicrobial properties used against bacteria, viruses, and fungi.	25	Structure-Activity Relationship (SAR)	Relationship between a molecule's chemical structure and its biological activity.
2	Quantitative Structure-Activity Relationship (QSAR)	Mathematical modeling of SAR to predict biological activities quantitatively.	14	Convolutional Neural Networks (CNN)	Deep learning models that analyze spatial patterns in data for feature extraction.	26	Recurrent Neural Networks (RNN)	Deep learning models designed for sequential data by retaining previous inputs.
3	Long Short-Term Memory (LSTM)	A type of RNN capable of learning long-term dependencies in sequential data.	15	Generative Adversarial Networks (GAN)	Models consisting of a generator and a discriminator for generating synthetic data.	27	Deep Neural Networks (DNN)	Deep learning models with multiple hidden layers to learn complex data patterns.
4	Machine Learning (ML)	A subset of AI that uses algorithms to learn patterns from data and make predictions.	16	Deep Learning (DL)	A field of ML using neural networks to model complex relationships in data.	28	Feature Engineering	The process of transforming raw data into meaningful features for ML models.
5	Feature Extraction	Extracting specific, relevant features from data to improve model performance.	17	Classifier	Algorithms used to categorize or predict classes in data.	29	Pharmacokinetic (PK)	The study of how drugs are absorbed, distributed, metabolized, and excreted.
6	Pharmacodynamic (PD)	The study of the effects and mechanisms of action of drugs on the body.	18	Polypeptide	Polymers made of amino acid chains, forming the basis of proteins.	30	Amino Acid	Organic compounds that are building blocks of proteins and peptides.
7	Proteins	Large biomolecules made of polypeptides that perform various functions in cells.	19	Hydrophilic	Molecules that attract water due to their polarity or charge.	31	Tumor-targeting peptides (TTPs)	Peptides designed to selectively target tumor cells.
8	Cell-penetrating peptides (CPPs)	Peptides capable of penetrating cell membranes to deliver cargo into cells.	20	Host defense peptides (HDPs)	Peptides that protect the host by neutralizing pathogens.	32	Lipopolysaccharides (LPS)	Components of the outer membrane of Gram-negative bacteria.
9	Lipoteichoic Acid (LTA)	A major component of Gram-positive bacterial cell walls.	21	Cationic Antimicrobial Peptides (CAPs)	Positively charged antimicrobial peptides targeting bacterial membranes.	33	Protein Transduction Domains (PTDs)	Another name for CPPs, known for their ability to translocate cellular membranes.

Table 1 (continued)

No.	Term	Explanation	No.	Term	Explanation	No.	Term	Explanation
10	Plasmid DNA	A DNA molecule used for gene transfer and expression in cells.	22	siRNA	Small interfering RNA used to silence gene expression.	34	Apoptosis	Programmed cell death that eliminates damaged or unnecessary cells.
11	SMART Model	A statistical model that simplifies and explains complex data.	23	Matthews Correlation Coefficient (MCC)	A metric for evaluating the quality of binary classifications.	35	Area Under the Receiver Operating Characteristic (AUROC)	A metric summarizing the performance of a binary classifier.
12	Protein Language Models (PLMs)	Advanced models for analyzing protein sequences in biological data.	24	Imbalanced Data	A situation where one class in a dataset is significantly smaller than the other.	36	Explainable AI (XAI)	AI that makes model decisions interpretable for humans.

Twenty different amino acids can be arranged in various sequences to form a specific primary structure, giving the protein its unique characteristics and functions (Watt 2006). Cyclic ACPs are a specialized class of peptides that form closed-chain, ring-like structures. These rings are often stabilized by disulfide bonds, which create cystine knots that provide enhanced stability compared to linear peptides (Huang et al. 2011). Recently, a novel cyclic peptide group named *diffusa cyclotide 1–3* was discovered in the leaves and roots of the white snake plant. These peptides have shown significant inhibitory effects on three types of prostate cancer cells and effectively inhibited cancer cell migration in in vitro experiments at a concentration as low as 0.05 micromolar (Hu et al. 2015). Currently, cyclic ACPs account for a substantial portion of the peptides being explored in clinical trials, largely due to their strong anticancer properties. These peptides have demonstrated impressive inhibitory effects on various cancer cells (Gaspar et al. 2013). For example, *H-10*, a newly identified cyclic pentapeptide, has shown concentration-dependent inhibition of mouse B16 melanoma cell growth, with an IC50 value of 39.68 micromolar. This indicates that the peptide was able to inhibit 50% of melanoma cell growth at this concentration. Remarkably, *H-10* displayed no significant toxicity toward human lymphocytes and rat aortic smooth muscle cells (Zhang et al. 2014).

Besides, cyclic peptides have emerged as a promising therapeutic modality due to their unique structural stability, high target selectivity, and diverse pharmacological applications. Over 40 cyclic peptide drugs have entered the market, with many others in various phases of clinical trials, including innovative approaches like LUNA18, an orally bioavailable cyclic peptide targeting KRAS, showcasing advancements in chemical optimization and drug delivery (Tanada et al. 2023). Recent developments in screening technologies, such as phage and mRNA display, have further expanded the scope of cyclic peptides, enabling the discovery of de novo ligands for challenging targets (You et al., 2024). Furthermore, their applications as linkers in therapeutic conjugates and their progression into Phase I trials for cancer and other diseases highlight their versatility and potential (Costa et al. 2023). These advancements underscore the vital role of cyclic peptides in modern drug development pipelines and their potential for addressing unmet clinical needs.

Overall, cyclic peptides are highly promising candidates for anticancer therapies due to their potent anticancer effects and lower toxicity compared to other peptides. Their unique structural stability and biological activity make them ideal models for improving and refining peptide-based medical treatments.

2.1.2 Secondary structure

The secondary structure refers to specific shapes that a peptide chain adopts due to hydrogen bonds between different parts of the chain (Lin and Pan 2001). These include Alpha-Helix, where, the peptide chain coils into a helical structure. Moreover, Beta-Sheets that is the peptide chains align into flat, parallel sheets. These configurations represent the secondary structures of peptides, which play a crucial role in their biological activities, especially in ACPs. Beta-peptides, for example, have been shown to adopt stable helical and sheet structures, relevant to pharmaceutical applications (Wu et al. 2008). The discovery of these structures has enhanced the understanding of protein complexity and folding (Choi et al. 2008).

Most naturally occurring ACPs possess α-helix structures with cationic properties, though some may have different structural arrangements. For example, rationally designed α-helical ACPs have shown selective anticancer activities (Hadianamrei et al. 2021). The helix is one of the most critical secondary structures in peptides and proteins, manifesting in various forms. In α-helical ACPs, the peptide chain is coiled into a short helix, the most common among ACPs. For instance, certain cationic amphipathic α-helical peptides have demonstrated significant anticancer activity, selectively targeting cancer cells over normal cells due to differences in membrane composition (Hadianamrei et al. 2021). The 310-helix involves hydrogen bonding between the carbonyl group of one amino acid and the amine group of another amino acid three positions ahead. This structure is less stable and less commonly observed but plays a role in the folding processes (Choi et al. 2008). The π-helix is stabilized by hydrogen bonds between amino acids five positions apart and is less common but important for protein functionality (Cooley et al. 2010).

Beta-sheet structures are the second most common configuration in ACPs. These peptides, such as defensins, often display anticancer properties and are structurally stabilized by disulfide bonds (Arenas et al. 2019). Some plant defensins, for example, have demonstrated potent anticancer properties (Wu et al. 2019).

Random coil structures often occur in peptides with high glycine or proline content, lacking a stable secondary structure. Some, like glycine-rich ACPs, exhibit significant therapeutic potential due to their interactions with immune cells (Hein et al. 2022; Hanaoka et al. 2016).

2.1.3 Tertiary structure

In proteins, the secondary structure folds into a three-dimensional configuration known as the tertiary structure, or the protein's 3D structure. This structure is stabilized by various types of chemical bonds, including ionic bonds (between NH3 + and COO- groups), hydrogen bonds, hydrophobic and hydrophilic interactions, and disulfide bonds (Godbey 2014). The tertiary structure results from the folding of the secondary structure and is referred to as a domain. Based on their tertiary structure, proteins can be categorized into two types:

1. **Fibrous Proteins**: These proteins have elongated structures that combine to form filaments or bundles, typically exhibiting hydrophobic properties.
2. **Globular Proteins**: These proteins are spherical in shape and generally hydrophilic (Engelking 2015).

The tertiary structure of proteins describes how the various structural elements are arranged in space. In this structure, α-helices can be oriented either parallel or perpendicular to one another. In essence, the tertiary structure reveals how different parts of the protein, such as helices, sheets, loops, and other segments, fold together to create the final 3D configuration of the protein (Feher 2017).

2.2 Quaternary structure

The quaternary structure refers to the organization of multiple polypeptide chains or protein subunits that assemble to form a larger, functional unit. Each of these subunits may have its own primary, secondary, and tertiary structures. In the quaternary structure, subunits are held together by non-covalent bonds such as hydrogen bonds and van der Waals forces. This arrangement allows proteins to function in a coordinated and efficient manner. For instance, hemoglobin, found in blood, is composed of four subunits that work together to transport oxygen. Each subunit maintains its own structure, and collectively, these subunits form the quaternary structure of the protein (see Fig. 2) (Feher 2017; Varadi et al. 2021).

2.3 Other peptides related to Cancer Treatment

Given the vast diversity of peptides used in cancer treatment, it is crucial to explore the types of peptides that are effective in this field. Generally, the peptides reviewed below can be categorized as ACPs. Some peptides possess specific anticancer properties and are referred to as ACPs. Additionally, certain AMPs that have amino acid sequences similar to ACPs may exhibit anticancer effects alongside their antimicrobial activities. Moreover, peptides such as tumor-targeting peptides (TTPs), cell-penetrating peptides (CPPs), hormones, and vaccines can also be considered subsets of ACPs in the context of cancer diagnosis and treatment. Consequently, these various categories of peptides can be examined and selected within the framework of ACPs.

2.3.1 Antimicrobial peptides (AMPs)

AMPs, also known as cationic antimicrobial peptides (CAPs), are named for the presence of numerous cationic residues in their structure (Furlong et al. 2008; Kim et al. 2018). These peptides are typically small (5 to 50 amino acids) and carry a positive charge due to high levels of arginine and lysine (Grantham 1974; Hancock and Sahl 2006; Wang et al. 2016).

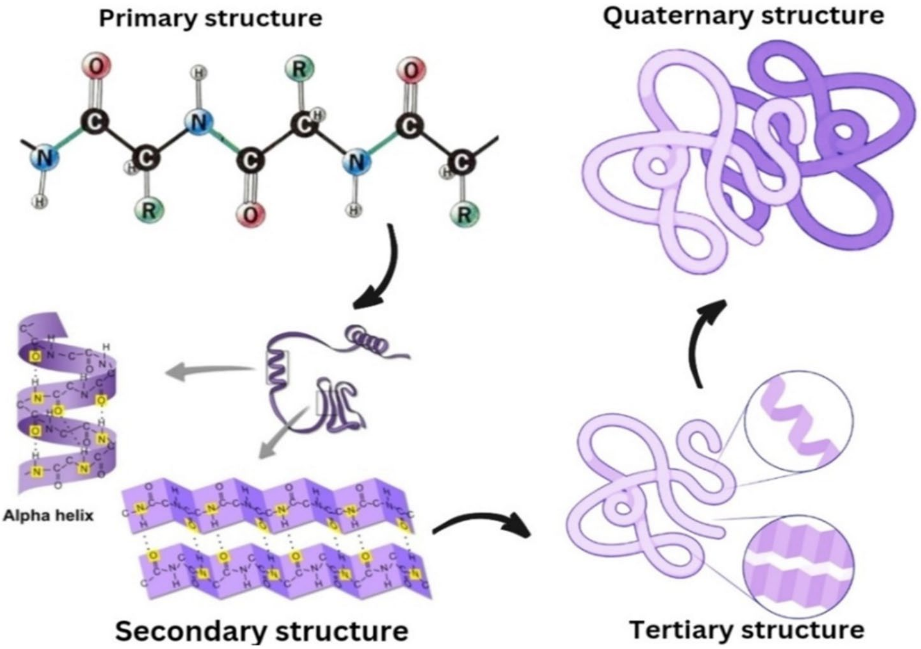

Fig. 2 Representation of peptide structures, illustrating the quaternary arrangement of subunits (Feher 2017; Varadi et al. 2021)

AMPs, often referred to as host defense peptides (HDPs), play a crucial role in the innate immune system of all living organisms (Wang et al. 2016). Over 1,500 antimicrobial peptides have been identified from various organisms, including plants, fungi, bacteria, and animals (Wang et al. 2016). Insects are also a significant source of AMPs due to their resistance to bacterial infections, attributed to their high AMP content. These peptides, generally less than 100 amino acids long and positively charged, exert antibacterial effects by disrupting microbial membranes. They must penetrate the thick peptidoglycan layer in Gram-positive bacteria and the outer lipopolysaccharide (LPS) layer in Gram-negative bacteria to act effectively (Cui et al. 2017).

AMPs exhibit significant diversity in their amino acid composition and secondary structures, which may include α-helix, β-sheet, extended helices, and loops. These peptides effectively interact with bacterial membranes, which carry a high negative charge, such as lipopolysaccharides (LPS) in Gram-negative bacteria or lipoteichoic acid (LTA) in Gram-positive bacteria. The antibacterial effectiveness of AMPs depends on their specific structure, which generally includes a cationic amphipathic pattern. In other words, AMPs must have positively charged regions to bind to negatively charged membranes, as well as hydrophobic regions to penetrate the membrane. The positive charge, hydrophobicity, and secondary structure of these peptides are critical for their antibacterial activity. A significant number of AMPs have already been discovered, with new peptides continuing to be added to the list (Wang et al. 2016). Insect-derived AMPs can be classified into three categories based on their structure or function: linear α-helical peptides without cysteine residues, globular β-sheet peptides stabilized by intramolecular disulfide bridges, and peptides unusually rich in proline and glycine (Wang et al. 2016). According to the Antimicrobial Peptide Database (APD), 2,619 peptides have been discovered so far (Wang et al. 2016). As concerns about microbial resistance grow, finding new disinfectants is essential, and various industries are increasingly focusing on this issue (Cui et al. 2017). Bacterial surfaces contain negatively charged molecules, making them prime targets for ACPs. These negative molecules are often lipids such as phosphatidylglycerol (PG), cardiolipin (CL), and phosphatidylserine (PS) (Deslouches and Di 2017).

2.3.2 Cell-penetrating peptides (CPPs)

Cell-penetrating peptides (CPPs), also known as protein transduction domains (PTDs), membrane-translocating peptides (MTPs), and Trojan peptides, have garnered significant attention due to their high cell permeability, ease of synthesis, sequence modifiability, and low toxicity (Nasiri et al. 2021; Raucher and Ryu 2015). These peptides, typically composed of 5 to 30 amino acids—mainly cationic residues—can cross tissue and cell membranes without specific receptor interactions, using either energy-dependent or energy-independent mechanisms (Desale et al. 2021). CPPs are amphipathic and non-toxic molecules (Desale et al. 2021).

Biological membranes typically restrict the passage of substances with a molecular weight exceeding 500 Daltons. However, CPPs provide an innovative and effective way to transport large molecules across these membranes. CPPs are short peptides, usually less than 40 amino acids long, and highly positively charged, primarily due to the presence of lysine or arginine residues. They can transport various biologically active materials, such as plasmid DNA, siRNA, and therapeutic proteins, across cellular membranes with minimal toxicity (Desale et al. 2021). CPPs act as molecular delivery vehicles capable of transporting a wide range of cargo molecules across the plasma membrane. They have numerous medical applications, including use as drug delivery agents in the treatment of various diseases, including cancer, and as contrast agents for cell labeling (Raucher and Ryu 2015).

2.3.3 Tumor-targeting peptides and tumor-homing peptides

Tumor-targeting peptides (TTPs) are small peptides that specifically bind to receptors expressed on the surface of tumor cells, facilitating the targeted delivery of therapeutic or diagnostic agents (Yang et al. 2008; Worm et al. 2020). These receptors, often overexpressed in cancerous tissues, interact with circulating molecules, such as peptides and antibodies, concentrating therapeutic agents in tumor cells while minimizing off-target effects (Ruoslahti 2017a).

Tumor-homing peptides (THPs), on the other hand, are designed to target specific regions within the tumor microenvironment or tumor-associated vasculature (Ruoslahti 2017b). They achieve this specificity by recognizing molecular motifs such as RGD (Arginine-Glycine-Aspartic acid) and NGR (Asparagine-Glycine-Arginine). The RGD motif binds to integrins expressed on tumor endothelial cells, while the NGR motif interacts with aminopeptidase N, a marker of tumor angiogenesis (Ruoslahti 2017a, b). These structural features allow THPs to selectively accumulate in tumors, enhancing therapeutic precision. Tumor-Selective Internalizing Peptides (TSIPs), often classified under THPs, share these capabilities and are typically short peptides (3–15 amino acids). While similar to ACPs in structure and tumor specificity, THPs are primarily utilized for targeted delivery, unlike ACPs, which exhibit cytotoxicity.

Advances in computational methodologies have revolutionized the identification and optimization of tumor-homing peptides (THPs), enabling researchers to predict and classify these peptides with remarkable accuracy. Several state-of-the-art tools have been developed to streamline this process. TumorHPD, a webserver leveraging support vector machines (SVMs), utilizes amino acid and dipeptide composition to predict THPs with high accuracy and user accessibility, providing a robust platform for researchers to analyze peptide features (Sharma et al. 2013). SCMTHP, another innovative tool, employs a scoring card-based approach, using machine learning algorithms trained on physicochemical properties and amino acid composition to identify THPs effectively (Charoenkwan et al. 2022). For more robust predictions, THPep integrates random forest classifiers with features such as amino acid composition, pseudo amino acid composition, and dipeptide composition, achieving reliable and versatile performance (Shoombuatong et al. 2019). Additionally, PLMTHP utilizes advanced protein language models (PLMs) within an ensemble learning framework to extract high-dimensional features, offering superior prediction capabilities by leveraging intricate peptide patterns and relationships (Chen et al. 2023). Finally, StackTHPred combines gradient boosting decision trees with stacking architectures to maximize prediction accuracy, presenting a highly effective tool for identifying THPs (Guan et al. 2023). Collectively, these computational tools have significantly enhanced the discovery pipeline for THPs by automating feature analysis and offering scalable solutions, thus bridging the gap between theoretical research and practical application in cancer-targeting peptide design. These tools analyze critical peptide characteristics such as motifs, charge distribution, and hydrophobicity. Their application enables researchers to accelerate the discovery and optimization of THPs while minimizing experimental costs. Integrating these tools into pipelines for peptide research ensures highly specific, effective, and scalable solutions for targeted therapies.

THPs and ACPs share structural and functional similarities, yet they differ significantly in their mechanisms of action and primary applications. ACPs are primarily designed to exert direct cytotoxic effects (Eliassen et al. 2006) on cancer cells by disrupting cell membranes or inhibiting cellular proliferation, making them potent agents for tumor eradication. In contrast, THPs function as highly specific carriers for therapeutic agents, such as chemotherapeutics, imaging probes, or nanoparticles, by targeting tumor-associated tissues and delivering these agents directly to the tumor microenvironment. This complementary relationship between ACPs and THPs offers significant potential for dual-function cancer therapies. ACPs can act as direct anticancer agents, targeting and killing cancer cells, while THPs enhance the precision and effectiveness of adjunctive treatments by improving drug delivery and minimizing systemic toxicity. Together, these two classes of peptides provide a versatile framework for developing multi-faceted cancer treatment strategies.

THPs have demonstrated extensive potential in clinical and experimental settings. They have been utilized for delivering chemotherapeutic agents like doxorubicin, cancer imaging probes, and nanoparticles. By binding to tumor vasculature and accumulating within the tumor microenvironment, THPs enhance the therapeutic index of anticancer agents while minimizing systemic toxicity (Ruoslahti 2017a, b). Additionally, THPs combined with nanocarriers have shown synergistic effects in drug delivery, leading to improved cancer imaging and therapeutic outcomes.

The integration of computational tools like TumorHPD, SCMTHP, and StackTHPred into experimental workflows bridges the gap between theoretical and practical applications, enabling researchers to address challenges of specificity and delivery in cancer therapy. The synergy between ACPs and THPs, coupled with advanced predictive methodologies, represents a significant leap toward multi-functional, peptide-based cancer-targeting therapies.

2.3.4 Peptides as Radiopharmaceutical Carriers

Peptides have numerous applications in cancer treatment. They can be used directly as cytotoxic agents against cancer cells or as carriers for toxic substances and radioisotopes that identify cancer cells. In peptide-based hormonal therapies, such as those for breast and prostate cancer, peptides have been extensively studied and applied (Schally and Nagy 2004). The conjugation of peptides to small molecules offers a novel and alternative method for developing peptide-based therapies with higher efficacy and safety. This approach is particularly important in oncology, where many therapeutic peptides have been developed (Schally and Nagy 2004). In the 1970s, this type of therapy was first introduced for melanotropin receptors, marking the beginning of the use of radioisotope-labeled peptides. In recent years, small radioisotope-labeled peptides have increasingly been used for diagnostic imaging and radionuclide therapy in nuclear oncology (Dash et al. 2015; Oyen et al. 2007). Peptide receptor radionuclide therapy (PRRT) is a targeted treatment that uses radioisotope-labeled peptides as biological carriers. These peptides deliver radiation specifically to cancer cells, minimizing damage to healthy tissues (Dash et al. 2015).

2.3.5 Anticancer peptides: Linear, Hybrid, and synthetic structures

Brevinin, a peptide consisting of 25 amino acids, was identified in the skin secretions of the giant-headed frog (*Limnonectes fujianensis*). This amphipathic, hydrophobic peptide features both α-helix and β-turn structures, allowing it to penetrate lipid bilayers of cell membranes (Arnab et al. 2023; Li et al. 2019). PR-39, a linear peptide rich in proline and arginine, belongs to the cathelicidin family. This peptide, comprising 39 amino acids, lacks secondary structure and is extracted from the small intestine and neutrophils of pigs. Hybrid ACPs are artificially constructed by combining different segments from other peptides. For example, the positively charged α-helix region from Cecropin A is combined with the hydrophobic α-helix region from peptides like Melittin, forming hybrid peptides with potentially enhanced anticancer properties. These hybrid peptides have demonstrated significant anticancer effects on lung cancer cell lines. The Melittin hybrid peptide exhibits low hemolytic activity (destruction of red blood cells), while the Magainin hybrid peptide also shows minimal hemolytic effects on red blood cells (Răileanu and Bacalum 2023). The flexible middle hinge region (Gly–Ile–Gly) plays a crucial role in anticancer activity, providing the necessary flexibility for effective interaction between the NH2-terminal α-helix region and the cell membrane. This flexibility allows the peptide to align parallel to the membrane, enabling the COOH-terminal α-helix region to effectively penetrate the cell membrane (Răileanu and Bacalum 2023).

Synthetic peptides are designed to penetrate cancer cell membranes without being degraded by serum enzymes. For instance, D-K4R2L9, a 15-amino acid synthetic peptide with diastereomeric and amphipathic properties, contains one-third D-type amino acids and includes leucine (Leu), lysine (Lys), and arginine (Arg) in its structure. This synthetic peptide has demonstrated anticancer activity against murine melanoma and human prostate cancer cell lines, showing potential in preventing lung cancer formation (Wu et al. 2022). Similarly, D-K6L9, another 15-amino acid diastereomeric and amphipathic peptide, contains D-Lys and D-Leu amino acids in one-third of its sequence. This peptide has proven to be particularly effective against human prostate cancer cells (Wu et al. 2022). In contrast, L-K6L9, a version of D-K6L9 made entirely from L-type amino acids, exhibits similar anticancer activity but also causes non-specific damage to fibroblasts and red blood cells (Wu et al. 2022).

Synthetic peptides such as Magainins A, B, and G—synthetic versions of the Magainin peptide—have demonstrated significant effects against lung cancer and drug-resistant tumor cells (Nabizadeh et al. 2023). Additionally, a hybrid peptide combining Cecropin and Magainin 2 has shown strong anticancer activity against various cancer cell types while causing minimal damage to red blood cells and fibroblasts (Nabizadeh et al. 2023). Synthetic versions of LfcinB peptides, which contain both cationic and hydrophobic regions, exhibit higher anticancer activity than their natural counterparts (Feng et al. 2000). For instance, peptides containing glutamic acid in mice demonstrated no anticancer activity, indicating that a high positive charge plays a critical role in their cytotoxic anticancer properties (Wu et al. 2022).

2.3.6 Membrane disruption by Anticancer Peptides

ACPs primarily interact with the surface of tumor cell membranes through electrostatic forces. Once enough peptides accumulate on the membrane, they can penetrate it, leading to membrane disruption and micellization. This process creates pores in the membrane, allowing the peptides to enter the tumor cell. Inside the cytosol, these peptides disrupt the mitochondrial membrane, causing the release of mitochondrial proteins like cytochrome c. Cytochrome c, a key apoptosis-inducing factor, initiates the aggregation of apoptotic protease activating factor-1 (Apaf-1), activates caspase-9, and converts procaspase-3 into caspase-3, which ultimately leads to programmed cell death (apoptosis) (Chiangjong et al. 2020).

Upon contact with the cell membrane, peptides induce changes in the membrane structure in less than 10 nanoseconds (Basith et al. 2018). Initially, they bind with phosphate groups in the membrane lipids, initiating membrane degradation (Wang et al. 2016). Research by Papo and colleagues has demonstrated that short peptides, resembling host defense peptides, selectively target cancer cells by binding to phosphatidylserine (PS) present on the surface of these cells. This binding disrupts membrane polarity, leading to cell death. Therefore, peptide-lipid interaction is a critical step for effective membrane disruption, although additional mechanisms may also contribute (Papo and Shai 2005).

Furthermore, studies have shown that peptide hydrophobicity is essential for anticancer activity. Increased hydrophobicity enhances the peptide's ability to form pores in cancer cell membranes, making them more effective at attacking cancer cells (Huang et al. 2011). However, this effect is only beneficial up to a certain point. If the hydrophobicity exceeds a threshold, the peptides not only become less effective against cancer cells but may also damage healthy cells, increasing toxicity (Glukhov et al. 2008) (see Fig. 3).

2.3.7 Formation of lipid-peptide domains

The SMART model (Soft Membranes Adapt and Respond, also Transiently), introduced by Bechinger (2015), explains the dynamic interaction between peptides and lipids. In this model, peptides and lipids mutually adjust their spatial structures and permeability, meaning that any alteration in one molecule affects the other. Peptides have the ability to penetrate bilayer lipid membranes, causing significant disruption and breakdown of the phospholipid layers. Essentially, in the SMART model, peptides and lipids continuously influence each other, with peptides exhibiting specific characteristics that can profoundly affect the membrane's structure and function (Bechinger 2015).

As the concentration of peptides increases, these molecules progressively disrupt the cell membrane's structure, ultimately leading to membrane breakdown. Although the mechanisms by which peptides disrupt membranes may differ, they all result in the leakage of intracellular contents and lead to cell death. Some researchers have proposed various models to explain these processes, including the cumulative channel model, peptide-induced lipid separation, leaky gaps, and peptide-induced non-lamellar phase formation (Bechinger 2015). Additionally, other models, such as in-plane diffusion and wormhole mechanisms, have been suggested (Wimley 2010) (see Fig. 4).

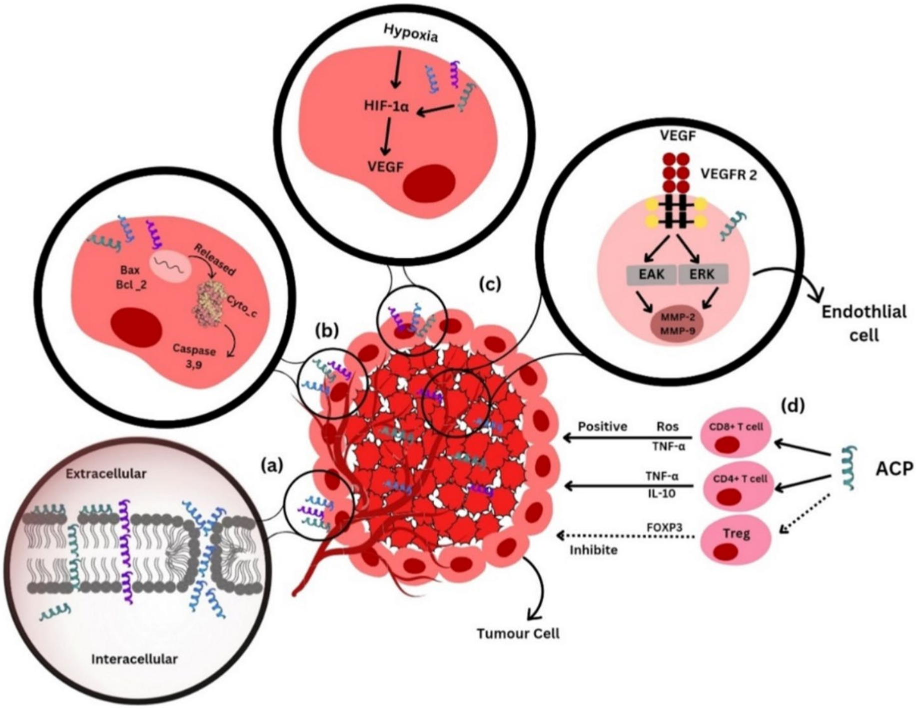

Fig. 3 Schematic representation of the anticancer mechanisms of ACPs: (a) disruption of the cell membrane structure, (b) induction of apoptosis, (c) inhibition of angiogenesis, and (d) immune modulation (Hilchie et al. 2019)

3 Machine learning for ACP Design

ML is transforming drug design, particularly in the identification and optimization of ACPs. By leveraging computational models, researchers can predict peptide sequences with high precision, reducing the need for traditional, costly experimental methods. A deep generative model has shown the ability to design novel and diverse functional peptides, addressing challenges like microbial resistance by producing effective candidates with high functional diversity (Mao et al. 2023a). Similarly, transformer-based generative models have demonstrated remarkable success in antiviral drug design, showcasing the potential of advanced architectures in generating optimized molecules for therapeutic purposes (Mao et al. 2023b). Moreover, comprehensive strategies for developing QSAR models provide a robust foundation for accurately predicting the bioactivity of molecules, further enhancing the reliability and scalability of AI-driven drug design pipelines (Mao et al. 2021). These advancements align closely with the themes discussed in this manuscript, emphasizing the transformative potential of ML in peptide design and cancer therapy. Moreover, recent advancements, including deep forest architecture and multi-domain transfer learning, have enhanced the accuracy of ACP identification by analyzing physicochemical properties and evolutionary data (Xu et al. 2024a, b; Yao et al. 2023). These techniques not only speed up ACP discovery

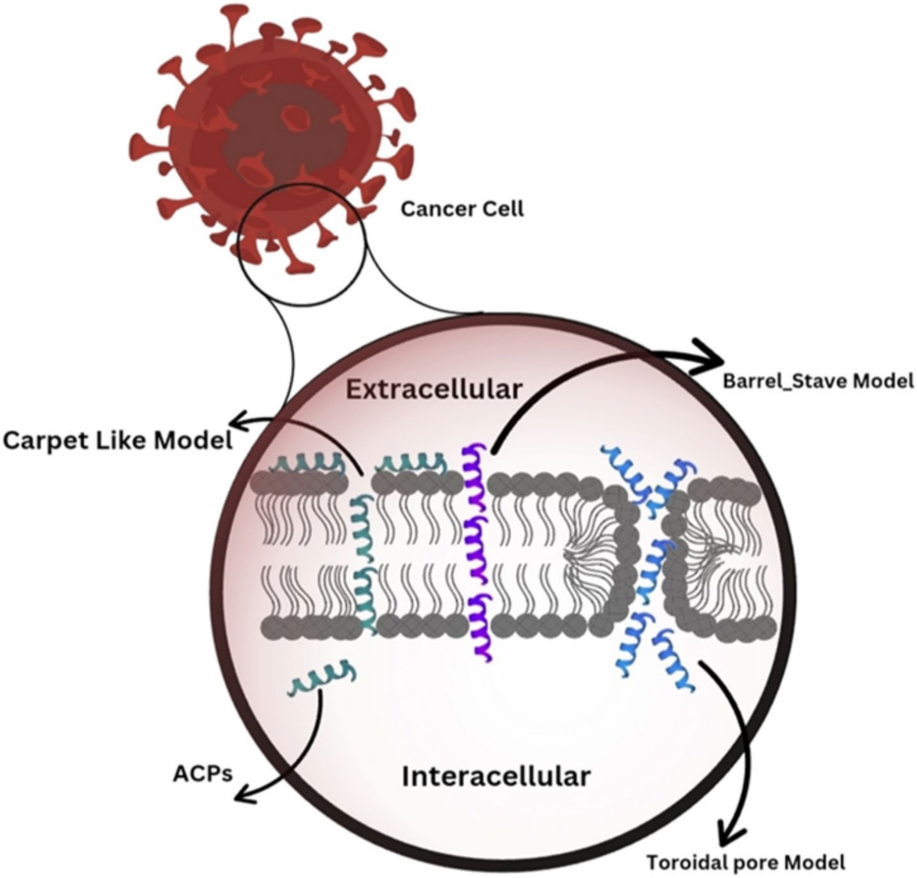

Fig. 4 Schematic representation of key models explaining the mechanisms of action of ACPs (Last et al. 2013; Borrelli et al. 2018)

but also improve their cancer-targeting specificity, making ML a crucial tool in peptide-based therapies and anticancer drug development.

3.1 Overview of machine learning in Drug Discovery

Structure-Activity Relationship (SAR) refers to the relationship between the chemical structure of a molecule and its biological activity. This concept is crucial in understanding how variations in peptide sequences influence their anticancer properties. One of the key applications of ML in drug discovery is predictive modeling, where algorithms are trained on large datasets to forecast the biological activity, toxicity, and pharmacokinetics of compounds (Chen et al. 2020). These models, such as QSAR, allow researchers to predict the efficacy and safety profiles of new drugs before entering costly experimental phases, reducing both time and cost in drug development (Dara et al. 2021). DL models have demonstrated remarkable success in identifying these targets, improving the likelihood of discovering effective drugs for complex diseases like cancer and neurodegenerative disorders (Sharma and Rani 2021).

ML techniques like graph-based neural networks have been used to model complex drug-target interactions, facilitating the identification of promising drug candidates for diseases like Alzheimer's and COVID-19 (Gaudelet et al. 2021). ML algorithms, driven by reinforcement learning and neural networks, enable the discovery of entirely new chemical entities optimized for binding affinities and selectivity towards biological targets (Pawar et al. 2023). The manual process often limits the discovery of peptides with optimal therapeutic profiles, as it is confined to known sequences and chemical spaces (Capecchi et al. 2021a, b). By training ML models on large datasets of known peptide sequences and their corresponding biological activities, researchers can predict the behavior of new, unseen peptides. This shift allows for the rational design of peptides with enhanced therapeutic potential, reducing reliance on laborious experimental methods (Boone et al. 2021a, b). Furthermore, these models enables the prediction of secondary and tertiary peptide structures, which play an essential role in the bioactivity and specificity of peptides (Janairo 2022).

In addition, ML models such as deep neural networks (DNNs) and generative adversarial networks (GANs) are now being used to generate entirely novel peptide sequences with desired characteristics, such as high binding affinity and low toxicity (Capecchi et al. 2021a, b). These generative models, which learn the underlying patterns in peptide datasets, can suggest new sequences that are optimized for specific drug targets or disease pathways. DNNs are advanced artificial neural networks with multiple hidden layers between the input and output layers. These architectures are capable of learning complex, nonlinear patterns from data, making them highly effective for tasks such as anticancer peptide classification and feature extraction. Notably, GANs are a class of deep learning models consisting of two neural networks—a generator and a discriminator—that compete with each other in a zero-sum game. GANs are highly effective in generating realistic synthetic data, making them invaluable for addressing data imbalance and augmenting datasets in anticancer peptide prediction. Thus, the GAN-based approaches not only accelerate the discovery process but also explore previously uncharted chemical spaces, making it possible to design peptides that might not be easily identified through ML methods (Lin et al. 2022). As a result, ML has emerged as a critical tool for the next generation of peptide therapeutics, offering higher efficiency and precision in the drug discovery pipeline (Pawar et al. 2023). By leveraging large datasets, ML tools reduce the time and resources needed for peptide screening and discovery, facilitating the identification of bioactive peptides with enhanced therapeutic potential (He et al. 2021). The complex nature of peptide-protein interactions makes ML a powerful tool for predicting peptide function, binding, and efficacy by analyzing sequence data and structural features, crucial for designing peptides with high specificity and minimal off-target effects (Guo et al. 2008; Ye et al. 2023; Shen et al. 2007).

Supervised learning is employed in peptide drug discovery by training models on labeled datasets to predict the behavior of new peptides, thereby facilitating the identification of promising drug candidates with specific properties, such as antimicrobial or anticancer activities (Boone et al. 2021a, b). Besides, Unsupervised learning techniques, like clustering, group peptides by properties, helping identify new therapeutic peptides and repurpose drugs for various applications (Janairo 2022).

As ML evolves, its application in peptide discovery expands, with emerging models combining supervised and unsupervised learning to optimize peptide drug design, predict bioactivity, stability, and toxicity, and develop safer, more effective therapeutics (Miao et al. 2021a, b).

Feature extraction, classification, and regression play a crucial role in peptide research by analyzing complex biological data to uncover patterns and predict peptide behavior. Feature extraction identifies key properties of peptide sequences, such as amino acid composition or structural motifs, that influence their bioactivity. Classification models help categorize peptides based on their functional properties (e.g., anticancer, antimicrobial) or structural characteristics, while regression models predict quantitative outcomes like binding affinity or stability. As shown in Fig. 5, the steps of ML-driven peptide design include data preprocessing, feature selection, model training, and validation, providing a comprehensive framework for generating optimized peptide sequences.

The performance of ML models in peptide design depends on the quality and quantity of training datasets. High-quality datasets with accurate annotations and diversity enable ML algorithms to learn the complex relationships between peptide sequences and their biological activities. Conversely, poor-quality or biased datasets can lead to inaccurate predictions and limit model generalizability. Curated databases, such as the APD, provide valuable collections of experimentally validated peptides and their activities, ensuring reliable model performance and enhancing the predictive power of ML models in peptide discovery.

3.2 Role of machine learning in ACP Prediction

As previously mentioned, traditional approaches to peptide discovery involve experimental testing, which can be labor-intensive and costly (Chen et al. 2021a, b, c; Basith et al. 2020). ML algorithms quickly analyze vast peptide libraries to predict potential anticancer activity, leveraging large datasets and extracting key biological and chemical characteristics to efficiently identify novel peptide sequences, with databases like APD and CancerPPD providing essential data for training predictive models such as neural networks and decision trees (Kaleem et al. 2022). Data augmentation techniques are employed to tackle the problem of limited labeled peptide data, improving the model's ability to generalize and predict new sequences more accurately (Chen et al. 2021a, b, c).

The APD and CancerPPD databases are crucial for ACP prediction. These repositories contain experimentally validated peptide sequences, which allow ML models to identify the underlying features of effective ACPs. AntiCP 2.0, a predictive model trained on these databases, utilizes ensemble ML methods and has shown high specificity and sensitivity in ACP predictions (Agrawal et al. 2020). Moreover, newer models like the one proposed by Yuan et al. (2023) also utilize ordinal positional encoding and have demonstrated significant improvements in ACP prediction performance (Yuan et al. 2023).

Several classification methods are employed to predict ACPs, including support vector machines (SVM), random forests, and deep neural networks. For example, Chen et al. (2021a, b, c) utilized CNN in their model xDeep-AcPEP, which demonstrated strong predictive accuracy for ACPs across various tumor types (Chen et al. 2021a, b, c). Additionally, the CACPP model, proposed by Yang et al. (2023), uses a Siamese network combined with CNN and contrastive learning to further improve ACP prediction by focusing solely on peptide sequence data (Yang et al. 2023).

Performance metrics like accuracy, sensitivity, specificity, and Matthews correlation coefficient (MCC) are commonly used to evaluate the effectiveness of ML models in

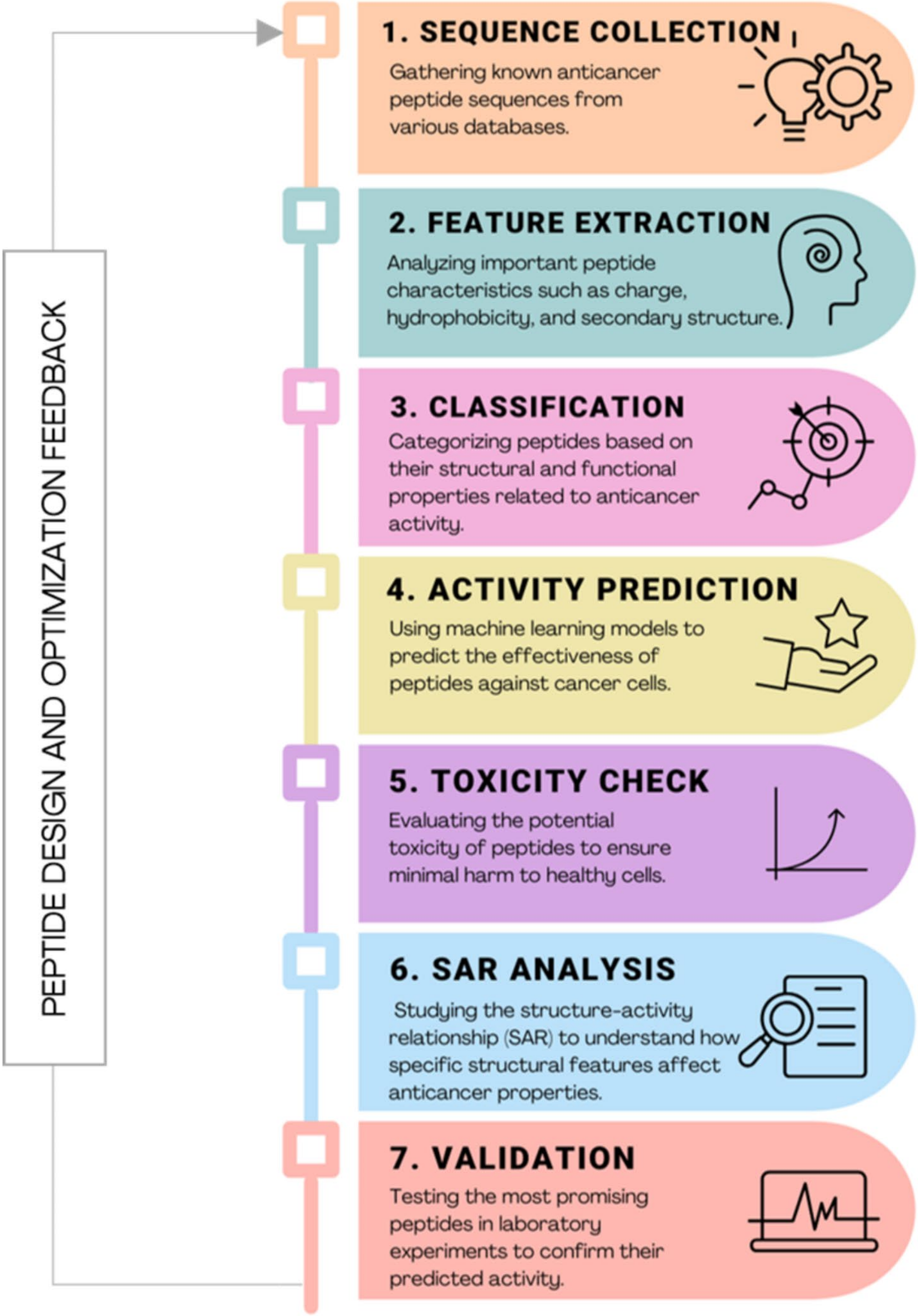
PEPTIDE DESIGN AND OPTIMIZATION FEEDBACK
1. SEQUENCE COLLECTION
Gathering known anticancer peptide sequences from various databases.
2. FEATURE EXTRACTION
Analyzing important peptide characteristics such as charge, hydrophobicity, and secondary structure.
3. CLASSIFICATION
Categorizing peptides based on their structural and functional properties related to anticancer activity.
4. ACTIVITY PREDICTION
Using machine learning models to predict the effectiveness of peptides against cancer cells.
5. TOXICITY CHECK
Evaluating the potential toxicity of peptides to ensure minimal harm to healthy cells.
6. SAR ANALYSIS
Studying the structure-activity relationship (SAR) to understand how specific structural features affect anticancer properties.
7. VALIDATION
Testing the most promising peptides in laboratory experiments to confirm their predicted activity.

Fig. 5 Schematic representation of the ML-driven peptide design and optimization process, illustrating key steps including data preprocessing, feature extraction, model training, validation, and the iterative feedback cycle for optimized peptide sequence generation

ACP prediction. For instance, AntiCP 2.0 achieved an MCC of 0.80 and an area under the receiver operating characteristic curve (AUROC) of 0.97, demonstrating its high predictive accuracy (Agrawal et al. 2020). Likewise, iACP-DRLF has been highlighted for its strong performance, achieving 94% accuracy by focusing on DL-based feature extraction (Lv et al. 2021).

DL models, particularly CNNs, GAN, and long short-term memory (LSTM) networks, have significantly outperformed other models. Examples of two architectures, CNN and GAN, are illustrated in Fig. 6a and b. These models are particularly effective in analyzing peptide structures and can significantly contribute to accurate classification and functional annotation. Similarly, Fig. 7 illustrates a schematic representation of one of the most well-known RNN architectures, the LSTM network, which is widely used for sequence classification tasks. This architecture offers unique capabilities, making it particularly effective in capturing and analyzing sequential patterns. For example, Chen et al. (2021a, b, c) introduced the xDeep-AcPEP model, which utilizes CNN with multitask learning to predict ACP activity across multiple tumor types. This model demonstrated higher accuracy and specificity compared to earlier models (Chen et al. 2021a, b, c). Similarly, Tao et al. (2023) proposed an augmented sample selection framework (ACPs-ASSF) to further enhance prediction accuracy by focusing on high-confidence samples (Tao et al. 2023). Moreover, Zhao et al. (2021) presented a hybrid model that incorporates 3D structural data to predict ACPs with high efficacy and low toxicity, significantly advancing the clinical application of ACPs (Zhao et al. 2021). Additionally, deep graphical representation techniques, as utilized by Yao et al. (2023), have shown great promise in identifying ACPs with improved robustness and generalizability (Yao et al. 2023).

3.3 Automated Design of Anticancer peptides

The physicochemical properties of ACPs are pivotal in determining their efficacy and mechanism of action. Key attributes such as hydrophobicity, charge distribution, amphipathic structures, and molecular weight significantly influence peptide interactions with cellular membranes. Hydrophobicity, for example, enhances a peptide's ability to integrate into lipid bilayers, which is crucial for its cytotoxic activity. Recent research by Neuhaus et al. (2023) highlights the importance of hydrophobic interactions in enhancing the anticancer potency of peptides, suggesting that optimizing these properties can lead to more effective therapeutic agents (Neuhaus et al. 2023). Moreover, the charge distribution of ACPs affects their binding affinity and selectivity towards cancer cells. For instance, cationic peptides tend to preferentially bind to the negatively charged membranes of cancer cells, leading to increased internalization and cytotoxicity. A study by Huang et al. (2021) used advanced computational models to show that optimizing charge and hydrophobicity in peptide design could enhance their therapeutic index against cancerous tissues while minimizing effects on healthy cells (Huang et al. 2021).

Amphipathic structures, which are characterized by distinct hydrophobic and hydrophilic regions, further augment the membrane-disruptive capabilities of ACPs. These structures facilitate the alignment of peptides with lipid bilayers, enhancing their penetration and subsequent therapeutic effects. Recent advancements in computational modeling have provided insights into how these structural properties correlate with bioactivity, supporting the rational design of novel peptides.

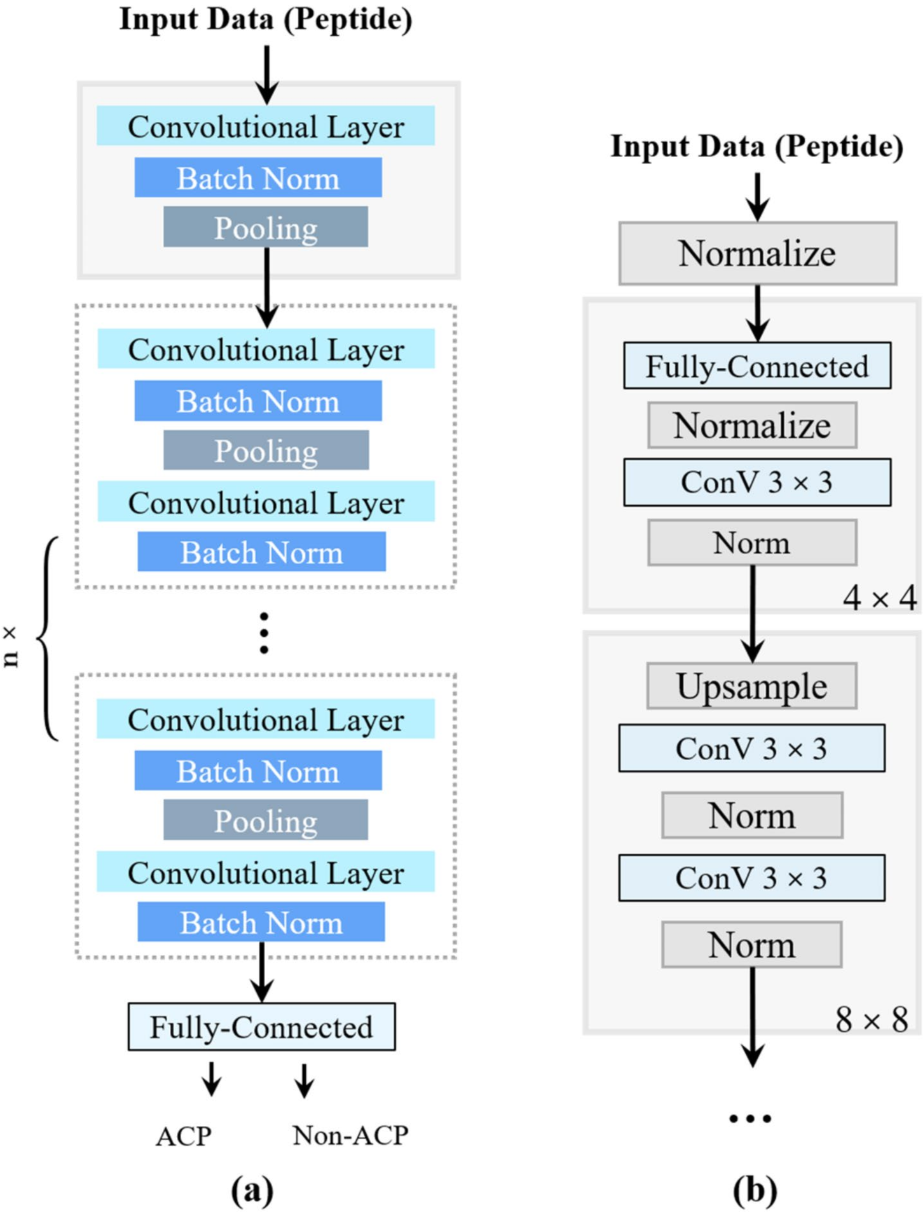

Fig. 6 Architectures of deep learning models used in peptide structure analysis. (a) simple 2D-CNN with multiple layers and (b) GAN are illustrated, showcasing their effectiveness in accurate classification and functional annotation of peptide sequences

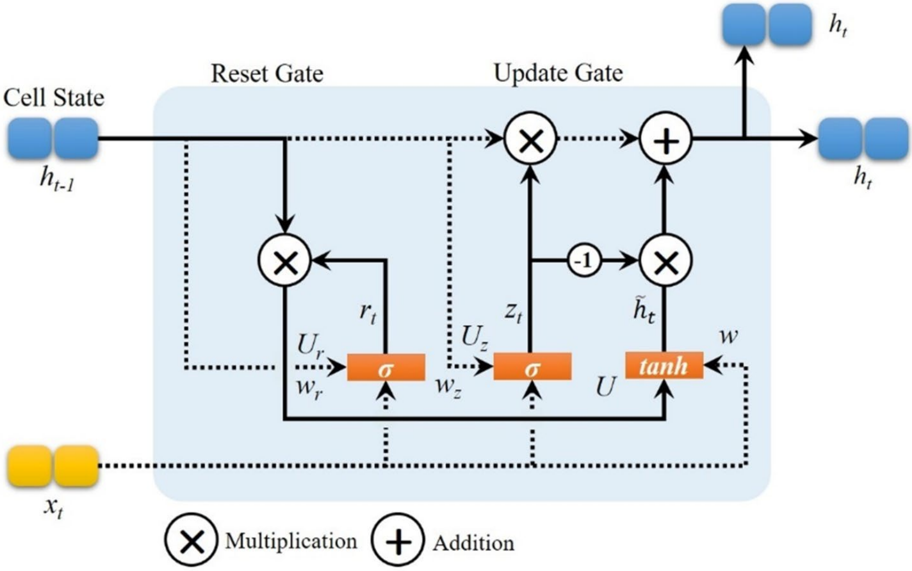

Fig. 7 Schematic representation of the LSTM architecture, a widely recognized RNN model known for its effectiveness in sequence classification tasks and its ability to capture complex temporal dependencies

3.3.1 Sequence patterns

The automated design of ACPs relies on precise preprocessing and Feature engineering of peptide sequences to enable robust computational modeling. Notably, Feature engineering refers to the process of transforming raw data into meaningful input features that enhance the performance of machine learning models. In the context of ACP prediction, this involves encoding peptide sequences into numerical representations, extracting relevant structural and physicochemical properties, and standardizing sequence lengths. Effective feature engineering is critical, as it allows models to identify patterns and relationships that are key to accurately predicting peptide activity.

As shown in Fig. 8, this process involves several key steps that ensure the input sequences are properly encoded, standardized, and enriched with meaningful features. These steps, including numerical representation, sequence length standardization, data normalization, feature matrix construction, and feature extraction, collectively form the foundation for machine learning or deep learning-based ACP design workflows. By systematically transforming raw peptide sequences into machine-readable formats, these steps facilitate the identification of functional patterns and properties critical to anticancer activity.

The structural properties of ACPs, particularly their secondary and tertiary conformations, play a significant role in their biological function. Features such as α-helices and β-sheets are common among effective ACPs and contribute to their mechanism of action. Research by Torres et al. (2020) indicates that the helicity of peptides can enhance their interaction with lipid membranes, thereby improving their anticancer activity (Torres et al. 2020). Moreover, the study of hydrophobic moments and amphiphilicity has shown that these structural features correlate strongly with the efficacy of ACPs. Dennison et al. (2021)

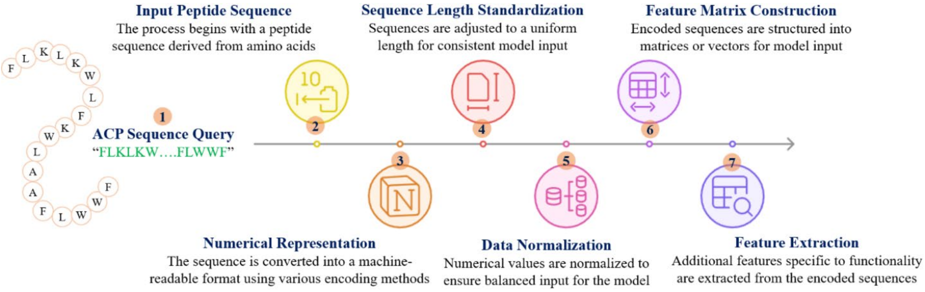
Input Peptide Sequence
The process begins with a peptide sequence derived from amino acids
Sequence Length Standardization
Sequences are adjusted to a uniform length for consistent model input
Feature Matrix Construction
Encoded sequences are structured into matrices or vectors for model input
1
ACP Sequence Query
"FLKLKW....FLWWF"
2
3
4
5
6
7
Numerical Representation
The sequence is converted into a machine-readable format using various encoding methods
Data Normalization
Numerical values are normalized to ensure balanced input for the model
Feature Extraction
Additional features specific to functionality are extracted from the encoded sequences

Fig. 8 Workflow for preprocessing peptide sequences in automated ACP design. Steps include input sequence handling, numerical encoding, length standardization, normalization, feature matrix construction, and feature extraction to prepare data for computational modeling

analyzed a database of ACPs and found that amphipathic properties significantly enhance membrane interaction and selectivity towards cancer cells, suggesting that careful consideration of structural characteristics is crucial in peptide design (Dennison et al. 2021).

The importance of structural stability in ACPs cannot be overstated, as it directly influences their therapeutic potential. Research by Mizejewski et al. (2021) demonstrated that the stability of peptide structures can be optimized through strategic amino acid substitutions, resulting in peptides that maintain their bioactivity while minimizing toxicity to healthy cells (Mizejewski et al. 2021).

3.3.2 Encoders

Fifty-five distinct protein sequence encoders have been developed to transform amino acid sequences into statistical vectors (Chen et al. 2007; Sokal and Thomson 2006). These encoders can be categorized into fourteen types based on the specific information they capture, including amino acid distribution, gap-based amino acid distribution, amino acid group distribution, autocorrelation, covariance, local-global context awareness, sequence order, binary representations, physicochemical properties, traditional networks, pre-trained deep neural networks, optimized physicochemical properties, substitution matrices, and Fourier transformation-based encoders.

Amino acid distribution encoders, such as Kmer (Bhasin and Raghava 2004), DPC (Bhasin and Raghava 2004), TPC (Bhasin and Raghava 2004), ANF (Shin et al. 2022), EAAC (Karami Fath et al. 2022), EGAAC, and DDE (Chidambaram et al. 2011), measure the frequency or proportion of individual amino acids or groups known as k-mers within a protein sequence. This type of encoder reflects the overall composition of amino acids, providing insights into the relative abundance or scarcity of specific k-mers.

Gap-based amino acid distribution encoders, including CKSAPP (López-Vallejo et al. 2011), Adaptive Skip Dipeptide Composition (ASDC) (Deng et al. 2023a, b), and segment protein sequences into bi-mers with defined gap values, capturing the distribution of these unique bi-mers. The gap value indicates the distance between paired amino acids, which influences the local context-aware representation of the protein sequence. Smaller gap values focus on short-range interactions, while larger values capture long-range associations.

Amino acid group distribution encoders, such as CTDC (Chen et al. 2015), CTDD (Chen et al. 2015), CTDT (Chen et al. 2015), GAAC (Zhou et al. 2018), GDPC (Zhou et al. 2018), GTPC (Zhou et al. 2018), KSCTriad (Zhou et al. 2018), and CTriad (DeVita and Chu 2008), categorize amino acids into groups based on specific physicochemical properties, including hydrophobicity, charge, or polarity. These encoders provide insights into the distribution of amino acid groups, illuminating the overall physicochemical characteristics of the sequence.

Autocorrelation encoders, such as NMBroto (Liu et al. 2017), assess the relatedness between amino acids or k-mers within a sequence by computing correlation coefficients based on their physicochemical properties. This analysis offers valuable information about pairwise interactions and dependencies, facilitating the identification of specific functional motifs. Covariance encoders, including auto-covariance (Liu et al. 2015), auto-crosscovariance (Dong et al. 2009; Liu et al. 2015), and bi-autocovariance (Liu et al. 2015), measure the joint variability of two amino acids or k-mers. Positive covariance suggests that when one amino acid exceeds its mean, the other is likely to do the same, whereas negative covariance indicates an inverse relationship. Unlike correlation encoders, covariance encoders focus solely on the direction of the relationship.

Local-global context-aware protein encoders, such as WSRC-local (Chou 2000), WDRC-global (Chou 2000), and WSRC-local-global (Chou 2000), analyze the composition and transitions of amino acids, providing essential information about distribution and changes across different segments of protein sequences. Sequence order encoders like PAAC (Chou 2001), APAAC (Chou 2004), and QSOrder (Chen et al. 2021a), emphasize both the distribution and order of amino acids based on varying distances, representing different levels of local or global interactions.

Binary encoders (Chen et al. 2008; Chou 2000) typically convert amino acid sequences into statistical vectors composed of 0s and 1s. Physicochemical properties and network-based encoders, such as AAIndex (Korde and Mahender 2012) and AESNN3 (Ng and Jordan 2001), substitute amino acids with pre-computed numerical values. Optimized physicochemical property encoders, like ZScale (Chen et al. 2012), leverage these properties to characterize amino acids, employing strategies such as PC, partial least squares (PLS), and multiple linear regression to filter out less informative properties.

Shanthappa and Melethadathil (2024) explored the evolutionary prediction of tRNA-encoded peptides (tREPs) as a novel approach to ACP design. Using computational tools and molecular docking, they identified peptide candidates with robust binding affinities to key cancer targets, such as the BCL2 protein. This study highlights the innovative use of tRNA-based peptide encoders for generating and screening new ACP candidates, showcasing their promise in expanding the chemical space for targeted cancer therapies.

Traditional network-based encoders, including complex networks and enhanced complex networks, represent protein sequences as graphs, where nodes represent amino acids and edges depict interactions. Fourier transformation-based sequence encoders, such as MappingClass-eiip-fourier and MappingClass-integerfourier, utilize Electron-Ion Interaction Potential values or integers to represent amino acids. By applying Fourier transformation, these encoders aim to uncover hidden patterns and trends within protein sequences. Conversely, substitution matrix-based encoders, like BLOSUM62 (Zhang 2012), generate matrices that score amino acid substitutions based on their frequencies in related protein sequences, providing a measure of similarity, with higher scores indicating a greater likelihood of occurrence in similar proteins.

3.3.3 Feature selection and reduction

Optimizing ML model performance through feature selection and reduction is critical in developing effective predictive models for ACPs. Feature engineering involves identifying the most relevant characteristics that contribute to a model's predictive capabilities. Techniques such as Principal Component Analysis (PCA) and recursive feature elimination (RFE) are employed to reduce dimensionality while retaining important features. A study by Akbar et al. (2021) employed two-level feature selection to improve the identification of ACPs. Their approach utilized K-space amino acid pairs and physicochemical properties to effectively reduce the dataset's complexity, leading to improved model performance (Akbar et al., 2021).

The integration of DL techniques has further enhanced the feature selection process. Lv et al. (2021) utilized deep representation learning features to significantly improve the classification accuracy of ACPs. Their research demonstrates how advanced ML techniques can uncover meaningful features from peptide sequences, thereby optimizing model performance and predictive power (Lv et al. 2021). Effective feature selection and reduction are essential for developing robust ML models for ACP prediction. By employing advanced techniques and focusing on the most impactful features, researchers can enhance the accuracy and reliability of predictive models, ultimately leading to the design of more effective therapeutic peptides.

Different feature selection methods in ML each have their own advantages and disadvantages. PCA reduces the dimensionality of data by transforming features into uncorrelated components, retaining important information, though it may reduce model interpretability (Peper et al. 2002). RFE effectively removes less important features iteratively, improving model performance, but can be computationally expensive for large datasets (Su et al. 2020). Filter methods, such as chi-square tests and correlation coefficients, are fast and simple, ranking features based on their statistical relevance, but they don't account for interactions between features (Kamalov and Thabtah 2017). Embedded methods, like LASSO and Ridge Regression, penalize less impactful features during model training, improving generalization, but may over-penalize and exclude important features (Muthukrishnan and Rohini 2016). DL-based methods, such as autoencoders, automatically discover hidden patterns and extract meaningful features from complex datasets, but require large amounts of data and high computational power (Song and Lu 2017).

3.3.4 Deep learning for ACP Design

In this section, we explore the applications of DL in ACP design, providing case studies and examples that highlight the advantages of these advanced techniques over traditional ML approaches. DL, a subset of ML, plays a pivotal role in ACP design, surpassing traditional ML models by handling complex peptide sequence predictions. DL models automatically extract meaningful patterns from data without relying on predefined features (Yu et al. 2020). DL has emerged as a powerful subset of ML, particularly in the field of ACP design. Traditional ML methods often rely on manual feature extraction and are limited in their ability to model complex biological sequences. In contrast, DL models, such as CNNs and RNNs, can automatically learn intricate patterns and dependencies within peptide sequences. These models significantly improve the accuracy of ACP prediction and generation. Furthermore, generative models like GANs and autoencoders have revolutionized the design of novel peptide sequences, offering new avenues for therapeutic peptide discovery.

Unlike traditional ML, which often requires manual feature extraction, DL can process raw peptide sequences, enabling it to capture more intricate relationships and improve ACP prediction accuracy (Yi et al. 2019a, b). CNNs are widely used in ACP research due to their ability to learn spatial hierarchies of patterns in peptide sequences, enhancing accuracy in peptide activity predictions (Chen et al. 2021a, b, c). RNNs, particularly those employing long short-term memory (LSTM), are effective in learning sequence dependencies in peptides, making them suitable for tasks like predicting peptide activity over time (Yu et al. 2020).

Generative models like GANs and autoencoders can design novel peptide sequences by learning from existing datasets, offering a promising avenue for discovering new ACPs (Grisoni et al. 2018). The DeepACP tool, using RNNs with LSTM, outperformed several traditional ML methods in accurately identifying ACPs, showcasing the advantages of DL in ACP research (Yu et al. 2020). xDeep-AcPEP, a CNN-based DL model, successfully predicted the biological activity of peptides against various tumor cells, demonstrating its utility in ACP research (Chen et al. 2021a, b, c).

Autoencoders have been used to generate new peptide sequences that show high anticancer activity, providing an efficient method for discovering novel therapeutic agents (Grisoni et al. 2018). Deep Forest architecture, which combines DL with graphical representation, has been applied to predict ACPs, demonstrating state-of-the-art performance on several datasets (Yao et al. 2023). Generative models have facilitated the design of new peptide sequences with enhanced selectivity against cancer cells, reducing toxicity and increasing efficacy (Grisoni et al. 2018).

ACP-DL, an LSTM-based framework, effectively predicts ACPs by leveraging high-efficiency feature representation methods (Yi et al. 2019a, b). CNNs, by extracting hierarchical features, enable better recognition of important patterns in peptide sequences, significantly improving ACP prediction (Chen et al. 2021a, b, c). BiLSTM networks, by learning bidirectional dependencies in peptide sequences, enhance the precision of ACP identification, making them a popular choice for ACP research (Yu et al. 2020). DL frameworks, such as TensorFlow and PyTorch, are extensively used in peptide research to develop and train models for ACP identification and generation, providing flexible and scalable solutions for researchers (Yi et al. 2019a, b). DL models, through CNNs, RNNs, and generative approaches, are revolutionizing ACP design by offering higher accuracy and enabling the creation of novel peptides, outperforming traditional ML models (Grisoni et al. 2018).

Lee and Shin (2024) discussed the effectiveness of contrastive learning by enhancing model performance through better representation of peptide features. Their model achieved superior results on five of six benchmark datasets compared to previous state-of-the-art approaches. By leveraging dual encoders as an alternative to data augmentation, the framework improved feature extraction efficiency and accuracy, highlighting the potential of contrastive learning to advance ACP discovery and streamline the identification of therapeutic candidates.

Kumar and Singh (2024) proposed a stacking ensemble framework that employs multiview learning and decision fusion to predict ACPs. By combining multiple 1D-CNN with a logistic regression meta-classifier, their method achieved a high classification accuracy of 95.52% and an AUC score of 0.971. This approach demonstrates how multiview learning can enhance the representation and integration of peptide sequence features, leading to improved prediction performance in ACP discovery.

Besides, Transformer models have emerged as powerful tools for ACP classification due to their ability to capture complex relationships within peptide sequences. Kilimci and Yalcin (2024) introduced ACP-ESM, a transformer-based framework that achieved a remarkable accuracy of 97.66% in ACP classification. By leveraging protein-oriented transformer models like ESM and integrating sequence-based features, the framework demonstrated state-of-the-art performance in identifying ACPs. This advancement underscores the potential of transformers to enhance precision in ACP discovery, making them a valuable addition to the computational toolkit for cancer peptide research.

3.3.5 Classifiers

In the prediction of ACPs based on classification problems, traditional ML classifiers are employed to construct two distinct predictive pipelines: binary classification and multi-label classification. In the binary classification pipeline, we evaluated the performance of the proposed CARE encoder using twelve widely recognized ML classifiers, including K-Nearest Neighbor (KNN) (Li et al. 2020), Support Vector Machine (SVM) (Deng et al. 2023a, b), Logistic Regression (LR) (Li et al. 2020), Extreme Gradient Boosting (XGB) (Deng et al. 2023a, b), Gradient Boosting (GB) (Deng et al. 2023a, b), Adaboost (AB) (Deng et al. 2023a, b), Decision Tree (DT) (Deng et al. 2023a, b), Bagging (BG) classifier (Deng et al. 2023a, b), Random Forest (RF) (Deng et al. 2023a, b), Extra Trees (ET) (Deng et al. 2023a, b), Gaussian Process (GP) (Deng et al. 2023a, b), and Naive Bayes (NB) (Deng et al. 2023a, b).

Figure 9 provides a comprehensive overview of peptide classification techniques, organized into two main categories: “Machine Learning Classifiers” and “Deep Learning Methods.” While numerous approaches are presented, certain techniques have emerged as particularly popular due to their superior accuracy and robustness in peptide-related applications. Among traditional machine learning methods, SVM, Random Forest, and Gradient Boosting Machines are the most commonly utilized. In the domain of deep learning, CNNs, LSTMs, and Transformers dominate as the preferred choices. These advanced methods have been widely adopted by researchers to achieve accurate peptide predictions and innovative designs, leveraging their exceptional ability to identify and model complex, nonlinear patterns in biological data.

Moreover, for multi-label classification, studies employed the proposed CARE encoder alongside the label powerset method and the Adaboost classifier to annotate the functional types of ACPs. As binary classifiers are not designed to handle multi-label peptide sequences, the label powerset method was applied as a data transformation strategy, converting each unique combination of functional types into a distinct class. Using the CARE encoder, discriminative statistical vectors were derived from raw sequences, and multi-label sequences were transformed into unique class sequences via the label powerset method. The Adaboost classifier was then utilized to perform functional type annotations on the transformed data.

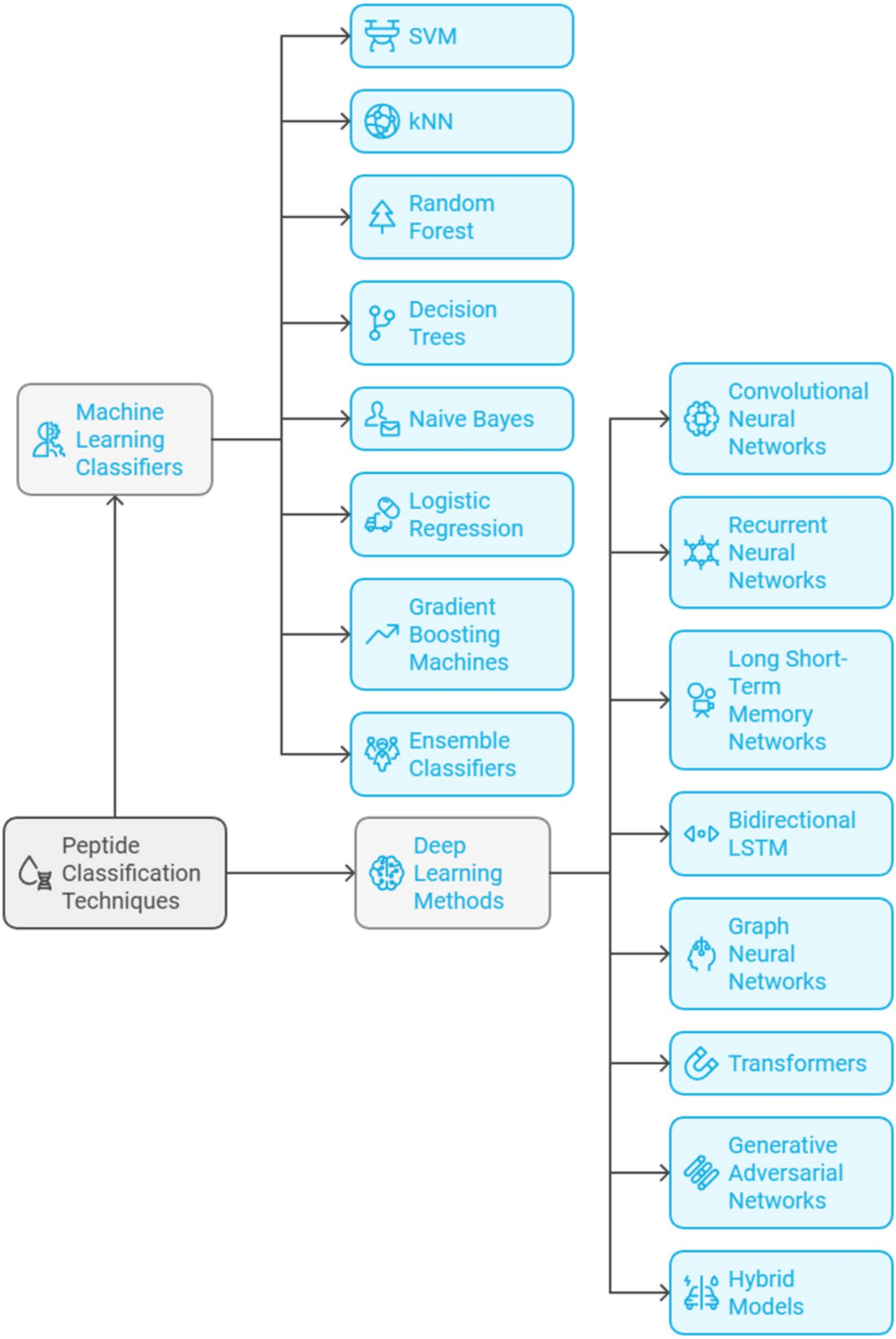
SVM
kNN
Random Forest
Decision Trees
Naive Bayes
Logistic Regression
Gradient Boosting Machines
Ensemble Classifiers
Machine Learning Classifiers
Peptide Classification Techniques
Deep Learning Methods
Convolutional Neural Networks
Recurrent Neural Networks
Long Short-Term Memory Networks
Bidirectional LSTM
Graph Neural Networks
Transformers
Generative Adversarial Networks
Hybrid Models

Fig. 9 A comprehensive classification of ML and DL methods for peptide analysis and design

4 Recent advances and comparison

Computational methods for ACP prediction have advanced significantly in recent years, with improvements in accuracy and efficiency driven by access to larger datasets and more powerful ML techniques. This section delves into recent methodologies, emphasizing innovations in feature encoding, algorithmic approaches, and overall model performance, and compares various strategies to better understand the evolution of ACP prediction and the trends shaping the field.

4.1 Protein Language models

Recent advancements in protein language models (PLMs) have dramatically transformed peptide design by allowing the prediction of peptide properties and optimizing sequences without explicit structural data. Models like PeptideBERT, based on transformers, have demonstrated substantial success in predicting key properties such as hemolysis, solubility, and non-fouling characteristics. PeptideBERT utilizes a pretrained model (ProtBERT) to offer precise predictions, enabling peptide design that avoids the challenges of experimental validation (Guntuboina et al. 2023). These models represent a breakthrough in peptide prediction, bypassing the need for direct structural characterization, which has traditionally hindered the peptide design process.

Protein language models have significantly enhanced the prediction of peptide properties, making them a powerful tool in early-stage peptide design. PeptideGPT extends these capabilities by generating novel peptide sequences with desired functional properties, such as enhanced solubility and specific activity against certain pathogens (Pham et al., 2024). These models learn from vast peptide sequence datasets, capturing relationships between sequence patterns and biological activities, thus enabling high-accuracy predictions. For instance, PeptideBERT excels at predicting hemolytic activity, which is crucial for identifying peptides that may induce red blood cell damage—a common issue in peptide drug development (Guntuboina et al. 2023).

One of the most promising aspects of protein language models is their ability to generate novel peptide sequences with specific properties. ProtGPT2, for instance, generates peptide sequences that align closely with natural protein patterns while maintaining their functional relevance (Pham et al., 2024). By exploring the latent space of protein sequences, these models provide a more robust and efficient way to generate peptides with particular structural and functional properties. This capability is especially critical in drug discovery, where creating novel peptides with optimized efficacy and reduced side effects is vital. Moreover, these models significantly improve the speed of peptide discovery by reducing the dependency on experimental methods for initial peptide screening.

Recent research has focused on enhancing protein language models by integrating structural data, which allows for more precise predictions of peptide stability and interaction potential. Models like AlphaFold have integrated sequence and structure prediction, showing excellent performance in predicting peptide conformation and stability (Motmaen et al., 2023). This integration is critical because it enables the design of peptides that not only exhibit the desired biological activity but also possess stable, therapeutically viable structures. The combination of structural insights with sequence prediction enhances the overall design process, making it more efficient and aligned with real-world applications.

The future of peptide design is closely tied to the continued development of protein language models. As these models evolve, integrating more structural and functional data, they are expected to streamline the design process, allowing for the rapid identification and optimization of peptides for a variety of therapeutic applications. By reducing the reliance on experimental methods, PLMs will drastically cut down the time and cost associated with peptide drug discovery. Furthermore, their ability to generate novel sequences and predict properties will enable the creation of peptides tailored for specific diseases and therapeutic needs, fostering advancements in personalized medicine (Liang et al. 2024a, b). These advancements could revolutionize the design of peptide-based therapies, offering precision and efficiency previously unattainable.

4.2 Key advances in ACP Prediction

Significant strides have been made in the development of ACP prediction models in recent years, fueled by advancements in ML and DL methodologies. By leveraging large peptide sequence datasets, innovative feature extraction techniques, and cutting-edge algorithms, these models enhance accuracy and uncover new therapeutic opportunities. Table 2 provides an overview of recent models that utilize diverse approaches, such as capsule networks, ensemble learning, and latent-space encoding, highlighting the key methodologies, datasets, and performance metrics.

In recent years, various ML and DL approaches for ACP prediction have yielded significant successes. DL methods like the LightGBM and GAT models introduced by Zhong and Deng (2024) have achieved an impressive accuracy of 92% and an AUC of 98.3%, highlighting the high predictive power of these models. By leveraging attention networks and focusing on protein features, these methods provide highly precise predictions for ACP classification. Similarly, the GAN and LSTM approach by Zhixing et al. (2024), which achieved over 90% accuracy, efficiently discovers new peptides. These methods excel due to their ability to handle large and complex datasets, making them particularly effective for cancer therapies, including breast and lung cancers.

However, some approaches face greater challenges. Traditional models like Naive Bayes and Random Forest, as used by Azad et al. (2024), despite achieving a high internal accuracy of 99%, struggle with overfitting when applied to external data due to the complexity of hybrid features. Additionally, models utilizing multiple feature extraction techniques, such as Danish et al. (2024), which applied SMOTE for dataset balancing, achieved an accuracy of 97.56% but encountered challenges in the complexity of integrating various methods. Overfitting occurs when a machine learning or deep learning model learns the training data too well, including its noise and outliers, leading to poor generalization on unseen or test data. This problem is especially prevalent in models trained on small or imbalanced datasets, where the model captures specific details of the training data that do not generalize to broader patterns. To mitigate overfitting, techniques such as regularization, dropout, cross-validation, and data augmentation are commonly employed in ACP prediction tasks.

While these models demonstrate high accuracy, they often require more computational resources and face difficulties in practical implementation due to their complexity. Models utilizing GANs (Generative Adversarial Networks), such as the approach by Zhixing et al. (2024), have demonstrated strong potential in ACP discovery.

GANs work by generating synthetic peptide sequences through adversarial training between a generator and a discriminator, allowing for the exploration of novel ACPs beyond traditional datasets. This method's key advantage is its ability to generate diverse and new peptide sequences, which traditional models may not discover, making it highly effective for peptide-based cancer therapies like those for breast and lung cancer. GANs offer the flexibility to identify new peptide structures, enhancing the overall efficiency and scope of ACP discovery. In contrast, ensemble learning methods, such as those employed by Liu et al. (2024) using PCA and SHAP (SHapley Additive exPlanations) models, rely on combining multiple learning algorithms to improve prediction accuracy and robustness. Ensemble methods are particularly useful in handling varied data and mitigating overfitting by integrating outputs from several models. However, while ensemble models provide high accuracy and explainability, they often lack the generative capabilities of DL approaches like GANs. Compared to DL, ensemble models are less flexible in discovering novel peptide sequences, but they excel in interpretability and handling large, complex datasets through ensemble-driven predictions. This makes ensemble methods highly reliable for existing ACP classification, whereas DL models, especially GANs, are better suited for discovering entirely new ACPs.

Moreover, Yao et al. (2024) introduced ACP-CapsPred, a framework that leverages capsule networks to not only classify ACPs but also predict their functional activities across cancer types with high accuracy (95.71%) and F1-scores (95.90%). Unlike traditional methods, capsule networks capture hierarchical relationships within peptide features and offer interpretability by identifying peptide regions responsible for anticancer activity.

As illustrated in the Table 2, various studies have utilized curated datasets such as CancerPPD, benchmark datasets, and independent collections. For instance, Yao et al. (2024) and Zhixing et al. (2024) utilized the CancerPPD database, which is a widely recognized resource for experimentally validated anticancer peptides. These datasets provide robust foundations for predictive models by ensuring that the training data is biologically relevant and well-annotated.

However, certain limitations remain. For example, while datasets like those used by Xu et al. (2024a, b; Song et al. (2024) include sufficient examples for training (861 ACPs and balanced classes), the representation of diverse peptide sequences across cancer types and functionalities is still limited. To mitigate this, Liang et al. (2024a, b) employed multimodal feature fusion and integrated multiple benchmark datasets, enabling models to learn from a broader range of peptide properties. Khan (2024) and Balaji et al. (2024) expanded their training data by combining peptides from databases such as PubChem, enriching the data diversity.

Data imbalance is a recurring challenge in ACP prediction, where the number of positive samples (ACPs) is often dwarfed by negative examples. This imbalance can bias models toward predicting non-ACPs. To address this issue, methods such as SMOTE (Synthetic Minority Oversampling Technique) have been applied. For example, Danish et al. (2024) specifically utilized SMOTE to create a balanced dataset, improving the model's ability to generalize across both ACP and non-ACP classes.

In addition, overfitting remains a concern, particularly with models relying on deep architectures such as CNNs and Transformers. Studies like those by Lee and Shin (2024) and Chen et al. (2024) implemented multitask learning and ensemble methods to enhance generalization and reduce overfitting. Furthermore, Niu et al. (2024) combined hybrid DL

Table 2 Overview of recent ML and DL methodologies used in ACP prediction

Authors (Year)	Methodology	Dataset	Cancer Type	Accuracy	Key Advantages	Limitations	Effectiveness
Azad et al. (2024)	ML (NB, RF, SVM, DT)	368 ACPs, 414 non-ACPs	Gastric	0.99 (Internal), 0.94 (External)	Speed, high accuracy	Hybrid feature complexity, overfitting	Valuable for gastric cancer therapy
Wang and Ma (2024)	ML (Dual embedding)	736 ACPs, 736 non-ACPs	Multiple tissue types	82.8%, AUC 89.5%	Strong generalization	Limited data size, overfitting	Advances in peptide discovery
Zhong and Deng (2024)	ML (LightGBM, GAT)	701 peptides	Multiple cancers	92%, AUC 98.3%	High accuracy, specificity	Complexity of feature integration	Significant advancement in ACP discovery
Yao et al. (2024)	Capsule networks	CancerPPD database	Multiple cancers	80.25%, 95.71%	Interpretability	Feature extraction complexity	Precise and interpretable
Liang et al. (2024ab	Multimodal feature fusion (Attention mechanism)	Multiple benchmark datasets	Various cancer cells	80.81%, 93.56%	Comprehensive peptide representation	Model complexity	Functional analysis of ACPs
Arif et al. (2024)	ML (PLM, Wavelet denoising)	861 ACPs, 861 non-ACPs	Breast, Lung, Melanoma	96.6% (ACP-main), 99% (ACPAlter), 98.3%	High accuracy, robustness	Computational complexity	Rapid ACP discovery
Danish et al. (2024)	Multiple feature extraction, SMOTE	Balanced dataset with SMOTE	Various cancers	97.56% (Benchmark), 95% (Independent)	Improved feature representation	Combining various feature methods	New peptide-based therapies
Khan (2024)	Deep latent-space encoding	ACP344, ACP740	Various cancers	94.18%, 95.1%	High classification accuracy	Model training complexity	Selective targeting of cancer cells
Karim et al. (2024)	Probabilistic feature fusion	2390 peptides	Various cancers	93.72%	High accuracy, robustness	Multiple encoding schemes complexity	Discovery of new ACPs
Yue et al. (2024)	DL (CNN, RNN, Bi-LSTM)	1786 T1DM, 756 T2DM peptides	Type 1 and Type 2 Diabetes	90.48%	High precision, efficiency	Computational resources	Antidiabetic peptide discovery
Song et al. (2024)	ML (BERT, BiGRU)	ACP606, ACP740, ACP240	Various cancers	85.25%, AUC 97.52%	Enhanced feature learning	Training multiple architectures	Peptide therapy advancement
Kaur et al. (2024)	Hybrid approach (Logistic regression, motif analysis)	1174 hormones, 1174 non-hormones	Hormonal diseases	AUROC 0.96, Accuracy 89.79%	High prediction accuracy	Model integration complexity	Peptide hormone therapy

Table 2 (continued)

Authors (Year)	Methodology	Dataset	Cancer Type	Accuracy	Key Advantages	Limitations	Effectiveness
Lee and Shin (2024)	DL (CNN, Transformer)	ACP-Mixed-80, 606–3210 samples	Breast cancer	Accuracy 85.25%	Enhanced feature learning	Training complexity	Breast cancer therapy
Chen et al. (2024)	DL (CNN, multitask learning)	CancerPPD, 592 peptides	Breast, Colon, Cervix, Lung, Skin, Prostate	MSE 0.1758, PCC 0.8086	Improved generalization	Small dataset	Multiple cancer therapies
Liu et al. (2024)	Ensemble learning (PCA, SHAP)	ACPfel, 4754 sequences	Breast, Lung	Accuracy 98.53%, AUC 0.9972	High accuracy, explainability	Model complexity	Targeted cancer therapies
Xu et al. (2024a, b)	Deep representation learning (BERT, BiLSTM)	861 ACPs, 861 non-ACPs	Breast, Lung, Colon, Skin	Accuracy 94.43%	High accuracy, no manual feature extraction	DL complexity	Advances ACP discovery
Zhixing et al. (2024)	GANs, LSTMs	CancerPPD	Breast, Lung	Exceeding 90%	Efficient discovery of ACPs	Model interpretability	Peptide-based cancer therapies
Karakaya and Kilimci (2024)	DL (FastText, BiLSTM)	ACPs250, Independent dataset	Breast, Lung, Colon, Prostate	92.50% (ACPs250), 96.15% (Independent)	High accuracy, semantic relationships	Hyperparameter tuning, costs	Peptide-based cancer therapies
Niu et al. (2024)	Hybrid DL (BERT, CNN)	ACP1, ACP2 datasets	Breast, Lung, Colon	AUC 0.9726, F1 0.9371	High classification performance	DL fine-tuning	Cancer therapy development
Balaji et al. (2024)	ML (LightGBM, RF, KNN)	10,000 compounds from PubChem	Breast, Lung, Colon	79%, AUC 0.88	High predictive performance	Model complexity, computational resources	Discovery of anticancer small molecules
Khawaja et al. (2024)	DL (CNN, MLP, GRU)	Leukemia sequences from UniProt	Leukemia	98.33%	High accuracy, handle complex patterns	Computational demands	Early leukemia detection
Xu et al. (2024a, b)	NLP (BERT, M3E)+ML	859 ACPs	Breast, Colon	93.85%, AUC 0.97	Capture semantic relationships	Model integration complexity	Advances in ACP discovery

techniques like BERT and CNN to achieve robust representation learning while mitigating overfitting risks through regularization techniques.

The methodologies summarized in the Table 2 reflect a growing emphasis on improving dataset quality and model robustness. Approaches like probabilistic feature fusion (Karim et al. 2024) and wavelet denoising (Arif et al. 2024) illustrate innovative ways to enhance data preprocessing and model accuracy. Moreover, advanced representation learning methods, such as deep latent-space encoding (Khan 2024) and GANs for synthetic data generation (Zhixing et al. 2024), showcase how models can address inherent limitations in dataset size and diversity.

Collectively, these efforts underline the critical role of dataset quality and diversity in shaping the performance of ML and DL models in ACP prediction. The integration of advanced data augmentation techniques, balanced datasets, and innovative feature engineering methods provides a clear pathway for addressing the challenges highlighted by the reviewer, ensuring models are accurate, generalizable, and effective in identifying anticancer peptides.

DCTPep is a comprehensive and open-access database for cancer therapy peptides, recently published by Sun et al. (2024) This dataset includes 6,214 entries covering traditional ACPs, cancer-targeting peptides, and clinically approved or investigational peptide drugs. The data were meticulously collected from research articles, patents, and other peptide databases, providing valuable insights into peptide-based cancer therapeutics. DCTPep stands out by offering detailed target annotations and covering a broader range of peptides compared to existing databases, making it a critical resource for designing and developing novel cancer therapy peptides.

4.3 Comparison of ACP Prediction methods

The prediction of ACP has become a focal point in bioinformatics due to its potential in therapeutic developments. As the complexity of peptides and the need for precision in prediction models increase, various ML and DL methods have been explored. Each model leverages different classifiers and feature encoding techniques to enhance accuracy and generalizability. Table 3 illustrates a comparison of recent ACP prediction methods, highlighting the methodologies, classifiers, and feature encoders employed in these studies. This comparative overview provides valuable insight into the diverse approaches undertaken in the field, setting the stage for a deeper analysis of their effectiveness and potential.

One of the most promising approaches identified is the voting-based ensemble learning method utilized in the ACP-ML model proposed by Bian et al. (2024). This method achieves a high level of accuracy, ranging from 90.89 to 92.57%, as it combines multiple feature extraction techniques such as CS-Pse-PSSM and CTDT, followed by rigorous feature selection processes. By aggregating the results from multiple classifiers, the model not only improves predictive accuracy but also enhances generalization across different datasets (Bian et al. 2024). This demonstrates the importance of ensemble techniques in handling the complexity of peptide data and maximizing prediction reliability.

Another innovative model is the GRDF model developed by Yao et al. (2023), which employs a deep forest architecture for ACP prediction. This model stands out for its integration of evolutionary information and graphical features of peptides, resulting in an accuracy of approximately 90.47%. The deep forest classifier, with its ability to capture hierarchical

Table 3 Comparative overview of ACP prediction methods, showcasing the variety of ML and DL classifiers, alongside the feature encoding techniques employed in each study

Ref.	Method	Classifier	Encoder of Features
Bian et al. (2024)	ACP-ML	Voting-based ensemble learning	CS-Pse-PSSM, CTDT, CTDC, PAAC, and DPC
Karim et al. (2024)	ANNprob-ACPs	Artificial neural network	CKSAAGP, word2vector, CTDT, CTDC, Quasi-sequence-order, PAAC, DPC, Cross-covariance, and AAC
Ghafoor et al. (2024)	CAPTURE	Adaboost	Transitional information, compositional, distributional, and Correlational
Deng et al. (2023a, b)	ACP-MLC	RF	C/T/D, AAINEX, TPC, DDE, BPF, and AAC
Yao et al. (2023)	GRDF	Deep forest	BP, Evolutionary information, Graphical features of peptides
Akbar et al. (2022)	cACP-DeepGram	Deep neural network	Word embedding
Arif et al. (2022)	StackACPred	SVM-RFE + CBR, LightGMB, stacking-based ensemble learning	PAAC, PsePSSM, and N-SegPSSM
Liang and Ma (2023)	iACP-GE	GBDT, ET	DMACA, EGAAC, BLOSUM62, and BC
Ahmed et al. (2021)	ACP-MHCNN	CNN	Evolutionary information, PHYC, and Sequential features
Akbar et al. (2020b)	cACP-2LFS	SVM	K-space amino acid pair
Akbar et al. (2020a)	cACP	SVM	Geary autocorrelation descriptor, Conjoint triad feature, and Quasi-sequence order
Yu et al. (2020)	DeepACP	RNN	Amino acid embedding
Agrawal et al. (2020)	AntiCP 2.0	SVM	BP, TC, DPC, and AAC
Schaduangrat et al. (2019)	ACPred	SVM, RF	Am-PAAC, PAAC, PHYC, DPC, and AAC
Akbar et al. (2017)	iACP-GAEnsC	Genetic algorithm-based ensemble learning	Reduce amino acid alphabet composition, g-Gap dipeptide composition, and Am-PAAC
Tyagi et al. (2013)	AntiCP	SVM	BP and AAC

feature representations, proves to be highly effective for complex biological sequences. This methodology showcases how novel feature integration can significantly boost the performance of ML models in bioinformatics (Yao et al. 2023).

Lastly, the StackACPred model proposed by Arif et al. (2022) combines SVM-RFE with LightGBM in a stacking-based ensemble learning framework. This multi-layered approach leverages various feature descriptors, including PAAC and PsePSSM, to optimize the prediction of ACPs. Although the specific accuracy for this model is not provided, its use of feature selection and stacking techniques underscores its sophistication in processing peptide data. This model exemplifies the shift towards more complex, multi-stage algorithms designed to improve both specificity and sensitivity in ACP prediction, emphasizing the ongoing evolution of predictive modeling in this domain.

4.4 Feature importance and selection techniques

Selecting informative and relevant features is crucial for improving the predictive accuracy of ML and DL models in ACP design. Effective feature selection reduces data dimensionality, eliminates redundancy, and prevents models from being overwhelmed by irrelevant information. This is particularly important in ACP prediction, where biological data can be complex and multifaceted. Techniques like RFE, PCA, and modern methods such as Representation Learning (RL) help models focus on key patterns, enhancing generalizability and performance.

The increasing sophistication in ACP prediction models also demands methods that can handle both the biological complexity of peptides and the computational efficiency of training these models. Methods like RL automatically extract high-level features without human intervention, a vital capability in scenarios where manual feature extraction may be impractical. On the other hand, traditional techniques like PCA offer interpretability and lower computational costs, making them suitable for simpler datasets. Understanding the trade-offs between these methods is crucial for building robust models that strike a balance between performance, complexity, and interpretability.

The datasets commonly used for predicting ACP typically consist of peptide sequences that are categorized into two groups: ACPs (Positive Class) and non-ACPs (Negative Class). These datasets provide crucial input for ML models aimed at predicting ACPs. Some of the well-known datasets in this domain include:

- **AntiCP Dataset**: This dataset has been utilized in numerous studies for training and evaluating ACP prediction models. It consists of validated ACPs collected from various biological databases, specifically identified for their anticancer properties.
- **CancerPPD Database**: Another key dataset, CancerPPD, compiles information about ACPs from a wide range of experimental sources. It serves as a critical reference for model development and validation in ACP research.

The four key criteria used for comparison in Figs. 10 and 11—interpretability, computational complexity, accuracy improvement, and flexibility—were carefully chosen based on their significance in the practical application of ML and DL models, especially in domains like ACP design. These criteria provide a comprehensive understanding of the strengths and weaknesses of different feature selection techniques. Moreover, five key feature selection methods were chosen for comparison in ACP design due to their distinct strengths. RFE

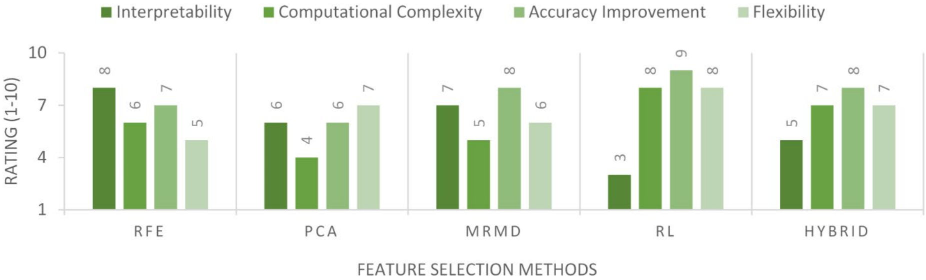

Fig. 10 Qualitative comparison of various feature selection techniques used in anticancer peptide (ACP) prediction models

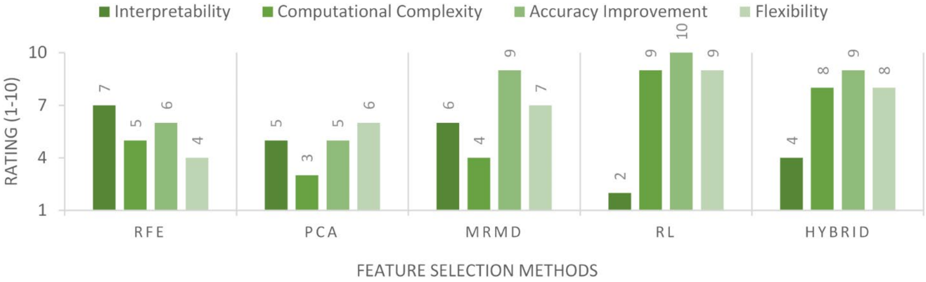

Fig. 11 Comparison of feature selection techniques under high-dimensional data conditions in ACP prediction

and PCA offer simplicity and high interpretability, making them ideal for projects requiring clear feature explanations. RL excels in automatically extracting high-level features from complex datasets, while Maximum Relevance Minimum Redundancy (MRMD) balances dimensionality reduction with feature relevance. Hybrid Methods combine multiple techniques to enhance accuracy and flexibility. This selection highlights the trade-offs between interpretability, accuracy, and computational efficiency in ACP prediction.

Interpretability refers to how easily the outputs of a model can be understood and explained based on the selected features. Simpler methods like RFE and PCA rank highly in interpretability because they allow clear insight into how specific features influence predictions. This makes these techniques particularly useful in contexts where transparency is crucial, such as in medical research or regulatory environments. On the other hand, more complex techniques like RL sacrifice interpretability for greater accuracy and flexibility, making them less accessible for those who need to understand the internal workings of the model.

The Y-axis in both charts represents a “Rating” scale from 1 to 10, where higher values indicate better performance across four key dimensions: Interpretability, Computational Complexity, Accuracy Improvement, and Flexibility. These ratings were assigned based on a qualitative assessment of each feature selection method’s relative strengths in these areas. For example, higher ratings for Accuracy Improvement suggest that the method significantly enhances predictive performance, while lower ratings in Computational Complexity imply that the method is more resource-efficient and quicker to train. This comparative analysis provides a clear visual framework for understanding how each method performs under different scenarios, guiding users to make informed decisions depending on their dataset and application needs.

Computational complexity, another vital criterion, measures the time and resources (CPU/GPU power) required to train the models. Techniques like PCA, which are computationally efficient, score lower on the complexity scale, making them suitable for projects with limited resources. In contrast, methods like RL require significantly more computational power, especially when implemented in deep neural networks. These techniques are designed to process vast amounts of data and learn intricate patterns, but they come with higher demands in terms of hardware and time. Accuracy improvement measures how much the selected features contribute to the overall performance of the model. RL excels in this aspect because it automatically extracts the most relevant features from the data, leading to better predictive performance, especially in large-scale, high-dimensional datasets. Lastly, flexibility refers to a model's ability to generalize across different types of data and adapt to new tasks. RL also stands out in this regard due to its DL architecture, which is versatile enough to handle various forms of data, from peptide sequences to images or textual data, making it particularly valuable in complex, multidisciplinary fields like bioinformatics.

The results illustrated in Figs. 10 and 11 highlight the inherent trade-offs between simplicity and performance across these feature selection techniques. RFE and PCA, known for their simplicity, score higher in interpretability and lower in computational complexity, making them ideal for projects where explainability and speed are priorities. These methods are suitable when computational resources are limited, and the focus is on producing interpretable models, even at the cost of some accuracy and flexibility. For example, RFE is commonly used in support vector machines (SVMs) because it simplifies feature selection while maintaining acceptable levels of accuracy. However, when the data is more complex, as is often the case in ACP design, these simpler methods show limitations in terms of their ability to capture all relevant information.

In contrast, more advanced techniques like RL and Hybrid Methods perform exceptionally well in terms of accuracy improvement and flexibility, particularly in Fig. 11, where the conditions are stricter due to high-dimensional data. These methods are ideal for complex problems such as bioinformatics or large-scale data analysis because they can automatically discover and process complex patterns in the data. However, these benefits come with the trade-off of higher computational complexity and lower interpretability. RL, in particular, requires extensive computational resources and is often used in conjunction with GPUs to speed up processing times. Despite these challenges, its ability to handle complex, high-dimensional data and automatically extract meaningful features makes RL a powerful tool for cutting-edge tasks like ACP prediction. This highlights the ongoing need to balance model complexity with interpretability and resource availability depending on the specific requirements of a project.

4.5 Impact of classifiers on Prediction Accuracy

The choice of classifier plays a pivotal role in determining the accuracy of ACP prediction models. While feature selection helps reduce the complexity and dimensionality of the data, the classifier is responsible for identifying patterns and relationships that can distinguish between anticancer and non-ACPs. Different classifiers have varying strengths depending on the nature of the data and the specific task at hand. From traditional ML approaches like SVM to more advanced DL models like CNN, the selection of the right classifier can significantly impact the model's performance. In this section, we will explore the impact of different classifiers on prediction accuracy, highlighting their benefits, limitations, and ideal use cases for ACP design.

five different classifiers, including SVM, Random Forest, CNN, RNN, and Ensemble Methods, have been chosen to compare their performance in predicting ACP. These classifiers were selected due to their diverse capabilities and applications, as each handle complex biological data, such as peptide sequences, differently. SVM and Random Forest are traditional models known for their simplicity and interpretability, while CNN and RNN are highly popular for processing complex sequential data in DL. Additionally, Ensemble Methods are valuable in complex projects because they combine multiple classifiers to reduce

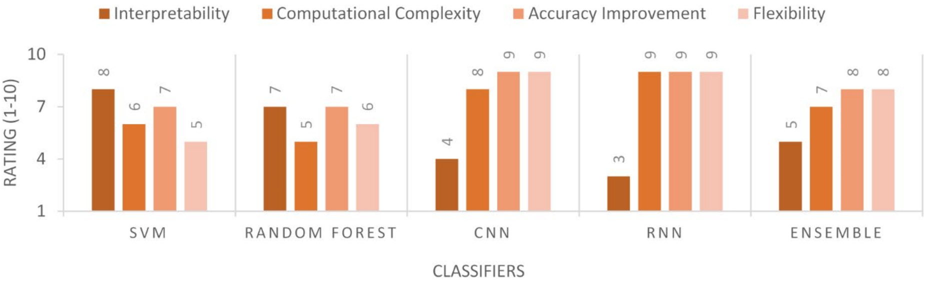

Fig. 12 Qualitative comparison of classifiers used in ACP prediction

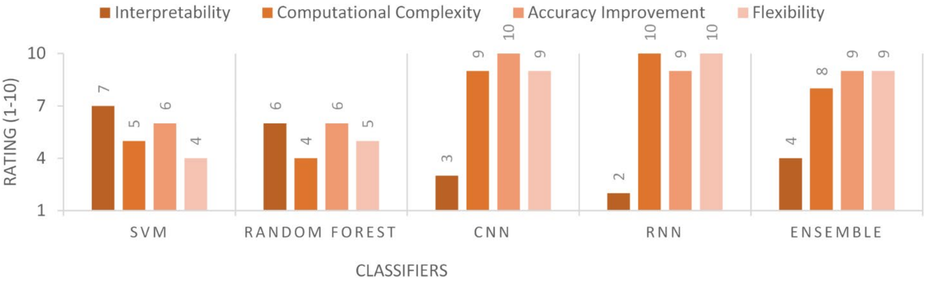

Fig. 13 Performance comparison of classifiers under high-dimensional data conditions in ACP prediction

prediction errors. This selection helps researchers evaluate different classifiers based on criteria such as accuracy, computational complexity, and flexibility.

In Fig. 12, the classifiers are compared based on four key criteria: interpretability, computational complexity, accuracy improvement, and flexibility. Traditional classifiers like SVM and Random Forest score highly in interpretability due to their simplicity and ease of explanation. These models are ideal for projects where transparency is crucial. On the other hand, CNN and RNN, although excelling in accuracy improvement and flexibility, score lower in interpretability due to their complex structures. Their computational complexity is also higher because they require more resources and processing time. Ensemble Methods perform relatively well across all categories but still involve greater complexity compared to traditional models.

It is important to note that the same dataset used in Sect. 4.3. is also employed here to evaluate the performance of these classifiers. The dataset includes peptide sequences and various features such as evolutionary and structural information, which are essential for training and evaluating the models. Thus, the results from this comparison offer valuable insights for researchers aiming to select the best classifier for ACP prediction tasks.

In Fig. 13, the classifiers are evaluated under stricter conditions with more complex and larger datasets. CNN and RNN continue to outperform other classifiers in accuracy improvement and flexibility, demonstrating their strength in learning complex patterns from biological data, such as peptide sequences. However, this comes at the cost of much higher computational complexity, which makes these models resource-intensive. This can be a challenge for projects with limited computational resources.

On the other hand, SVM and Random Forest show weaker performance under these challenging conditions, particularly in terms of accuracy. These classifiers are better suited for simpler, smaller datasets. Ensemble Methods continue to perform well in more complex scenarios, striking a balance between complexity and accuracy. This analysis highlights that selecting the right classifier depends on the nature of the data and available resources. In complex biological projects, DL models such as CNN and RNN often deliver superior results.

Moreover, in Table 4a comprehensive comparison of three widely employed methods in ACP prediction: CNN, RNN, and GAN. Each of these models serves distinct roles and brings unique strengths to the table, addressing different challenges in this specialized domain. CNNs are particularly effective for extracting structural features from peptide sequences, excelling in tasks where identifying spatial patterns is crucial. This makes CNNs an ideal choice for feature-based ACP activity prediction. However, their limitations become apparent when handling sequential dependencies, which restrict their applicability in dynamic datasets. Moreover, their computational cost is relatively high due to the intensive processing required by convolutional layers, especially when working with high-resolution or complex data.

RNNs, on the other hand, are well-suited for modeling sequential dependencies within peptide datasets, enabling them to capture temporal relationships critical for understanding the dynamic behavior of ACPs. This makes RNNs particularly effective in time-series analysis or tasks involving evolving peptide interactions. Despite their strengths, RNNs encounter challenges in handling long-term dependencies and often require longer training times to achieve convergence. Nonetheless, their ability to uncover patterns in sequential data positions them as a valuable tool in ACP prediction, particularly for datasets with time-dependent properties.

Table 4 Comparative analysis of CNN, RNN, and GAN for ACP prediction, summarizing their primary roles, strengths, weaknesses, accuracy, and key evaluation criteria

Criteria	CNN	RNN	GAN
Primary Role in ACP Prediction	Feature extraction from peptide sequences	Modeling sequential dependencies in peptide data	Generating synthetic peptide datasets
Common Use Cases in ACP	ACP activity prediction based on structural features	Dynamic behavior prediction for peptide activity	Improving dataset diversity for ACP model training
Accuracy	High accuracy for feature-based predictions	Moderately high, particularly for sequential data	Variable; depends on data quality and tuning
Handling Imbalanced Data	Moderate	Moderate	Very High
Interpretability	High	Medium	Low
Sensitivity to Hyperparameters	Low	Medium	High
Strengths	Excels at identifying spatial patterns; strong performance in feature extraction	Effective at capturing time-dependent relationships in sequences	Produces realistic synthetic peptides; ideal for augmenting datasets
Weaknesses	Limited effectiveness for sequential dependencies; computationally intensive for high-resolution data	Struggles with long-term dependencies; slower convergence	Prone to instability during training; computationally expensive

GANs bring an entirely different capability to the field by generating synthetic peptide datasets, which can significantly address the issue of data imbalance—a common challenge in ACP research. By creating realistic synthetic peptides, GANs enhance the diversity of training datasets, boosting the performance of predictive models. However, their effectiveness comes at the cost of complexity; GANs are highly sensitive to hyperparameter configurations, often prone to instability during training, and computationally demanding. Despite these challenges, their utility in augmenting datasets and exploring new peptide configurations makes them indispensable for advancing ACP discovery.

The evaluation of these methods was carried out across multiple criteria, including accuracy, interpretability, sensitivity to hyperparameters, and their ability to handle imbalanced data. Accuracy was qualitatively assessed based on each model's suitability for specific tasks, with CNNs achieving high accuracy for feature-based predictions, RNNs delivering moderately high performance for sequential data, and GANs exhibiting variable accuracy depending on data quality and model tuning. Handling imbalanced data was a critical metric where GANs excelled, while CNNs and RNNs provided moderate solutions. Interpretability was highest for CNNs, owing to their relatively simple architecture, while RNNs and GANs presented challenges due to their complex, often opaque designs. Sensitivity to hyperparameters revealed GANs as the most demanding, requiring precise adjustments for stability, followed by RNNs, with CNNs being the least sensitive.

This analysis underscores the importance of aligning model selection with the specific objectives and challenges of the research. CNNs are best suited for tasks focused on spatial feature extraction, RNNs excel in modeling dynamic behaviors, and GANs offer unmatched advantages in augmenting datasets. Together, these insights provide a structured framework for researchers to make informed decisions when employing these methods, individually or in combination, to push the boundaries of ACP discovery and enhance therapeutic innovations in cancer treatment.

Table 5 presents a detailed quantitative comparison of three advanced classifiers—CNN, RNN, and GAN—evaluated across three widely used datasets: CancerPPD, APD3, and a Benchmark dataset. In this study, we focused on three advanced deep learning models—CNN, RNN, and GAN—as they effectively address key challenges in ACP prediction, including structural complexity, sequential dependencies, and data imbalance. CNNs excel at extracting spatial features from encodings like PSSM, RNNs model sequential relationships in peptide sequences using DPC and Word Embedding, and GANs enhance dataset

Table 5 Quantitative comparison of CNN, RNN, and GAN classifiers for ACP prediction across three datasets (CancerPPD, APD3, and Benchmark datasets)

Dataset	Classifier	Encoding Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUROC
CancerPPD	CNN	PSSM, Graphical Features	92.45	90.34	93.21	91.75	0.956
	RNN	DPC, Word Embedding	89.78	88.12	90.15	89.12	0.938
	GAN	Word Embedding	88.21	86.47	88.98	87.71	0.915
APD3	CNN	PSSM, Graphical Features	90.56	89.10	91.67	90.37	0.941
	RNN	DPC, Word Embedding	88.23	87.45	89.14	88.29	0.924
	GAN	Word Embedding	87.11	85.67	87.94	86.79	0.905
Benchmark	CNN	PSSM, Graphical Features	91.82	90.73	92.78	91.74	0.949
	RNN	DPC, Word Embedding	90.01	89.14	90.89	89.96	0.936
	GAN	Word Embedding	88.67	87.34	89.12	88.22	0.923

diversity by generating synthetic sequences. These models were chosen for their ability to automatically learn features, offering scalability and superior performance compared to traditional machine learning methods. The simulations were conducted using carefully fine-tuned hyperparameters for each model to ensure fair and meaningful comparisons. Parameters such as learning rates, hidden layers, dropout rates, and optimization techniques were optimized to achieve the best performance. Encoding methods were selected based on the specific strengths of each model. For instance, CNN was paired with PSSM and Graphical Features, which effectively capture structural and spatial patterns in peptide sequences. RNN utilized DPC and Word Embedding to model sequential dependencies, while GAN employed Word Embedding to enhance dataset diversity.

The classifiers were evaluated using a standardized 5-fold cross-validation protocol, and their performance was assessed through key metrics, including Accuracy, Precision, Recall, F1-score, and AUROC. CNN consistently delivered the highest performance across all datasets, achieving an Accuracy of 92.45%, F1-score of 91.75%, and AUROC of 0.956 on CancerPPD. These results underscore CNN's superior ability to extract spatially rich features from encodings like PSSM. RNN demonstrated robust sequential modeling capabilities, particularly excelling in Recall (90.15%), making it highly effective for capturing temporal relationships within peptide data. However, its slightly lower F1-score and AUROC compared to CNN highlight its challenges in handling long-term dependencies. GAN showed its unique strength in addressing data imbalance by generating synthetic sequences, improving the training process. Despite this, its performance in direct prediction tasks, such as Accuracy and AUROC, was slightly lower compared to CNN and RNN, reflecting its reliance on synthetic data quality and sensitivity to hyperparameter configurations.

The analysis in Table 5 highlights CNN as the most effective classifier for ACP prediction, particularly for datasets with rich structural features. Its ability to model spatial dependencies and process high-dimensional encodings makes it a reliable choice for a wide range of datasets. RNN, while slightly behind CNN in overall performance, remains a strong alternative for sequence-based tasks, especially when paired with encodings like DPC and Word Embedding. GAN, on the other hand, proved invaluable as a complementary tool for augmenting datasets and addressing imbalance challenges, despite its lower predictive metrics compared to CNN and RNN. These findings emphasize the importance of matching classifiers with suitable encoding methods to maximize their potential. Overall, CNN offers the best performance for primary prediction tasks, RNN excels in sequential modeling, and GAN adds value by enhancing dataset quality, providing a robust framework for advancing anticancer peptide prediction.

4.6 Practical applications

In this section, we explore several groundbreaking studies that exemplify the real-world applications of ML and DL in the design of ACPs and their transformative role in cancer treatment. As shown in Fig. 14, these technologies have moved beyond theoretical concepts, now being actively applied to identify effective peptide candidates and expedite the drug discovery process, leading to tangible improvements in both efficacy and efficiency. Accordingly, Fig. 14 illustrates the chronological progression of significant studies that demonstrate how ML and DL models have been applied in real-world cancer treatment scenarios, with concrete examples of their impact on peptide discovery, drug activity pre-

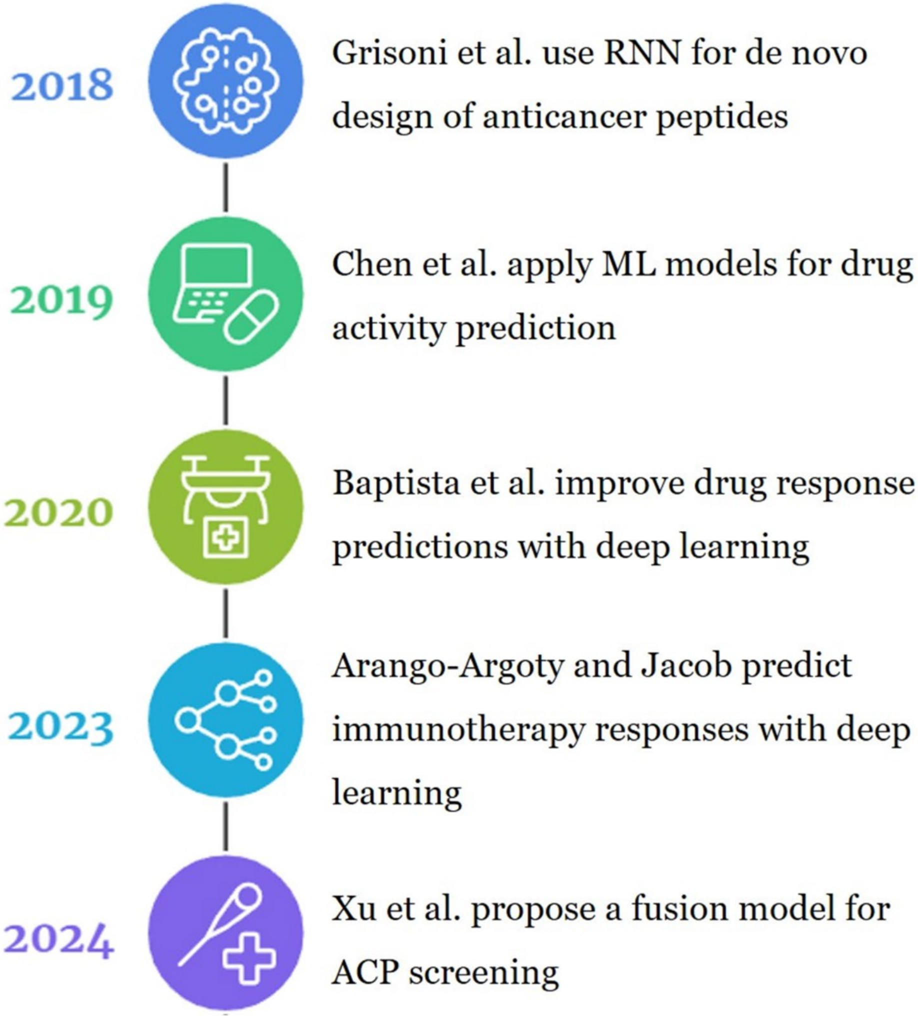
2018
Grisoni et al. use RNN for de novo design of anticancer peptides
2019
Chen et al. apply ML models for drug activity prediction
2020
Baptista et al. improve drug response predictions with deep learning
2023
Arango-Argoty and Jacob predict immunotherapy responses with deep learning
2024
Xu et al. propose a fusion model for ACP screening

Fig. 14 Examples of how ML and DL models have been applied in cancer treatment, including drug activity prediction, anticancer peptide design, and immunotherapy response prediction, demonstrating real-world advancements in cancer therapies

diction, and immunotherapy responses. The following examples highlight how ML and DL are being successfully utilized in clinical settings and experimental treatments, contributing significantly to the advancement of cancer therapies.

4.6.1 Case studies from Cancer Treatment

Several studies have highlighted the significant role of ML and DL in cancer treatment and peptide design. Chen et al. (2019) applied ML models, including Random Forest, AdaBoost, and Gradient Boosting, along with deep learning techniques, to predict drug activity and design peptides for cancer treatment. Their study involved molecular docking and peptide optimization experiments. Similarly, Mohammadzadeh-Vardin et al. (2024) explored how deep learning models, particularly autoencoders, predict cancer drug responses based on multi-omics data, significantly enhancing drug repurposing efforts. Baptista et al. (2021) demonstrated the use of deep learning to improve drug response predictions in cancer treatment, showing how these models outperform traditional machine learning approaches. Grisoni et al. (2018) used a RNN for the de novo design of anticancer peptides, tested on human breast cancer cells, illustrating how ML can generate peptides with anticancer activity.

4.6.2 Predictive models and their applications

Salam et al. (2024) introduced a deep learning model using a two-dimensional convolutional neural network (2D CNN) to predict anticancer peptides with remarkable precision. Leveraging peptide sequences annotated with anticancer activity, the model achieved an impressive AUC-ROC value of 0.91, indicating its robust performance in distinguishing active peptides. This deep learning approach outperforms traditional methods by capturing the intricate spatial patterns within peptide sequences, significantly enhancing prediction accuracy. The study effectively demonstrates the practical potential of DL in cancer therapy, offering a reliable tool for identifying promising peptide candidates for clinical application.

In a similar vein, Danish et al. (2024) proposed an innovative ML framework that integrates various peptide encoding methods to predict ACPs with 97.56% accuracy. Their model outperformed others in identifying novel ACPs, showcasing the strength of ML in peptide discovery. Additionally, their work explores how advanced technologies, such as the metaverse, could revolutionize cancer treatment by combining multi-modal data for drug discovery. By enhancing the selectivity of ACPs, this approach paves the way for more personalized and effective cancer therapies.

4.6.3 AI-Driven clinical trials in Cancer

Arango-Argoty and Jacob (2023) utilized deep learning models to predict patient responses to immunotherapy, leveraging real-world evidence (RWE) to assess cancer treatment efficacy and predict survival outcomes. Their model underscores the potential of deep learning in bridging the gap between preclinical studies and clinical applications, thereby improving decision-making processes for immunotherapy in cancer patients. This study highlights the growing importance of AI in clinical oncology, offering a powerful tool to personalize and optimize cancer treatment strategies.

Bhattarai et al. (2024) conducted a comprehensive analysis of several ML and DL models, including CNNs and SVMs, for predicting ACPs. Their research focused on tools like ACPpred and AI4ACP, which predict ACPs for cancer therapies. The study demonstrated that these models not only surpass traditional approaches in terms of accuracy but also provide valuable insights into optimizing peptide-based therapies for clinical use. By improving predictive precision, these AI tools are critical in the discovery of novel, highly efficient anticancer peptides.

In another notable real-world application, Yao et al. (2023) introduced a novel deep learning framework combining deep graphical representations with deep forest architecture for ACP prediction. This model achieved 94.10% accuracy on a standard dataset, surpassing traditional ACP prediction models in both accuracy and robustness. Its interpretability allows researchers to better understand the key features of peptides that contribute to their anticancer properties, making it particularly valuable for peptide discovery, even in smaller datasets.

Xu et al. (2024a, b) proposed a fusion model that integrates both traditional ML and advanced DL techniques to enhance ACP screening. Their integrated approach achieved significant improvements in prediction accuracy, with AUC values as high as 0.97. Tested with in vitro cell experiments, the model successfully identified ACPs with proven antitumor efficacy, highlighting its practical applicability in drug discovery. This study exemplifies how the combination of traditional ML methods and advanced DL architectures can lead to more accurate predictions and more effective cancer therapies.

Therefore, CNNs and GANs have emerged as leading techniques in this field, each with unique contributions. For instance, Salam et al. (2024) developed a 2D CNN model for ACP prediction, achieving an AUC-ROC of 0.91, which significantly outperformed traditional methods by accurately capturing spatial patterns within peptide sequences (Salam et al. 2024). Bhattarai et al. (2024) analyzed several ML and DL models, including CNNs and SVMs, showcasing the effectiveness of ACPred and AI4ACP tools in accurately identifying ACPs for cancer therapy (Bhattarai et al. 2024). Furthermore, Zhao et al. (2024) introduced dsAMPGAN, a GAN-based model that generates novel antimicrobial peptides, underscoring GANs' potential to explore new chemical spaces and generate clinically relevant peptides (Zhao et al. 2024). These examples demonstrate how ML and DL frameworks are driving breakthroughs in ACP discovery, offering faster and more precise approaches to identify potential peptide therapeutics for cancer treatment. Platforms such as EDGE and dsAMPGAN represent the next frontier in peptide-based drug discovery. These platforms integrate DL methods to predict peptide-HLA binding and generate novel anticancer peptides. Their success in preclinical studies and clinical applications underscores the transformative potential of AI in developing targeted cancer therapies (Zhao et al. 2024). AI is also reshaping the design and execution of clinical trials. By leveraging large datasets and biomarker analysis, AI-driven methods enhance patient selection and drug repurposing for cancer therapies. These advancements have significantly increased trial success rates by aligning treatments with patients' specific genetic and clinical profiles (Xiao, 2021).

5 Challenges and limitations

Despite the promising results achieved in recent years with ML and DL models for predicting ACPs, several challenges and limitations remain that hinder their full potential. These obstacles span various aspects of the process, including data availability, feature selection, model complexity, and interpretability.

(a) Data Imbalance: A common challenge in ACP prediction is the imbalance in datasets. Positive samples of ACPs are far fewer than negative ones, making it difficult to train models effectively (Cai et al. 2021). This imbalance can lead to overfitting and reduced predictive accuracy.

(b) Feature Representation: One of the biggest hurdles is identifying the most relevant features to accurately distinguish ACPs from other peptides (Wei et al. 2019). Some methods rely heavily on hand-crafted features, which can limit the ability to generalize to new data (Lv et al. 2021).

(c) Computational Complexity: Advanced techniques such as representation learning and ensemble models often come at a high computational cost. These methods require more resources for both training and real-time application, limiting their accessibility in resource-constrained settings (Boopathi et al. 2019).

(d) Interpretability: Complex DL models, such as those using RNN and CNN, offer high accuracy but are often criticized for their lack of transparency. It can be challenging to explain how these models arrive at their predictions (Lv et al. 2021).

(e) Limited Generalization: Many models trained on specific datasets fail to generalize well to other datasets or unseen data, which is a significant challenge in ACP prediction. This limitation necessitates the use of diverse datasets or advanced methods like data augmentation (Chen et al. 2021a, b, c).

(f) Data Augmentation Issues: Although data augmentation helps increase sample size, poorly implemented techniques can introduce noise, which degrades model performance. Models like ACP-DA attempt to mitigate these issues but face challenges in creating reliable augmented data (Chen et al. 2021a, b, c).

(g) Manual Feature Selection: Methods that rely on manual feature selection, such as Recursive Feature Elimination (RFE) or Principal Component Analysis (PCA), may miss important interactions in high-dimensional peptide data, thus limiting accuracy improvements (Charoenkwan et al. 2021).

(h) Small Dataset Size: Most ML methods struggle when applied to small datasets, which are common in ACP research due to the difficulty and expense of peptide experiments. Models like ACP-DL attempt to address this but still face challenges when datasets are too small (Yi et al. 2019a, b).

(i) Overfitting: The complexity of models can sometimes lead to overfitting, where the model performs well on training data but poorly on unseen data. DL models are particularly susceptible to this problem (Kaleem et al. 2022).

(j) High Dimensionality: The large number of possible features, such as amino acid composition and physicochemical properties, can overwhelm ML models, leading to poor performance if feature selection is not carefully performed (Deng et al. 2023a, b).

(k) Bias in Datasets: Biases in available datasets can lead to models that are not fully representative of the broader range of peptide sequences. This issue is particularly problematic when datasets are manually curated (Agrawal et al. 2020).

(l) Feature Fusion Complexity: Combining multiple types of features, such as sequence-based and physicochemical descriptors, is challenging. ACPred-Fuse attempts to address this issue but encounters difficulties in optimizing the fusion process (Rao et al., 2020).

(m) Prediction Speed: As models become more complex, the speed at which they can make predictions often decreases. This slow speed can be a barrier to real-time or large-scale applications in ACP discovery (Yuan et al. 2023).

(n) Model Evaluation: The lack of standardized evaluation metrics across different studies makes it difficult to compare the performance of different models. Different studies often use varying accuracy metrics, leading to inconsistent benchmarks (Yao et al. 2023).

6 Future directions and opportunities

To address the challenges in ACP prediction, several specific and actionable future directions leveraging both DL and ML are promising:

1. Advanced Data Augmentation Techniques: One highly actionable research direction is the development of GAN-based synthetic data generation pipelines to tackle small and imbalanced datasets. For example, researchers could design experiments to validate the effectiveness of GAN-generated sequences by comparing their predicted anticancer activity with experimental results. Additionally, combining GANs with active learning frameworks could help iteratively refine these datasets by including only the most informative and experimentally validated peptides.

Hypothesis GAN-generated synthetic peptides enhance prediction accuracy and reduce overfitting when incorporated into training data.

2. Automated and Dynamic Feature Selection: While current models rely on static or manual feature selection methods, future work can focus on dynamic feature selection frameworks using reinforcement learning or attention-based mechanisms. For instance, a reinforcement learning agent could be designed to evaluate the contribution of each feature to model performance in real time. Experimental designs could test the effectiveness of this dynamic approach by benchmarking it against traditional techniques such as PCA.

Hypothesis Reinforcement learning-based feature selection improves model interpretability and generalization compared to static approaches like PCA.

3. Multi-Modal Data Integration: A highly specific avenue for future research is the integration of multi-modal data, including 3D structural data, physicochemical properties, and patient-specific biomarkers. For example, DL models could be trained on datasets combining peptide sequences with structural and clinical data, and their performance could be compared with single-modality models. A potential experimental design could involve testing multi-modal models on diverse cancer cell lines to validate their robustness.

Hypothesis Multi-modal DL models outperform single-modality models by capturing more comprehensive patterns in ACP prediction.

4. Transfer Learning and Pretrained Models: Transfer learning offers a tangible and testable solution for limited labeled data. Models pretrained on large datasets, such as general protein sequences or antimicrobial peptides, can be fine-tuned for ACP-specific tasks. Researchers could design experiments to systematically evaluate the impact of different pretraining datasets and architectures, such as CNNs, RNNs, or Transformers, on ACP prediction accuracy.

Hypothesis Pretraining on large protein-related datasets significantly enhances performance on small ACP datasets compared to training from scratch.

5. Generative Models for Novel ACP Design: Generative models, such as GANs and Variational Autoencoders (VAEs), hold immense potential for designing entirely new ACPs.

Future research could involve training these models on existing peptide datasets and testing the generated sequences experimentally for anticancer activity. A systematic evaluation of the generated peptides' properties (e.g., activity, stability, toxicity) could provide actionable insights into their viability as therapeutic agents.

Hypothesis GANs and VAEs can design novel ACPs with experimentally validated anticancer properties, expanding the diversity of potential therapeutic candidates.

6) Integration of ACP Prediction Models into the Drug Development Pipeline: A critical future direction involves integrating ACP prediction models into the drug discovery pipeline to streamline processes like lead compound identification, optimization, and validation. Predictive models can prioritize peptides with high anticancer potential, reducing experimental effort and costs. Aligning computational predictions with experimental workflows and ensuring peptides meet pharmacokinetic (PK) and pharmacodynamic (PD) requirements will be essential for clinical application. Additionally, combining predictive models with experimental data through multi-modal datasets can enhance reliability and create a feedback loop for model refinement. Embedding these tools into drug development could accelerate peptide-based therapies, making them more efficient and impactful.

Hypothesis Integrating predictive ACP models into the drug discovery pipeline reduces experimental attrition rates and accelerates the identification of clinically viable peptides with optimized PK and PD properties.

7. Explainable AI (XAI) in ACP Prediction: A critical focus of future research should be on explainable AI (XAI) techniques that enhance the transparency of DL models. For instance, researchers could develop XAI methods to visualize the contribution of individual amino acids or peptide features to the model's predictions. Experimental designs could include testing whether these insights align with known biological mechanisms or experimental data.

Hypothesis XAI techniques improve trust and interpretability in DL predictions, enabling broader clinical adoption.

8. Reinforcement Learning for Peptide Optimization: Reinforcement learning (RL) can be a transformative tool for designing optimized peptides. For example, RL algorithms can be trained to iteratively design peptides by selecting amino acids that maximize anticancer activity while minimizing toxicity. Experimental designs could involve validating these optimized peptides in vitro and in vivo.

Hypothesis RL-based peptide optimization produces peptides with superior anticancer activity and minimal toxicity compared to traditional approaches.

9. Development of Benchmarking Frameworks: Another actionable direction is the creation of standardized benchmarking frameworks for evaluating ML and DL models in ACP prediction. These frameworks could include curated datasets, standardized evaluation metrics (e.g., Matthews Correlation Coefficient, AUROC), and reproducible experimental proto-

cols. This would provide a uniform basis for comparing new methods and encourage reproducibility in the field.

Hypothesis Standardized benchmarks enable fair and reproducible comparisons of ML and DL models, accelerating progress in ACP prediction.

7 Conclusion

This paper presents a thorough and comprehensive review of the application of machine learning (ML) and deep learning (DL) methodologies for the identification and design of anticancer peptides (ACPs). By systematically analyzing the potential of these advanced computational approaches, the study highlights how they have contributed to significant advancements in peptide-based therapies, improving their precision, efficacy, and applicability in clinical and experimental settings. Through an in-depth exploration of state-of-the-art models—such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Generative Adversarial Networks (GANs)—the paper provides a detailed understanding of the capabilities and limitations of these methodologies. Additionally, their performance was evaluated across three widely recognized datasets: CancerPPD, APD3, and a Benchmark dataset.

A key contribution of this work lies in the detailed quantitative comparison of CNN, RNN, and GAN classifiers, supported by rigorous simulations with carefully fine-tuned hyperparameters to ensure fairness and reproducibility. The results demonstrate that CNNs consistently deliver superior performance in extracting spatial and structural features, achieving the highest accuracy and predictive reliability. RNNs, while slightly behind CNNs in overall performance, excel in modeling sequential dependencies, leveraging encodings such as DPC and Word Embedding to capture temporal relationships effectively. GANs, on the other hand, play a pivotal role in addressing the common challenge of data imbalance by generating realistic synthetic peptide sequences, thereby enhancing the robustness of training data. These findings emphasize the critical role of integrating model-specific encoding techniques—such as PSSM for CNNs and Word Embedding for RNNs and GANs—to maximize their potential in ACP prediction tasks. Beyond model performance, the paper addresses critical challenges in ACP research, including the need to overcome data imbalance, reduce overfitting, and enhance model interpretability. It also provides actionable recommendations for future research, such as the Attention networks, Transformers, and Ensemble learning frameworks to increase the transparency of DL models, the integration of multi-modal datasets combining sequence, structural, and clinical data for enhanced predictive power, and the application of transfer learning techniques to address the challenges posed by limited labeled datasets. These directions underscore the importance of combining advanced DL techniques with innovative data strategies to further advance ACP prediction. Finally, this review synthesizes the current state-of-the-art in computational ACP design and provides a structured roadmap for future research and innovation. By presenting a detailed, performance-driven analysis of the capabilities of ML and DL models, this paper contributes to advancing our understanding of their roles in ACP discovery. It also serves as a foundation for accelerating the development of peptide-based cancer therapies, fostering precision medicine approaches that hold immense potential for improving patient outcomes.

These insights provide a valuable resource for researchers seeking to address the ongoing challenges in ACP research and design.

Author contributions Khosro Rezaee and Hossein Eslami collaboratively contributed to the conceptualization and design of the study. Khosro Rezaee was primarily responsible for data collection and initial drafting of the manuscript, focusing on gathering relevant literature and organizing the foundational sections of the paper. Hossein Eslami supervised the project and provided critical guidance throughout the manuscript preparation, particularly in the analysis and interpretation of machine learning and deep learning techniques applied to anticancer peptide design. He also led the revisions process, addressing reviewers' comments to ensure the manuscript adhered to the journal's high standards.The division of responsibilities ensured a rigorous and detailed examination of the topic, with both authors actively contributing to the integration of reviewers' feedback and refinement of the manuscript. Both authors reviewed and approved the final version of the manuscript, and they take full responsibility for the integrity and accuracy of the work presented.

Funding This research did not receive any external funding.

Data availability All data used in this research, including data from external sources and data generated during the study, is publicly available and included in this article and its supplementary files.

Declarations

Competing interests The authors declare no competing interests.

Conflict of interest The authors declare no conflicts of interest.

Ethical approval Since the study utilized publicly available and well-established datasets from external sources, where ethical considerations were already addressed during data collection, no further ethical approval or participant consent was required for this specific research.

References

Agrawal P, Bhagat D, Mahalwal M, Sharma N, Raghava G (2020) AntiCP 2.0: an updated model for predicting anticancer peptides. Brief Bioinform

Ahmed S, Muhammod R, Khan ZH et al (2021) Acp-mhcnn: an accurate multi-headed deep-convolutional neural network to predict anticancer peptides. Sci Rep 11:23676

Akbar S, Hayat M, Iqbal M et al (2017) Iacp-gaensc: evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space. Artif Intell Med 79:62–70

Akbar S, Rahman AU, Hayat M et al (2020a) CACP: classifying anticancer peptides using discriminative intelligent model via Chou's 5-step rules and general pseudo components. Chemom Intel Lab Syst 196:103912

Akbar S, Hayat M, Tahir M, Chong K (2020b) cACP-2LFS: classification of anticancer peptides using sequential discriminative model of KSAAP and two-level feature Selection Approach. IEEE Access 8:131939–131948

Akbar S, Ashraf G, Sadiq M et al (2021) A comprehensive review on anticancer peptides and their mechanisms of action. Anticancer Agents Med Chem 21(3):336–355 Available from: link

Akbar S, Hayat M, Tahir M et al (2022) Cacp-deepgram: classification of anticancer peptides via deep neural network and skip-grambased word embedding model. Artif Intell Med 131:102349

Aljabery F, Shabo I, Gimm O, Jahnson S, Olsson H (2018) The expression profile of p14, p53 and p21 in tumour cells is associated with disease-specific survival and the outcome of postoperative chemotherapy treatment in muscle-invasive bladder cancer. Urol Oncol 36(12):530e7–530e16

Anand U, Dey A, Chandel AKS, Sanyal R, Mishra A, Pandey DK et al (2023) Cancer chemotherapy and beyond: current status, drug candidates, associated risks and progress in targeted therapeutics. Genes Dis 10(4):1367–1401

Arango-Argoty G, Jacob E (2023) Enhancing the utilization of deep learning to predict patient response in small immunotherapy cohorts using real-world data. Cancer Res 83(7Supplement):1174

Arenas JL, Kaffy J, Ongeri S (2019) Peptides and peptidomimetics as inhibitors of protein-protein interactions involving β-sheet secondary structures. Curr Opin Chem Biol 52:157–167

Arif M, Ahmed S, Ge F et al (2022) Stackacpred: prediction of anticancer peptides by integrating optimized multiple feature descriptors with stacked ensemble approach. Chemom Intel Lab Syst 220:104458

Arif M, Musleh S, Fida H, Alam T (2024) PLMACPred prediction of anticancer peptides based on protein language model and wavelet denoising transformation. Sci Rep 14(1):16992

Arnab MK, Hasan M, Islam MM (2023) An insight into the structure-activity relationship of antimicrobial peptide brevinin. Jordan J Pharm Sci 16(4):815–829

Azad H, Akbar MY, Sarfraz J, Haider W, Riaz MN, Ali GM (2024) Ghazanfar S G-ACP: a machine learning approach to the prediction of therapeutic peptides for gastric cancer. J Biomol Struct Dynamics 27:1–4

Balaji PD, Selvam S, Sohn H, Madhavan T (2024) MLASM: machine learning based prediction of anticancer small molecules. Mol Diversity 30:1–9

Baptista D, Ferreira PG, Rocha M (2021) Deep learning for drug response prediction in cancer. Brief Bioinform 22(1):360–379

Basith S, Manavalan B, Shin TH, Lee G, IGHBP (2018) Computational identification of growth hormone-binding proteins from sequences using extremely randomized tree. Comput Struct Biotechnol J 16:412–420

Basith S, Manavalan B, Shin T, Lee D, Lee G (2020) Evolution of machine learning algorithms in the prediction and design of anticancer peptides. Curr Protein Pept Sci 21(12):1236–1248

Bechinger B (2015) The SMART model: soft membranes adapt and respond, also transiently, in the presence of antimicrobial peptides. J Pept Sci 21(5):346–355

Bhasin M, Raghava GPS (2004) Classification of nuclear receptors based on amino acid composition and dipeptide composition. J Biol Chem 279(22):23262–23266

Bhattarai S, Tayara H, Chong KT (2024) Advancing peptide-based Cancer Therapy with AI: In-Depth analysis of state-of-the-art AI models. J Chem Inf Model 14

Bian J, Liu X, Dong G et al (2024) Acp-ml: a sequence-based method for anticancer peptide prediction. Comput Biol Med 170:108063

Bidwell IIIGL, Raucher D (2009) Therapeutic peptides for cancer therapy. Part I–peptide inhibitors of signal transduction cascades. Expert Opin Drug Deliv 6(10):1033–1047

Boman HG, Nilsson I, Rasmuson B (1972) Inducible antibacterial defense system in Drosophila. Nature 237:232–235

Boone K, Wisdom C, Camarda KV, Spencer P, Tamerler C (2021a) Combining genetic algorithm with machine learning strategies for designing potent antimicrobial peptides. BMC Bioinformatics 22:41 Available here

Boone RA, Baxter S, Luksza M (2021b) Combining algorithm and machine learning strategies to accelerate peptide-based drug Discovery. J Chem Inf Model 61(4):1667–1675

Boopathi V, Subramaniyam S, Malik A, Lee G, Manavalan B, Yang D-C, mACPpred (2019) A support vector machine-based meta-predictor for identification of anticancer peptides. Int J Mol Sci 20(8):1964

Borrelli A, Tornesello AL, Tornesello ML, Buonaguro FM (2018) Cell penetrating peptides as molecular carriers for anti-cancer agents. Molecules 23(2):295

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68(6):394–424

Cai Z, Yin Y, Shen C, Wang J, Yin X, Chen Z et al (2018) Comparative effectiveness of preoperative, postoperative and perioperative treatments for resectable gastric cancer: a network meta-analysis of the literature from the past 20 years. Surg Oncol 27(4):563–574

Cai Y, He Y, Song J, Li W, Zuo Y, Zheng W et al (2021) An active semi-supervised model for improving the identification of anticancer peptides. Front Bioeng Biotechnol 9:683478

Capecchi A, Cai X, Personne H, Köhler T, van Delden C, Reymond JL (2021a) Machine learning designs non-hemolytic antimicrobial peptides. Chem Sci 12:9221–9232 Available here

Capecchi A, Reiher M, Rothlisberger U (2021b) Machine learning designs peptides: Exploring machine learning models in peptide generation and their applications. ACS Chemical Biology.;16(6): 935–944. Available from: Consensus

Charoenkwan P, Chiangjong W, Lee VS, Nantasenamat C, Hasan MM, Shoombuatong W (2021) Improved prediction and characterization of anticancer activities of peptides using a novel flexible scoring card method. Sci Rep 11(1):3017

Charoenkwan P, Chiangjong W, Nantasenamat C, Moni MA, Lio' P, Manavalan B, Shoombuatong W (2022) SCMTHP: a new approach for identifying and characterizing tumor-homing peptides using estimated propensity scores of amino acids. Pharmaceutics 14(1):122

Chen K, Kurgan L, Ruan J (2007) Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs. BMC Struct Biol 7(1):25

Chen K, Jiang Y, Du L, Kurgan L (2008) Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs. J Comput Chem 30(1):163–172

Chen Y-Z, Chen Z, Gong Y-A, Ying G (2012) SUMOhydro: a novel method for the prediction of sumoylation sites based on hydrophobic properties. PLoS ONE. 7(6)

Chen W, Tran H, Liang Z-Y, Lin H, Zhang L (2015) Identification and analysis of the N(6)-methyladenosine in the Saccharomyces cerevisiae transcriptome. Sci Rep 5(1):13859

Chen JQ, Chen HY, Dai WJ, Lv QJ, Chen CY (2019) Artificial intelligence approach to find lead compounds for treating tumors. J Phys Chem Lett 10(15):4382–4400

Chen H, Li F, Wang L, Jin Y, Chi C-H, Kurgan L et al (2020) Systematic evaluation of machine learning methods for identifying human-pathogen protein-protein interactions. Brief Bioinform 22(3):1–NA

Chen J, Cheong HH, Siu SW (2021a) xDeep-AcPEP: deep learning method for anticancer peptide activity prediction based on convolutional neural network and multitask learning. J Chem Inf Model 61(8):3789–3803

Chen X, Zhang W, Yang X, Li C, Chen H (2021b) ACP-DA: improving the prediction of anticancer peptides using data augmentation. Front Genet 12

Chen XG, Zhang W, Yang X, Li C, Chen H (2021c) Acp-da: improving the prediction of anticancer peptides using data augmentation. Front Genet 12:698477

Chen L, Qin Y, Yao X (2023) PLMTHP: an ensemble framework for tumor-homing peptide prediction using protein language models. Brief Bioinform 24(1):bbad045

Chiangjong W, Chutipongtanate S, Hongeng S (2020) Anticancer peptide: physicochemical property, functional aspect and trend in clinical application. Int J Oncol 57(3):678–696

Chidambaram M, Manavalan R, Kathiresan K (2011) Nanotherapeutics to overcome conventional cancer chemotherapy limitations. J Pharm Pharm Sci 14:67

Choi SH, Guzei I, Spencer LC, Gellman S (2008) Crystallographic characterization of helical secondary structures in alpha/beta-peptides with 1:1 residue alternation. J Am Chem Soc 130(20):6544–6550

Chou K-C (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem Biophys Res Commun 278(2):477–483

Chou K-C (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43(3):246–255

Chou K-C (2004) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21(1):10–19

Cooley RB, Arp DJ, Karplus PA (2010) Evolutionary origin of a secondary structure: pi-helices as cryptic but widespread insertional variations of alpha-helices that enhance protein functionality. J Mol Biol 404(2):232–246

Costa L, Sousa E, Fernandes C (2023) Cyclic peptides in pipeline: what future for these great molecules? Pharmaceuticals 16(7):996

Cui H, Zhang C, Li Y, Hu C (2017) Targeting calcium signaling in cancer therapy. Acta Pharm Sin B 7(1):3–17

Danish S, Khan A, Dang LM, Alonazi M, Alanazi S, Song HK, Moon H (2024) Metaverse Applications in Bioinformatics: a machine learning Framework for the discrimination of anti-cancer peptides. Information 15(1):48

Dara S, Dhamercherla S, Jadav SS, Babu CM, Ahsan MJ (2021) Machine learning in drug discovery: a review. Artif Intell Rev 55(1):1947–1999. https://doi.org/10.1007/s10462-021-10058-3

Dash A, Chakraborty S, Pillai MR, Knapp FF Jr (2015) Peptide receptor radionuclide therapy: an overview. Cancer Biother Radiopharm 30(2):47–71

Deng H, Ding M, Wang Y, Li W, Liu G, Tang Y (2023a) ACP-MLC: a two-level prediction engine for identification of anticancer peptides and multi-label classification of their functional types. Comput Biol Med 158:106844

Deng Y, Ma S, Li J, Zheng B, Lv Z (2023b) Using the random forest for identifying key physicochemical properties of amino acids to discriminate anticancer and non-anticancer peptides. Int J Mol Sci 24(13):10854

Dennison SR, Pouny Y, Nitzan N et al (2021) Analysis of the characteristics of anticancer peptides identifies molecular features Associated with their potency. Mol Diversity 25:331–345

Desale K, Kuche K, Jain S (2021) Cell-penetrating peptides (CPPs): an overview of applications for improving the potential of nanotherapeutics. Biomater Sci 9(4):1153–1188

Deslouches B, Di YP (2017) Antimicrobial peptides with selective antitumor mechanisms: Prospect for anticancer applications. Oncotarget 8:46635–46651

DeVita VT Jr, Chu E (2008) A history of cancer chemotherapy. Cancer Res 68(21):8643–8653

Dong Q, Zhou S, Guan J (2009) A new taxonomy-based protein fold recognition approach based on auto-cross-covariance transformation. Bioinformatics 25(20):2655–2662

Eghtedari M, Jafari Porzani S, Nowruzi B (2021) Anticancer potential of natural peptides from terrestrial and marine environments: a review. Phytochemistry Lett 42:87–103

Eliassen LT, Berge G, Leknessund A, Wikman M, Lindin I, Løkke C et al (2006) The antimicrobial peptide, lactoferricin B, is cytotoxic to neuroblastoma cells in vitro and inhibits xenograft growth in vivo. Int J Cancer 119(3):493–500

Engelking LR (2015) Protein structure. In: Engelking LR (ed) Textbook of Veterinary physiological Chemistry, 3rd edn. Academic, Boston, pp 18–25

Feher JJ (2017) Quantitative human physiology: an introduction, 2nd edn. Academic, San Diego, pp 360–380

Feng Z-P, Zhang C-T (2000) Prediction of membrane protein types based on the hydrophobic index of amino acids. J Protein Chem 19(4):269–275

Fuchs JA, Grisoni F, Kossenjans M, Hiss JA, Schneider G (2018) Lipophilicity prediction of peptides and peptide derivatives by consensus machine learning. MedChemComm 9:1538–1546

Furlong SJ, Ridgway ND, Hoskin DW (2008) Modulation of ceramide metabolism in T-leukemia cell lines potentiates apoptosis induced by the cationic antimicrobial peptide bovine lactoferricin. Int J Oncol 32(3):537–544

Gaspar D, Veiga AS, Castanho MA (2013) From antimicrobial to anticancer peptides. A review. Front Microbiol 4:294

Gaudelet T, Day B, Jamasb AR, Soman J, Regep C, Liu G et al (2021) Utilising graph machine learning within drug discovery and development. Brief Bioinform 22(6):1–18

Ghafoor H, Asim MN, Ibrahim MA et al (2024) Capture: comprehensive anti-cancer peptide predictor with a unique amino acid sequence encoder. Comput Biol Med 176:108538

Glukhov E, Burrows LL, Deber CM (2008) Membrane interactions of designed cationic antimicrobial peptides: the two thresholds. Biopolymers 89(5):360–371

Godbey W (2014) Proteins. In: Godbey W (ed) Biotechnology and its applications. Cambridge University Press, Cambridge, pp 9–33

Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185(4154):862–864

Grisoni F, Neuhaus CS, Gabernet G, Müller AT, Hiss JA, Schneider G (2018) Designing anticancer peptides by constructive machine learning. ChemMedChem 13(13):1300–1302

Guan Y, Yao W, Zhou C et al (2023) StackTHPred: identifying tumor-homing peptides through ensemble learning. Artif Intell Med 140:102457

Guntuboina C, Das A, Mollaei P, Kim S, Farimani A (2023) PeptideBERT: a Language Model based on transformers for peptide property prediction. J Phys Chem Lett 14(24):10427–10434

Guo Y, Yu L, Wen Z, Li M (2008) Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic Acids Res 36(9):3025–3030

Hadianamrei R, Tomeh MA, Brown S, Wang J, Zhao X (2021) Rationally designed short cationic α-helical peptides with selective anticancer activity. J Colloid Interface Sci 607(Pt 1):488–501

Hanaoka Y, Yamaguchi Y, Yamamoto H, Ishii M, Nagase T, Kurihara H, Akishita M, Ouchi Y (2016) In Vitro and in vivo anticancer activity of human β-Defensin-3 and its mouse Homolog. Anticancer Res 36(11):5999–6004

Hancock RE, Sahl HG (2006) Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. Nat Biotechnol 24(12):1551–1557

He B, Gao S, Liu W, Liu J, Su Z (2021) Accelerating peptide discovery through information-based design and high-throughput screening. Comput Struct Biotechnol J 19:1099–1110

Hein MJA, Kvansakul M, Lay FT, Phan TK, Hulett M (2022) Defensin–lipid interactions in membrane targeting: mechanisms of action and opportunities for the development of antimicrobial and anticancer therapeutics. Biochem Soc Trans 50:423–437

Henninot A, Collins JC, Nuss JM (2018) The current state of peptide drug discovery: back to the future? J Med Chem 61(4):1382–1414

Hilchie AL, Vale R, Zemlak TS, Hoskin DW (2019) Generation of a hematologic malignancy-selective membranolytic peptide from the antimicrobial core (RRWQWR) of bovine lactoferricin. Exp Mol Pathol 95:192–198

Hu E, Wang D, Chen J, Tao X (2015) Novel cyclotides from Hedyotis diffusa induce apoptosis and inhibit proliferation and migration of prostate cancer cells. Int J Clin Exp Med 8(4):4059–4065

Huang YB, Wang XF, Wang HY, Liu Y, Chen Y (2011) Studies on mechanism of action of anticancer peptides by modulation of hydrophobicity within a defined structural framework. Mol Cancer Ther 10(3):416–426

Huang K-Y, Tseng Y, Kao H-J, Chen C-H, Yang H-H, Weng S (2021) Identification of subtypes of anticancer peptides based on sequential features and Physicochemical Properties. Sci Rep 11:93124

Jafari A, Babajani A, Sarrami Forooshani R, Yazdani M, Rezaei-Tavirani M (2022) Clinical applications and anticancer effects of antimicrobial peptides: from bench to bedside. Front Oncol 12:819563

Janairo JI (2022) A machine learning classification model for gold-binding peptides. ACS Omega 7(16):14069–14073

Kaleem H, Rukhsar S, Khalid MN (2022) Anticancer peptides prediction: a deep learning approach. J Comput Biomedical Inf 10(4):1367–1401

Kamalov F, Thabtah F (2017) A feature selection method based on ranked vector scores of features for classification. Annals of Data Science

Karakaya O, Kilimci ZH (2024) An efficient consolidation of word embedding and deep learning techniques for classifying anticancer peptides: FastText+BiLSTM. PeerJ Comput Sci 10:e1831

Karami Fath M, Babakhaniyan K, Zokaei M, Yaghoubian A, Akbari S, Khorsandi M et al (2022) Anti-cancer peptide-based therapeutic strategies in solid tumors. Cell Mol Biol Lett 27(1):33

Karim T, Shaon MS, Sultan MF, Hasan MZ, Kafy AA (2024) ANNprob-ACPs: a novel anticancer peptide identifier based on probabilistic feature fusion approach. Comput Biol Med 169:107915

Kaspar AA, Reichert JM (2013) Future directions for peptide therapeutics development. Drug Discov Today 18(17–18):807–817

Kaur D, Arora A, Vigneshwar P, Raghava GP (2024) Prediction of peptide hormones using an ensemble of machine learning and similarity-based methods. Proteomics 27:2400004

Khan S (2024) Deep-representation-learning-based classification strategy for anticancer peptides. Mathematics 12(9):1330

Khawaja SA, Farooq MS, Ishaq K, Alsubaie N, Karamti H, Montero EC, Alvarado ES, Ashraf I (2024) Prediction of leukemia peptides using convolutional neural network and protein compositions. BMC Cancer 24(1):900

Kilimci ZH, Yalcin M (2024) ACP-ESM: a novel framework for classification of anticancer peptides using protein-oriented transformer approach. Preprint at arXiv:2401.02124

Kim MK, Oh SW, Lim JY, Jeon EY, Shin JH, Cho S (2018) Antibacterial and antibiofilm activity and mode of action of Magainin 2 against drug-resistant Acinetobacter baumannii. Int J Mol Sci 19(10):3041

Kim KJ, Kim KJ, Choi J, Kim NH, Kim SG (2023) Linear association between radioactive iodine dose and second primary malignancy risk in thyroid cancer. JNCI J Natl Cancer Inst 115(6):695–702

Korde V, Mahender CN (2012) Text classification and classifiers: a survey. Int J Artif Intell Appl 3(2):85

Kumar A, Singh D (2024) Multiview and decision fusion in stacking ensemble to predict anti-cancer peptides. In 2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0 (pp. 1–7)

Langan RA, Boyken SE, Ng AH, Samson JA, Dods G, Westbrook AM et al (2019) De novo design of bioactive protein switches. Nature 572:205–210

Last NB, Schlamadinger DE, Miranker AD (2013) A common landscape for membrane-active peptides. Protein Sci 22(7):870–882

Lau JL, Dunn MK (2018) Therapeutic peptides: historical perspectives, current development trends, and future directions. Bioorg Med Chem 26:2700–2707

Lee B, Shin D (2024) Contrastive learning for enhancing feature extraction in anticancer peptides. Brief Bioinform 25(3):bbae220

Li W, Joshi MD, Singhania S, Ramsey KH, Murthy AK (2014) Peptide vaccine: Progress and challenges. Vaccines 2:515–536

Li B, Lyu P, Xie S, Qin H, Pu W, Xu H et al (2019) LFB: a novel antimicrobial brevinin-like peptide from the skin secretion of the Fujian large-headed frog, Limnonectes fujianensis. Biomolecules 9(7):242

Li Q, Zhou W, Wang D, Wang S, Li Q (2020) Prediction of anticancer peptides using a low-dimensional feature model. Front Bioeng Biotechnol 8:31

Liang Y, Ma X (2023) Iacp-ge: accurate identification of anticancer peptides by using gradient boosting decision tree and extra tree. SAR QSAR Environ Res 34:1–19

Liang PY, Huang X, Duran T, Wiemer AJ, Bai J (2024a) Exploring Latent Space for Generating Peptide Analogs Using Protein Language Models. In IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2024 Dec 3 (pp. 842–847)

Liang X, Zhao H, Wang J, MA-PEP: (2024b) A novel anticancer peptide prediction framework with multimodal feature fusion based on attention mechanism. Protein Sci 33(4):e4966

Lin Z, Pan X-M (2001) Accurate prediction of protein secondary structural content. J Protein Chem 20(3):217–220

Lin E, Lin CH, Lane HY (2022) De novo peptide and protein design using generative adversarial networks: an update. J Chem Inf Model 62(4):761–774

Liu B, Fang L, Long R, Lan X, Chou K-C (2015) iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics 32(3):362–369

Liu R, Li X, Lam KS (2017) Combinatorial chemistry in drug discovery. Curr Opin Chem Biol 38:117–126

Liu M, Wu T, Li X, Zhu Y, Chen S, Huang J, Zhou F, Liu H (2024) ACPPfel: explainable deep ensemble learning for anticancer peptides prediction based on feature optimization. Front Genet 15:1352504

López-Vallejo F, Caulfield T, Martínez-Mayorga K, Giulianotti MA, Nefzi A, Houghten RA, Medina-Franco JL (2011) Integrating virtual screening and combinatorial chemistry for accelerated drug discovery. Comb Chem High Throughput Screen 14(6):475–487

Luo Q, Zhang L, Luo C, Jiang M (2019) Emerging strategies in cancer therapy combining chemotherapy with immunotherapy. Cancer Lett 454:191–203

Lv Z, Cui F, Zou Q, Zhang L, Xu L (2021) Anticancer peptides prediction with deep representation learning features. Brief Bioinform 22(3)

Manavalan B, Shin TH, Kim MO, Lee G, AIPpred (2018) Sequence-based prediction of anti-inflammatory peptides using random forest. Front Pharmacol 9:276

Mao J, Akhtar J, Zhang X, Sun L, Guan S, Li X, Chen G, Liu J, Jeon HN, Kim MS, No KT (2021) Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models. Iscience.;24(9)

Mao J, Guan S, Chen Y, Zeb A, Sun Q, Lu R, Dong J, Wang J, Cao D (2023a) Application of a deep generative model produces novel and diverse functional peptides against microbial resistance. Comput Struct Biotechnol J 21:463– 71

Mao J, Wang J, Zeb A, Cho KH, Jin H, Kim J, Lee O, Wang Y, No KT (2023b) Transformer-based molecular generative model for antiviral drug design. J Chem Inf Model 64(7):2733–2745

Miao J, Descoteaux M, Lin YS (2021a) Structure prediction of cyclic peptides by molecular dynamics + machine learning. Chem Sci 12:14927–14936

Miao J, Xu Y, Sun S, Yang Y (2021b) Peptide structure prediction via molecular dynamics and machine learning algorithms. J Chem Inf Model 61(10):5297–5310

Micale N, Scarbaci K, Troiano V, Ettari R, Grasso S, Zappala M (2014) Peptide-based proteasome inhibitors in anticancer drug design. Med Res Rev 34:1001–1069

Mizejewski G, Eisele L, Maccoll R (2021) Anticancer versus antigrowth activities of three analogs of the growth-inhibitory peptide: relevance to physicochemical properties. Anticancer Res 41(4B):2071–2077

Mohammadzadeh-Vardin T, Ghareyazi A, Gharizadeh A, Abbasi K, Rabiee HR (2024) DeepDRA: drug repurposing using multi-omics data integration with autoencoders. PLoS ONE 19(7):e0307649

Motmaen A, Dauparas J, Baek M, Abedi MH, Baker D, Bradley P (2023) Peptide-binding specificity prediction using fine-tuned protein structure prediction networks. Proc Natl Acad Sci 120(9):e2216697120

Muthukrishnan R, Rohini R (2016) LASSO: A feature selection technique in predictive modeling for machine learning. IEEE International Conference on Advances in Computer Applications

Nabizadeh S, Rahbarnia L, Nowrozi J, Farajnia S, Hosseini F (2023) Rational design of hybrid peptide with high antimicrobial property derived from Melittin and Lasioglossin. J Biomol Struct Dyn 1–9

Nasiri F, Atanaki FF, Behrouzi S, Kavousi K, Bagheri M (2021) CpACpP: in silico cell-penetrating anticancer peptide prediction using a novel bioinformatics framework. ACS Omega 6(30):19846–19859

Neuhaus CS, Gabernet G, Steuer C, Root K, Hiss JA, Zenobi R, Schneider G (2023) Deconstructing the potency and cell-line selectivity of membranolytic anticancer peptides. ChemBioChem 24(6):1518–1532

Ng A, Jordan M (2001) On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. Adv Neural Inf Process Syst

Niu Y, Li Z, Chen Z, Huang W, Tan J, Tian F, Yang T, Fan Y, Wei J, Mu J (2024) Efficient screening of pharmacological broad-spectrum anti-cancer peptides utilizing advanced bidirectional encoder representation from transformers strategy. Heliyon 10(9)

Ortega-García MB, Mesa A, Moya EL, Rueda B, Lopez-Ordoño G, García JA et al (2020) Uncovering tumour heterogeneity through PKR and nc886 analysis in metastatic colon cancer patients treated with 5-FU-based chemotherapy. Cancers 12(2):379

Oyen WJG, Bodei L, Giammarile F, Maecke HR, Tennvall J, Luster M et al (2007) Targeted therapy in nuclear medicine—current status and future prospects. Ann Oncol 18(11):1782–1792

Papo N, Shai Y (2005) Host defense peptides as new weapons in cancer treatment. Cell Mol Life Sci 62(7–8):784–790

Pawar GM, Patil S, Tarase D, Kshirsagar V (2023) Recent advancements in machine learning for drug discovery and design. Future Med Chem 15(2):189–203

Payandeh Z, Noori E, Khalesi B, Mard-Soltani M, Abdolalizadeh J, Khalili S (2018) Anti-CD37 targeted immunotherapy of B-Cell malignancies. Biotechnol Lett 40(10):1459–1466

Peper F, Noda H, Shirazi MN (2002) Determination of principal components in data. Elsevier

Pham TL, Saurav JR, Omere AA, Heyl CJ, Nasr MS, Reynolds CT, Veerla JP, Shang HH, Jaworski J, Ravenscraft A, Buonomo JA (2024) Peptide sequencing via protein language models. In Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (pp. 1–8)

Quijano-Rubio A, Yeh HW, Park J, Lee H, Langan RA, Boyken SE et al (2021) De novo design of modular and tunable protein biosensors. Nature 591:482–487

Raffatellu M (2018) Learning from bacterial competition in the host to develop antimicrobials. Nat Med 24:1097–1103

Răileanu M, Bacalum M (2023) Cancer Wars: revenge of the AMPs (antimicrobial peptides), a New Strategy against Colorectal Cancer. Toxins.;15

Rao B, Zhou C, Zhang G, Su R, Wei L (2020) ACPred-Fuse: fusing multi-view information improves the prediction of anticancer peptides. Brief Bioinform 21(5):1846–1855

Raucher D, Ryu JS (2015) Cell-penetrating peptides: strategies for anticancer treatment. Trends Mol Med 21(9):560–570

Raucher D, Moktan S, Massodi I, Bidwell Iii GL (2009) Therapeutic peptides for cancer therapy. Part II–cell cycle inhibitory peptides and apoptosis-inducing peptides. Expert Opin Drug Deliv 6(10):1049–1064

Ruoslahti E (2017a) Peptides as targeting elements and tissue penetration devices for nanoparticles. Adv Mater 29(36):1605471

Ruoslahti E (2017b) Tumor penetrating peptides for improved drug delivery. Adv Drug Deliv Rev 110–111:3–12

Rusiecka I, Gągało I, Kocić I (2022) Cell-penetrating peptides improve pharmacokinetics and pharmacodynamics of anticancer drugs. Tissue Barriers 10(1):1965418

Salam A, Ullah F, Amin F, Khan IA, Villena EG, de la Castilla AK (2024) Torre I. Efficient prediction of anticancer peptides through deep learning. PeerJ Comput Sci 10:e2171

Saxena M, van der Burg SH, Melief CJ, Bhardwaj N (2021) Therapeutic cancer vaccines. Nat Rev Cancer 21(6):360–378

Schaduangrat N, Nantasenamat C, Prachayasittikul V et al (2019) Acpred: a computational tool for the prediction and analysis of anticancer peptides. Molecules 24:1973

Schally AV, Nagy A (2004) Chemotherapy targeted to cancers through tumoral hormone receptors. Trends Endocrinol Metab 15(7):300–310

Schneider G, Wrede P (1994) The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site. Biophys J 66(2):335–344

Schweizer F (2009) Cationic amphiphilic peptides with cancer-selective toxicity. Eur J Pharmacol 625:190–194

Shanthappa PM, Melethadathil N (2024) Exploring Novel Anticancer Peptides: Evolutionary Prediction of Potential tRNA-Encoded Peptides for Targeted Therapies. In International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE) 2024 Jan 24 (pp. 1–6)

Sharma A, Rani R (2021) Machine learning applications in anti-cancer drug discovery. Intell Healthc 1(3):101–116

Sharma A, Kapoor P, Gautam A, Chaudhary K, Kumar R, Chauhan JS, Tyagi A, Raghava GP (2013) Computational approach for designing tumor homing peptides. Sci Rep 3(1):1607

Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K et al (2007) Predicting protein-protein interactions based only on sequences information. Proc Natl Acad Sci USA 104(11):4337–4341

Shewach DS, Kuchta RD (2009) Introduction to cancer chemotherapeutics. Chem Rev 109(7):2859–2861

Shin MK, Jang B-Y, Bu K-B, Lee S-H, Han D-H, Oh JW et al (2022) De novo design of AC-P19M, a novel anticancer peptide with apoptotic effects on lung cancer cells and anti-angiogenic activity. Int J Mol Sci 23(24):15594

Shoombuatong W, Schaduangrat N, Nantasenamat C (2018) Unraveling the bioactivity of anticancer peptides as deduced from machine learning. EXCLI J 17:734

Shoombuatong W, Schaduangrat N, Pratiwi R et al (2019) THPep: a machine learning-based approach for predicting tumor-homing peptides. Sci Rep 9(1):17456

Sokal RR, Thomson BA (2006) Population structure inferred by local spatial auto-correlation: an example from an amerindian tribal population. Am J Phys Anthropol 129(1):121–131

Song X, Lu H (2017) Regression embedded feature selection with application to fMRI analysis. Proceedings of the 31st AAAI Conference on Artificial Intelligence

Song H, Lin X, Zhang H, Yin H (2024) ACP-ESM2: the prediction of anticancer peptides based on pre-trained classifier. Comput Biol Chem 110:108091

Soon TN, Chia AYY, Yap WH, Tang YQ (2020) Anticancer mechanisms of bioactive peptides. Protein Pept Lett 27:823–830

Su R, Liu X, Wei L (2020) MinE-RFE: determine the optimal subset from RFE by minimizing the subset-accuracy-defined energy. Briefings in Bioinformatics

Sun X, Liu Y, Ma T, Zhu N, Lao X, Zheng H (2024) DCTPep, the data of cancer therapy peptides. Sci Data 11(1):541

Tanada M, Tamiya M, Matsuo A, Chiyoda A, Takano K, Ito T, Irie M, Kotake T, Takeyama R, Kawada H, Hayashi R (2023) Development of orally bioavailable peptides targeting an intracellular protein: from a hit to a clinical KRAS inhibitor. J Am Chem Soc 145(30):16610–16620

Tao H, Shan S, Fu H, Zhu C, Liu B (2023) An augmented sample selection framework for prediction of anticancer peptides. Molecules 28(18):6680

Tesauro D, Accardo A, Diaferia C, Milano V, Guillon J, Ronga L et al (2019) Peptide-based drug-delivery systems in biotechnological applications: recent advances and perspectives. Molecules 24:351

Torres MD, Chen Y, Zhang L, Wang Y (2020) Wasp venom peptide Polybia-MP1 derivatives display membrane disruption and cytotoxicity against Cancer cells. Sci Rep 10(1):13592

Tyagi A, Kapoor P, Kumar R et al (2013) In silico models for designing and discovering novel anticancer peptides. Sci Rep 3

Uhlig T, Kyprianou TD, Martinelli FG, Oppici CA, Heiligers D, Hills D et al (2014) The emergence of peptides in the pharmaceutical business: from exploration to exploitation. EuPA Open Proteom 4:58–69

Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G et al (2021) AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 50:D1

Verbeke F, de Craemer S, Debunne N, Janssens Y, Wynendaele E, van de Wiele C et al (2017) Peptides as quorum sensing molecules: measurement techniques and obtained levels in vitro and in vivo. Front Neurosci 11:183

Wang S, Ma B (2024) Anti-cancer peptides identification and activity type classification with protein sequence pre-training. IEEE J Biomedical Health Inf

Wang G, Li X, Wang Z (2016) APD3: the antimicrobial peptide database as a tool for research and education. Nucleic Acids Res 44

Watt PM (2006) Screening for peptide drugs from the natural repertoire of biodiverse protein folds. Nat Biotechnol 24(2):177–183

Wei L, Zhou C, Su R, Zou Q (2019) PEPred-Suite: improved and robust prediction of therapeutic peptides using adaptive feature representation learning. Bioinformatics 35(21):4272–4280

Wimley WC (2010) Describing the mechanism of antimicrobial peptide action with the interfacial activity model. ACS Chem Biol 5(10):905–917

Worm DJ, Els-Heindl S, Beck-Sickinger AG (2020) Targeting of peptide-binding receptors on cancer cells with peptide-drug conjugates. Pept Sci 112(3):e24171

Wu Y-D, Han W, Wang D-P, Gao Y, Zhao Y-L (2008) Theoretical analysis of secondary structures of beta-peptides. Acc Chem Res 41(10):1418–1427

Wu C, Gao R, Zhang Y, De Marinis Y (2019) PTPD: predicting therapeutic peptides by deep learning and word2vec. BMC Bioinform 20(1):1–8

Wu CL, Chih YH, Hsieh HY, Peng KL, Lee YZ, Yip B et al (2022) High level expression and purification of Cecropin-like antimicrobial peptides in Escherichia coli. Biomedicines 10(6):1351

Xiao Q, Zhang F, Xu L, Yue L, Kon OL, Zhu Y, Guo T (2021) High-throughput proteomics and AI for cancer biomarker discovery. Adv Drug Deliv Rev 176:113844

Xie M, Liu D, Yang Y (2020) Anti-cancer peptides: classification, mechanism of action, reconstruction and modification. Open Biol 10:200004

Xu M, Pang J, Ye Y, Zhang Z (2024a) Integrating Traditional Machine Learning and Deep Learning for Precision screening of anticancer peptides: a Novel Approach for efficient drug Discovery. ACS Omega 9(14):16820–16831

Xu X, Li C, Yuan X, Zhang Q, Liu Y, Zhu Y, Chen T (2024b) ACP-DRL: an anticancer peptides recognition method based on deep representation learning. Front Genet 15:1376486

Yang W, Luo D, Wang S, Wang R, Chen R, Liu Y et al (2008) TMTP1, a novel tumor-homing peptide specifically targeting metastasis. Clin Cancer Res 14(17):5494–5502

Yang K, Xu J, Liu Q, Li J, Xi Y (2019) Expression and significance of CD47, PD1 and PDL1 in T-cell acute lymphoblastic lymphoma/leukemia. Pathol Res Pract 215(2):265–271

Yang X, Jin J, Wang R, Li Z, Wang Y, Wei L (2023) CACPP: a contrastive learning-based siamese network to identify anticancer peptides based on sequence only. J Chem Inf Model 64(7):2807–2816

Yao L, Li W, Zhang Y, Deng J, Pang Y, Huang Y et al (2023) Accelerating the discovery of anticancer peptides through deep forest architecture with deep graphical representation. Int J Mol Sci 24(5):4328

Yao L, Xie P, Guan J, Chung CR, Zhang W, Deng J, Huang Y, Chiang YC, Lee TY (2024) ACP-CapsPred: an explainable computational framework for identification and functional prediction of anticancer peptides based on capsule network. Brief Bioinform 25(5):bbae460

Ye Y, Du Y, Xu L, Tang W, Zhao Y, Guo J et al (2023) Machine learning advances in predicting peptide-protein interactions for drug discovery. Brief Bioinform 24(2):bbac516

Yi HC, You Z, Zhou X, Cheng L, Li X, Jiang T et al (2019a) ACP-DL: a deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation. Mol Therapy - Nucleic Acids 17:1–9

Yi HC, You ZH, Zhou X, Cheng L, Li X, Jiang TH, Chen ZH (2019b) ACP-DL: a deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation. Mol Therapy-Nucleic Acids 17:1–9

You S, McIntyre G, Passioura T (2024) The coming of age of cyclic peptide drugs: an update on discovery technologies. Expert Opin Drug Discovery 15:1–3

Yu L, Jing R, Liu F, Luo J, Li Y (2020) DeepACP: a novel computational approach for accurate identification of anticancer peptides by deep learning algorithm. Mol Therapy - Nucleic Acids 22:862–870

Yuan Q, Chen K, Yu Y, Le NQ, Chua MC (2023) Prediction of anticancer peptides based on an ensemble model of deep learning and machine learning using ordinal positional encoding. Brief Bioinform 24(1):bbac630

Yue J, Xu J, Li T, Li Y, Chen Z, Liang S, Liu Z, Wang Y (2024) Discovery of potential antidiabetic peptides using deep learning. Comput Biol Med 180:109013

Zhang Y (2012) Support vector machine classification algorithm and its application. In: Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14–16, 2012. Proceedings, Part II 3. Springer; pp. 179–86

Zhang G, Liu S, Liu Y, Wang F, Ren J, Gu J, Zhou K, Shan B (2014) A novel cyclic pentapeptide, H-10, inhibits B16 cancer cell growth and induces cell apoptosis. Oncol Lett 8(6):248–252

Zhang C, Yang M, Ericsson AC (2019) Antimicrobial peptides: potential application in liver cancer. Front Microbiol 10:1257

Zhao Y, Wang S, Fei W, Feng Y, Shen L, Yang X et al (2021) Prediction of anticancer peptides with high efficacy and low toxicity by hybrid model based on 3D structure of peptides. Int J Mol Sci 22(11):5630

Zhao M, Zhang Y, Wang M, Ma LZ (2024) dsAMP and dsAMPGAN: deep learning networks for antimicrobial peptides Recognition and Generation. Antibiotics 13(10):948

Zhixing ZH, Hua DE, Yun TA (2024) Applications and challenges of artificial intelligence in the development of anticancer peptides. J China Pharm Univ 55(3):347–356

Zhong L, Li Y, Xiong L, Wang W, Wu M, Yuan T et al (2021) Small molecules in targeted cancer therapy: advances, challenges, and future perspectives. Signal Transduct Target Ther 6(1):201

Zhong G, Deng L, Acpscanner (2024) Prediction of anticancer peptides by integrated machine learning methodologies. J Chem Inf Model 64(3):1092–1104

Zhou C, Wang C, Liu H, Zhou Q, Liu Q, Guo Y et al (2018) Identification and analysis of adenine N6-methylation sites in the rice genome. Nat Plants 4(8):554–563