ChemComm

ROYAL SOCIETY OF CHEMISTRY

HIGHLIGHT

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Chem. Commun.*, 2026, **62**, 672

Peptide-based drug design using generative AI

Srinivasan Ekambaram[a] and Nikolay V. Dokholyan*[ab]

Peptide-based therapeutics have emerged as a significant treatment strategy, offering high specificity and tunable pharmacokinetics. Recent advances in Artificial Intelligence (AI) have shifted the focus towards structure prediction, generative design, and interaction modelling, significantly accelerating drug design and discovery. Deep learning architectures, including graph neural networks, transformers, and diffusion models, have facilitated the generation of novel sequences for the target of interest, although predicting the solubility, immunogenicity, and toxicity of these sequences remains a challenge. Innovations in peptide chemistry, such as cyclization, stapling, non-canonical amino acids, and nanoparticle formulations, help overcome the hurdles of bioavailability and permeation. These chemical approaches, combined with developments in autonomous peptide synthesis and high-throughput screening, have considerably reduced discovery timelines from years to months. Clinically, this progress is apparent in the growing number of approved peptide drugs for metabolic disorders, oncology, and medical imaging. Here, we review recent progress in peptide-based drug design using AI, focusing on generative architectures and interactions. We then examine AI-driven screening and delivery optimization for these peptide-based discoveries. Finally, we discuss the current limitations, practical challenges, and future direction with particular emphasis on data quality and autonomous drug discovery.

Received 29th August 2025,
Accepted 4th December 2025

DOI: 10.1039/d5cc04998a

rsc.li/chemcomm

[a] *Department of Neurology, University of Virginia, Charlottesville, VA 22901, USA. E-mail: dokh@virginia.edu*
[b] *Departments of Neuroscience, Pharmacology, Biomedical Engineering, Microbiology, Immunology, & Cancer Biology, University of Virginia, Charlottesville, VA 22901, USA*

Srinivasan Ekambaram

Srinivasan Ekambaram is a postdoctoral fellow in Prof. Nikolay Dokholyan's lab at the University of Virginia. He earned his MS (2017) and PhD (2021) in Computational Biology from the Vellore Institute of Technology, Vellore, under the supervision of Dr R. Rajasekaran. Following his doctoral studies, he served as an Institute Postdoctoral Fellow in Dr Anand Srivastava's group at the Indian Institute of Science, Bengaluru. His research focuses on designing peptide-based inhibitors for disordered proteins, developing deep learning models for protein structure and function, and applying large language models (LLMs) for precision medicine.

Nikolay V. Dokholyan

Dr Nikolay Dokholyan received his PhD in Physics from Boston University in 1999 under the supervision of Dr H. Eugene Stanley. He was then an NIH NRSA Fellow at Harvard University in Dr Eugene Shakhnovich's lab, focusing on protein folding, design, and evolution. Dr Dokholyan held faculty positions at the University of North Carolina at Chapel Hill (2002–2017) and Penn State College of Medicine (2018–2025) before joining the University of Virginia. His research involves developing computational methods to engineer molecular structures and applying these tools to understand human diseases and create therapeutic strategies. Dr Dokholyan has held several endowed positions, and has been elected to be a Fellow of the American Physical Society (2012), the American Association for the Advancement of Science (2019), and the American Institute for Medical and Biological Engineering (2022).

Introduction

Peptides have emerged as a compelling class of therapeutic interventions for various disorders.[1] Since the initial use of insulin, the first peptide-based drug from the early 20th century, numerous modified peptides, such as cyclic and stapled forms, have been developed.[2,3] These peptides exhibit enhanced potency on target interventions with unique features, including specificity, tunable half-life, and structural modifications.[4] The approval for the peptide-based drugs has shown a consistent increase from 2020 to 2024 for various diseases, especially diabetes, obesity, and cancer.[5] Peptides are also used as diagnostic markers in various disorders for imaging purposes, aiding clinical studies.[6]

Peptides are versatile molecules whose moderate size allows them to effectively target complex interactions, such as protein–protein interactions and binding to G protein-coupled receptors (GPCRs), with higher binding affinity.[7,8] From the chemical perspective, the synthesizability and modifications, such as cyclization, non-canonical amino acids, stapling, and PEGylation, make it more suitable for therapeutic purposes.[9–12] However, the drawbacks of the peptides are the poor oral bioavailability and vulnerability to proteolysis with limited tissue penetration.[13,14] These factors urge increased dosage requirements and careful design for delivery and administration. Overcoming these negative factors is vital to progressing clinical drug development and discovery.

In recent years, science, technology, engineering, and medicine (STEM) fields have seen a transformative development in artificial intelligence (AI).[15] For instance, protein folding, considered a significant challenge in protein studies, has been emulated with high accuracy using the Alphafold models.[16] AI has been applied in various fields to unravel the hidden patterns and address the experimental difficulties, paving the way towards novel studies and discoveries.[17,18] These foundations of AI have provided the capability for rational peptide design and docking studies. The AI-based models, such as ESM2,[19] deep learning models,[20] Generative Adversarial Networks (GANs),[21] diffusion models,[22] and Variational Autoencoders (VAEs),[23] are applied to computationally design the peptide sequences for the specific target of interest.[24] AI-based tools that predict key drug-like parameters such as toxicity, solubility, and permeability provide additional support for drug development by vastly reducing the experimental cost.[25] Unifying these characteristic features using AI could aid in accelerating peptide-based drug discovery with greater success.

Numerous studies over the past five years report the integration of AI into peptide-based drug discovery. A recent review by Zheng *et al.* has listed progress in peptide discovery, synthesis, and clinical translation. At the same time, Xiao *et al.* provided a broad summary of peptide therapeutics, delivery platforms, and market trends, highlighting the evolving role of AI.[2] Developments in peptide–protein interaction modeling, such as AfCycDesign for cyclic peptides, establish the effective use of computational methods to design cyclic peptides.[26] These developments highlight a fundamental innovation in the field, where the power of experimental and computational methods converges to unravel novel therapeutic discoveries.

This review aims to provide a comprehensive and critical analysis of the current developments in AI-driven peptide-based therapeutics. We summarize the AI-based peptide prediction and modelling studies designed using the generative frameworks and their applications. We then delve into the AI-driven screening and delivery optimization for these peptide-based discoveries. We summarize the innovations of these peptide-based drugs from the clinical and translational perspective. Finally, we review the current limitations, practical challenges, and future direction with particular emphasis on the data quality and autonomous drug discovery.

Computational approaches in traditional peptide design

The traditional peptide drug discovery process has focused on empirical approaches such as screening large combinatorial libraries using phage display, bacterial display, and mammalian cell surface systems.[27,28] These approaches express the peptide libraries on the protein or cell surfaces to identify the binding proteins through an iterative process. Phage display, in particular, has been widely utilized in peptide discovery that expresses the peptide variant as a genetic fusion to bacteriophage proteins, developing a direct link between the displayed peptide and its encoding DNA.[29] The phase display libraries from the New England Biolabs have become a dominant tool that facilitates epitope mapping, protein–protein contact mapping, and identification of bioactive peptides. However, recent studies using next-generation sequencing showed a significant variation in the phage display libraries, including the bias in amino acid and over-representation of stop codons.[30] Hence, advanced computational tools are needed to complement the traditional experimental methods (Fig. 1).

Structure-based drug design

Structure-based computational approaches have been central to peptide design, using protein structures to guide rational peptide optimization. Molecular docking appeared as a primary technique for modeling peptide–protein interactions, providing detailed information about binding modes and interactions.[31] Traditional molecular docking approaches, however, encountered significant challenges when applied to peptides due to their intrinsic flexibility. Initial methods typically treated peptides as rigid entities, limiting their applicability to highly flexible peptide sequences.[32] The development of flexible docking algorithms, such as CABS-dock,[33] Hpepdock,[34] and PatchMAN,[35] represented a significant advancement by incorporating conformational flexibility of both peptide and protein molecules. For instance, CABS-dock utilizes a coarse-grained protein model that enables the search for peptide conformational space while retaining computational efficiency. Reports from various studies have used these tools in multiple systems, including peptide docking to G-protein-coupled receptors

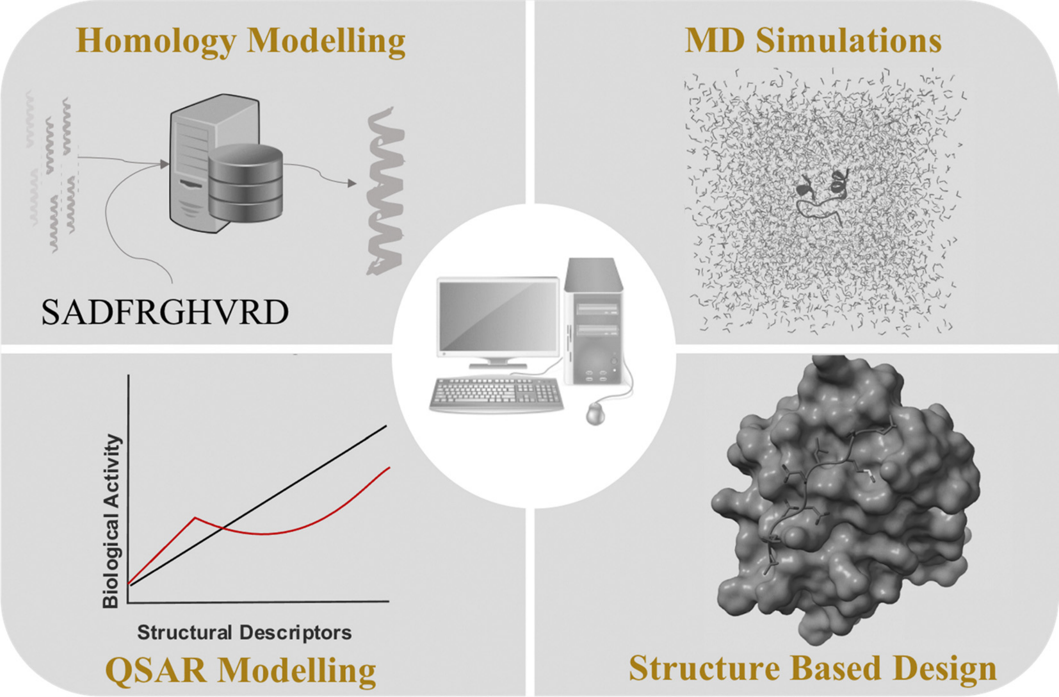

Fig. 1 **Traditional approaches in peptide-based drug design**. This figure outlines the conventional computational workflow for identifying and optimizing peptide therapeutics. The process begins with homology modeling, which aids in predicting the three-dimensional structure of a peptide. Followed by screening of large peptide libraries for candidates with desirable properties. Quantitative structure–activity relationship (QSAR) modeling is applied to establish correlations between molecular features and the biological activity of the screened peptides. Molecular docking is further employed to predict the preferred binding orientation of a peptide with its target protein, providing insights into molecular recognition. Finally, molecular dynamics (MD) simulations capture atomistic details of peptide–protein interactions, stability, and conformational changes, offering critical guidance for lead optimization.

(GPCRs) and short linear motifs (SLiMs) in protein–protein interactions.[36]

Homology modeling and structure prediction

Homology modeling was widely used in peptide design when experimental structural data are unavailable. This computational method predicts three-dimensional peptide structures, assuming the folding pattern remains identical for peptides with similar sequences. However, the model accuracy relies mainly on the protein structure availability and the degree of sequence similarity between target and template peptides, with a minimum threshold of 30% sequence identity. Conventional tools such as I-TASSER have been valuable for discovering novel antimicrobial peptides and differentiating their anti-tumoral properties.[37] While tools like Backbone-based Rotamer Library (BbRL) and PEP-FOLD have enabled prediction of peptide structures ranging from 5 to 50 amino acids, facilitating both *de novo* prediction and biased prediction when interaction sites are known.[38]

Virtual screening and library design

Virtual screening offers a cost-effective strategy to identify promising peptide candidates from vast computational libraries and has become a cornerstone of structure-based virtual screening (SBVS).[39] SBVS utilizes protein structures to screen and design peptides with high affinity and specificity for relevant targets. Recent advances have aided in the development of ultra-large virtual peptide libraries containing billions of sequences available for screening. For instance, a study described a *de novo* design strategy using directed mutation-driven high-throughput virtual screening (HTVS) to develop vast virtual libraries, supposedly expanding from 10^4 initial scaffolds to 10^{14} library members through iterative mutation.[40] The development of virtual libraries has also extended to include non-natural amino acids. Various studies have reported that approximately 380 000 readily synthesizable non-natural amino acids exist, demonstrating vast chemical diversity compared to the 20 natural amino acids. Virtual screening approaches using these expanded libraries have successfully identified peptides with significantly higher predicted affinities than natural sequences.[41]

Quantitative structure–activity relationship (QSAR) modeling

QSAR approaches are critical in drug discovery, bridging applications from small-molecule development to peptide optimization. However, they have had limited success compared to traditional pharmaceutical applications. QSAR studies in peptide research have explored various bioactive properties, including antioxidant activity, antimicrobial effects, and enzyme inhibition.[42] A comprehensive review of QSAR applications to food protein-derived bioactive peptides revealed that while some studies successfully identified structural requirements for specific bioactivities, differences across analyses were common, possibly due to dataset quality and descriptor selection issues.[43] The limited availability of high-quality experimental data for peptide–protein interactions has constrained the development of robust QSAR models. Recent advances have introduced machine learning-guided QSAR approaches tailored explicitly for peptides. The streaMLine program represents a noteworthy example, systematically exploring peptide libraries through rigorous design, synthesis, screening, and ML-driven analysis. Researchers successfully screened 2688 peptides using this approach and identified multiple options for designing stable and potent GLP-1R agonists.[44]

Molecular dynamics (MD) simulations

Molecular dynamics (MD) simulations play a pivotal role in computational studies for peptide-based drug design, providing an in-depth understanding of peptide–protein interactions and dynamics at the atomistic level, which is essential for drug design.[45–47] Numerous studies have reported the use of MD simulations to rationally design the inhibitory peptides against various disorders.[48–55] Advanced MD simulation techniques, such as enhanced sampling and replica-exchange, have further enabled the study of cyclic and intrinsically disordered peptides.[56–58] By elucidating binding mechanisms through broad views of molecular interactions, MD simulations can identify probable binding poses, calculate binding affinities, and determine the thermodynamic and kinetic parameters governing peptide–target interactions.[59–61] MD simulations have also been utilized significantly in anti-microbial peptide design, revealing membrane disruption mechanisms and physicochemical properties that enhance activity against resistant bacterial strains.[62–64] Our group has developed discrete molecular dynamics (DMD), which employs discrete energetic potentials to reach microsecond-scale timescales for complex biomolecular systems efficiently.[65,66] DMD has been extensively validated and applied to explore protein folding dynamics, aggregation mechanisms, and prion-like conformational conversions.[67–69] It has also provided insights into the molecular mechanisms of polypeptide aggregation in human diseases[70] and the effects of macromolecular crowding on folding cooperativity.[71] More recently, DMD has been successfully applied to rational peptide design, including the development of therapeutic peptides against various disorders.[72–75] Together, these studies demonstrate the versatility of DMD in uncovering atomistic mechanisms of peptide aggregation, protein–peptide interactions, and structure-based peptide design. With the current advancements in computational resources that support MD simulations of large systems over millisecond timescales, they accelerate the advantage in peptide drug discovery.

Accurate prediction of the peptide target interactions and binding affinity *via* MD simulations is dependent on the choice of force field and selection. The AMBER force field (ff14SB, ff19SB) is widely used for protein–peptide complex systems, capturing the conformational dynamics essential for binding.[76] A recent study by Miao *et al.* suggested that among seven force fields tested for cyclic peptide simulations, RSFF2 + TIP3P, RSFF2C + TIP3P, and Amber14SB + TIP3P showed the highest accuracy, recapitulating NMR-derived structures for 10 peptides, while others performed significantly worse.[77] CHARMM36m and Amber ff99SB-disp force fields are shown to provide better efficiency in capturing the range of dynamics from intrinsically disordered peptides, which can accurately capture experimental calculations.[78–80] Furthermore, the MM-PBSA and MM-GBSA methods provide estimates of binding affinities by computing the enthalpic contribution from force field energies and the entropic penalty from conformational sampling.[81]

Despite significant advances, traditional computational approaches face several persistent challenges. The inherent flexibility of peptides makes accurate structure prediction and docking computationally demanding. Limited availability of experimental structural data for peptide–protein complexes constrains model training and validation.[82] Additionally, traditional methods often struggle with non-canonical amino acids and chemical modifications commonly used in therapeutic peptides. Incorporating cyclization, stapling, and other structural modifications remains a significant challenge for conventional modeling approaches.

AI-enabled structure prediction for peptide-based design

Over the past decade, AI has revolutionized the field of peptide-based drug discovery. From the Nobel Prize-winning development of AlphaFold to the upcoming quantum-based generative models, AI has enhanced research approaches towards peptide structure prediction, interaction, and therapeutic design (Fig. 2).

AlphaFold for structure prediction

AlphaFold, introduced by DeepMind, made a breakthrough in computational structural biology.[16] The AlphaFold, released in early 2020, demonstrated astonishing accuracy in the protein structure prediction competition CASP, with a median backbone accuracy of 0.96 Å RMSD, compared to 2.8 Å for competing methods. This development showed immediate applications towards peptide research as the AlphaFold database now holds 200 million protein structures with vast data information. The recent development of AlphaFold3 has displayed the capability of predicting protein–peptide interactions.[83] Independent validation studies on the 588 peptide sequences with varying amino acid ranges (10–40) showed that AlphaFold could predict the peptides with better accuracy.[84] However, the flexible regions and intrinsically disordered peptides still show reduced reliability.

Complementary structure prediction platforms

RoseTTAFold is an alternative to AlphaFold, with a neural network architecture that simultaneously considers sequence patterns, amino acid interactions, and three-dimensional structure.[85] RoseTTAFold revealed precise value in modeling protein complexes and extended to handle nucleic acid–protein

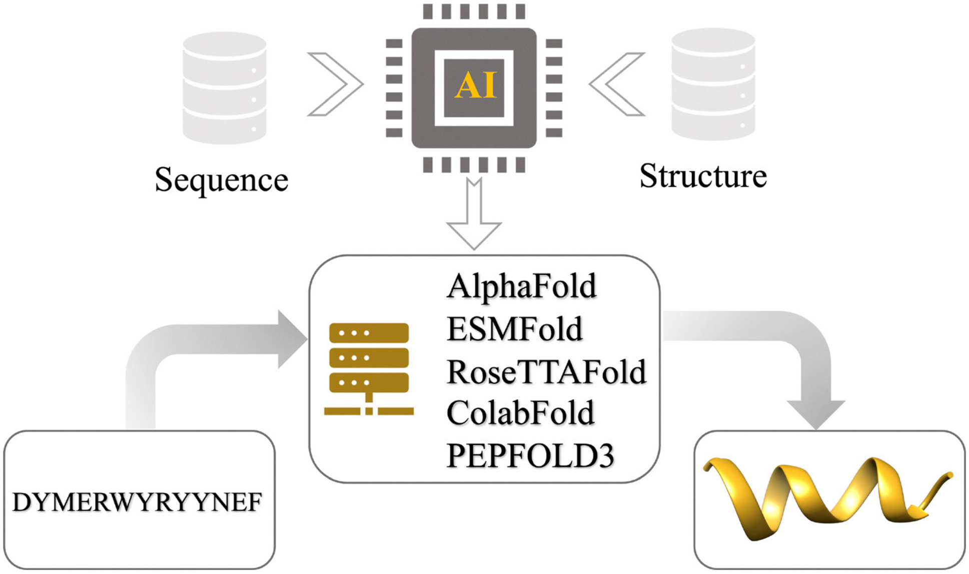

Fig. 2 **Artificial Intelligence in peptide structure prediction**. The prediction of peptide structures has shown a transformative shift with the integration of artificial intelligence (AI) and advanced computational methods. Deep learning–based models such as AlphaFold, RoseTTAFold, and ESMFold are trained on vast datasets of protein sequences and experimentally determined structures. As a result, they predict the 3D structure of novel peptide sequences with exceptional accuracy. This innovation has redefined peptide modeling, opening new avenues for drug discovery, protein engineering, and therapeutic design.

interactions through RoseTTAFoldNA.[86] The RoseTTAFold all-atom (AA) model can accept a wide range of ligands and covalently modified amino acids, which is particularly significant for peptide design, as it enables modeling of non-natural amino acids and chemical modifications commonly used in therapeutic peptides. ColabFold has made high-quality protein structure prediction accessible to the broader research community by combining AlphaFold 2 with accelerated homology search algorithms. The platform achieves 40–60 fold faster sequence searching than standard AlphaFold 2 pipelines while maintaining prediction accuracy.[87] ColabFold supports both homo and heteromeric complex prediction, with specific handling of peptide–protein interactions. ESMFold, an embedding-based predictor leveraging protein language models, has exhibited applicability for peptide sequences, with significant hits capturing short-range interactions relevant to secondary structure formation.[19]

Beyond these predictors, particular AI frameworks dedicated to peptide modeling are also emerging. For instance, PEP-FOLD3 employs a *de novo* fragment assembly strategy integrated with neural network-based scoring functions to capture low-energy peptide conformations.[38] The refinement of PepFold's pipeline has extended its use for linear and cyclic peptides, including post-translational modifications. These methods employ reinforcement learning strategies to iteratively refine peptide conformations against experimentally derived constraints such as NMR or cryo-EM data.

Recently, the AfCycDesign framework adapted AlphaFold2 with cyclic positional encoding, enabling accurate structure prediction and sequence redesign of cyclic peptides.[26] Notably, several *de novo* designs yielded experimental confirmation with RMSDs under 1 Å, and served as effective scaffolds for nanomolar peptide binders against targets like MDM2. These results underscore the growing capability of AI models to generate and validate stable peptide architectures, with significant implications for macrocyclic therapeutic development.

Deep learning approaches for peptide–protein interaction prediction

Deep learning has transformed peptide–protein interaction (PPI) prediction by aiding to that capture complex binding patterns. Modern neural network architectures, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer models, have shown better performance over traditional machine learning approaches. These sophisticated models can simultaneously predict binary peptide–protein interactions and identify critical binding residues, providing multi-level insights that accelerate peptide drug discovery processes (Fig. 3).

Convolutional neural networks (CNNs) for sequence-based prediction

Convolutional neural networks are the foundational architectures for peptide–protein interaction prediction, particularly effective at capturing local sequence patterns and motifs for binding interactions.[88] PepCNN represents a significant advancement, incorporating structural and sequence-based information from primary protein sequences to predict peptide binding residues.[89] Recent developments have extended CNN applications to multi-level peptide–protein interaction prediction.

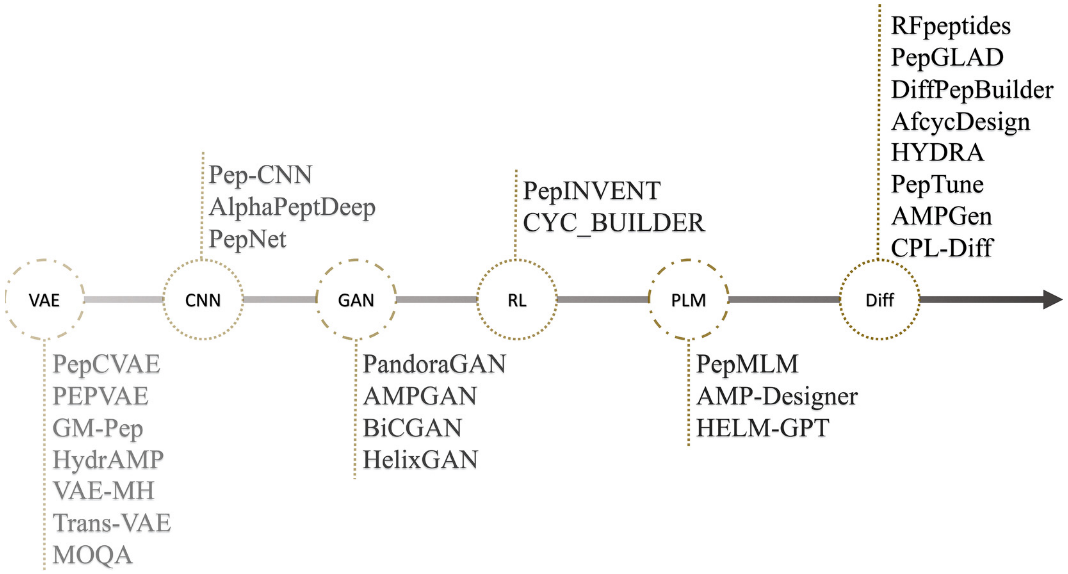

Fig. 3 **Deep learning architectures for predicting protein–peptide interactions**. The prediction of protein–peptide interactions, a key step in therapeutic discovery, has been significantly advanced by applying diverse deep learning architectures. Convolutional neural networks (CNNs) and Recurrent neural networks (RNNs) have been widely used to extract features from sequence data, while Graph neural networks (GNNs) capture the complex three-dimensional topology of molecular structures. Generative models, such as variational autoencoders (VAEs), enable the design of novel peptides with tailored properties. More recently, transformers and protein language models (PLMs) have provided exceptional insights into molecular context, achieving state-of-the-art predictive performance.

The CAMP (comprehensive analysis of multi-level peptide–protein interactions) framework employs CNN modules to extract hidden contextual features from both peptides and proteins.[90] The framework incorporates multi-channel architectures to separately process numerical and categorical features, addressing inconsistencies that arise from multi-source feature integration. This design enables simultaneous binary interaction prediction and peptide-binding residue identification, providing comprehensive insights into interaction mechanisms. Thus, CNNs efficiently capture local sequence motifs through hierarchical feature extraction, enabling fast inference for high-throughput peptide screening with minimal preprocessing requirements. However, their limited capacity to model long-range dependencies and inherent lack of 3D structural awareness constrain accurate prediction of binding interactions and conformational properties.

Transformer architectures and attention mechanisms

Transformer-based models have revolutionized peptide–protein interaction prediction by using attention mechanisms to capture long-range dependencies and identify critical interaction patterns. The cross-TCR-interpreter model utilizes transformer architectures in predicting T-cell receptor-peptide-major histocompatibility complex (TCR-pMHC) interactions. ABTrans represents another innovative transformer application, specifically designed for predicting interactions between amyloid-β peptides and antibodies.[91] PepNN advances transformer applications by enabling sequence- and structure-based predictions through graph attention layers.[92] The model incorporates multi-head reciprocal attention layers that simultaneously update embeddings of both peptides and proteins, learning interactions between residues involved in binding. More recently, TPepPro,[93] a transformer-based model trained on 19 187 peptide–protein complex pairs, achieved 85.5% accuracy with an AUC of 0.922. The model combines local protein sequence feature extraction with global protein structure feature extraction, demonstrating the power of attention mechanisms in capturing long-range interactions. Transformer models leverage self-attention mechanisms to capture long-range sequence dependencies and global context, achieving state-of-the-art performance across diverse peptide prediction tasks with highly parallelizable training. Nonetheless, their quadratic computational complexity with respect to sequence length, substantial memory requirements, and dependence on large training datasets limit their practical deployment for resource-constrained applications and small peptide datasets prone to overfitting.

Graph neural networks for structural interaction modeling

Graph neural networks model protein–protein interfaces and peptide binding sites by representing molecular structures as graphs with nodes representing residues or atoms and edges, capturing spatial relationships.[94] The TP-LMMSG model significantly advances therapeutic peptide prediction, incorporating hierarchical multi-scale residual networks with pre-trained language model embeddings.[95] This approach achieves remarkable performance across antimicrobial peptide (AMP), antiviral peptide (AVP), and anticancer peptide (ACP) prediction tasks while reducing preprocessing time by over seven-fold compared to traditional graph learning models. Research utilizing graph convolutional networks (GCN) and graph attention networks (GAT) shows that structural information combined with sequence features considerably enhances prediction performance.[96] These models construct protein graphs from protein files within threshold distances, creating residue contact networks that capture spatial relationships critical for interaction prediction.

Integrating protein language models with graph neural networks represents a substantial improvement. By using pre-trained language models like SeqVec and ProtBERT to generate node features directly from protein sequences, these hybrid approaches eliminate the need for domain-specific features while maintaining the structural knowledge provided by graph representations.[97] This combination leverages evolutionary information learned by language models and spatial constraints captured by graph architectures. Hence, graph neural networks naturally represent molecular topology and 3D spatial relationships, enabling explicit modeling of atom-level interactions critical for accurate binding affinity and protein–peptide complex prediction. Their computational expense for large molecular systems, strict requirement for high-quality structural inputs, and limited generalization to novel structural motifs outside training distributions present significant practical limitations.

Recurrent neural networks and LSTM applications

Long short-term memory (LSTM) networks are used for peptide property prediction and sequence generation since the network captures long-range interactions in amino acid sequences.[98] The use of LSTM networks to generate antimicrobial peptides exhibits their generative capabilities.[99] Recent implementations have combined LSTM networks with other architectural components to improve performance. Models combining convolutional layers with bidirectional LSTM capture local patterns and long-range dependencies simultaneously. This hybrid approach is remarkably effective for T-cell receptor–epitope binding prediction, where local sequence motifs and global sequence context contribute to binding specificity.[100] However, slow sequential training that precludes parallelization, vanishing gradient problems affecting long sequences, and inferior performance compared to Transformer architectures on contemporary benchmarks have diminished their adoption in modern peptide design workflows.

Protein language models and BERT-based approaches

Pre-trained protein language models have transformed peptide–protein interaction prediction by providing contextualized representations of amino acid sequences. BERT-based models, such as PepBCL, reveal substantial advantages over traditional hand-crafted features. This approach is particularly effective for peptide-binding residue prediction, where only approximately 5.4% of residues interact.[101] PeptideBERT represents a specialized application of transformer language models to peptide property prediction.[102] The model's success demonstrates that language models can effectively capture the relationship between sequence and function for therapeutic peptide development. The SWING (sliding window interaction grammar) framework extends protein language model applications with contrastive learning for protein–peptide interaction.[103] Thus, protein language models pre-trained on vast sequence databases capture evolutionary and functional patterns, enabling powerful transfer learning with strong zero-shot and few-shot prediction capabilities that substantially reduce experimental data requirements. However, their sequence-only representations fundamentally ignore structural dynamics, require computationally intensive fine-tuning, and may inadequately represent rare peptide classes underrepresented in natural protein training data.

Variational autoencoders for peptide generation and optimization

Variational autoencoders are used in peptide design and generation by learning continuous latent representations of sequence space. The VAE-MH (variational autoencoder with Metropolis–Hastings) represents the first deep learning-based model explicitly designed for peptide extension. This approach learns from protein–protein interactions and conducts focused searches for optimal sequences rather than random exploration.[104]

Applications to antimicrobial peptide generation showcase VAE effectiveness in functional peptide design. VAE models trained on known antimicrobial peptides can generate novel sequences retaining critical features such as high hydrophobic moment in alpha-helical peptides while exploring sequence variations.[105] Recent developments have widened VAE applications to protein conformational exploration. This approach reveals VAE utility beyond sequence generation, extending to structural modeling and conformational analysis for understanding peptide–protein interactions. Hence, VAEs offer continuous latent spaces, enabling smooth interpolation and property-guided optimization, within probabilistic frameworks that quantify uncertainty and facilitate efficient training on moderately sized datasets. Posterior collapse frequently reduces latent space informativeness, generated sequences often exhibit limited diversity, and fine-grained control over specific structural features remains challenging.

Generative AI for peptide design

Generative AI has shown a transformative approach in peptide-based drug discovery, presenting the ability to design novel peptide sequences with therapeutic properties. Unlike traditional methods, which rely heavily on exhaustive screening, generative models learn the underlying patterns of protein sequences and structures from vast biological datasets. By leveraging architectures such as variational autoencoders (VAE), generative adversarial networks (GAN), diffusion models, and protein language models (PLM), these systems can design peptides that are not only diverse but also optimized for stability, specificity, and efficacy (Fig. 4). This paradigm accelerates discovery while reducing cost and experimental workload.

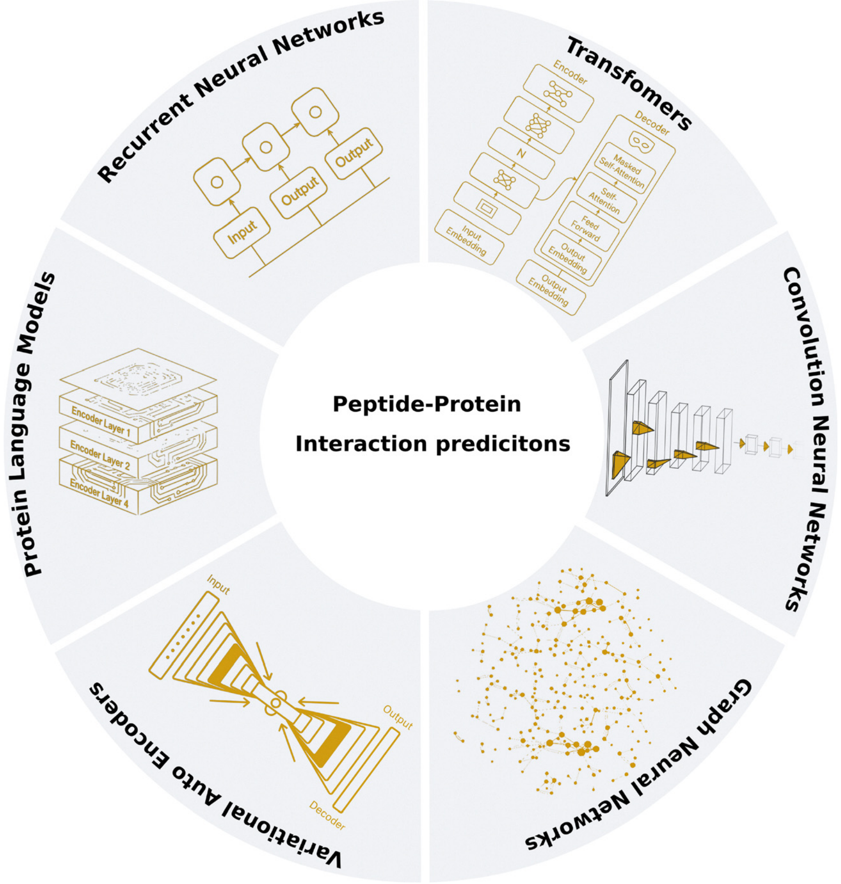

Fig. 4 **Advanced generative AI tools for peptide discovery**. This schematic illustrates the diverse AI models that generate and refine novel peptides with therapeutic potential. Variational autoencoders (VAE) and convolutional neural networks (CNN) enable the generation of diverse peptide candidates, while graph adversarial Networks (GAN) incorporate structural context into design. Reinforcement learning (RL) is applied for goal-directed optimization, guiding peptides toward desired properties. More advanced approaches, including Protein language models (PLM) and diffusion models (Diff), are expanding the field further by learning the fundamental rules of protein structure and function, enabling the creation of highly specific and compelling peptide therapeutics.

Variational autoencoders and generative adversarial networks

The application of generative AI to peptide design has progressed considerably over the past decade. VAEs are one such model used for antimicrobial peptide discovery, with frameworks like PepVAE representing semi-supervised learning approaches that jointly model labeled and unlabeled peptide sequences.[106] These studies have shown that VAE-generated peptides can be experimentally validated, with some novel antimicrobial peptides discovered within 48 days of computational prediction. The HydrAMP system advances VAE applications through conditional variational autoencoders (cVAE) that unravel antimicrobial properties from other latent features.[107] This approach leverages parameter-controlled creativity to generate diverse peptides for multiple tasks, including unrestrained generation and analogue design.

Generative adversarial networks (GANs) have applications in peptide design. PandoraGAN specifically targets bioactive antiviral peptides using modified GAN architectures adapted from natural language processing.[108] AMPGAN v2 represents a significant advancement, employing bidirectional conditional GANs (BiCGAN) for rational AMP design.[109] HelixGAN addresses structural peptide design by generating *de novo* left-handed and right-handed alpha-helix structures at the atomic level.[110] The model employs gradient-based latent space optimization to match exact conformations of selected hotspot residues. However, GANs face challenges, including mode collapse and training instability, limiting their broader adoption.

Diffusion models for backbone and interface generation

Diffusion-based models have become noteworthy in the development of peptide *de novo* and interface design in recent years. Remarkably, RFdiffusion showed that denoising the generative trajectories can aid in forming the backbone to fit the target and yield high-affinity binders.[111] That was comprehensive to helical peptides using partial diffusion refinement of the ligands, yielding picomolar binding validated experimentally.[112] Meanwhile, RFpeptides adapts the diffusion framework to macrocyclic peptides to generate cyclic backbones that match the protein pockets, aiding stability and affinity.[113] Similarly, all-atom peptide generation with the geometric latent diffusion (PepGLAD) conditioning on the binding site geometry to produce sequences and 3D structures, benchmarking the test sets.[114] Further, ProteinMPNN, a message-passing neural network for sequence design from structure, can redesign a scaffold to bind a peptide epitope, significantly improving affinity in experimental assays when paired with Rosetta-based assembly.[115]

DiffPepBuilder, an SE(3)-equivariant diffusion model trained on a synthetic dataset (PepPC-F), enables *de novo* design of peptide binders co-optimizing sequence and conformation, presenting stabilizing disulfide bonds.[116] In comparative benchmarks, DiffPepBuilder exceeded AfDesign and RFdiffusion with ProteinMPNN scaffolding in producing diverse, structurally faithful binding peptides with improved free energy profiles. Besides, HYDRA represents another innovative diffusion-based approach, combining target-aware amino acid residue generation with binding affinity maximization.[117] This hybrid model generates high-quality, diverse, and stable peptide binders tailored to target receptor proteins.

PepTune, introduced in 2025, uses a masked diffusion language model guided by Monte Carlo Tree Guidance (MCTG) to optimize for binding affinity, solubility, permeability, hemolysis, and non-fouling characteristics simultaneously.[118] This modular, search-guided framework represents a powerful step forward in optimizing multiple biophysical properties in one model. Specialized application of antimicrobial peptides has yielded compelling results. AMPGen integrates an autoregressive diffusion generator with evolutionary information from MSAs, a discriminative XGBoost screen for physicochemical features, and an LSTM-based scorer for target specificity.[119] Experimental validation showed that of 40 *de novo* peptides synthesized, over 80 percent displayed antibacterial activity, and these sequences were absent from existing AMP databases, demonstrating novelty and efficacy. Beyond AMPs, general-purpose diffusion frameworks are emerging. CPL-Diff, a mask-controlled diffusion model, allows length tuning during peptide generation, offering flexibility in sequence design that accommodates therapeutic constraints like half-life or membrane binding.[120] MMCD (multimodal contrastive diffusion) integrates both sequence and structure information in a contrastive diffusion learning framework, boosting generation quality across antimicrobial, anticancer, diversity, and docking metrics.[121] Besides, their computational expense requiring 200–1000 denoising steps, slow inference limiting throughput, dependence on high-quality structural training data, and practical deployment challenges without GPU infrastructure constrain widespread accessibility.

Flow matching models for peptide design

Flow matching models represent a new class of continuous generative architecture that learn to map the prior distribution into complex distributions of biologically relevant structures. Unlike diffusion models that rely on static denoising, flow matching uses deterministic ordinary differential equations to transfer data points from latent to molecular space, achieving high efficiency.[122,123] This framework supports flexible generations, allowing for the rapid sampling of diverse structural motifs, such as alpha helices, beta sheets, and cyclic peptides. Flow matching enables large-scale screening of candidate peptides within seconds and could incorporate constraints such as secondary structure propensity.

ProtFlow[124] leverages compressed protein language model embeddings to efficiently generate peptides and antibodies in a single inference step, while HelixFlow[125] extends this capability to full-atom, SE(3)-equivariant helical peptide design, including pocket-specific inpainting for receptor-guided binder generation. Recently developed NCFlow is able to incorporate arbitrary uses of the non-canonical amino acid in protein design.[126] However, these lack induced fit modeling, thus requiring further refinement with docking or MD simulations to ensure the validity. Despite these limitations, flow matching offers a powerful tool for early-stage screening and the rapid generation of peptide libraries.

Comparatively, diffusion models are more preferable when structural fidelity and experimental validation are prioritized, while flow-matching methods are ideally suited for rapid, large-scale exploration and early-stage design under limited computational resources. In the near future, the use of hybrid architectures with diffusion and flow matching is likely to determine the next generation of peptides, as supported by equivariant modeling, attention mechanisms, and guided optimizations.

BindCraft AlphaFold2-based one-shot functional design

BindCraft represents a distinct paradigm in peptide design, diverging from the diffusion and flow-based models by leveraging AlphaFold2 to optimize the peptide-target complex through iterative confidence maximization directly.[127] BindCraft co-folds both the target and binder, enabling induced-fit modeling that captures the conformational adaptation upon binding. Initially, BindCraft was tested on miniprotein (60–200 amino acids) binders, but recent studies have demonstrated its performance on short peptides, achieving nanomolar binding (65–650 nM) against MDM2 and WDR5.[128] The streamlined multimer protocol allows users to define peptide length, binding hotspots, and filtering stringency, optimizing efficiency for short peptides that primarily function as binding interfaces. Its main limitation lies in structural diversity, as current implementations favor alpha-helical motifs, restricting applicability to targets with helical recognition interfaces. Despite its computational intensity, BindCraft's interpretability and capacity for target-specific sequence refinement make it a top tool for precision-driven binder design.

Protein language models for target-conditioned peptide generation

Protein language models (PLMs) enable the generation of target-conditioned peptides without explicit structural templates. PepMLM fine-tunes ESM-2 with a masking strategy that conditions on the target protein sequence, reconstructing the cognate peptide region and yielding binders validated both *in silico* and experimentally, including degradation assays.[129] Recent reports show peptides designed directly from protein sequences, underscoring that sequence-only conditioning can discover functional binders at scale when paired with robust structural inspection.[130] These PLM-centric generators are attractive for proteome-wide scans, and they interface naturally with downstream structure models to eliminate false positives and select pose-consistent candidates for synthesis. Similarly, AMP-designer, an LLM-based model, emphasizes speed and potency.[131] Within 48 days, it produced 18 AMP candidates, of which nearly all were active *in vitro*. Two exhibited high efficacy, low hemotoxicity, robust plasma stability, and strong *in vivo* performance, suppressing bacterial load by ∼100-fold in murine lung infection models. This performance highlights the feasibility of rapid, AI-driven design-to-validation loops even in low-data or strain-specific contexts. Therefore, target-conditioned protein language models integrate target protein information to generate binder-specific sequences, leveraging evolutionary co-variation patterns for rational, context-aware peptide design that accounts for binding partner characteristics. The scarcity of paired peptide-target training data, suboptimal conditioning mechanisms, and inheritance of biases from pre-training on natural protein sequences limit their current effectiveness for *de novo* therapeutic peptide discovery.

Backbone-to-sequence decoding with (ligand) MPNN

Generative workflows commonly split shape from sequence: diffusion proposes peptide/receptor geometries, and a decoder assigns sequences that stabilize the interface. ProteinMPNN remains a standard for fast, high-quality sequence design on fixed backbones, with strong experimental validation across diverse scaffolds.[115] LigandMPNN extends this concept for chemically rich environments by conditioning sequence design on atomic context, including small molecules, metals, and nucleotides, thereby supporting peptides and receptors that bind cofactors or include bound ligands in the pocket.[132] In peptide design, the pair forms a practical core to propose peptide conformations against a target surface, decode sequences with (Ligand)MPNN to optimize packing and polar satisfaction, and then re-score with AlphaFold-multimer to check pose stability and side-chain realism before experimental selection.

Reinforcement learning and search over constrained chemistries

Reinforcement learning (RL) improves generative models by imposing design constraints, including cyclization, incorporating non-canonical residues, and integrating cell-penetrating motifs. Recent RL frameworks assemble and score peptides in closed chemical spaces, for example, CYC_BUILDER for target-specific cyclic peptides using Monte Carlo tree search (MCTS) and bond-forming actions to optimize binding while ensuring synthetically viable cyclization.[133] Broader RL approaches, including graph-attention and prior-guided RL, efficiently explore peptide sequence/structure space, balancing novelty and physicochemical properties.[134] Diffusion or PLM proposals can act as priors, while rewards integrate docking scores and heuristic developability metrics. This RL layer is especially valuable for macrocycles and stapled designs, where combinatorial chemistry explodes and naive generators drift from synthesizable, bioactive regions of sequence space. Consequently, RL frameworks enable simultaneous multi-objective optimization balancing efficacy, safety, and drug-likeness, and incorporating iterative experimental feedback for continuous model improvement. Careful reward function design proves challenging and susceptible to sample-inefficient training demands extensive computational simulations, and training instability, particularly with sparse rewards, impedes reliable convergence.

Non-canonical residues and macrocyclic design

Generative models increasingly target non-natural amino acids and macrocycles to address protease stability and permeability. PepINVENT reports a sequence generator explicitly beyond natural amino acids, enabling exploration of peptidomimetic design spaces relevant to oral bioavailability and serum stability.[135] Diffusion pipelines specialized for macrocycles generate cyclic backbones shaped to receptor pockets, a crucial advance because ring closure and side-chain geometry strongly constrain feasible sequences.[113] On the validation side, AlphaFold3 (AF3) allows complex modeling with modified residues and ligands, which helps vet non-canonical designs and flag steric or coordination issues before synthesis. Modeling of cyclic peptides with AF3 is emerging, indicating improved fidelity for unnatural chemistries and providing a standardized route for pose confirmation across modified peptide classes.[136] Thus, incorporation of non-canonical amino acids and macrocyclic constraints dramatically expands chemical diversity beyond the 20 natural amino acids, enhancing protease resistance, metabolic stability, and enabling novel binding modes inaccessible to linear peptides. Limited training data for non-standard modifications, increased synthetic complexity and cost, unpredictable pharmacokinetic profiles, and less-established regulatory approval pathways present substantial barriers to clinical translation.

Programmable generative models and conditioning (Chroma & beyond)

Programmable generators such as Chroma sample structures and sequences under explicit constraints, symmetry, topology, or shape, can then be conditioned toward epitope presentation or pocket complementarity, offering scaffolds that support short peptide motifs or present peptide-like surfaces for binding.[137] In peptide design, these scaffolds serve as templates: a peptide motif is embedded or recognized by the generated protein surface, after which MPNN-style decoding and AF-based validation refine the complex. Together with PLM-conditioned peptide generators (PepMLM) and diffusion-based macrocycle design, such programmable models create an end-to-end toolkit.

Despite remarkable progress, several challenges persist in AI-enabled peptide design. Current models struggle with highly flexible regions and intrinsically disordered peptides. The limited availability of high-quality experimental data for peptide–protein complexes also constrains model training and validation. Optimization remains computationally expensive, and navigating these limits for multiple properties remains a challenge. There are also concerns about novelty *versus* acceptability as systems like AMPGen show promise, but whether such peptides scale in manufacturability and clinical safety remains unknown. Future designs are inclining toward foundation models, self-driving lab, and richer structural conditioning to address these.

Integration of rational design with generative AI, genetic algorithms, simulations, and experiments

Current peptide drug discovery approaches employ hybrid strategies that combine rational design with generative AI and physics-based simulations. This integrated method enhances the strategies by providing a deeper mechanistic understanding through rational design. Meanwhile, AI enables rapid predictions across vast, complex spaces with atomistic calculations from MD simulations, offering deeper insights to advance experimental studies.[138,139]

Chang *et al.* employed a latent diffusion model (AMP-diffusion), fine-tuned on antimicrobial peptide sequences using protein language model embeddings. From 50 000 generated candidates, 46 top peptides were synthesized and experimentally validated, demonstrating broad-spectrum antibacterial activity against various pathogens.[140] Chen *et al.* developed deep learning models combining support vector machines with AlphaFold2 structure predictions to design cysteine- and lysine-stapled peptides. Experimentally validated stapled peptides showed broad-spectrum antimicrobial activity (MIC 2–8 μg mL^{-1}), excellent serum stability, and minimal hemolytic activity, successfully integrating computational prediction with rational stapling modifications.[141] Wang *et al.* employed a large language model-based foundation model generating candidates from extensive peptide sequence databases. Of the 18 synthesized peptides, 17 exhibited antimicrobial activity against ESKAPE pathogens, with five AMPs achieving MICs of 4–16 μg mL^{-1} against multidrug-resistant Gram-negative bacteria and demonstrating a reduction of ~99% in bacterial load in mouse pneumonia models.[142] Ortega *et al.* combined RNN-LSTM and GAN architectures with helical wheel analysis to generate 6004 peptide candidates. From 12 synthesized peptides, 9 achieved MIC values below 10 μM against various pathogens, with OrP1M, OrP9M, and VeP1 demonstrating MIC as low as 2 μM, and six peptides showing anticancer activity against MCF-7 cells.[143]

Combined VAE with Metropolis–Hastings sampling to generate peptide extensions for β-catenin inhibitors, using MD simulations for binding pose refinement and MM/GBSA affinity calculations. Experimentally validated peptides achieved IC_{50} values of 0.03 μM (β-catenin), demonstrating the effectiveness of iterative VAE fine-tuning with MD-guided selection.[144] Zhao *et al.* developed a conditional denoising VAE framework that integrates a transformer architecture with guidance on physicochemical properties for the generation of AMPs. The model incorporated denoising techniques to address data sparsity and enhance generalization, enabling the generation of AMPs with preserved desirable properties and reduced hemolytic activity while maintaining antimicrobial efficacy.[145] Wang *et al.* employed a transformer-based VAE for the latent diffusion model, integrating MD simulations to validate membrane-binding stability and antibacterial mechanisms.[146] Das *et al.* utilized a VAE trained on 1.7 million peptide sequences combined with MD simulations in explicit membrane environments to analyze conformational stability and antimicrobial mechanisms. Generated and experimentally validated two novel AMPs within 48 days, demonstrating rapid processing in the drug discovery process while maintaining potent activity and low toxicity[147]

The most successful peptide discovery programs implement active learning cycles where experimental data continuously refines computational models. After initial computational design and synthesis, experimental assays (binding affinities, antimicrobial activity, cytotoxicity, pharmacokinetics) provide ground-truth data. These results retrain predictive models, update reward functions in reinforcement learning frameworks, and refine physics-based force field parameters, progressively improving prediction accuracy and reducing experimental attrition rates.

Therapeutic screening and translational optimization

Transforming peptides from laboratory discoveries into viable therapeutics requires addressing critical challenges, including toxicity, immunogenicity, solubility, and delivery. AI has emerged as a sophisticated tool across each domain, offering unparalleled insights into peptide optimization (Fig. 5).

Toxicity and safety assessment

Modern toxicity prediction has evolved from simple sequence-based approaches to sophisticated multimodal architectures. ToxGIN represents a significant advancement, utilizing graph isomorphism networks (GIN) to integrate structural information with sequence data for peptide toxicity prediction.[148] The model performs better (F1 score = 0.83, AUROC = 0.91) by representing peptide structures as graphs where amino acids serve as nodes and spatial interactions as edges, incorporating ESM-2 embeddings and physicochemical properties. Similarly, tAMPer demonstrates the power of structure-aware prediction by combining ESM-2 protein language model embeddings with ColabFold-predicted structures.[149] The multimodal framework extracts structural features using graph neural networks while capturing sequential dependencies through recurrent networks, achieving 91.7% AUROC and establishing strong correlations with experimental HC50 values.

Physicochemical properties prediction

Immunogenicity assessment has advanced significantly through transformer and attention-based architectures. UnifyImmun represents a unified framework that simultaneously

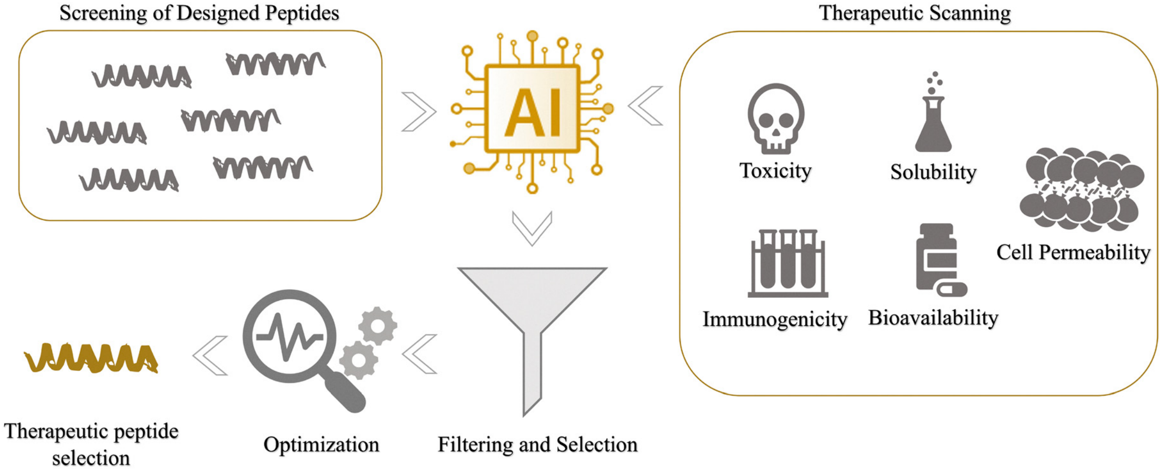

Fig. 5 **AI-driven therapeutic optimization in peptide drug design**. This figure illustrates the use of artificial intelligence to streamline the optimization of novel peptide therapeutics. Predictive models enable the rapid screening of candidates across key properties, including toxicity, solubility, membrane permeability, and immunogenicity. By integrating these assessments into a computational design framework, AI facilitates efficient multi-parameter optimization, guiding the selection of peptides with improved efficacy and an increased likelihood of clinical success.

predicts peptide binding to HLA and TCR molecules using cross-attention mechanisms.[150] The model employs virtual adversarial training to enhance generalizability, achieving superior performance on multiple test sets and outperforming state-of-the-art methods by over 10% on large-scale COVID-19 datasets. Besides, peptide solubility prediction has benefited from advanced neural architectures addressing oral and parenteral delivery requirements. MahLooL,[98] implemented as a serverless bidirectional RNN model, achieves 70% overall accuracy in solubility prediction while excelling for short peptides (18–50 amino acids) with 91.3% accuracy and 0.95 AUROC. The model incorporates simulated gastric conditions, making it valuable for oral delivery assessments.

Stapled peptide development

Hydrocarbon-stapled peptides represent a promising therapeutic class requiring specialized computational approaches due to their unique structural constraints.[151] StaPep, introduced in 2024, provides the first comprehensive toolkit for stapled peptide structure prediction and feature extraction.[152] The Python-based platform supports multiple stapling chemistries and achieved 0.85 AUC for cell penetration prediction using a curated dataset of 201 stapled *versus* 384 linear peptides. Recent deep learning advances have significantly improved stapled peptide design capabilities. A recent study applying CNN and LSTM models to antimicrobial stapled peptides achieved perfect accuracy on test sets by combining AlphaFold structural predictions with sequence-based descriptors. Experimental validation confirmed robust bacterial inhibition, low hemolysis, and excellent serum stability for designed cysteine- and lysine-stapled peptides.[153]

Integrated screening platforms

Comprehensive screening platforms now integrate multiple prediction capabilities. Peptipedia v2.0 is a user-friendly database containing over 100 000 peptide sequences from 70+ databases, incorporating 90+ binary classification models for biological activity prediction. The platform combines traditional sequence analysis with machine learning predictions, enabling comprehensive candidate evaluation.[154] Top-ML shows interpretable machine learning approaches for anticancer peptide classification, employing topological features to achieve state-of-the-art performance while providing interpretable insights into structure–activity relationships.[155] The Extra-Trees classifier-based model outperforms existing deep learning approaches on benchmark datasets while offering greater transparency in prediction rationale.

Clinical and translational context

Clinical success in the peptide therapies has been transformative across multiple areas. For example, Tirzepatide was approved by the FDA for type 2 diabetes,[156] and Zepbound for obesity.[157] Various peptide-based therapies, such as Survodutide (Phase 3) for obesity[158] and Glepaglutide (registration phase) for short bowel syndrome,[159] have been under clinical trial. Besides, radiolabeled peptides such as Pluvicto, approved in 2022 for metastatic castration resistant prostate cancer therapy.[160] Recent clinical successes on the oral formulation of Semaglutide using the absorption enhancer SNAC displayed a practical oral medicine, expanding patient access and adherence while preserving pharmacodynamic benefit.[161] Constrained peptides and stapled designs that enhance cell permeability and helicity are subjected to oncology trials, as demonstrated by ALRN-6924. This stabilized α-helical peptide targets the p53 regulatory axis and has shown tolerability and pharmacodynamic activity in early-phase studies.[162]

Peptide delivery is now clinically validated for selected molecules using permeation enhancers and promoters. At the same time, other formulations include the use of nanoparticles and conjugate strategies to address the target specificity, and intracellular administration has been tested in clinical trials.[163] The growing use of non-canonical amino acids and peptidomimetics augments the stability and membrane permeability of the peptide-based therapeutics. However, these modifications pose a significant challenge since they could alter the metabolic pathways and safety profile.[164] The development of translation has seen significant advances with AI.[165] The generative structure prediction and design with potentially peptide candidate selection and prediction towards the physicochemical properties have accelerated the peptides-based drug discovery pipeline.[166,167] Nevertheless, translating the peptide design from computational platform to preclinical and FDA approval requires a coordinated timeline and clinical strategy that, upon aligning for the well-suited study, could pave the way towards more compelling solutions for drug discovery.

Bridging computational predictions and therapeutic efficacy: critical gaps and emerging solutions

AI-driven peptide design has yielded notable outcomes, including predicted binding affinities in the nanomolar range and remarkable structural accuracy. Numerous computational tools spanning deep learning architectures, generative models, and therapeutic screening platforms have been developed in recent years to address distinct phases of peptide drug discovery (Table 1). However, despite this architectural diversity and experimental successes, a fundamental disconnect persists between these *in silico* predictions and therapeutic efficacy in clinical testing.[168] We discuss some of the critical parameters that serve as crucial barriers and emerging solutions to bridge the gap between computational and clinical perspectives.

Table 1 Comprehensive overview of AI-based tools for peptide-based drug discovery

Tool/model	Architecture/type	Primary application
PepCNN	Convolutional neural network	Peptide-binding residue prediction
CAMP	Multi-channel CNN	Multi-level peptide–protein interaction prediction
Cross-TCR-interpreter	Transformer	TCR-pMHC interaction prediction
ABTrans	Transformer	Peptide–antibody interaction
PepNN	Graph attention network	Sequence & structure-based PPI prediction
TPepPro	Transformer	Peptide–protein interaction prediction
TP-LMMSG	Hierarchical Multi-scale residual network + PLM	Therapeutic peptide prediction (AMP/AVP/ACP)
PepBCL	BERT-based	Peptide-binding residue prediction
PeptideBERT	BERT-based	Peptide property prediction
VAE-MH	VAE + Metropolis–Hastings	Peptide extension design
PepVAE	Semi-supervised VAE	Antimicrobial peptide generation
HydrAMP	Conditional VAE	Antimicrobial peptide design
PandoraGAN	GAN (LeakGAN-based)	Antiviral peptide generation
AMPGAN v2	Bidirectional conditional GAN	Antimicrobial peptide design
HelixGAN	GAN for structural design	De novo alpha-helix structural design
RFdiffusion	Diffusion model	Protein/peptide backbone generation
RFpeptides	Diffusion model for cyclic peptides	Cyclic peptide design
PepGLAD	Geometric latent diffusion	All-atom peptide generation
ProteinMPNN	Message-passing neural network	Sequence design from fixed backbones
DiffPepBuilder	SE(3)-equivariant diffusion	De novo peptide binder design
HYDRA	Target-aware diffusion	Target-aware peptide generation
PepTune	Masked diffusion + MCTS guidance	Multi-objective peptide optimization
AMPGen	Autoregressive diffusion + XGBoost + LSTM	Antimicrobial peptide generation
CPL-Diff	Mask-controlled diffusion	Length-tunable peptide generation
MMCD	Multimodal contrastive diffusion	Multimodal peptide generation
ProtFlow	Flow matching on compressed PLM embeddings	Fast peptide/antibody generation
HelixFlow	SE(3)-equivariant flow matching	Full-atom helical peptide design
NCFlow	Flow matching with non-canonical residues	Non-canonical amino acid incorporation
BindCraft	AlphaFold2-based hallucination	One-shot peptide binder design
PepMLM	Masked language model (fine-tuned ESM-2)	Target-conditioned peptide generation
AMP-Designer	Large language model	Rapid AMP design
LigandMPNN	Message-passing neural network (with ligands)	Sequence design with ligand context
CYC_BUILDER	Monte Carlo tree search + RL	Target-specific cyclic peptide design
PepINVENT	Sequence generator for non-canonical AAs	Non-natural amino acid exploration
AlphaFold3	Structure prediction (modified residues)	Complex modeling with modified residues
Chroma	Programmable generative model	Constrained protein/peptide generation
AMP-Diffusion	Latent diffusion model + PLM	Broad-spectrum AMP generation
ToxGIN	Graph isomorphism Network	Peptide toxicity prediction
tAMPer	Multimodal GNN + RNN (ESM-2 + ColabFold)	Structure-aware toxicity prediction
UnifyImmun	Transformer + Cross-attention	Immunogenicity prediction (HLA/TCR)
MahLooL	Bidirectional RNN	Peptide solubility prediction
StaPep	Structure prediction + Feature extraction	Stapled peptide structure prediction
Peptipedia v2.0	Database + ML models	Comprehensive peptide screening
Top-ML	Classifier	Anticancer peptide classification

The metrics-reality gap

Current AI-based models demonstrate considerable efficiency in predicting the binding affinities, structural stabilities, and AMP (toxic vs nontoxic) binary classification properties. However, the therapeutic efficacy is multifactorial, determined by various biological processes that extend far beyond simple parameters and features.[169] While AI/ML models achieve an AUROC value exceeding 0.95 in determining the AMPs, they may provide limited transferability in *in vivo* efficacy against distinct bacterial strains under varying physiological conditions, such as pH and ionic strength.[170] This disparity highlights the importance of bridging computational model development with experimental outcomes.

Off-target effects and selectivity

While computational predictions show success in estimating affinity for the designated targets, they also account for binding to off-targets. A critical gap that needs to be considered, which could lead to adverse effects.[171] In therapeutic peptide development, selectivity is relatively important; however, most computational screening pipelines prioritize binding over adequately assessing specificity.[172] Thus, complex models still struggle to distinguish between the highly selective binders and off-targets that would lead to toxicity in clinical testing.

Metabolic stability and pharmacokinetics

The half-life of peptides in physiological conditions often depends on a complex network of degradation pathways.[173] The current AI-based models solely rely on the predominance of *in vivo* assay data collected under controlled pH and other factors, with limited resemblance to the dynamic enzymatic environment. Recent studies have integrated the enzymatic cleavage site prediction and transfer learning approach to augment model performance,[174] but these approaches remain constrained by the shortage of pharmacokinetic data. Thus, a persistent gap still prevails between the predicted structural stability and the observed serum half-life, which continues to hinder the translation of computationally designed peptides into viable clinical candidates.

Immunogenicity and immune system interactions

The immunogenic response poses one of the most significant yet least predictable challenges in peptide-based drug development.[175] Although deep learning-based models for predicting T-cell epitopes and MHC binding have made substantial progress,[176] they fail to capture the full complexity of adaptive immune responses. For instance, various factors, such as peptide aggregation propensity, PTSMs, and patient-specific HLA, influence immunogenicity that current models cannot fully capture.[177] While discrepancies regarding progress towards the immune response or failed peptides are minimally reported. This prompts the model to focus solely on predictions that are overly optimistic and have a limited capacity to identify novel designs.

Physiological and microenvironmental factors

Peptide activity is highly dependent on various factors, including pH, ionic strength, lipid composition, and the presence of serum proteins.[178] AMPs that demonstrate potent activity against specific bacterial strains in standardized laboratories may lose efficacy in distinct microenvironments or when exposed to competitive host defense peptides.[179] Similarly, anticancer peptides optimized for binding to target proteins often exhibit reduced activity when competing with serum proteins or navigating the extracellular compartments.[180] These failures of the designed peptides often emerge from the limited training datasets. The scarcity of the training datasets spanning diverse microenvironments limits the model's generalizability and predictive power.

ADMET prediction limitations

Machine learning and AI-based studies addressing ADMET predictions have focused on critical gaps through reinforcement learning frameworks that jointly predict various endpoints, such as membrane permeability and stability.[181] Despite these advances in the AI/ML framework, the current models still remain constrained by biased training datasets and poor generalization to novel peptides. Critical properties such as bioavailability and drug–drug interactions, as well as the dose–response relationship, still remain inadequately modeled, which hinders the accurate prediction of clinical translatability and optimal dose regimen.[182] ADMET prediction will represent a significant opportunity until these critical gaps are addressed through experimental data collection and an improved modeling framework in computational peptide design.

Data quality and bias limitations

The quality of the training dataset represents a critical yet underappreciated limitation in the AI design.[183] Currently, most datasets in repositories are evaluated under narrow experimental conditions, such as specific strains, pH levels, ionic strengths, and defined cell lines. Models trained on those datasets achieve good performance but fail to yield better results on different experimental models or physiological conditions. Furthermore, the reported data from articles are biased towards positive results, which often overshadow the negative results and inactive compounds.[184–186]

Practical barriers to accessibility and equity

With the current developments in AI-based tools, another important factor to be monitored is the use of infrastructure for training computational models. This might be a barrier to upcoming scientific discoveries, as the size of the datasets is increasing exponentially.[187] Training advanced models demands high-performance computing resources and specialized software environments, thereby forming a hurdle to model development. The disparity in resources has also been observed in the research ecosystems, where the expert AI models remain concentrated in a small, well-resourced group. Such barriers may hinder innovations in peptide therapeutics by limiting participation from research groups with limited resources in developing countries.[188] Additionally, cloud-based platforms offer an alternative to this barrier, but they still incur costs that accumulate exponentially for intensive calculations. Furthermore, data privacy concerns persist when working with proprietary compounds, and the long-term use of these platforms remains uncertain.[189]

Risks associated with generative AI in peptide design

Current advances in generative deep-learning architectures for peptide design enable rapid development of novel peptides. However, these models tend to introduce a set of restraints that limit the translation from computational design to clinically effective therapeutics. For instance, one of the key factors is the ''hallucination'' where LLM models generate peptide sequences that are biophysically unstable and synthetically infeasible.[190] Moreover, the generative AI models frequently struggle to represent intrinsically disordered peptides (IDPs) or sequences that adopt ensemble states. This occurs due to training dataset bias, which heavily relies on rigid structures, thereby failing to capture the dynamic flexibility required for many biological interactions.[191] Most importantly, the development of peptides using generative AI, when misused, poses a significant risk to biosecurity and raises concerns about dual use. These could lead to the development of peptides with biologically harmful properties. A recent study demonstrated that AI can develop peptides that bind to the toxic substance while diverging sufficiently in their primary sequence to evade existing screening protocols.[192] This vulnerability poses a significant oversight in current biosafety frameworks and needs improved governance strategies.

Emerging solutions to bridge the gap between computational and experimental studies

Bridging the gap between computational and experimental outcomes requires a paradigm shift in the AI model development, training, and validation. With the current development of optimizations, multiple-objective training, and physics-based machine learning models ensure mechanical soundness. Further, incorporating feedback and continuous adaptation approaches for computational validation would advance AI-driven peptide design. Here, we report some recent advances in AI that could aid in bridging the gap between computational and experimental studies.

Agentic AI: adaptive multi-agent frameworks

A new generation of AI agents has emerged that differs from the conventional predictive models. Rather than operating as static pattern recognizers trained on a fixed dataset, agentic AI systems dynamically incorporate new experimental data and adapt to unexpected outcomes and reasons across diverse knowledge domains without requiring complete retraining.[193] These systems employ multi-agent architectures, where AI components cooperate with distinct expertise to collaborate autonomously and work together as a cohesive system.[194]

The ProtAgents framework demonstrates this approach by deploying multiple large language-based model agents, including a knowledge retrieval specialist, a structural analysis module, and a physics simulation engine that communicates through a structured protocol.[195] When designing peptides, these agents collaboratively integrate literature knowledge, predict structures, simulate binding interactions, and evaluate the outcomes. Critically, these agents can incorporate experimental results in real-time, adjusting their design strategy without retraining from scratch.

The Sparks system has further extended these concepts by implementing paired generation and reflection agents that autonomously brainstorm hypotheses, design computational experiments, execute simulations, and iteratively refine designs based on the outcomes.[196] Sparks has successfully identified unknown structure–property relationships in the protein mechanism, thus suggesting that these agentic systems can generate significant insights rather than merely interpolating existing knowledge. Besides the ability of the agentic system to reason about trade-off balancing conflicts, such as binding affinity *versus* toxicity or stability *versus* permeability, this shows a significant advance over the optimization approach.[197]

Chain-of-thought reasoning models

Conventional deep learning networks operate as black boxes, making predictions without providing any justification for their decisions.[198] The chain of thought reasoning models decompose complex design problems into logical steps, making their decision-making process interpretable.[199] This provides adequate information when the prediction fails experimentally, which aids the researchers in tracing back the reasoning chain to identify the missing considerations.

The PepThink-R1 model demonstrates this approach for cyclic peptide optimization by combining chain-of-thought reasoning with supervised fine-tuning and reinforcement learning. PepThink-R1 improved peptide sequences and explained the reasoning behind the modifications that enhance stability.[200] This enables the researchers to critically evaluate the AI suggestion and make informed decisions about the predicted candidate peptides before the experimental validation.

The ether0 model, which was trained on 24 billion parameters specifically for chemistry using reinforcement learning on more than 600 000 experimentally grounded problems spanning synthesizability, blood–brain barrier permeability, and metabolic activity. This model was designed to reason through multistep chemical problems using deterministic verification of correct vs incorrect solutions. The model's performance excels in specialized chemistry models and aligns with human experts in molecular design.[201]

The PRefLexOR framework demonstrates the preference-based recursive language models that can train AI systems to reason through multi-step scientific problems with explicit reflection and corrections.[202] PRefLexOR introduces thinking tokens that mark reasoning phases, enabling models to evaluate the reasoning strategy and regenerate improved solutions through refined cycles. Hence, the training methodology in these reasoning models represents a paradigm shift. Rather than relying on the pretraining datasets, these systems learn to reason through the problem using reinforcement learning. Thus, for peptide drug design specifically, chain-of-thought reasoning provides critical advantages in overcoming the barriers of data scarcity and the knowledge gap.

Physics-based data-driven models

The recent advancement in hybrid frameworks that combine the mechanistic insight of MD simulations with the scalability of ML. The physics-based force field developed in the MD simulations accurately captures the properties of proteins and peptides, including atomic interactions and thermodynamic behavior, which are often inaccessible to data-driven methods constrained by limited datasets. In turn, the ML models aid in extracting the complex interactions and patterns from the MD simulation data.[203,204] Recently, physics-informed training, where the MD-generated conformational ensembles and trajectories serve as training inputs, enables the complete pattern to be derived from simulation-derived data.[205] Active learning workflows further enhance efficiency by allowing ML models to validate top predictions and refine model accuracy.[206] This iterative, closed-loop integration not only preserves physical rigor but also accelerates discovery, offering a balanced and pragmatic path toward more predictive and experimentally relevant peptide design.

Ethical considerations in AI-driven peptide drug design

The development of AI in drug discovery raises a key challenge in terms of transparency, accountability, and bias.[207] Due to the limited interpretability of the model and the complexity of the training data, which tends to yield specific positive results, this raises concerns about the therapeutic outcomes. On the other hand, the misuse of AI-based technology for harmful purposes underscores the need for the development of stringent government policies and robust ethical oversigh.[208] In the absence of a harmonized regulatory framework, the need to develop transparency and sustainability with ethically responsible AI is a key factor to consider in the development of peptide-based rugs, with future directions toward clinical translation.

Challenges, opportunities, and future perspectives

Peptide-based therapeutics stand at the forefront, where AI-driven innovations accelerate drug discovery and development. Yet, the AI-based models still face significant hurdles in design and development. Data quality is a vital factor that provides the foundation for the AI models. Databases such as APd3 and CAMPR3 are curated for specific species, structural classes (alpha-helix, beta-sheet), and assay conditions. However, these databases are limited to non-canonical amino acids, stapled peptides, and D-amino acid modifications that have been proven to exhibit clinical importance. Model interpretability presents a significant challenge, as the need to explain AI models in drug discovery studies becomes essential for both scientific and regulatory acceptance.

The critical challenge in the AI-based peptide design is the limited use of empirical knowledge from chemistry and physics-based principles. Most models currently developed rely on statistical associations with the training dataset rather than thermodynamic and physicochemical properties that determine key properties, such as folding and proteolytic stability. Various studies have highlighted that peptide efficacy is strongly dependent on the physicochemical landscape, and that purely data-driven models frequently fail to capture in determining peptide efficacy.[209] However, emerging hybrid frameworks are bridging this gap by integrating molecular simulation data and quantum chemistry-derived potentials. These physics-informed strategies have demonstrated robustness for complex targets, including macrocyclic peptides, by effectively filtering ‘hallucinations’ and unrealistic conformational states.[210,211] Thus, embedding first-principles physics into generative pipelines represents a criterion for the future of reliable peptide-based therapeutics.

The integration of advanced computational approaches into peptide-based drug discovery offers immense opportunities. The use of deep learning models shows potential to enhance predictive accuracy in peptide stability, pharmacokinetics, and interactions, thereby facilitating rational design to improve efficacy. Integrating multiple AI approaches through ensemble methods and consensus scoring represents a promising direction for future research. From a scientific viewpoint, peptide applications are rapidly expanding beyond therapeutics to include biomaterials, diagnostics, and synthetic biology. However, this therapeutic expansion requires addressing the current challenges and the future directions. We can emphasize that the integration of algorithmic innovations and experimental validation in these domains accelerates the efficiency of the drug discovery pipeline, bridging the gap towards clinical and industrial implementations.

Author contributions

S. E. drafted the manuscript; N. V. D. reviewed and edited the manuscript and provided guidance and supervision throughout the work. All authors have approved the manuscript.

Conflicts of interest

There are no conflicts to declare.

Data availability

No primary research results, software, or code have been included, and no new data were generated or analyzed as part of this review.

Acknowledgements

We acknowledge support from the NIH grant 1R35GM134864 (to N. V. D.)

References

1 L. Wang, N. Wang, W. Zhang, X. Cheng, Z. Yan, G. Shao, X. Wang, R. Wang and C. Fu, *Signal Transduction Targeted Ther.*, 2022, 7, 48.

2 W. Xiao, W. Jiang, Z. Chen, Y. Huang, J. Mao, W. Zheng, Y. Hu and J. Shi, *Signal Transduction Targeted Ther.*, 2025, **10**, 74.

3 D. A. Scott and C. H. Best, *Ind. Eng. Chem.*, 1925, **17**, 238–240.

4 K. Fosgerau and T. Hoffmann, *Drug Discovery Today*, 2015, **20**, 122–128.

5 O. Al Musaimi, *J. Pept. Sci.*, 2024, **30**, e3627.

6 M. Shabsigh and L. A. Solomon, *Chem. Biomed. Imaging*, 2024, **2**, 615–630.

7 S. Gupta, N. Azadvari and P. Hosseinzadeh, *BioDesign Res.*, 2022, **2022**, 9783197.

8 L. R. Hoegen Dijkhof, T. K. E. Rönkkö, H. C. von Vegesack, J. Lenzing and A. S. Hauser, *Briefings Bioinf.*, 2025, **26**, bbaf186.

9 F. Lu, X. Zhang, Y. Geng, H. Wang and J. Tang, *Chem. Commun.*, 2024, **60**, 7942–7945.

10 C. Wynne and R. B. P. Elmes, *Sens. Diagn.*, 2024, **3**, 987–1013.

11 L. Lombardi, V. D. Genio, F. Albericio and D. R. Williams, *Chem. Rev.*, 2025, **125**, 7099–7166.

12 Y. Du, L. Li, Y. Zheng, J. Liu, J. Gong, Z. Qiu, Y. Li, J. Qiao and Y.-X. Huo, *Appl. Environ. Microbiol.*, 2022, **88**, e01617–22.

13 Q. Zhu, Z. Chen, P. K. Paul, Y. Lu, W. Wu and J. Qi, *Acta Pharm. Sin. B*, 2021, **11**, 2416–2448.

14 C. Lamers, *Future Drug Discovery*, 2022, **4**, FDD75.

15 J. Southworth, K. Migliaccio, J. Glover, J. Glover, D. Reed, C. McCarty, J. Brendemuhl and A. Thomas, *Comput. Educ.: Artif. Intell.*, 2023, **4**, 100127.

16 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Nature*, 2021, **596**, 583–589.

17 Y. Xu, X. Liu, X. Cao, C. Huang, E. Liu, S. Qian, X. Liu, Y. Wu, F. Dong, C.-W. Qiu, J. Qiu, K. Hua, W. Su, J. Wu, H. Xu, Y. Han, C. Fu, Z. Yin, M. Liu, R. Roepman, S. Dietmann, M. Virta, F. Kengara, Z. Zhang, L. Zhang, T. Zhao, J. Dai, J. Yang, L. Lan, M. Luo, Z. Liu, T. An, B. Zhang, X. He, S. Cong, X. Liu, W. Zhang, J. P. Lewis, J. M. Tiedje, Q. Wang, Z. An, F. Wang, L. Zhang, T. Huang, C. Lu, Z. Cai, F. Wang and J. Zhang, *Innovation*, 2021, **2**, 100179.

18 M. Elahi, S. O. Afolaranmi, J. L. Martinez Lastra and J. A. Perez Garcia, *Discov. Artif. Intell.*, 2023, **3**, 43.

19 Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido and A. Rives, *Science*, 2023, **379**, 1123–1130.

20 I. H. Sarker, *SN Comput. Sci.*, 2021, **2**, 420.

21 I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, *arXiv*, 2014, preprint, arXiv:1406.2661, DOI: 10.48550/arXiv.1406.2661.

22 L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui and M.-H. Yang, *arXiv*, 2024, preprint, arXiv:2209.00796, DOI: 10.48550/arXiv.2209.00796.

23 D. P. Kingma and M. Welling, *FNT in Machine Learning*, 2019, **12**, 307–392.

24 M. Goles, A. Daza, G. Cabas-Mora, L. Sarmiento-Varón, J. Sepúlveda-Yañez, H. Anvari-Kazemabad, M. D. Davari, R. Uribe-Paredes, Á. Olivera-Nappa, M. A. Navarrete and D. Medina-Ortiz, *Briefings Bioinf.*, 2024, **25**, bbae275.

25 S. Singh, H. Gupta, P. Sharma and S. Sahi, *Artif. Intell. Chem.*, 2024, **2**, 100039.

26 S. A. Rettie, K. V. Campbell, A. K. Bera, A. Kang, S. Kozlov, Y. F. Bueso, J. De La Cruz, M. Ahlrichs, S. Cheng, S. R. Gerben, M. Lamb, A. Murray, V. Adebomi, G. Zhou, F. DiMaio, S. Ovchinnikov and G. Bhardwaj, *Nat. Commun.*, 2025, **16**, 4730.

27 C.-H. Wu, I.-J. Liu, R.-M. Lu and H.-C. Wu, *J. Biomed. Sci.*, 2016, **23**, 8.

28 A. T. Tucker, S. P. Leonard, C. D. DuBois, G. A. Knauf, A. L. Cunningham, C. O. Wilke, M. S. Trent and B. W. Davies, *Cell*, 2018, **172**, 618–628.

29 W. Jaroszewicz, J. Morcinek-Orłowska, K. Pierzynowska, L. Gaffke and G. Węgrzyn, *FEMS Microbiol. Rev.*, 2022, **46**, fuab052.

30 A. B. Sloth, B. Bakhshinejad, M. Jensen, C. Stavnsbjerg, M. B. Liisberg, M. Rossing and A. Kjaer, *Viruses*, 2022, **14**, 2402.

31 A. Mondal, L. Chang and A. Perez, *QRB Discov.*, 2022, **3**, e17.

32 M. Ciemny, M. Kurcinski, K. Kamel, A. Kolinski, N. Alam, O. Schueler-Furman and S. Kmiecik, *Drug Discovery Today*, 2018, **23**, 1530–1537.

33 M. Kurcinski, M. Pawel Ciemny, T. Oleniecki, A. Kuriata, A. E. Badaczewska-Dawid, A. Kolinski and S. Kmiecik, *Bioinformatics*, 2019, **35**, 4170–4172.

34 P. Zhou, B. Jin, H. Li and S.-Y. Huang, *Nucleic Acids Res.*, 2018, **46**, W443–W450.

35 A. Khramushin, Z. Ben-Aharon, T. Tsaban, J. K. Varga, O. Avraham and O. Schueler-Furman, *Proc. Natl. Acad. Sci. U. S. A.*, 2022, **119**, e2121153119.

36 A. E. Badaczewska-Dawid, S. Kmiecik and M. Koliński, *Briefings Bioinf.*, 2021, **22**, bbaa109.

37 C. Agoni, R. Fernández-Díaz, P. B. Timmons, A. Adelfio, H. Gómez and D. C. Shields, Molecular Modelling in Bioactive Peptide Discovery and Characterisation, *Biomolecules*, 2025, **15**(4), 524, DOI: 10.3390/biom1504052, (accessed August 21, 2025).

38 A. Lamiable, P. Thévenet, J. Rey, M. Vavrusa, P. Derreumaux and P. Tufféry, *Nucleic Acids Res.*, 2016, **44**, W449–W454.

39 M. Vincenzi, F. A. Mercurio and M. Leone, *Int. J. Mol. Sci.*, 2024, **25**, 1798.

40 B. Xue, R. Li, Z. Cheng and X. Zhou, *ACS Cent. Sci.*, 2024, **10**, 2111–2118.

41 K. N. Amarasinghe, L. De Maria, C. Tyrchan, L. A. Eriksson, J. Sadowski and D. Petrović, *J. Chem. Inf. Model.*, 2022, **62**, 2999–3007.

42 J. Verma, V. M. Khedkar and E. C. Coutinho, *Curr. Top. Med. Chem.*, 2010, **10**, 95–115.

43 A. B. Nongonierma and R. J. FitzGerald, *RSC Adv.*, 2016, **6**, 75400–75413.

44 J. C. Nielsen, C. Hjorringgaard, M. M. Nygaard, A. Wester, L. Elster, T. Porsgaard, R. B. Mikkelsen, S. Rasmussen, A. N. Madsen, M. Schlein, N. Vrang, K. Rigbolt and L. S. Dalbooge, *J. Med. Chem.*, 2024, **67**, 11814–11826.

45 M. De Vivo, M. Masetti, G. Bottegoni and A. Cavalli, *J. Med. Chem.*, 2016, **59**, 4035–4061.

46 H. Geng, F. Chen, J. Ye and F. Jiang, *Comput. Struct. Biotechnol. J.*, 2019, **17**, 1162–1170.

47 E. A. Proctor and N. V. Dokholyan, *Curr. Opin. Struct. Biol.*, 2016, **37**, 9–13.

48 E. Srinivasan and R. Rajasekaran, *J. Neurol. Sci.*, 2019, **405**, 116425.

49 A. Tuan Do, T. Hai Nguyen, M. Quan Pham, H. Truong Nguyen, N. Phuoc Long, V. Van Vu, H. T. Thu Phung and S. Tung Ngo, *RSC Adv.*, 2025, **15**, 12866–12875.

50 F. Moraca, I. Vespoli, D. Mastroianni, V. Piscopo, R. Gaglione, A. Arciello, M. D. Nisco, S. Pacifico, B. Catalanotti and S. Pedatella, *RSC Med. Chem.*, 2024, **15**, 2286–2299.

51 G. Chandrasekhar, E. Srinivasan, S. Nandhini, G. Pravallika, G. Sanjay and R. Rajasekaran, *J. Biomol. Struct. Dyn.*, 2023, **0**, 1–12.

52 O. Dagliyan, E. A. Proctor, K. M. D'Auria, F. Ding and N. V. Dokholyan, *Structure*, 2011, **19**, 1837–1845.

53 M. M. Gomari, S. S. Arab, S. Balalaie, S. Ramezanpour, A. Hosseini, N. V. Dokholyan and P. Tarighi, *Proteins*, 2024, **92**, 76–95.

54 J. Hao, A. W. R. Serohijos, G. Newton, G. Tassone, Z. Wang, D. C. Sgroi, N. V. Dokholyan and J. P. Basilion, *PLoS Comput. Biol.*, 2008, **4**, e1000138.

55 F. Ding and N. V. Dokholyan, *PLoS Comput. Biol.*, 2006, **2**, e85.

56 J. Clayton, L. Baweja and J. Wereszczynski, *Methods Mol. Biol.*, 2022, **2405**, 151–167.

57 W. Wang, *Phys. Chem. Chem. Phys.*, 2021, **23**, 777–784.

58 R. Appadurai, J. Nagesh and A. Srivastava, *Nat. Commun.*, 2021, **12**, 958.
59 V. Salmaso, M. Sturlese, A. Cuzzolin and S. Moro, *Structure*, 2017, **25**, 655–662.
60 J. Ouyang, Y. Sheng and W. Wang, *Cells*, 2022, **11**, 4016.
61 C. Agoni, R. Fernández-Díaz, P. B. Timmons, A. Adelfio, H. Gómez and D. C. Shields, *Biomolecules*, 2025, **15**, 524.
62 B. Senthilkumar, D. Meshach Paul, E. Srinivasan and R. Rajasekaran, *J. Cluster Sci.*, 2017, **28**, 2549–2563.
63 P. C. Sekar, D. M. Paul, E. Srinivasan and R. Rajasekaran, *J. Mol. Model.*, 2021, **27**(1), 10.
64 P. C. Sekar, E. Srinivasan, G. Chandrasekhar, D. M. Paul, G. Sanjay, S. Surya, N. S. A. R. Kumar and R. Rajasekaran, *J. Mol. Model.*, 2022, **28**, 128.
65 F. Ding, D. Tsao, H. Nie and N. V. Dokholyan, *Structure*, 2008, **16**, 1010–1018.
66 E. A. Proctor and N. V. Dokholyan, *Curr. Opin. Struct. Biol.*, 2016, **37**, 9–13.
67 S. Peng, F. Ding, B. Urbanc, S. V. Buldyrev, L. Cruz, H. E. Stanley and N. V. Dokholyan, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2004, **69**, 041908.
68 S. Barton, R. Jacak, S. D. Khare, F. Ding and N. V. Dokholyan, *J. Biol. Chem.*, 2007, **282**, 25487–25492.
69 F. Ding, J. M. Borreguero, S. V. Buldyrey, H. E. Stanley and N. V. Dokholyan, *Proteins*, 2003, **53**, 220–228.
70 S. D. Khare and N. V. Dokholyan, *Curr. Protein Pept. Sci.*, 2007, **8**, 573–579.
71 D. Tsao and N. V. Dokholyan, *Phys. Chem. Chem. Phys.*, 2010, **12**, 3491–3500.
72 G. Chandrasekhar and R. Rajasekaran, *Int. J. Pept. Res. Ther.*, 2021, **27**, 1555–1575.
73 P. C. Sekar and R. Rajasekaran, *Int. J. Pept. Res. Ther.*, 2021, **27**, 1043–1056.
74 J. Hao, A. W. R. Serohijos, G. Newton, G. Tassone, Z. Wang, D. C. Sgroi, N. V. Dokholyan and J. P. Basilion, *PLoS Comput. Biol.*, 2008, **4**, e1000138.
75 M. M. Gomari, S. S. Arab, S. Balalaie, S. Ramezanpour, A. Hosseini, N. V. Dokholyan and P. Tarighi, *Proteins*, 2024, **92**, 76–95.
76 C. Coppa, A. Bazzoli, M. Barkhordari and A. Contini, *J. Chem. Inf. Model.*, 2023, **63**, 3030–3042.
77 J. Miao, A. P. Ghosh, M. N. Ho, C. Li, X. Huang, B. L. Pentelute, J. D. Baleja and Y.-S. Lin, *J. Phys. Chem. B*, 2024, **128**, 5281–5292.
78 M. J. Amundarain, A. Vietri, V. I. Dodero and M. D. Costabel, *J. Phys. Chem. B*, 2023, **127**, 2407–2417.
79 P. Robustelli, S. Piana and D. E. Shaw, *Proc. Natl. Acad. Sci. U. S. A.*, 2018, **115**, E4758–E4766.
80 J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmüller and A. D. MacKerell, *Nat. Methods*, 2017, **14**, 71–73.
81 T. Hou, J. Wang, Y. Li and W. Wang, *J. Chem. Inf. Model.*, 2011, **51**, 69–82.
82 G. Rossino, E. Marchese, G. Galli, F. Verde, M. Finizio, M. Serra, P. Linciano and S. Collina, *Molecules*, 2023, **28**, 7165.
83 J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis and J. M. Jumper, *Nature*, 2024, **630**, 493–500.
84 E. F. McDonald, T. Jones, L. Plate, J. Meiler and A. Gulsevin, *Structure*, 2023, **31**, 111–119.
85 M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read and D. Baker, *Science*, 2021, **373**, 871–876.
86 M. Baek, R. McHugh, I. Anishchenko, H. Jiang, D. Baker and F. DiMaio, *Nat. Methods*, 2024, **21**, 117–121.
87 M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov and M. Steinegger, *Nat. Methods*, 2022, **19**, 679–682.
88 M. Mou, Z. Zhang, Z. Pan and F. Zhu, *Research*, 2025, **8**, 0615.
89 A. Chandra, A. Sharma, I. Dehzangi, T. Tsunoda and A. Sattar, *Sci. Rep.*, 2023, **13**, 20882.
90 Y. Lei, S. Li, Z. Liu, F. Wan, T. Tian, S. Li, D. Zhao and J. Zeng, *Nat. Commun.*, 2021, **12**, 5465.
91 K. Koyama, K. Hashimoto, C. Nagao and K. Mizuguchi, *Front. Bioinform.*, 2023, **3**, 1274599.
92 O. Abdin, S. Nim, H. Wen and P. M. Kim, *Commun. Biol.*, 2022, **5**, 503.
93 X. Jin, Z. Chen, D. Yu, Q. Jiang, Z. Chen, B. Yan, J. Qin, Y. Liu and J. Wang, *Bioinformatics*, 2024, **41**, btae708.
94 N. Pancino, C. Gallegati, F. Romagnoli, P. Bongini and M. Bianchini, *Int. J. Mol. Sci.*, 2024, **25**, 5870.
95 N. Chen, J. Yu, L. Zhe, F. Wang, X. Li and K.-C. Wong, *Briefings Bioinf.*, 2024, **25**, bbae308.
96 K. Jha, S. Saha and H. Singh, *Sci. Rep.*, 2022, **12**, 8360.
97 K. Jha, S. Karmakar and S. Saha, *Sci. Rep.*, 2023, **13**, 5663.
98 M. Ansari and A. D. White, *J. Chem. Inf. Model.*, 2023, **63**, 2546–2553.
99 J. Mao, S. Guan, Y. Chen, A. Zeb, Q. Sun, R. Lu, J. Dong, J. Wang and D. Cao, *Comput. Struct. Biotechnol. J.*, 2023, **21**, 463–471.
100 A. Papanikolaou, V. Sivtsov, E. Zereik, E. Ruggiero, C. Bonini and F. Bonsignorio, *bioRxiv*, 2025, preprint, DOI: 10.1101/2025.03.18.643937.
101 R. Wang, K. Nakai and L. Wei, *Methods Mol. Biol.*, 2025, **2941**, 269–278.
102 C. Guntuboina, A. Das, P. Mollaei, S. Kim and A. Barati Farimani, *J. Phys. Chem. Lett.*, 2023, **14**, 10427–10434.
103 J. C. Siwek, A. A. Omelchenko, P. Chhibbar, S. Arshad, A. Rosengart, I. Nazarali, A. Patel, K. Nazarali, J. Rahimikollu, J. S. Tilstra, M. J. Shlomchik, D. R. Koes, A. V. Joglekar and J. Das, *Nat. Methods*, 2025, **22**, 1707–1719.
104 S. Chen, T. Lin, R. Basu, J. Ritchey, S. Wang, Y. Luo, X. Li, D. Pei, L. B. Kara and X. Cheng, *Nat. Commun.*, 2024, **15**, 1611.
105 S. N. Dean and S. A. Walper, *ACS Omega*, 2020, **5**, 20746–20754.
106 S. N. Dean, J. A. E. Alvarez, D. Zabetakis, S. A. Walper and A. P. Malanoski, *Front. Microbiol.*, 2021, **12**, 725727.
107 P. Szymczak, M. Możejko, T. Grzegorzek, R. Jurczak, M. Bauer, D. Neubauer, K. Sikora, M. Michalski, J. Sroka, P. Setny, W. Kamysz and E. Szczurek, *Nat. Commun.*, 2023, **14**, 1453.
108 S. Surana, P. Arora, D. Singh, D. Sahasrabuddhe and J. Valadi, *SN Comput. Sci.*, 2023, **4**, 607.
109 C. M. Van Oort, J. B. Ferrell, J. M. Remington, S. Wshah and J. Li, *J. Chem. Inf. Model.*, 2021, **61**, 2198–2207.
110 X. Xie, P. A. Valiente and P. M. Kim, *Bioinformatics*, 2023, **39**, btad036.
111 J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek and D. Baker, *Nature*, 2023, **620**, 1089–1100.
112 S. Vázquez Torres, P. J. Y. Leung, P. Venkatesh, I. D. Lutz, F. Hink, H.-H. Huynh, J. Becker, A. H.-W. Yeh, D. Juergens, N. R. Bennett, A. N. Hoofnagle, E. Huang, M. J. MacCoss, M. Expòsit, G. R. Lee, A. K. Bera, A. Kang, J. De La Cruz, P. M. Levine, X. Li, M. Lamb, S. R. Gerben, A. Murray, P. Heine, E. N. Korkmaz, J. Nivala, L. Stewart, J. L. Watson, J. M. Rogers and D. Baker, *Nature*, 2024, **626**, 435–442.
113 S. A. Rettie, D. Juergens, V. Adebomi, Y. F. Bueso, Q. Zhao, A. N. Leveille, A. Liu, A. K. Bera, J. A. Wilms, A. Üffing, A. Kang, E. Brackenbrough, M. Lamb, S. R. Gerben, A. Murray, P. M. Levine, M. Schneider, V. Vasireddy, S. Ovchinnikov, O. H. Weiergräber, D. Willbold, J. A. Kritzer, J. D. Mougous, D. Baker, F. DiMaio and G. Bhardwaj, *bioRxiv*, 2024, preprint, DOI: 10.1101/2024.11.18.622547.
114 X. Kong, Y. Jia, W. Huang and Y. Liu, *arXiv*, 2024, preprint, arXiv:2402.13555, DOI: 10.48550/arXiv.2402.13555.
115 J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King and D. Baker, *Science*, 2022, **378**, 49–56.
116 F. Wang, Y. Wang, L. Feng, C. Zhang and L. Lai, *J. Chem. Inf. Model.*, 2024, **64**, 9135–9149.

117 V. S. R, S. Choudhuri and B. Ghosh, *J. Chem. Inf. Model.*, 2024, **64**, 6912–6925.
118 S. Tang, Y. Zhang and P. Chatterjee, arXiv, 2025, arXiv:2412.17780v4, , DOI: 10.48550/2412.17780v4.
119 S. Jin, Z. Zeng, X. Xiong, B. Huang, L. Tang, H. Wang, X. Ma, X. Tang, G. Shao, X. Huang and F. Lin, *Commun. Biol.*, 2025, **8**, 839.
120 Z. Luo, A. Geng, L. Wei, Q. Zou, F. Cui and Z. Zhang, *Adv. Sci.*, 2025, **12**, 2412926.
121 Y. Wang, X. Liu, F. Huang, Z. Xiong and W. Zhang, *arXiv*, 2024, preprint, arXiv:2312.15665, DOI: 10.48550/arXiv.2312.15665.
122 Z. Li, Z. Zeng, X. Lin, F. Fang, Y. Qu, Z. Xu, Z. Liu, X. Ning, T. Wei, G. Liu, H. Tong and J. He, *arXiv*, 2025, preprint, arXiv:2507.17731, DOI: 10.48550/arXiv.2507.17731.
123 Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel and M. Le, *arXiv*, 2023, preprint, arXiv:2210.02747, DOI: 10.48550/arXiv.2210.02747.
124 Z. Kong, Y. Zhu, Y. Xu, H. Zhou, M. Yin, J. Wu, H. Xu, C.-Y. Hsieh, T. Hou and J. Wu, *arXiv*, 2025, preprint, arXiv:2504.10983, DOI: 10.48550/arXiv.2504.10983.
125 X. Xie, P. A. Valiente, J. Kim, J. S. Lee and P. M. Kim.
126 J. S. Lee and P. M. Kim, *bioRxiv*, 2025, preprint, DOI: 10.1101/2025.07.31.667780.
127 M. Pacesa, L. Nickel, C. Schellhaas, J. Schmidt, E. Pyatova, L. Kissling, P. Barendse, J. Choudhury, S. Kapoor, A. Alcaraz-Serna, Y. Cho, K. H. Ghamary, L. Vinué, B. J. Yachnin, A. M. Wollacott, S. Buckley, A. H. Westphal, S. Lindhoud, S. Georgeon, C. A. Goverde, G. N. Hatzopoulos, P. Gönczy, Y. D. Muller, G. Schwank, D. C. Swarts, A. J. Vecchio, B. L. Schneider, S. Ovchinnikov and B. E. Correia, *Nature*, 2025, **646**, 483–492.
128 M. Filius, T. Patsos, H. Minee, G. Turco, J. Liu, M. Gnatzy, R. S. M. Rooth, A. C. H. Liu, R. D. T. Ta, I. H. A. Rijk, S. Ziani, F. J. Boxman and S. J. Pomplun, *bioRxiv*, 2025, preprint, DOI: 10.1101/2025.07.23.666285.
129 T. Chen, M. Dumas, R. Watson, S. Vincoff, C. Peng, L. Zhao, L. Hong, S. Pertsemlidis, M. Shaepers-Cheu, T. Z. Wang, D. Srijay, C. Monticello, P. Vure, R. Pulugurta, K. Kholina, S. Goel, M. P. DeLisa, R. Truant, H. C. Aguilar and P. Chatterjee, *arXiv*, 2024, preprint, arXiv:2310.03842v3, DOI: 10.48550/2310.03842v3.
130 S. Bhat, K. Palepu, L. Hong, J. Mao, T. Ye, R. Iyer, L. Zhao, T. Chen, S. Vincoff, R. Watson, T. Z. Wang, D. Srijay, V. S. Kavirayuni, K. Kholina, S. Goel, P. Vure, A. J. Deshpande, S. H. Soderling, M. P. DeLisa and P. Chatterjee, *Sci. Adv.*, 2025, **11**, eadr8638.
131 J. Wang, J. Feng, Y. Kang, P. Pan, J. Ge, Y. Wang, M. Wang, Z. Wu, X. Zhang, J. Yu, X. Zhang, T. Wang, L. Wen, G. Yan, Y. Deng, H. Shi, C.-Y. Hsieh, Z. Jiang and T. Hou, *Sci. Adv.*, 2025, **11**, eads8932.
132 J. Dauparas, G. R. Lee, R. Pecoraro, L. An, I. Anishchenko, C. Glasscock and D. Baker, *Nat. Methods*, 2025, **22**, 717–723.
133 F. Wang, T. Zhang, J. Zhu, X. Zhang, C. Zhang and L. Lai, Reinforcement Learning-Based Target-Specific *De Novo* Design of Cyclic Peptide Binders, *J. Med. Chem.*, 2025, **68**(16), 17287–17302, DOI: 10.1021/acs.jmedchem.5c00789, (accessed August 22, 2025).
134 Q. Wang, X. Hu, Z. Wei, H. Lu and H. Liu, Reinforcement learning-driven exploration of peptide space: accelerating generation of drug-like peptides, *Briefings Bioinf.*, 2024, **25**(5), bbae444.
135 G. Geylan, J. P. Janet, A. Tibo, J. He, A. Patronov, M. Kabeshov, W. Czechtizky, F. David, O. Engkvist and L. De Maria, *Chem. Sci.*, 2025, **16**, 8682–8696.
136 C. Zhang, W. Wang, N. Zhu, Z. Cao, Q. Mao, C. Zhu, C. Zhang, J. Guo and H. Duan, *bioRxiv*, 2025, preprint, DOI: 10.1101/2025.05.24.655528.
137 J. B. Ingraham, M. Baranov, Z. Costello, K. W. Barber, W. Wang, A. Ismail, V. Frappier, D. M. Lord, C. Ng-Thow-Hing, E. R. Van Vlack, S. Tie, V. Xue, S. C. Cowles, A. Leung, J. V. Rodrigues, C. L. Morales-Perez, A. M. Ayoub, R. Green, K. Puentes, F. Oplinger, N. V. Panwar, F. Obermeyer, A. R. Root, A. L. Beam, F. J. Poelwijk and G. Grigoryan, Illuminating protein space with a programmable generative model, *Nature*, 2023, **623**(7989), 1070–1078, DOI: 10.1038/s41586-023-06728-8, (accessed August 22, 2025).
138 N. Qayyum, H. Seo, N. Khan, A. Manan, R. Ramachandran, M. Haseeb, E. Kim and S. Choi, *Int. J. Biol. Macromol.*, 2025, **316**, 144652.
139 L. Chang, A. Mondal, B. Singh, Y. Martínez-Noa and A. Perez, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2024, **14**, e1693.
140 M. D. T. Torres, T. Chen, F. Wan, P. Chatterjee and C. de la Fuente-Nunez, *bioRxiv*, 2025, preprint, DOI: 10.1101/2025.01.31.636003.
141 R. Chen, Y. You, Y. Liu, X. Sun, T. Ma, X. Lao and H. Zheng, *Microb. Biotechnol.*, 2025, **18**, e70121.
142 J. Wang, J. Feng, Y. Kang, P. Pan, J. Ge, Y. Wang, M. Wang, Z. Wu, X. Zhang, J. Yu, X. Zhang, T. Wang, L. Wen, G. Yan, Y. Deng, H. Shi, C.-Y. Hsieh, Z. Jiang and T. Hou, *Sci. Adv.*, 2025, **11**, eads8932.
143 A. Mesa, A. Orrego, J. W. Branch-Bedoya, C. Mera-Banguero and S. Orduz, *Curr. Microbiol.*, 2025, **82**, 379.
144 S. Chen, T. Lin, R. Basu, J. Ritchey, S. Wang, Y. Luo, X. Li, D. Pei, L. B. Kara and X. Cheng, *Nat. Commun.*, 2024, **15**, 1611.
145 W. Zhao, K. Hou, Y. Shen and X. Hu, *Bioinformatics*, 2025, **41**, btaf069.
146 Y. Wang, M. Song, F. Liu, Z. Liang, R. Hong, Y. Dong, H. Luan, X. Fu, W. Yuan, W. Fang, G. Li, H. Lou and W. Chang, *Sci. Adv.*, 2025, **11**, eadp7171.
147 P. Das, T. Sercu, K. Wadhawan, I. Padhi, S. Gehrmann, F. Cipcigan, V. Chenthamarakshan, H. Strobelt, C. dos Santos, P.-Y. Chen, Y. Y. Yang, J. P. K. Tan, J. Hedrick, J. Crain and A. Mojsilovic, *Nat. Biomed. Eng.*, 2021, **5**, 613–623.
148 Q. Yu, Z. Zhang, G. Liu, W. Li and Y. Tang, *Briefings Bioinf.*, 2024, **25**, bbae583.
149 H. Ebrahimikondori, D. Sutherland, A. Yanai, A. Richter, A. Salehi, C. Li, L. Coombe, M. Kotkoff, R. L. Warren and I. Birol, *Protein Sci.*, 2024, **33**, e5076.
150 C. Yu, X. Fang, S. Tian and H. Liu, *Nat. Mach. Intell.*, 2025, **7**, 278–292.
151 Y. Li, M. Wu, Y. Fu, J. Xue, F. Yuan, T. Qu, A. N. Rissanou, Y. Wang, X. Li and H. Hu, *Pharmacol. Res.*, 2024, **203**, 107137.
152 Z. Wang, J. Wu, M. Zheng, C. Geng, B. Zhen, W. Zhang, H. Wu, Z. Xu, G. Xu, S. Chen and X. Li, *J. Chem. Inf. Model.*, 2024, **64**, 9361–9373.
153 R. Chen, Y. You, Y. Liu, X. Sun, T. Ma, X. Lao and H. Zheng, *Microb. Biotechnol.*, 2025, **18**, e70121.
154 G. Cabas-Mora, A. Daza, N. Soto-García, V. Garrido, D. Alvarez, M. Navarrete, L. Sarmiento-Varón, J. H. Sepúlveda Yañez, M. D. Davari, F. Cadet, Á. Olivera-Nappa, R. Uribe-Paredes and D. Medina-Ortiz, *Database*, 2024, **2024**, baae113.
155 J. Z. E. Tan, J. Wee, X. Gong and K. Xia, *J. Chem. Inf. Model.*, 2025, **65**, 4232–4242.
156 D. Dahl, Y. Onishi, P. Norwood, R. Huh, R. Bray, H. Patel and Á. Rodríguez, *JAMA*, 2022, **327**, 534–545.
157 L. J. Aronne, N. Sattar, D. B. Horn, H. E. Bays, S. Wharton, W.-Y. Lin, N. N. Ahmad, S. Zhang, R. Liao, M. C. Bunck, I. Jouravskaya and M. A. Murphy, and SURMOUNT-4 Investigators, *JAMA*, 2024, **331**, 38–48.
158 S. Wharton, C. W. le Roux, M. N. Kosiborod, E. Platz, M. Brueckmann, A. M. Jastreboff, S. Ajaz Hussain, S. D. Pedersen, L. Borowska, A. Unseld, I. M. Kloer and L. M. Kaplan, *Obesity*, 2025, **33**, 67–77.
159 P. B. Jeppesen, T. Vanuytsel, S. Subramanian, F. Joly, G. Wanten, G. Lamprecht, M. Kunecki, F. Rahman, T. S. S. Nielsen, M. Berner-Hansen, U.-F. Pape and D. F. Mercer, *Gastroenterology*, 2025, **168**, 701–713.
160 U. Hennrich and M. Eder, *Pharmaceuticals*, 2022, **15**, 1292.
161 C. Solis-Herrera, M. P. Kane and C. Triplitt, *Clin. Diabetes*, 2024, **42**, 74–86.
162 V. Guerlavais, T. K. Sawyer, L. Carvajal, Y. S. Chang, B. Graves, J.-G. Ren, D. Sutton, K. A. Olson, K. Packman, K. Darlak, C. Elkin, E. Feyfant, K. Kesavan, P. Gangurde, L. T. Vassilev, H. M. Nash, V. Vukovic, M. Aivado and D. A. Annis, *J. Med. Chem.*, 2023, **66**, 9401–9417.
163 S. Kim, Y. H. No, R. Sluyter, K. Konstantinov, Y. H. Kim and J. H. Kim, *Coord. Chem. Rev.*, 2024, **500**, 215530.
164 D. Wang, F. Yin, Z. Li, Y. Zhang and C. Shi, *J. Nanobiotechnol.*, 2025, **23**, 305.
165 N. Nissan, M. C. Allen, D. Sabatino and K. K. Biggar, *Biomolecules*, 2024, **14**, 1303.
166 O. Bayley, E. Savino, A. Slattery and T. Noël, *Matter*, 2024, **7**, 2382–2398.
167 Y. N. Talluri, S. K. Sankaranarayanan, H. C. Fry and R. Batra, *Sci. Adv.*, 2025, **11**, eadt9466.
168 T. D. D. Kazmirchuk, C. Bradbury-Jost, T. A. Withey, T. Gessese, T. Azad, B. Samanfar, F. Dehne and A. Golshani, *Genes*, 2023, **14**, 1194.

169 M. Z. Gładysz, M. Stevanoska, M. K. Włodarczyk-Biegun and A. Nagelkerke, *Adv. Drug Delivery Rev.*, 2022, **184**, 114183.
170 R. Yang, X. Ma, F. Peng, J. Wen, L. W. Allahou, G. R. Williams, J. C. Knowles and A. Poma, *Biotechnol. Adv.*, 2025, **81**, 108570.
171 L. Xie, L. Xie and P. E. Bourne, *Curr. Opin. Struct. Biol.*, 2011, **21**, 189–199.
172 A. Mondal, B. Singh, R. H. Felkner, A. D. Falco, G. Swapna, G. T. Montelione, M. J. Roth and A. Perez, *Angew. Chem., Int. Ed.*, 2024, **63**, e202405767.
173 J. Pei, X. Gao, D. Pan, Y. Hua, J. He, Z. Liu and Y. Dang, *Curr. Res. Food Sci.*, 2022, **5**, 2162–2170.
174 X. Tan, Q. Liu, Y. Fang, S. Yang, F. Chen, J. Wang, D. Ouyang, J. Dong and W. Zeng, *Briefings Bioinf.*, 2024, **25**, bbae350.
175 K. Achilleos, C. Petrou, V. Nicolaidou and Y. Sarigiannis, *J. Pept. Sci.*, 2025, **31**, e70016.
176 S. Y. Seo and J.-K. Rhee, *Bioinformatics*, 2025, **41**, i125–i132.
177 M. Puig and S. Shubow, *Front. Immunol.*, 2025, **16**, 1608401.
178 B. Ha Gan, J. Gaynord, S. M. Rowe, T. Deingruber and D. R. Spring, *Chem. Soc. Rev.*, 2021, **50**, 7820–7880.
179 C. G. Starr, J. He and W. C. Wimley, *ACS Chem. Biol.*, 2016, **11**, 3391–3399.
180 G. Ghaly, H. Tallima, E. Dabbish, N. Badr ElDin, M. K. Abd El-Rahman, M. A. A. Ibrahim and T. Shoeib, *Molecules*, 2023, **28**, 1148.
181 S. Hashemi, P. Vosough, S. Taghizadeh and A. Savardashtaki, *Heliyon*, 2024, **10**(22), e40265.
182 M. Venkataraman, G. C. Rao, J. K. Madavareddi and S. R. Maddi, *ADMET DMPK*, 2025, **13**, 2772.
183 L. Alzubaidi, J. Bai, A. Al-Sabaawi, J. Santamaría, A. S. Albahri, B. S. N. Al-dabbagh, M. A. Fadhel, M. Manoufali, J. Zhang, A. H. Al-Timemy, Y. Duan, A. Abdullah, L. Farhan, Y. Lu, A. Gupta, F. Albu, A. Abbosh and Y. Gu, *J. Big Data*, 2023, **10**, 46.
184 F. Wan, F. Wong, J. J. Collins and C. de la Fuente-Nunez, *Nat. Rev. Bioeng.*, 2024, **2**, 392–407.
185 K. Sidorczuk, P. Gagat, F. Pietluch, J. Kała, D. Rafacz, L. Bąkała, J. Słowik, R. Kolenda, S. Rödiger, L. C. H. W. Fingerhut, I. R. Cooke, P. Mackiewicz and M. Burdukiewicz, *Briefings Bioinf.*, 2022, **23**, bbac343.
186 C. A. Brizuela, G. Liu, J. M. Stokes and C. de la Fuente-Nunez, *Microb. Biotechnol.*, 2025, **18**, e70072.
187 L. Bornmann, R. Haunschild and R. Mutz, *Humanit. Soc. Sci. Commun.*, 2021, **8**, 224.
188 D. Deshpande, K. Chhugani, T. Ramesh, M. Pellegrini, S. Shiffman, M. S. Abedalthagafi, S. Alqahtani, J. Ye, X. Shirley Liu, J. T. Leek, A. Brazma, R. A. Ophoff, G. Rao, A. J. Butte, J. H. Moore, V. Katritch and S. Mangul, *Cell*, 2024, **187**, 4449–4457.
189 Y. Joly, S. O. M. Dyke, B. M. Knoppers and T. Pastinen, *Cell*, 2016, **167**, 1150–1154.
190 H. Li, L. Lv, H. Cao, Z. Liu, Z. Yan, Y. Wang, Y. Tian, Y. Li and L. Yuan, *arXiv*, 2025, preprint, arXiv:2504.12314, DOI: 10.48550/arXiv.2504.12314.
191 A. Abbaszadeh and A. Shahlaee, *arXiv*, 2025, preprint, arXiv:2508.18446, DOI: 10.48550/arXiv.2508.18446.
192 B. J. Wittmann, T. Alexanian, C. Bartling, J. Beal, A. Clore, J. Diggans, K. Flyangolts, B. T. Gemler, T. Mitchell, S. T. Murphy, N. E. Wheeler and E. Horvitz, *Science*, 2025, **390**, 82–87.
193 U. Nisa, M. Shirazi, M. A. Saip and M. S. M. Pozi, *J. Autom. Intell.*, 2025, DOI: 10.1016/j.jai.2025.08.003.
194 S. Hosseini and H. Seilani, *Array*, 2025, **26**, 100399.
195 A. Ghafarollahi and M. J. Buehler, *Digital Discovery*, 2024, **3**, 1389–1409.
196 A. Ghafarollahi and M. J. Buehler, *arXiv*, 2025, preprint, arXiv:2504.19017, DOI: 10.48550/arXiv.2504.19017.
197 A. Ünlü, P. Rohr and A. Celebi, *arXiv*, 2025, preprint, arXiv:2508.03444, DOI: 10.48550/arXiv.2508.03444.
198 C. Rudin, *Nat. Mach. Intell.*, 2019, **1**, 206–215.
199 J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le and D. Zhou, *arXiv*, 2023, preprint, arXiv:2201.11903, DOI: 10.48550/arXiv.2201.11903.
200 R. Wang, H. Zhang, T. Nguyen, S. Feng, H.-W. Pang, X. Yu, L. Xiao and P. Z. Zhang, *arXiv*, 2025, preprint, arXiv:2508.14765, DOI: 10.48550/arXiv.2508.14765.
201 S. M. Narayanan, J. D. Braza, R.-R. Griffiths, A. Bou, G. Wellawatte, M. C. Ramos, L. Mitchener, S. G. Rodriques and A. D. White, *arXiv*, 2025, preprint, arXiv:2506.17238, DOI: 10.48550/arXiv.2506.17238.
202 M. J. Buehler, *npj Artif. Intell.*, 2025, **1**, 4.
203 C. M. Sha, J. Wang and N. V. Dokholyan, *Briefings Bioinf.*, 2024, **25**, bbad456.
204 O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko and K.-R. Müller, *Chem. Rev.*, 2021, **121**, 10142–10186.
205 G. P. P. Pun, R. Batra, R. Ramprasad and Y. Mishin, *Nat. Commun.*, 2019, **10**, 2339.
206 A. Venkatraman, M. A. Wilson and D. Montes de Oca Zapiain, *npj Comput. Mater.*, 2025, **11**, 24.
207 A. L. Boudi, M. Boudi, C. Chan and F. B. Boudi, *Cureus*, 2024, **16**(11), e74495.
208 D. Bloomfield, J. Pannu, A. W. Zhu, M. Y. Ng, A. Lewis, E. Bendavid, S. M. Asch, T. Hernandez-Boussard, A. Cicero and T. Inglesby, *Science*, 2024, **385**, 831–833.
209 S. Zhai, T. Liu, S. Lin, D. Li, H. Liu, X. Yao and T. Hou, *Drug Discovery Today*, 2025, **30**, 104300.
210 D. de Raffele and I. M. Ilie, *Chem. Commun.*, 2024, **60**, 632–645.
211 S. Yang, J. Ren, W. Gao, L. Cao and S. Ling, *npj Soft Matter*, 2025, **1**, 4.