ELSEVIER

Artificial intelligence in peptide-based drug design

Silong Zhai [1,2], Tiantao Liu [1], Shaolong Lin [1], Dan Li [2], Huanxiang Liu [1], Xiaojun Yao [1,*], Tingjun Hou [2,*]

[1] Faculty of Applied Science, Macao Polytechnic University, 999078, Macao
[2] College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, Zhejiang, China

Protein–protein interactions (PPIs) are fundamental to a variety of biological processes, but targeting them with small molecules is challenging because of their large and complex interaction interfaces. However, peptides have emerged as highly promising modulators of PPIs, because they can bind to protein surfaces with high affinity and specificity. Nonetheless, computational peptide design remains difficult, hindered by the intrinsic flexibility of peptides and the substantial computational resources required. Recent advances in artificial intelligence (AI) are paving new paths for peptide-based drug design. In this review, we explore the advanced deep generative models for designing target-specific peptide binders, highlight key challenges, and offer insights into the future direction of this rapidly evolving field.

Keywords: peptide design; artificial intelligence; protein-peptide interactions; deep generative models; protein–protein interactions

Introduction

PPIs have a pivotal role in biological processes, such as signal transduction, cellular metabolism, and molecular transport, making them potential targets for drug discovery.[(p1),(p2)] However, the large and shallow nature of PPI interfaces poses significant challenges to designing small molecules that target PPIs with high binding affinity and specificity. Although antibodies effectively recognize these interfaces, their poor membrane permeability limits their applications against intracellular targets. By contrast, peptides offer a balance between molecular flexibility and rigidity, allowing them to bind PPIs with high affinity and specificity. Their small size, coupled with favorable properties, such as ease of synthesis, low toxicity, and minimal immunogenicity, further enhance their clinical potential. For example, glucagon-like peptide 1 (GLP-1) analogs have been successfully used to regulate metabolism in diabetes treatment.[(p3)] These advantages position peptide therapeutics as promising candidates for previously 'undruggable' targets, providing a more effective and biologically natural path for future drug development.

The development of peptide therapeutics traces back almost a century to the isolation of insulin (Figure 1a).[(p4)] However, early efforts were impeded by the complexities of synthesis and purification.[(p5)] Significant breakthroughs occurred during the 1960s with the introduction of solid-phase peptide synthesis (SPPS)[(p6)] and during the 1980s with the advent of recombinant technologies.[(p7)] These advancements revolutionized peptide production, greatly accelerating research and commercialization efforts.[(p8)] Currently, more than 100 peptide-based drugs have been approved by the US Food and Drug Administration (FDA), with many more in development, targeting a range of applications, including immunosuppression, antimicrobial and antiviral therapies, and cancer treatment.[(p9)]

Despite these advancements, peptide drug development has largely relied on natural products or their derivatives, with *de*

* *Corresponding authors. Yao, X. (xjyao@mpu.edu.mo), Hou, T. (tingjunhou@zju.edu.cn).*

https://doi.org/10.1016/j.drudis.2025.104300

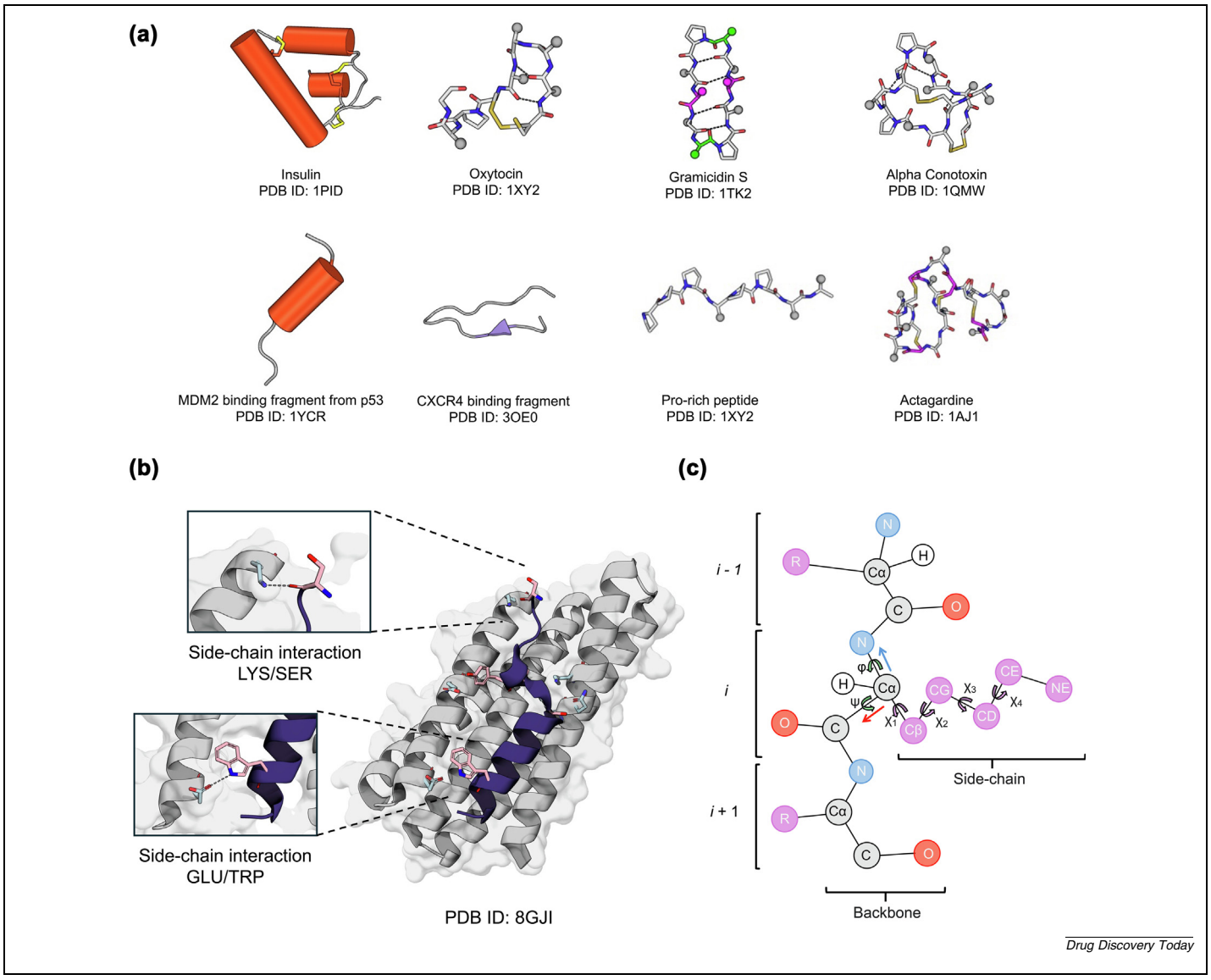

FIGURE 1

Structural diversity and therapeutic applications of peptides targeting protein–protein interactions (PPI). **(a)** Diverse structures of insulin and various natural peptides. Crosslinks are represented by sticks, with magenta highlighting non-canonical amino acids (NCAAs) and green marking D-amino acids; spheres indicate the positions of side-chain C_β atoms and dashed lines denote hydrogen bonds. **(b)** A bioactive helical peptide (glucagon) bound to a protein designed by RFdiffusion,[p12],[p13] with hydrogen bonds highlighted. **(c)** The backbone atoms and side-chain atoms for each residue. Abbreviation: PDB, Protein Data Bank. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

novo peptide design posing a considerable challenge. A major bottleneck stems from the structural characterization of peptides, because their intrinsic flexibility enables them to adopt various conformations, complicating accurate modeling and prediction. Moreover, the scarcity of structural data on protein–peptide complexes, coupled with the incorporation of non-canonical amino acids (NCAAs) and unconventional cyclic structures, further complicates the design process (Figure 1a). Traditional computational approaches, such as molecular docking and molecular dynamics (MD) simulations, are often constrained by their high computational costs, limiting their applicability across a range of design scenarios.[p10],[p11].

AI, particularly deep learning (DL), has emerged as a transformative tool in peptide design, and can process intricate structural data and capture nonlinear patterns with exceptional precision. We start this review with an overview of key data sets related to protein–peptide interactions (PpIs), emphasizing advances in predicting these interactions and modeling protein-peptide complex structures. Next, we summarize methods for target-specific peptide design, demonstrating their practical applications through illustrative case studies that highlight their importance in drug discovery. We also provide a comprehensive overview of recent advances in deep generative models for peptide design. Finally, we discuss the current challenges faced by AI in this field and

conclude by outlining future directions, offering insights and direction for advancing next-generation peptide therapeutics.

Protein–peptide interactions

It has been estimated that 15–40% of all PPIs within cells are mediated by PpIs,[p14] and a comprehensive understanding of PpIs is essential for advancing peptide-based therapeutics. Here, we provide an overview of key data sets related to PpIs, prediction methods for PpIs, and structural modeling approaches for protein–peptide complexes. These advances establish a fundamental structural framework for data-driven and AI-powered peptide design, empowering researchers to develop target-specific peptides more effectively.

Protein–peptide interaction databases

Structural information about PpIs is essential for understanding the properties and functional mechanisms of peptides. PpI databases can be classified into three main types based on their construction purpose: (i) large-scale general data sets; (ii) peptide-specific data sets that include benchmarks; and (iii) customized data sets originating from research studies. Table 1 provides examples of type (ii) and (iii) data sets with >300 entries.

In recent years, DL has gained significant traction in peptide science, empowering tasks such as peptide identification, property prediction, and peptide generation,[p15] fueled by the expanding structural data sourced from repositories, including the Protein Data Bank (PDB).[p16] However, the scarcity of high-quality structural data poses a significant challenge to training large-scale models, thereby hindering advances in peptide-based predictions. Unlike proteins, peptides exhibit high conformational flexibility, which complicates their structural characterization. Most protein–peptide complexes in Table 1 are sourced from PDB, often with significant redundancy and predominantly short peptides, which limits structural diversity and conformational space coverage. In addition, some databases are outdated, with inaccessible websites, with only a few being actively updated alongside PDB. Overall, the availability of high-quality structural data for peptide complexes remains

TABLE 1

Protein–peptide interaction databases

Name	Type	Description and features	Length	Size	Year	Web server	Refs
CPSet	2	Protein-cyclic peptide complex data set sourced from PDB	5–20 residues	493	2024	https://github.com/huifengzhao/CPSet	(p23)
ProPedia	2	Comprehensive data set of experimental protein–peptide complexes, for peptides ranging from 2 to 50 residues	2–50 residues	19 813	2021	https://bioinfo.dcc.ufmg.br/propedia	(p24)
PepBDB	2	Curated structural database of biological peptide-mediated interactions	<50 residues	13 299	2018	https://huanglab.phys.hust.edu.cn/pepbdb	(p25)
PepSite	2	Recently determined, refined X-ray structures	3–20 residues	405	2012	https://pepsite2.russelllab.org (updated on 2021)	(p26),(p27)
PepX	2	Peptide <10 residues from PDB, divided into 505 unique protein-peptide interface clusters.	5–35 residues	1431	2009	https://pepx.switchlab.org (unavailable)	(p28)
PepPC-F/PepPC	3	Buried interfacial peptide fragments with their corresponding binding proteins; total of 232 helical peptides and 3600 nonhelical peptides	8–30 residues	14 897/3832	2024	https://github.com/YuzheWangPKU/DiffPepBuilder	(p29)
$PpI[S/A]_{BM}$	3	Similar data set to $PpI[S/A]_{DS}$ with different scope of K_d and pK_d.	4–28 residues	356	2024	–	(p30)
PPI-Affinity	3	Binding affinity data expressed as dissociation (K_d) or inhibition (K_i) constants	3–29 residues	1149	2022	https://protdcal.zmb.uni-due.de/PPIAffinity (unavailable)	(p31)
SPRINT-Str	3	Protein–peptide complexes filtered for peptides with <0 residues, clustered by 30% sequence identity	<30 residues	1241	2018	https://sparks-lab.org/server/SPRINT-Str (unavailable)	(p32)
PixelDB	3	Nonredundant, high-resolution structures of protein–peptide complexes, filtered to minimize impact of crystal packing on peptide conformation	5–50 residues	1966	2017	https://github.com/KeatingLab/PixelDB	(p33)
PepBind	3	Collection of protein–peptide complex data from PDB, featuring structural, sequence, and experimental information for peptides <35 residues	≤35 residues	3100	2013	https://pepbind.bicpu.edu.in (unavailable)	(p34)

limited compared with small molecules, posing challenges for data-driven research. Importantly, studies have indicated that protein loop regions share structural and dynamic similarities with peptides, making these regions a valuable resource for conformational analysis and benchmarking peptide-modeling methods.[p17]

In protein structure prediction, researchers use alternative sequence data, such as the ~300 million sequences in UniProt,[p18] to overcome the limitations of structural data sets. To enhance prediction accuracy, AlphaFold2[p19] (AF2) leverages multiple sequence alignments (MSAs) and ESMFold[p20] captures coevolutionary signals, highlighting the importance of integrating both sequence and structural information. Another effective solution lies in the use of virtual structural data sets. High-precision structure prediction models can rapidly generate large-scale structural data, facilitating the creation of extensive databases, such as the AlphaFold Protein Structure Database[p21] (AFDB). Furthermore, these databases support rapid screening and the customization of task-specific data sets. Incorporating virtual structures has been demonstrated to significantly improve model performance, marking a crucial advance in structural prediction and drug design.[p22] However, future efforts should focus on improving the success rate of peptide design using expansive external databases to maximize their utility and impact.

Protein–peptide interaction prediction

Accurate prediction of PpIs can guide peptide optimization through effective amino acid substitutions and backbone modifications. Researchers have developed various computational methods to identify peptide-binding residues on protein surfaces. One example is PepBind,[p35] which operates on the premise that protein-binding residues are static and independent of the peptide involved. Nonetheless, different peptides can interact with the same protein through different binding modes, leading to diverse residue interactions, which limits the applicability of PepBind in complex biological systems.

For accurate and efficient prediction of PpIs, it is necessary to integrate both sequence and structural data. InterPep,[p36] a structure-based model, applies random forest algorithm and hierarchical clustering to predict the most likely peptide-binding sites on proteins. However, its dependency on 3D structural data and peptide sequences restricts its applicability to proteins with resolved structures.

To overcome these limitations, Lei *et al.*[p37] introduced CAMP, a DL framework that can simultaneously predicts PpIs and identifies key binding residues within peptides. By combining convolutional neural networks (CNNs) with self-attention mechanisms, CAMP efficiently extracts both local and global features, enabling it to not only predict PpIs, but also identify critical binding sites.

Similarly, Abdin *et al.*[p38] proposed PepNN, a parallel prediction model that integrates sequence and structural data. PepNN takes protein structures and peptide sequences as inputs and generates residue-level scores to evaluate the probability of peptide binding. The model features two unique architectures: PepNN-Struct, which captures structural context using graph attention layers, and PepNN-Seq, which focuses on sequence-based predictions.

Despite the extensive exploration of machine learning (ML) and DL for predicting PpIs, the Molecular Surface Interaction Fingerprinting (MaSIF) framework[p39] represents a transformative, generalized approach that bridges the understanding of PPIs and broader protein–ligand interactions. By harnessing geometric DL, MaSIF directly deciphers interaction fingerprints from protein molecular surfaces, revealing complex patterns with ligands, peptides, and other proteins. Together, these developments are providing deeper insights into PpIs and offering more accurate and efficient tools for peptide-based drug design.

Protein–peptide complex structure prediction

Accurate prediction of protein–peptide complex structures is valuable for effective peptide design. As a primary tool in this endeavor, molecular docking can predict peptide binding modes by optimizing molecular conformations, orientations, and positions on the potential energy surface. Protein–peptide docking methods are broadly categorized into template-based and template-free approaches. Template-based docking uses known complex structures for predictions and performs well in specific tasks, but its applicability is limited by the availability and diversity of templates. By contrast, template-free docking does not require prior structural information, making it more versatile for a broader range of targets, including those without resolved structures. As a result, template-free docking has become a major research focus. Within this category, methods are further divided into local docking (e.g., DynaDock[p40] and Rosetta FlexPepDock[p41]) and global docking (e.g., PIPER-FlexPepDock[p42] and HPEPDOCK[p43]). Furthermore, tools such as AutoDock CrankPep[p44] (ADCP) support flexible cyclic peptide modeling, thereby offering new possibilities for the development of peptide-based therapeutics.

Despite these advances, modeling and scoring protein–peptide complexes remain challenging. Unlike small molecules, peptides exhibit high conformational flexibility and can adopt various structures that adapt dynamically to their chemical environments. It is possible that peptides are disordered when unbound but stabilize into specific conformations upon interacting with proteins. In addition, peptide-target binding commonly relies on water-mediated hydrogen bonds, and modeling interfacial water molecules adds another layer of complexity to this challenge.[p45]

Extensive studies indicate that existing docking methods often struggle to accurately capture the native conformations of peptides,[p23],[p53] constrained by their intrinsic flexibility and the limitations of scoring algorithms.[p54] Nevertheless, MD simulations offer invaluable insights into the thermodynamics, kinetics, and mechanistic details of protein–peptide binding and dissociation. However, the reliability of these simulations depends on the accuracy of the physical models used and the effectiveness of sampling energy landscapes. Unfortunately, achieving exhaustive sampling remains computationally infeasible with existing resources.[p11]

DL provides advanced solutions that transcend traditional limitations by directly learning scoring criteria from data, eliminating the requirement for explicit conformational enumeration. Table 2 presents common tools for biomolecular structure prediction, highlighting representative AI-driven methods cap-

TABLE 2

Common DL-based tools for biomolecular structure prediction

Name	Description and features	Year	Open source	Web server	Refs
AF3/AFM	Cutting-edge AI model by DeepMind that predicts protein structures with near-experimental accuracy, using diffusion-based architecture to model complex biomolecular systems, including proteins, nucleic acids, small molecules, and ions	2024	Inference	https://github.com/google-deepmind/alphafold3	(p46)
RFAA/RoseTTAFold	Biomolecular structure prediction neural network that can predict broad range of biomolecular assemblies, including proteins, nucleic acids, small molecules, covalent modifications, and metals	2024	Inference	https://github.com/baker-laboratory/RoseTTAFold-All-Atom	(p47)
Chai-1	Multi-modal foundation model for molecular structure prediction that performs at state-of-the-art across variety of benchmarks; enables unified prediction of proteins, small molecules, DNA, RNA, glycosylations, and more	2024	Inference	https://github.com/chaidiscovery/chai-lab	(p48)
HelixFold3	Replicates capabilities of AF3 in biomolecular structure prediction, achieving accuracy on par with AF3 for predicting structures of proteins, nucleic acids, and conventional ligands	2024	Inference	https://github.com/PaddlePaddle/PaddleHelix	(p49)
Protenix	Trainable PyTorch reproduction of AF3	2024	Trainable	https://github.com/bytedance/Protenix	(p50)
Boltz-1	SOTA open-source model to predict biomolecular structures containing combinations of proteins, RNA, DNA, and other molecules; also supports modified residues, covalent ligands, and glycans, as well as conditioning prediction on specified interaction pockets or contacts	2024	Trainable	https://github.com/jwohlwend/boltz	(p51)
OpenFold	Trainable, memory-efficient, and GPU-friendly PyTorch reproduction of AF2	2024	Trainable	https://github.com/aqlaboratory/openfold	(p52)

able of predicting protein-peptide complex structures. For example, AF captures structural physics from coevolutionary signals using MSA features, enabling the prediction of atomic-level 3D structures. This approach is effective for modeling peptides up to 40 amino acids with well-defined secondary structures and limited flexibility,[p55] and also supports the prediction of protein–peptide complex structures.[p56] Using accurate prediction of protein–peptide complex structures, Mondal *et al.*[p57] introduced an AF Competition Binding Assay[p58] pipeline to identify the most likely binding polypeptides from peptide libraries, aiding the study of PPIs, epitope identification, and design of high-affinity binding epitopes.

The recently developed AF3[p46] and RFAA[p47] support all-atom modeling, enabling precise structure prediction for protein–ligand complexes, including peptides. These advances not only enhance prediction accuracy and stability, but also introduce innovative strategies for peptide design. However, comprehensive benchmarks on peptide-related structures remain limited. To address this, it is essential to resolve data-quality and cleaning issues, incorporate cutting-edge models, and to develop tailored approaches specifically for short peptides (5–30 residues), which are often overlooked by conventional MSA-based methods.

MSA is crucial for protein structure prediction, but its application in protein–peptide structure prediction is limited because of the shorter length and lesser evolutionary conservation of peptides compared with proteins. AF3 shows potential in modeling non-canonical modifications, such as modified peptides and macrocycles, through its support for user-defined Chemical Component Dictionaries (CCDs). However, the success rate for these structures remains uncertain, necessitating further research to assess its capabilities, especially in predicting protein–peptide complexes. To enhance the performance of AF3 in this area, systematic studies are essential, along with improvements in preprocessing tools, docking functions, structure accuracy measures, and data sets.

Despite the scarcity of literature on AF3-based peptide modeling, initial studies highlight its potential in protein–peptide complex prediction. For example, Manshour *et al.*[p59] evaluated AFM, ColabFold (CF), and AF3 using a benchmark data set of 60 protein–peptide complexes. AF3 generated high-quality structures with fewer models compared with AFM, which relied on a larger model pool. However, the performance of AF3 was limited by its small model pool, accessible solely via a web server, emphasizing the trade-off between model pool size and computational resources in protein–peptide complex prediction.

As DL tools evolve and data sets expand, future evaluation pipelines are expected to improve, especially for nonstandard and macrocyclic peptides. Although AF3 shows considerable promise in PpI modeling, its application to modified peptides and macrocycles requires further validation. With advances in evaluation methods and technologies, the role of AF3 in peptide-based drug design is expected to expand significantly.

Target-specific peptide drug design

Despite progress in rational peptide design and combinatorial chemistry,[p60],[p61] existing methods continue to struggle with achieving a balance between efficiency and accuracy. A key obstacle is that many functional peptides exist in disordered states or can freely transition between multiple conformations, complicating the design process. Designing peptide binders for protein targets is challenging because of the need to accurately

predict their optimal bound conformations. Moreover, the lack of secondary structures, typically found in proteins, introduces additional constraints to peptide design.

ML-based approaches have significantly advanced the field of *de novo* protein design, with diffusion-based generative methods becoming increasingly central to modern design pipelines[p62] (Figure 2a). The boundary between peptide design models and protein design has become blurred, because many models now extensively use methodologies originally devised for proteins. For the sake of consistency and clarity, the well-established classification framework from protein design was adopted.[p63] For a foundational understanding, readers may refer to recent literature on diffusion models in structural biology.[p64] Here, we focus on two widely used strategies in *de novo* peptide binder design: hallucination-based methods and structure–sequence co-design methods (Figure 2b).

Hallucination-based methods

Various hallucination methods for protein design have been developed, all aimed at generating novel sequences that fold into stable, unseen structures. These methods optimize random sequences using structure prediction algorithms, such as AF and iterative techniques, including Markov chain Monte Carlo (MCMC), guided by folding-aware loss functions[p65],[p66] (Figure 2b). A well-established framework for designing peptides targeting PPIs is ColabDesign,[p67] a *de novo* protein design pipeline powered by fold-based models. Building upon this, Kosugi and Ohue[p68] introduced a solubility-aware extension that incorporates a solubility loss function based on amino acid solubility indices. This refinement improved the ColabDesign binder hallucination protocol, enhancing the solubility of the generated sequences by weighting the solubility loss function.

Bryant and Elofsson[p69] later introduced EvoBind, an advanced framework for peptide binder design that integrates multiple computational tools to streamline the design process. The framework initially uses Foldseek[p70] to generate seed structures, followed by ESM-IF1[p71] for inverse folding to create sequences that align with the predicted backbone structures. These protein–peptide complexes are then evaluated using AF to ensure binding stability and accuracy. Notably, EvoBind generates successful binders with interface RMSD $\leq$ 2 Å for 185 (6.5%) heteromeric and 42 (3.6%) homomeric protein interfaces, significantly outperforming ProteinMPNN,[p72] which achieves 18 (1.5%) successful designs from the same 100 samples. Here, ProteinMPNN is developed for protein sequence design as AF is designed for protein structure prediction. This DL-based algorithm predicts amino acid sequences for specified protein backbones and is widely used as a benchmark in the field of protein inverse folding.[p10],[p11]

The newly introduced EvoBind2[p73] revolutionizes peptide binder design by using only the amino acid sequence of the target protein, without requiring the prior knowledge of binding sites, templates, or binder lengths, making it suitable for novel targets. A key issue in peptide design is avoiding adversarial

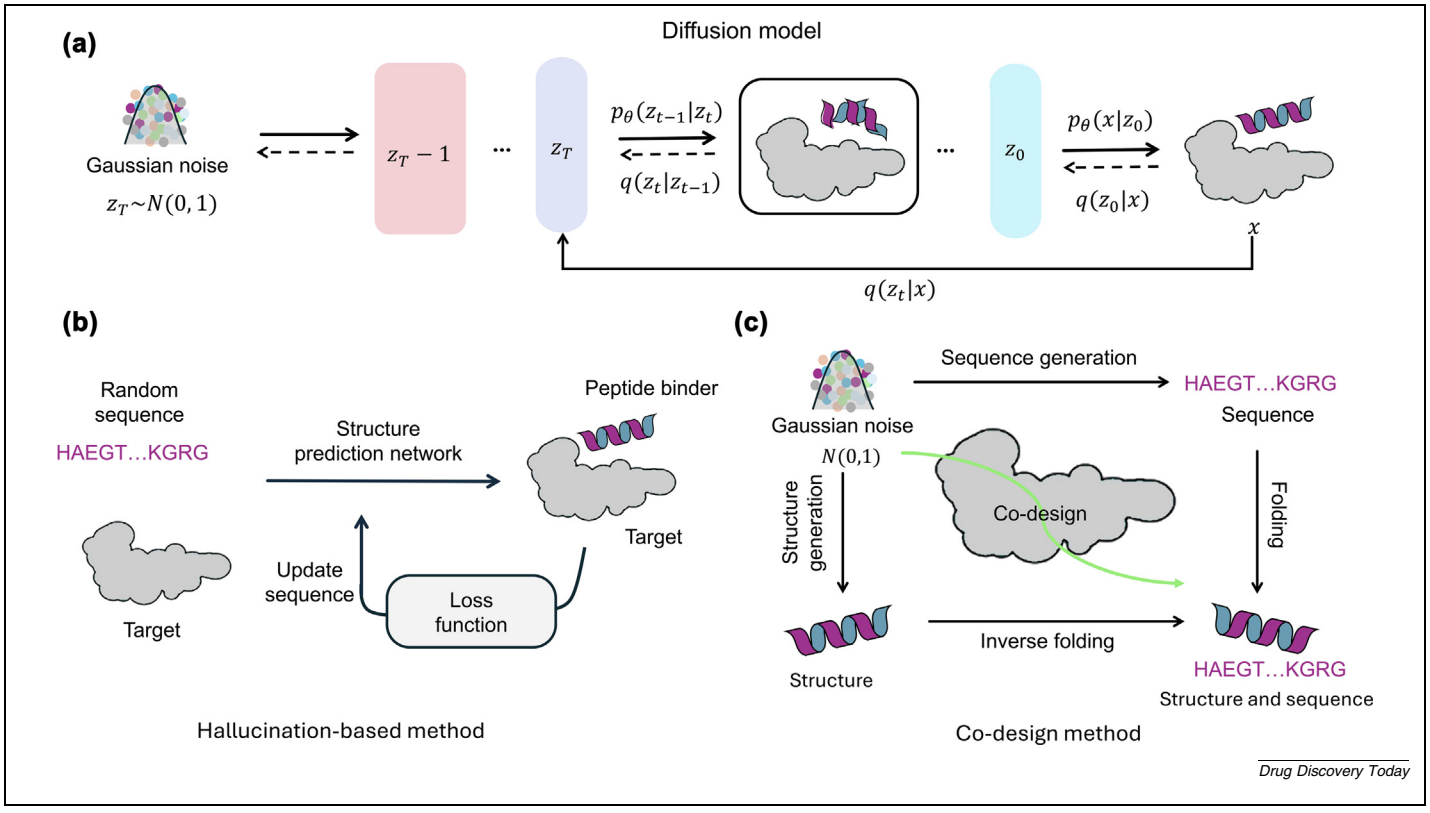

FIGURE 2

The diffusion model and the peptide design pipeline. **(a)** Diffusion model for target-specific peptide binder design. Peptide binder design with **(b)** the hallucination-based and **(c)** co-design methods.

sequences with high predicted local difference distance test (pLDDT) scores but weak binding. To address this, integrating AFM for validation significantly reduces false positives and triples the success rate, ensuring more reliable designs. EvoBind2 also supports cyclic peptide input through a cyclic offset.[p74],[p75] Experimental validation demonstrated that the designed peptides achieved binding affinities ranging from 5.7 μM to 0.26 nM for cyclic peptides and from 7.9 μM to 19 nM for linear peptides, with success rates of 75% and 46%, respectively.

Iterative optimization methods can be enhanced with reinforcement learning (RL) for more effective results. Wang *et al.*[p76] introduced EvoPlay, a self-play RL framework designed to optimize protein sequences for specific functional or structural goals. Both EvoBind and EvoPlay aim to design high-affinity peptide binders, but EvoPlay offers a more efficient and robust solution by integrating RL with look-ahead Monte Carlo Tree Search (MCTS). This ability to balance exploration and exploitation gives EvoPlay an advantage over EvoBind, which is prone to local optima and higher computational costs. EvoPlay has been successfully applied to design peptide binders and optimize proteins, such as GFP and PAB1, for enhanced function, highlighting its broad potential in tackling various protein-engineering challenges.

Hallucination stands out for its simplicity and flexibility, enabling novel peptide design tasks by integrating new loss functions into existing structure-prediction models.[p77] Many advances in protein structure design are transferable to peptide research, allowing for precise exploration of peptide-specific structural and sequence spaces. For example, Verkuil[p78] applied sequence-based methods, leveraging language models trained solely on sequence data to explore a broader space of natural proteins beyond conventional structure-based approaches. Another example of enhanced sequence optimization is provided by Frank *et al.*,[p79] who hypothesized that gradient descent-based hallucination could be improved by relaxing the constraints of discrete (i.e., physically realistic) protein sequence space. This relaxed sequence optimization (RSO) approach offers notable benefits, including greater designability, broader applicability across diverse design challenges, and scalability to proteins of different sizes.

Exciting progress has been made in peptide inverse folding. Models, such as ProteinMPNN, a message-passing encoder–decoder, have a key role in structure-based design by predicting amino acid sequences that fold into desired structures. However, when applied to peptides, they often generate repetitive sequences that fail to match the reference structure. To overcome this, Park *et al.*[p80] fine-tuned ProteinMPNN using direct preference optimization (DPO), introducing two major improvements: online diversity regularization and domain-specific priors. These enhancements not only promote more diverse sequence generation, but also ensure high structural fidelity. When conditioned on OpenFold-generated[p52] structures, their method achieves state-of-the-art similarity scores, improving the baseline of ProteinMPNN by over 8% and increasing sequence diversity by up to 20% without compromising structural accuracy.

A major advantage of this framework is its ability to leverage advancements in protein design for rapid adaptation to peptides. By building on the structure prediction network, it facilitates precise exploration of broad structure and sequence spaces, thereby enabling the generation of high-quality peptide structures for effective design, optimization, and future applications.

Sequence-structure co-design methods

Existing DL methods for peptide design fall into structure-based and sequence-based approaches. Structure-based design generates a peptide backbone first, followed by a compatible sequence, leveraging detailed structural information but suffering from constraints resulting from limited and biased data sets. By contrast, sequence-based methods directly generate sequences, allowing for generalization to broader domains or disordered regions, thus often resulting in noisier predictions because of limited structural guidance. To overcome these limitations, emerging co-generation models integrate sequence and structure reasoning throughout the design process, improving accuracy and consistency, and enabling the design of peptides with complex conformations and dynamic properties (Figure 2c).

An outstanding co-design model, DiffPepBuilder, was developed by Wang *et al.*[p29] This model uses an SE(3)-equivariant diffusion architecture, incorporating protein language model (pLM)[p81] embeddings and positional encodings as node features, while using a distogram to encode edge information. It converts 3D coordinates into local reference frames, which interact via a Cross Update Module. The multitask decoder then outputs translational and rotational scores, predicted residue types, torsion angles, and residue entropies. To enhance peptide stability and binding potency, the authors introduced an SSBuilder module within DiffPepBuilder to strategically design disulfide bonds. MD simulations on 30 validated peptide binders confirmed that disulfide bonds increased peptide rigidity and improved binding performance. Comparative studies on three biological targets demonstrated that DiffPepBuilder outperformed ColabDesign and RFdiffusion (with ProteinMPNN) in terms of recall, interface quality, and structural diversity.

Considering the non-conserved nature of peptide backbones, co-designing both peptide sequence and structure remains particularly challenging. Models addressing this challenge often rely on all-atom representations to capture subtle side-chain interactions. In this regard, DiffPepBuilder encodes side-chain atoms of each residue using frames parameterized within the SE(3) manifold, providing flexible handling of varying atom types and counts across different residues.

PepGLAD, introduced by Kong *et al.*,[p82], tackles two major challenges in peptide design: the intricacies of full-atom geometry and the variability of binding conformations. This geometric latent diffusion model leverages a VAE to encode residues of varying sizes into fixed-dimensional latent spaces, improving the efficiency of diffusion processes. In addition, through receptor-specific affine transformations, it aligns peptide 3D coordinates within a common space, boosting the generalization capabilities of the model. PepGLAD enhances peptide sequence–structure co-design diversity by 18%, *in silico* success rates by 8%, and recovery of reference binding conformations by 26%.

Beyond diffusion models, flow-matching frameworks have emerged as powerful tools in peptide design. PepFlow, developed by Li *et al.*,[p83] is a multimodal generative model based on the flow-matching framework. It captures residue backbone orienta-

tions and side-chain dynamics crucial to PpIs by representing rigid backbones in SE(3) space and encoding side-chain angles on high-dimensional tori. The peptide sequence is represented as categorical distributions over the probability simplex. By learning joint distributions across these modalities via flows and vector fields on their corresponding manifolds, PepFlow excels in tasks such as fix-backbone sequence design and side-chain packing. For sequence-structure co-design, PepFlow excels in metrics including geometry (e.g., AAR, RMSD), energy (e.g., stability, affinity), and diversity. However, it lags behind RFdiffusion in terms of designability. Despite this limitation, extensive benchmarking underscores the robust performance of PepFlow and significant potential to advance computational peptide design methodologies.

Sequence-based methods

The success of structure-based peptide binder design, as discussed above, typically relies on high-resolution co-crystal structures for accurate modeling. However, high-quality structural data remain scarce, and structure-based methods are inherently limited by the static nature of such data, diminishing their effectiveness for proteins that undergo dynamic structural transitions. Notably, this approach faces significant challenges when applied to disordered or unstable proteins, such as certain transcription factors, because of their crystallization difficulties and tendency to adopt multiple conformations.[p84]

To address these challenges, pLMs have introduced a sequence-based paradigm for target-specific peptide binder design. These large-scale models are trained on vast protein sequence data sets, capturing not only key physicochemical properties, but also higher-order structural features. Among these, ESM2 stands out for its robustness, leveraging masked language modeling (MLM) tasks to predict protein functions, design antibodies, and even predict protein structures.[p20] Building on the advances in pLMs, *de novo* peptide design methods have emerged, proving particularly valuable for proteins lacking reliable structural data or those previously considered 'undruggable'.

PepMLM[p85] is a representative pLM-based method specifically designed for generating peptides targeting protein sequences. It functions by placing a contiguous mask at the C terminus of a target protein sequence, representing the peptide yet to be generated, and then uses ESM2 to reconstruct the masked region, producing high-affinity peptide binders. Empirical results demonstrate that PepMLM achieves a hit rate exceeding 38%, significantly outperforming RFdiffusion. In addition, when integrated into a ubiquitin-binding antibody (ubiAb) system, it shows promising potential in degrading intrinsically disordered proteins, such as TRIM8.

For applications requiring even higher specificity, especially when targeting discrete motifs, moPPIt offers a motif-specific PPI-targeting algorithm.[p86] Central to this approach is BindEvaluator, a transformer-based model that interpolates between two pLM embeddings using multi-headed self-attention, prioritizing local motif features. Trained on over 510 000 annotated PPI data points, BindEvaluator achieves an impressive test AUC of >0.94, which increases above 0.96 when fine-tuned on protein–peptide pairs. By combining BindEvaluator with PepMLM and a genetic optimization step, moPPIt generates peptides that selectively bind key residues on a target protein. Notably, moPPIt extends beyond known targets, successfully accommodating previously unexplored structured or disordered proteins, offering a robust solution for dynamic or structurally elusive targets.

Another sequence-based framework for designing target-specific peptides, Cut&CLIP, was introduced by Palepu *et al.*[p87] This method integrates pretrained protein embeddings with contrastive learning to design peptides that not only bind to target proteins, but also induce degradation via an E3 ubiquitin ligase domain. By jointly encoding both proteins and candidate peptides, the model captures essential similarities between known protein–peptide pairs. Experimental validation demonstrated that fusing the generated peptides with ubiAb constructs consistently led to the degradation of pathogenic proteins in human cells, highlighting the effectiveness of this framework for peptide-mediated protein degradation.

Collectively, pLMs are proving to be highly promising complementary tools to structure-based peptide design. This success marks a significant milestone in sequence-level peptide design, significantly expanding the potential for programmable proteome editing and novel strategies against traditionally 'undruggable' targets. When combined with insights into protein degradation mechanisms such as ubiAb, these models open exciting new avenues for precise protein control and the development of next-generation therapeutics. As pLMs continue to evolve, we can anticipate further breakthroughs in *de novo* peptide design, programmable protein editing, and targeted drug development.

Evaluation metrics

In peptide design, two primary evaluation metrics are commonly used: self-consistency and diversity. Self-consistency assesses the alignment of generated sequences with their corresponding backbone structures, with methods such as ProteinMPNN serving as the standard.[p64] Diversity measures the ability of the model to generalize beyond the training data by calculating backbone RMSD or TM-scores[p88] through alignment with structural data from sources such as the PDB or AF. However, the absence of standardized benchmarks complicates comparisons across different peptide design studies and models for antibodies or protein binders. Although experimental validation serves as a gold standard for some self-consistency assessments, most novel co-design models still heavily rely on MD simulations.

Binding energy is key metric for assessing the stability of target–peptide interactions and selecting top peptides for further analysis. Rosetta binding energy[p89] is commonly used to rank the generated peptides. Confidence scores from AF also serve as proxies for binding affinity, assisting peptide ranking.[p59] The interface predicted template modeling (ipTM) score is effective for evaluating PPIs and is similarly applied to rank peptides and assess protein–peptide structure predictions.[p90] However, the ipTM score has limitations, particularly its reliance on full-length binding partners, which can introduce errors, especially in sequences with unstructured or flexible regions. These issues are more pronounced in PpIs, where binding interfaces are typically short and flexible. To address these challenges, the actual

interface pTM (actifpTM)[p91] score was introduced, focusing on confident interface residues for a more accurate measure of interaction confidence, minimizing the influence of unstructured or flexible regions.

Looking forward, future research should prioritize establishing comprehensive data sets and benchmarks specifically tailored to peptide design. An integrated peptide design platform would encompass extensive data-driven model training, thorough evaluations, and experimental validations. Such a platform would lay the groundwork for the development of more effective peptide-based therapeutics.

Challenges and future perspectives

Early drug development was dominated by the Lipinski's rule-of-five, favoring small molecules with molecular weights <500 Da, clogP <5, fewer than ten H-bond acceptors, and fewer than five H-bond donors to ensure favorable oral bioavailability.[p92] This guideline initially cast doubt on the feasibility of larger molecules, such as proteins and peptides, as therapeutic agents.[p93] Despite natural peptides having crucial roles in regulating membrane receptors and secretory proteins, their therapeutic potential is limited by poor thermal stability, rapid protease degradation, low binding affinity, and short half-lives, ultimately leading to low oral bioavailability.

To overcome these limitations, expanding the peptide chemical space through innovative strategies has become essential. For example, the introduction of NCAAs, *N*-methylation, and advanced cyclization techniques offers new avenues to enhance peptide stability and efficacy. Among these strategies, cyclic peptides stand out in medicinal chemistry,[p94] attracting significant attention for their ability to constrain conformational flexibility, thereby reducing degradation, instability, and poor membrane permeability, which are key limitations of peptide-based therapeutics.[p95]

Despite advancements in AI-driven models, existing structure-based models for cyclic peptide design predominantly focus on conventional cyclization strategies, such as N-to-C terminal cyclization or disulfide bond formation. Unconventional methods, such as thioether cyclization, remain underexplored. The recent development of RFpeptides,[p96] based on RFdiffusion, enables the design of macrocyclic peptide binders for diverse protein targets. By incorporating cyclic relative position encodings and leveraging ProteinMPNN for sequence design, RFpeptides can create high-affinity binders with exceptional accuracy. This approach greatly enhances design efficiency, offering a powerful tool for therapeutic and diagnostic applications.

Furthermore, the integration of NCAAs into DL-based models has yet to be fully realized, limiting the chemical diversity available for peptide design. Future work could integrate these strategies into design frameworks by leveraging approaches such as the use of CCD by AF3 for modification data inputs. This would enhance the versatility and applicability of diffusion-based models, enabling more diverse and effective cyclic peptide designs. Although sequence-based peptide drug development offers broader applications across various therapeutic contexts, PepINVENT[p97] exemplifies this progress by using SMILES notation to represent peptide molecules. This approach captures non-standard components, including NCAAs and unconventional cyclic structures. Designed with the needs of pharmaceutical researchers in mind, it bridges computational methods with real-world drug discovery.

Concluding remarks

In summary, the future of peptide design depends on our ability to understand and translate peptide structure–function relationships into computational frameworks. Although AI has already demonstrated significant potential in peptide drug design, no AI-assisted peptide drugs have yet been approved by the US Food and Drug Administration (FDA). Most FDA-approved peptide drugs still rely on traditional drug discovery methods. However, with the progression of AI-driven techniques, this field is moving beyond traditional physics-based strategies toward data-driven approaches, unlocking new opportunities for innovative drug discovery.

For target specific peptide binder design, although binding affinity is the key parameter in the initial phase in therapeutic development, high-affinity peptides must also satisfy pharmacokinetics. Future models should take a multi-objective approach, balancing efficacy, stability, and drug-like properties, to bridge computational predictions with real-world applications and push forward peptide-based drug discovery.

An equally important consideration is peptide binder specificity. High-affinity binders are of limited practical use if they do not achieve high on-target specificity. Although structure-based models often focus on optimizing binding energy, they might overlook the step of validating whether the binder can selectively target its intended protein. Future structure-based approaches should not prioritize structural accuracy alone; they must also consider multiple objectives, such as specificity, to minimize off-target effects.

Peptide-based drug design is progressing, especially for well-characterized receptors. New users should start with versatile models, such as AF3, whereas advanced users might explore diffusion-based or flow-matching DL models for specialized tasks. However, challenges persist, particularly in peptide flexibility, nonstandard modifications, and computational efficiency. Furthermore, combining molecular simulations with AI offers a synergistic approach to enhance peptide design strategies. Incorporating experimental validation is essential to ensure the accuracy, reliability, and real-world applicability of computational methods. This integrated approach will advance the development of peptide-based therapeutics with both high specificity and efficacy.

CRediT authorship contribution statement

Silong Zhai: Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **Tiantao Liu:** Writing – original draft. **Shaolong Lin:** Writing – original draft. **Dan Li:** Writing – review & editing. **Huanxiang Liu:** Writing – review & editing, Writing – original draft. **Xiaojun Yao:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Tingjun Hou:** Writing – review & editing, Supervision, Resources, Project administration, Conceptualization.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported financially by National Key R&D Program of China (2024YFA1307501), Natural Science Foundation of Zhejiang Province of China (LD22H300001), Science and Technology Development Fund, Macau, SAR (No. 0030/2024/RIA1), and Macao Polytechnic University (No. RP/FCA-15/2023). The manuscript was approved by Macao Polytechnic University with the submission code s/c fca.0dba.1174.0. To advance AI-driven peptide drug discovery, we have established a continuously updated GitHub repository (https://github.com/zhaisilong/awesome-peptide), which curates the latest and most impactful research in the field.

References

1. Lu H et al. Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. *Signal Transduct Target Ther*. 2020;5:213.
2. Zhang J, Durham J, Cong Q. Revolutionizing protein–protein interaction prediction with deep learning. *Curr Opin Struct Biol*. 2024;85, 102775.
3. Drucker DJ, Nauck MA. The incretin system: glucagon-like peptide-1 receptor agonists and dipeptidyl peptidase-4 inhibitors in type 2 diabetes. *Lancet*. 2006;368:1696–1705.
4. Banting FG, Best CH, Collip JB, Campbell WR, Fletcher AA. Pancreatic extracts in the treatment of diabetes mellitus. *Can Med Assoc J*. 1922;12:141.
5. Sharma K, Sharma KK, Sharma A, Jain R. Peptide-based drug discovery: current status and recent advances. *Drug Discov Today*. 2023;28, 103464.
6. Merrifield RB. Solid phase peptide synthesis. I. The synthesis of a tetrapeptide. *J Am Chem Soc*. 1963;85:2149–2154.
7. Johnson IS. Human insulin from recombinant DNA technology. *Science*. 1983;219:632–637.
8. Muttenthaler M, King GF, Adams DJ, Alewood PF. Trends in peptide drug discovery. *Nat Rev Drug Discov*. 2021;20:309–325.
9. Chen Z, Wang R, Guo J, Wang X. The role and future prospects of artificial intelligence algorithms in peptide drug development. *Biomed Pharmacother*. 2024;175, 116709.
10. Chang L, Mondal A, Singh B, Martínez-Noa Y, Perez A. Revolutionizing peptide-based drug discovery: advances in the post-AlphaFold era. *Wires Comput Mol Sci*. 2024;14:e1693.
11. Mondal A, Chang L, Perez A. Modelling peptide–protein complexes: docking, simulations and machine learning. *QRB Discov*. 2022;3:e17.
12. Vázquez Torres S et al. De novo design of high-affinity binders of bioactive helical peptides. *Nature*. 2024;626:435–442.
13. Watson JL et al. De novo design of protein structure and function with RFdiffusion. *Nature*. 2023;620:1089–1100.
14. London N, Raveh B, Schueler-Furman O. Druggable protein–protein interactions – from hot spots to hot segments. *Curr Opin Chem Biol*. 2013;17:952–959.
15. Wan F, Kontogiorgos-Heintz D, de la Fuente-Nunez C. Deep generative models for peptide design. *Digit Discov*. 2022;1:195–208.
16. Berman HM et al. The Protein Data Bank. *Nucleic Acids Res*. 2000;28:235–242.
17. Pelay-Gimeno M, Glas A, Koch O, Grossmann TN. Structure-based design of inhibitors of protein–protein interactions: mimicking peptide binding epitopes. *Angew Chem Int Ed*. 2015;54:8896–8927.
18. UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res*. 2023;51:D523–D531.
19. Jumper J et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583–589.
20. Lin Z et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*. 2023;379:1123–1130.
21. Varadi M et al. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res*. 2024;52:D368–D375.
22. Ma W et al. Enhancing protein function prediction performance by utilizing AlphaFold-predicted protein structures. *J Chem Inf Model*. 2022;62:4008–4017.
23. Zhao H et al. Comprehensive evaluation of 10 docking programs on a diverse set of protein–cyclic peptide complexes. *J Chem Inf Model*. 2024;64:2112–2124.
24. Martins PM et al. Propedia: a database for protein–peptide identification based on a hybrid clustering algorithm. *BMC Bioinf*. 2021;22:1.
25. Wen Z, He J, Tao H, Huang SY. PepBDB: a comprehensive structural database of biological peptide–protein interactions. *Bioinformatics*. 2019;35:175–177.
26. Petsalaki E, Stark A, García-Urdiales E, Russell RB. Accurate prediction of peptide binding sites on protein surfaces. *PLOS Comput Biol*. 2009;5, e1000335.
27. Trabuco LG, Lise S, Petsalaki E, Russell RB. PepSite: prediction of peptide-binding sites from protein surfaces. *Nucleic Acids Res*. 2012;40:W423–W427.
28. Vanhee P et al. PepX: a structural database of non-redundant protein–peptide complexes. *Nucleic Acids Res*. 2010;38(Suppl. 1):D545–D551.
29. Wang F, Wang Y, Feng L, Zhang C, Lai L. Target-specific de novo peptide binder design with DiffPepBuilder. *J Chem Inf Model*. 2024;64:9135–9149.
30. Wang S, Ye H, Shang S, Li Z, Peng Y, Zhou P. A structure-based data set of protein-peptide affinities and its nonredundant benchmark: potential applications in computational peptidology. *Curr Med Chem*. 2024;31:4127–4137.
31. Romero-Molina S et al. PPI-Affinity: a web tool for the prediction and optimization of protein–peptide and protein–protein binding affinity. *J Proteome Res*. 2022;21:1829–1841.
32. Taherzadeh G, Zhou Y, Liew AWC, Yang Y. Structure-based prediction of protein–peptide binding regions using Random Forest. *Bioinformatics*. 2018;34:477–484.
33. Frappier V, Duran M, Keating AE. PixelDB: protein–peptide complexes annotated with structural conservation of the peptide binding mode. *Protein Sci*. 2018;27:276–285.
34. Das AA, Sharma OP, Kumar MS, Krishna R, Mathur PP. PepBind: a comprehensive database and computational tool for analysis of protein–peptide interactions. *Genom Proteom Bioinf*. 2013;11:241–246.
35. Zhao Z, Peng Z, Yang J. Improving sequence-based prediction of protein–peptide binding residues by introducing intrinsic disorder and a consensus method. *J Chem Inf Model*. 2018;58:1459–1468.
36. Johansson-Åkhe I, Mirabello C, Wallner B. Predicting protein–peptide interaction sites using distant protein complexes as structural templates. *Sci Rep*. 2019;9:4267.
37. Lei Y et al. A deep-learning framework for multi-level peptide–protein interaction prediction. *Nat Commun*. 2021;12:5465.
38. Abdin O, Nim S, Wen H, Kim PM. PepNN: a deep attention model for the identification of peptide binding sites. *Commun Biol*. 2022;5:503.
39. Gainza P et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods*. 2020;17:184–192.
40. Antes I. DynaDock: a new molecular dynamics-based algorithm for protein–peptide docking including receptor flexibility. *Proteins Struct Funct Bioinforma*. 2010;78:1084–1104.
41. London N, Raveh B, Cohen E, Fathi G, Schueler-Furman O. Rosetta FlexPepDock web server: high resolution modeling of peptide–protein interactions. *Nucleic Acids Res*. 2011;39(Suppl. 2):W249–W253.
42. Alam N, Goldstein O, Xia B, Porter KA, Kozakov D, Schueler-Furman O. High-resolution global peptide-protein docking using fragments-based PIPER–FlexPepDock. *PLOS Comput Biol*. 2017;13, e1005905.
43. Zhou P, Jin B, Li H, Huang SY. HPEPDOCK: a web server for blind peptide–protein docking based on a hierarchical algorithm. *Nucleic Acids Res*. 2018;46:W443–W450.
44. Zhang Y, Sanner MF. Docking flexible cyclic peptides with AutoDock CrankPep. *J Chem Theory Comput*. 2019;15:5161–5168.
45. Malde AK, Hill TA, Iyer A, Fairlie DP. Crystal structures of protein-bound cyclic peptides. *Chem Rev*. 2019;119:9861–9914.
46. Abramson J et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*. 2024;630:493–500.
47. Krishna R et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science*. 2024;384, eadl2528.
48. Chai Discovery et al. Chai-1: decoding the molecular interactions of life Published online October 11, 2024. *bioRxiv*. 2024. https://doi.org/10.1101/2024.10.10.615955.

49. Liu L et al. Technical report of HelixFold3 for biomolecular structure prediction Published online August 30, 2024. *arXiv*. 2024. https://doi.org/10.48550/arXiv.2408.16975.
50. Chen X et al. Protenix: advancing structure prediction through a comprehensive AlphaFold3 reproduction Published online January 11, 2025. *bioRxiv*. 2025. https://doi.org/10.1101/2025.01.08.631967.
51. Wohlwend J et al. Boltz-1 democratizing biomolecular interaction modeling Published Online November 20, 2024. *bioRxiv*. 2024. https://doi.org/10.1101/2024.11.19.624167.
52. Ahdritz G et al. OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nat Methods*. 2024;21:1514–1524.
53. Weng G et al. Comprehensive evaluation of fourteen docking programs on protein–peptide complexes. *J Chem Theory Comput*. 2020;16:3959–3969.
54. Ciemny M et al. Protein–peptide docking: opportunities and challenges. *Drug Discov Today*. 2018;23:1530–1537.
55. McDonald EF, Jones T, Plate L, Meiler J, Gulsevin A. Benchmarking AlphaFold2 on peptide structure prediction. *Structure*. 2023;31:111–119.
56. Tsaban T, Varga JK, Avraham O, Ben-Aharon Z, Khramushin A, Schueler-Furman O. Harnessing protein folding neural networks for peptide–protein docking. *Nat Commun*. 2022;13:176.
57. Mondal A et al. A computational pipeline for accurate prioritization of protein-protein binding candidates in high-throughput protein libraries. *Angew Chem Int Ed*. 2024;63, e202405767.
58. Chang L, Perez A. Ranking peptide binders by affinity with AlphaFold. *Angew Chem Int Ed*. 2023;62, e202213362.
59. Manshour N, Ren JZ, Esmaili F, Bergstrom E, Xu D. Comprehensive evaluation of AlphaFold-Multimer, AlphaFold3 and ColabFold, and Scoring Functions in predicting protein–peptide complex structures Published Online November 11, 2024. *bioRxiv*. 2024. https://doi.org/10.1101/2024.11.11.622992.
60. Vanhee P, van der Sloot AM, Verschueren E, Serrano L, Rousseau F, Schymkowitz J. Computational design of peptide ligands. *Trends Biotechnol*. 2011;29:231–239.
61. Gupta S, Azadvari N, Hosseinzadeh P. Design of protein segments and peptides for binding to protein targets. *BioDesign Res*. 2022;2022, 9783197.
62. Chu AE, Lu T, Huang PS. Sparks of function by de novo protein design. *Nat Biotechnol*. 2024;42:203–215.
63. Wang C, Alamdari S, Domingo-Enrich C, Amini A, Yang KK. Towards deep learning sequence-structure co-generation for protein design Published online October 2, 2024. *arXiv*. 2024. https://doi.org/10.48550/arXiv.2410.01773.
64. Yim J, Stärk H, Corso G, Jing B, Barzilay R, Jaakkola TS. Diffusion models in protein structure and docking. *Wires Comput Mol Sci*. 2024;14:e1711.
65. Anishchenko I et al. De novo protein design by deep network hallucination. *Nature*. 2021;600:547–552.
66. Wicky BIM et al. Hallucinating symmetric protein assemblies. *Science*. 2022;378:56–61.
67. Sokrypton OS, Rettie S, Favor A, Batra H, Amani K. sokrypton/ColabDesign: v1.1.2. Published online August 2024. https://github.com/sokrypton/ColabDesign [Accessed January 15, 2025].
68. Kosugi T, Ohue M. Solubility-aware protein binding peptide design using AlphaFold. *Biomedicines*. 2022;10:1626.
69. Bryant P, Elofsson A. Peptide binder design with inverse folding and protein structure prediction. *Commun Chem*. 2023;6:229.
70. van Kempen M et al. Fast and accurate protein structure search with Foldseek. *Nat Biotechnol*. 2024;42:243–246.
71. Hsu C et al. Learning inverse folding from millions of predicted structures Published online March 10, 2022. *bioRxiv*. 2022. https://doi.org/10.1101/2022.04.10.487779.
72. Dauparas J et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science*. 2022;378:49–56.
73. Li Q, Vlachos EN, Bryant P. Design of linear and cyclic peptide binders of different lengths from protein sequence information Published online October 12, 2024. *bioRxiv*. 2024. https://doi.org/10.1101/2024.06.20.599739.
74. Rettie SA et al. Cyclic peptide structure prediction and design using AlphaFold Published Online February 26, 2023. *bioRxiv*. 2023. https://doi.org/10.1101/2023.02.25.529956.
75. Kosugi T, Ohue M. Design of cyclic peptides targeting protein–protein interactions using AlphaFold. *Int J Mol Sci*. 2023;24:13257.
76. Wang Y et al. Self-play reinforcement learning guides protein engineering. *Nat Mach Intell*. 2023;5:845–860.
77. Wang J et al. Scaffolding protein functional sites using deep learning. *Science*. 2022;377:387–394.
78. Verkuil, R. et al., Language models generalize beyond natural proteins. Published online December 22, 2022. http://dx.doi.org/10.1101/2022.12.21.521521.
79. Frank C et al. Scalable protein design using optimization in a relaxed sequence space. *Science*. 2024;386:439–445.
80. Park R et al. Improving inverse folding for peptide design with diversity-regularized direct preference optimization Published online October 25, 2024. *arXiv*. 2024. https://doi.org/10.48550/arXiv.2410.19471.
81. Ruffolo JA, Madani A. Designing proteins with language models. *Nat Biotechnol*. 2024;42:200–202.
82. Kong X, Huang W, Liu Y. Full-atom peptide design with geometric latent diffusion Published online February 21, 2024. *arXiv*. 2024. https://doi.org/10.48550/arXiv.2402.13555.
83. Li J et al. Full-atom peptide design based on multi-modal flow matching Published online June 2, 2024. *arXiv*. 2024. https://doi.org/10.48550/arXiv.2406.00735.
84. Das P, Matysiak S, Mittal J. Looking at the disordered proteins through the computational microscope. *ACS Cent Sci*. 2018;4:534–542.
85. Chen T, Pertsemlidis S, Chatterjee P. PepMLM: target sequence-conditioned generation of peptide binders via masked language modeling. In: *ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design*. https://openreview.net/forum?id=p6fz0rq7zu [Accessed January 15, 2025].
86. Chen T, Zhang Y, Chatterjee P. moPPIt: de novo generation of motif-specific binders with protein language models Published online August 1, 2024. *bioRxiv*. 2024. https://doi.org/10.1101/2024.07.31.606098.
87. Palepu K et al. Design of peptide-based protein degraders via contrastive deep learning Published online May 24, 2022. *bioRxiv*. 2022. https://doi.org/10.1101/2022.05.23.493169.
88. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins Struct Funct Bioinforma*. 2004;57:702–710.
89. Alford RF et al. The Rosetta All-Atom Energy Function for macromolecular modeling and design. *J Chem Theory Comput*. 2017;13:3031–3048.
90. Wee J, Wei GW. Benchmarking AlphaFold3's protein–protein complex accuracy and machine learning prediction reliability for binding free energy changes upon mutation Published online June 6, 2024. *arXiv*. 2024. https://doi.org/10.48550/arXiv.2406.03979.
91. Varga JK, Ovchinnikov S, Schueler-Furman O. actifpTM: a refined confidence metric of AlphaFold2 predictions involving flexible regions Published online December 20, 2024. *arXiv*. 2024. https://doi.org/10.48550/arXiv.2412.15970.
92. Lipinski CA. Drug-like properties and the causes of poor solubility and poor permeability. *Curr Dir Drug Discov Rev Mod Tech*. 2000;44:235–249.
93. Craik DJ, Fairlie DP, Liras S, Price D. The future of peptide-based drugs. *Chem Biol Drug Des*. 2013;81:136–147.
94. Vinogradov AA, Yin Y, Suga H. Macrocyclic peptides as drug candidates: recent progress and remaining challenges. *J Am Chem Soc*. 2019;141:4167–4181.
95. Hill TA, Shepherd NE, Diness F, Fairlie DP. Constraining cyclic peptides to mimic protein structure motifs. *Angew Chem Int Ed*. 2014;53:13020–13041.
96. Rettie S et al. Accurate de novo design of high-affinity protein binding macrocycles using deep learning Published online November 18, 2024. *bioRxiv*. 2024. https://doi.org/10.1101/2024.11.18.622547.
97. Geylan G et al. PepINVENT: Generative peptide design beyond the natural amino acids Published Online September 21, 2024. *arXiv*. 2024. https://doi.org/10.48550/arXiv.2409.14040.