OXFORD

Briefings in Bioinformatics, 2024, **25(4)**, bbae275
https://doi.org/10.1093/bib/bbae275
Review

Peptide-based drug discovery through artificial intelligence: towards an autonomous design of therapeutic peptides

Montserrat Goles [1,2], Anamaría Daza [3], Gabriel Cabas-Mora [1], Lindybeth Sarmiento-Varón [4], Julieta Sepúlveda-Yañez [5], Hoda Anvari-Kazemabad [1], Mehdi D. Davari [6], Roberto Uribe-Paredes [1], Álvaro Olivera-Nappa [3], Marcelo A. Navarrete [4,7,*], David Medina-Ortiz [1,3,*]

[1]Departamento de Ingeniería en Computación, Universidad de Magallanes, Av. Pdte. Manuel Bulnes 01855, 6210427, Punta Arenas, Chile
[2]Departamento de Ingeniería Química, Biotecnología y Materiales, Universidad de Chile, Beauchef 851, 8370456, Santiago, Chile
[3]Centre for Biotechnology and Bioengineering, CeBiB, Universidad de Chile, Beauchef 851, 8370456, Santiago, Chile
[4]Centro Asistencial de Docencia e Investigación, CADI, Universidad de Magallanes, Av. Los Flamencos 01364, 6210005, Punta Arenas, Chile
[5]Facultad de Ciencias de la Salud, Universidad de Magallanes, Av. Pdte. Manuel Bulnes 01855, 6210427, Punta Arenas, Chile
[6]Department of Bioorganic Chemistry, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120, Halle, Germany
[7]Escuela de Medicina, Universidad de Magallanes, Av. Pdte. Manuel Bulnes 01855, 6210427, Punta Arenas, Chile

*Corresponding authors. E-mail: marcelo.navarrete@umag.cl; david.medina@cebib.cl

Abstract

With their diverse biological activities, peptides are promising candidates for therapeutic applications, showing antimicrobial, antitumour and hormonal signalling capabilities. Despite their advantages, therapeutic peptides face challenges such as short half-life, limited oral bioavailability and susceptibility to plasma degradation. The rise of computational tools and artificial intelligence (AI) in peptide research has spurred the development of advanced methodologies and databases that are pivotal in the exploration of these complex macromolecules. This perspective delves into integrating AI in peptide development, encompassing classifier methods, predictive systems and the avant-garde design facilitated by deep-generative models like generative adversarial networks and variational autoencoders. There are still challenges, such as the need for processing optimization and careful validation of predictive models. This work outlines traditional strategies for machine learning model construction and training techniques and proposes a comprehensive AI-assisted peptide design and validation pipeline. The evolving landscape of peptide design using AI is emphasized, showcasing the practicality of these methods in expediting the development and discovery of novel peptides within the context of peptide-based drug discovery.

Keywords: protein engineering; therapeutic peptides; deep learning; large language models; deep generative models; peptide-base drug discovery; artificial intelligence

Montserrat Goles is a Master's student in Departamento de Ingeniería Química, Biotecnología y Materiales, Universidad de Chile. Her research interests include bioinformatics, machine learning strategies, and computational biology.
Anamaría Daza is an young researcher at Centre for Biotechnology and Bioengineering, CeBiB, Universidad de Chile. Her research interests include cell culture, cell and gene therapy, heterologous protein expression, characterisation and purification, and extremophile enzymes for industrial applications.
Gabriel Cabas-Mora is an Assistant researcher at Departamento de Ingeniería en Computación, Universidad de Magallanes. His main research interest is machine learning applied to protein engineering and molecular biology.
Lindybeth Sarmiento-Varón is a Post-doctoral researcher at Centro Asistencial de Docencia e Investigación (CADI), Universidad de Magallanes. Her research interests include the immune system's interactions post-infection and the development of biomarkers.
Julieta Sepúlveda-Yañez is a researcher at Centro Asistencial de Docencia e Investigación, CADI, Universidad de Magallanes. She has experience in molecular biology and bioinformatics and specialises in studying mutagenic mechanisms in B cells.
Hoda Anvari-Kazemabad holds a master's degree in Software System Engineering from RWTH Aachen University. Her research interests encompass explainable AI (XAI) in protein engineering and the autonomous design of therapeutic peptides.
Mehdi D. Davari leads the Computational Chemistry Group at the Leibniz Institute of Plant Biochemistry in Halle, Germany. His research advances protein and enzyme design by unravelling the complex relationships between protein sequences, structures, and dynamics. It functions through computational techniques, including multiscale simulations and artificial intelligence.
Roberto Uribe-Paredes is an Associated Professor at Departamento de Ingeniería en Computación, Universidad de Magallanes. His research interests include Information Retrieval, High-Performance Computing, bioinformatics, and genomic analysis.
Álvaro Olivera-Nappa is an Full researcher at Centre for Biotechnology and Bioengineering, CeBiB, Universidad de Chile. His research interests include protein engineering and computational biology.
Marcelo A. Navarrete is a researcher and group leader at Centro Asistencial de Docencia e Investigación, CADI, Universidad de Magallanes. His research interests include computational immunology and oncogenomics in lymphoid malignancies.
David Medina-Ortiz is an Assistant Professor at Departamento de Ingeniería en Computación, Universidad de Magallanes. His research interests include machine learning algorithms, deep learning architectures, and de novo protein design.
Received: February 8, 2024. **Revised:** April 23, 2024. **Accepted:** June 4, 2024
© The Author(s) 2024. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.
For commercial re-use, please contact journals.permissions@oup.com

Introduction

Engineered peptides exhibit wide-ranging therapeutic effects, such as antimicrobial, antitumour, antithrombotic and immunomodulatory activity [1]. Peptide drugs offer potential advantages over traditional small-molecule drugs, including increased selectivity, affinity, efficacy, safety and reduced toxicity and immunogenicity [2]. However, the brief half-life, limited oral bioavailability and susceptibility to plasma degradation of peptide drugs remain a challenge for their widespread clinical application [3].

The growing availability of numerous peptide sequences and their functions have driven the development of computational tools to store and share peptide information[4]. General and activity-specific databases have been established and are publicly accessible [5, 6]. Besides, artificial intelligence (AI) methods and predictive systems through machine learning (ML) have substantially evolved in recent years, particularly in protein engineering applications [7].

The rapid evolution of AI has marked the beginning of a new era of sophisticated ML algorithms in biomedicine, significantly impacting drug discovery [8]. Integration of ML has taken on a multifaceted dimension, including classifier methods, predictive systems and generative approaches [9].

Classifier methods are crucial for identifying diverse peptide activities, from biological functions to categorical properties [5]. In contrast, predictive models prove their utility by estimating numerical properties, such as binding affinities in protein–peptide interactions, evaluating toxicity and predicting inhibitory activity [10].

Deep generative models (DGMs) have emerged as tools for the design of therapeutic peptides [11]. In particular, generative adversarial networks (GANs), variational autoencoder (VAE) and diffusion models facilitate the generation of novel peptide sequences aimed at specific objectives [9, 12]. However, automated or assisted peptide design still faces challenges, such as the optimization of peptide processing, the establishment of informative representation strategies and the meticulous development and validation of predictive models.

This work describes peptides and their different properties and usabilities for biotechnology and therapeutic applications. It also presents traditional strategies for training ML models. Subsequently, it discusses the potential applications of DGMs as a foundational tool for designing peptides with therapeutic potential. Besides, a comparative analysis of the most common DGMs strategies is addressed. This work also proposes a comprehensive AI-assisted peptide design and validation pipeline. Finally, different challenges are discussed concerning achieving an autonomous therapeutic peptide design supported by AI approaches.

Peptides and their biotechnology applications

Peptides exhibit unique biochemical and therapeutic attributes and can be synthesized or obtained from natural sources. This section aims to offer a perspective on peptides, exploring their multifaceted biological functions, moonlighting attributes and applications [13, 14].

Peptide overview, synthesis process, function biological activities and moonlighting effects

Peptides consist of short chains of amino acids, with a molecular weight ranging from 0.5 to 10 kDa [15], and a length ranging from a couple of amino acids up to 100 [16]. Their role in biological processes is diverse, including functions as structural components, enzymatic inhibitors, hormones, host defence molecules and neurotransmitters [11]. Moreover, peptides can also act as cell surface receptors [17] and play an essential role in drug delivery applications [17, 18].

These macromolecules can acquire secondary structures, commonly forming α-helix and β sheets structural patterns [19], but also more complex structures [20]. Lasso peptides, for example, have a unique 3D structure where the C-terminus threads through an N-terminal macrolactam ring in a right-handed conformation, which provides stability against chemical, thermal and proteolytic degradation [21].

Peptides can be isolated from natural sources [1], including venoms [22], food products [23] and marine organisms [24]. Alternatively, peptides can be produced using recombinant techniques and chemical synthesis [25, 26].

In the recombinant approach, organisms like *E. coli*, *S. cerevisiae* and *P. pastoris* are used to produce heterologous peptides using the host cell protein production machinery [14]. In contrast, chemical procedures allow synthesis automation for large-scale industrial production [27]. Common chemical techniques are solid-phase peptide synthesis [11] and liquid-phase peptide synthesis [28].

Peptides can exhibit different biological activities with multifaceted properties. Figure 1A summarizes the main biological activities reported. As therapeutic properties, peptides play a role as antimicrobial, antitoxin and anticancer [2]. In the cosmetic industry, peptides can be used for wound healing and anti-aging activities. Peptides can also facilitate the molecular binding of proteins, DNA and RALF molecules. Peptides with immunological, neurological, signalling, drug delivery, propeptide, taste and cell–cell communication properties have been reported [5, 29–31]. Finally, peptides can also exhibit toxic effects, with reports of cytotoxic, allergen, endotoxin and neurotoxin characteristics [32–34].

Evolution and adaptability have moulded peptides into exhibiting two or more concurrent biological activities [35]. This unique and promiscuous attribute is called moonlighting activity. Unlike conventional peptides or proteins with defined functions, moonlighting peptides can acquire different activities in various cellular contexts or environments [35, 36].

Figure 1B describes the moonlighting for the primary biological activities reported in the Peptipedia database [5]. Therapeutic peptides have been demonstrated to present toxic properties. Cell–cell communication peptides also play therapeutic and signalling roles. The moonlighting effect was also analyzed for specific therapeutic peptides, including antimicrobial, metabolic, and anti-cancer. (See section S1 of Supplementary Material for more details). Moreover, specific antimicrobial peptides (AMPs) with antifungal, antibacterial and antiviral biological activities exhibit moonlighting properties. The moonlighting effect is also a property for antibacterial peptides, existing peptides with anti-gram (+) and anti-gram (-) activity (See section S1 of Supplementary Material for more details).

Leveraging moonlighting peptides as therapeutics presents potential advantages over traditional small-molecule drugs. The ability to combine multiple therapeutic functions into a single drug could alleviate the treatment burden for patients. Furthermore, such peptides could simultaneously target various disease-related pathways [35]. However, their multifunctional nature may also give rise to specificity issues, and their design complexity is compounded by the necessity to understand these unique structure–function relationships.

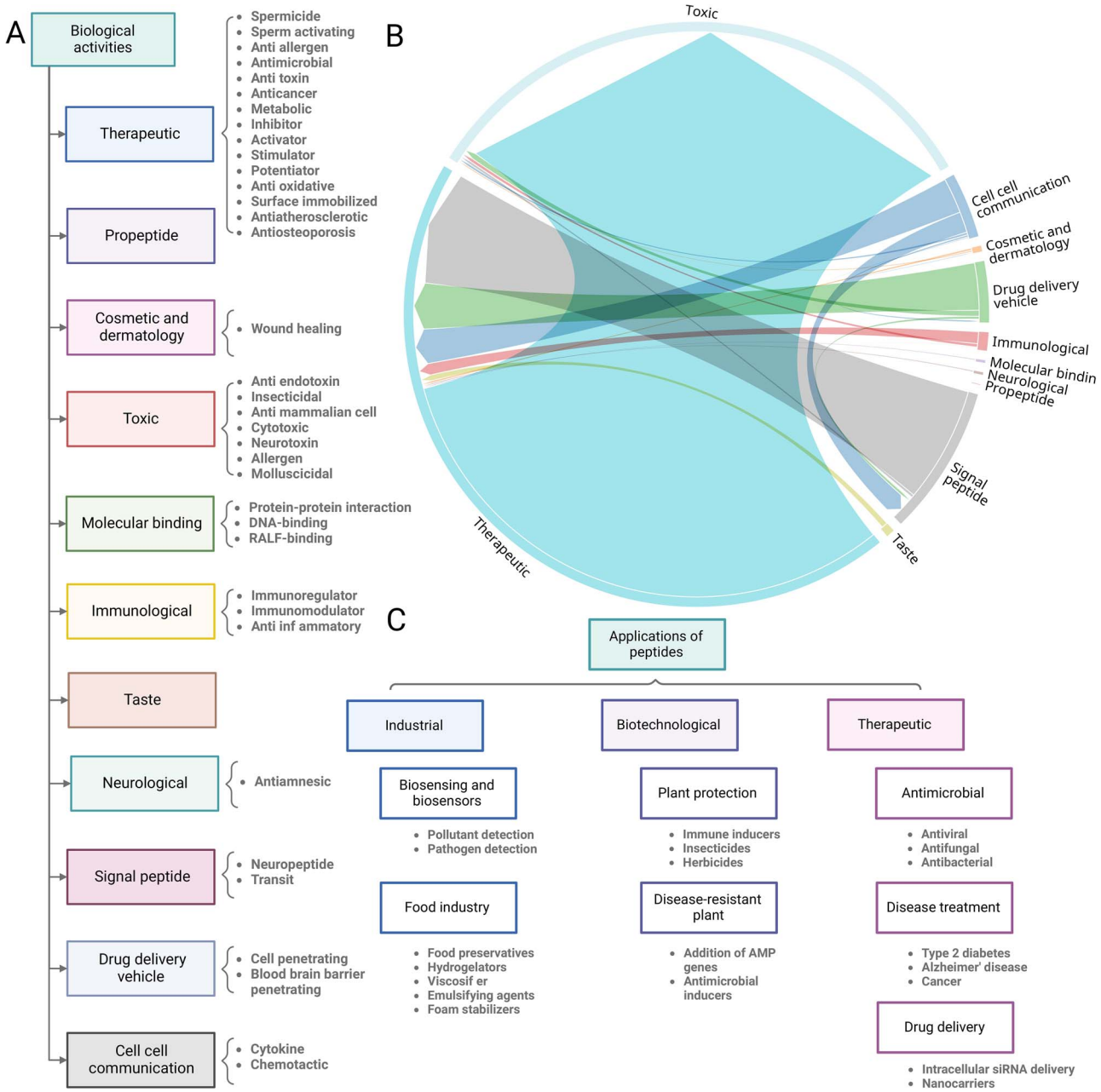

Figure 1. **Biological activities of peptides, moonlight activity and main applications**. **A.** Functional biological activities were grouped into 10 categories, including therapeutic, neurological drug delivery vehicle, sensorial, immunological, molecular binding, propeptide, signal peptide, transit and other activities (colour boxes) with different categories and subcategories. A total of 96 activities considering properties, reported activity and experimental validations have been generated in our previous work [5]. **B.** Moonlight evaluation of the 10 main activities reported in [5]. Therapeutic peptides are highly related to propeptides, drug delivery vehicles and sensorial peptides. Neurological peptides are also highly related to drug delivery vehicles and molecular binding. In contrast, sensorial peptides have a low relation with immunological peptides. **C.** Summary of peptide applications, considering industrial, biotechnology and therapeutic applications.

Biotechnology and therapeutic applications of peptides

Peptides exhibit many natural functions and offer significant potential for diverse applications. Peptide synthesis with non-canonical or artificial residues further expands their utility (e.g. by modifying residues in cationic AMPs to enhance proteolysis resistance) [37]. Figure 1C provides an overview of the primary applications of peptides in industry, biotechnology and therapeutics.

In the industrial sector, peptides have been applied as biosensors and used explicitly for pollutant and pathogen detection [38]. Peptides have also been employed in the food industry as food preservatives, hydrogelators, foam stabilizers and emulsifying agents [39]. For biotechnology applications, peptides have been used for plant protection and as immune inducers, playing pivotal roles as insecticidal and herbicides [40–42].

In therapeutic applications, peptides have generated antimicrobial drugs, including antifungal, antiviral and antibacterial [43]. Examples of therapeutic peptides are (i) WK2, designed to combat multi-drug resistant *Salmonella* [44], (ii) Teicoplanin a semisynthetic peptide for treating severe infections [45] and (iii) insulin and semaglutide for diabetes treatment [46].

Another relevant type of peptide for therapeutic applications are the cell-penetrating peptides (CPPs). These peptides can penetrate cell membranes, reach intracellular targets and facilitate drug delivery [47]. For example, the cyclic and amphipathic peptide $[WR]_5$ has been used for the intracellular delivery of small interfering RNA and the enhancement of curcumin uptake [48].

Other peptides have the potential to improve neurodegenerative disorders like Alzheimer's and Parkinson's disease [49]. The P110 peptide inhibits Dynamin-related protein 1 (Drp1), a crucial regulator of mitochondrial fission, offering stable preservation of dopaminergic neurons in Parkinson's disease [50]. Under stress, activated Drp1 translocates to mitochondria, leading to excessive mitochondrial fission and dopaminergic neuronal death.

ML strategies to develop predictive models for peptides

In recent years, there has been a substantial increase in the number of peptides documented in the literature [5]. This sequence abundance has, in turn, stimulated interest in applying computational biology techniques to analyse peptide sequences, predict biological activities, calculate physicochemical properties and assist in peptide design [51, 52]. However, applying AI and ML methodologies presents various challenges, from data collection and processing to predictive model validation to selecting an appropriate training strategy. This section reviews the classical approaches of data-driven methods for building predictive models and discusses the ML-based models for peptide tasks.

ML definitions and main characteristics

ML is a sub-field of AI focused on algorithms and statistical models based on the idea that computers can analyse and learn from data patterns and use that knowledge to make predictions, classify objects or solve various problems [53].

Four learning types are currently applied to solve computational biology problems: (i) supervised learning, (ii) non-supervised learning, (iii) reinforcement learning and (iv) generative learning [54] (See Fig. 2A and section S2 of Supplementary Materials for more details). In addition to the different learning types, two main focuses have been implemented for developing models using ML strategies, classic ML approaches and deep learning (DL) architectures [55].

DL approaches have been efficient for image recognition, natural language processing and speech recognition [56]. In biotechnology and bioinformatics, DL has allowed the development of predictive models to estimate protein–protein interactions, classification systems for unknown enzymes through recognition of enzyme commission numbers [57] and predictive systems for protein structures such as RoseTTAFold [58] or AlphaFold [59].

The applications of transfer learning approaches by applying foundational trained models to biology context have demonstrated usability to develop efficient predictive models [60]. Moreover, the combination of transfer learning with semi-supervised strategies has facilitated the implementation of predictive models to address the data scarcity in protein engineering and fitness landscapes [61].

A classic data-driven pipeline to develop predictive models through ML techniques for protein engineering tasks

There are four main steps for developing predictive models for peptides using ML strategies and protein sequences or structures: collecting and processing datasets, numerical representation strategies, training and validating predictive models and evaluating performances using classic metrics. Figure 2B summarizes the standard strategies and schematizes a classic data-driven pipeline.

During the collecting process, different sources could be employed to obtain protein sequences or structures depending on the objective of the predictive model. Usually, generic databases like Peptipedia [5], SATPdb [62] or LAMP2 [63] are used for collecting AMPs. In contrast, specialized databases like quorum sensing signalling peptides [64], anti-angiogenic peptides predictors [65] and bacteriocin peptides [66] are employed for collecting specific information.

Before training predictive models, numerical representation strategies must be applied to process peptide sequences. Current approaches favour the use of learning representations, due to their ability to catch functional and structural information from the context sequence [67, 68]. However, baseline methods like One Hot encoder, feature engineering and physicochemical encoders could also be applied (See section S2 of Supplementary Material for more details).

Finally, for training models, a classic pipeline could be applied [69], including a standard division of the dataset in training and testing partitions, a $k-$fold cross-validation technique to prevent overfitting and the application of classic metrics to evaluate and compare the performance of the predictive models (See section S2 of Supplementary Materials for more details).

Current methods, strategies and implemented models for peptide design

Predictive models have been developed to explore and analyse peptide sequences, with a shared focus on tasks such as AMP classification, including ensemble learning process [70] and DL architectures [71] to train the classification models. Besides, antiviral classification systems applying Random Forest, support vector machine and DL methods have been successfully implemented [72–75].

Classification models for therapeutic applications have been implemented, including anti-inflammatory detection peptides [76], anticancer peptides identification [77] and CPPs recognition [78].

Pharmacological properties for therapeutic peptides have also been the focus of developing predictive models. The prediction of half maximal inhibitory concentration (IC_{50}) [79], estimating free radical scavenging activity and chelating properties [80] and half-life in blood [81] have been addressed by applying ML approaches.

These models not only facilitate the exploration of unknown sequences but may also contribute to the design of peptides with desirable properties. Typically, they integrate DGM strategies to advance sequence generation learning supported by computational property classification methods. Nevertheless, despite this progress, a unified pipeline for training predictive models in functional classification tasks remains needed.

Towards an autonomous peptide-based drug discovery

In this section, we will discuss the integration of ML and AI techniques in developing peptide-based drug discovery systems. First, we will show ML methods to support drug discovery; then, different deep generative strategies will be explained and discussed, evaluating different advantages and disadvantages. Next,

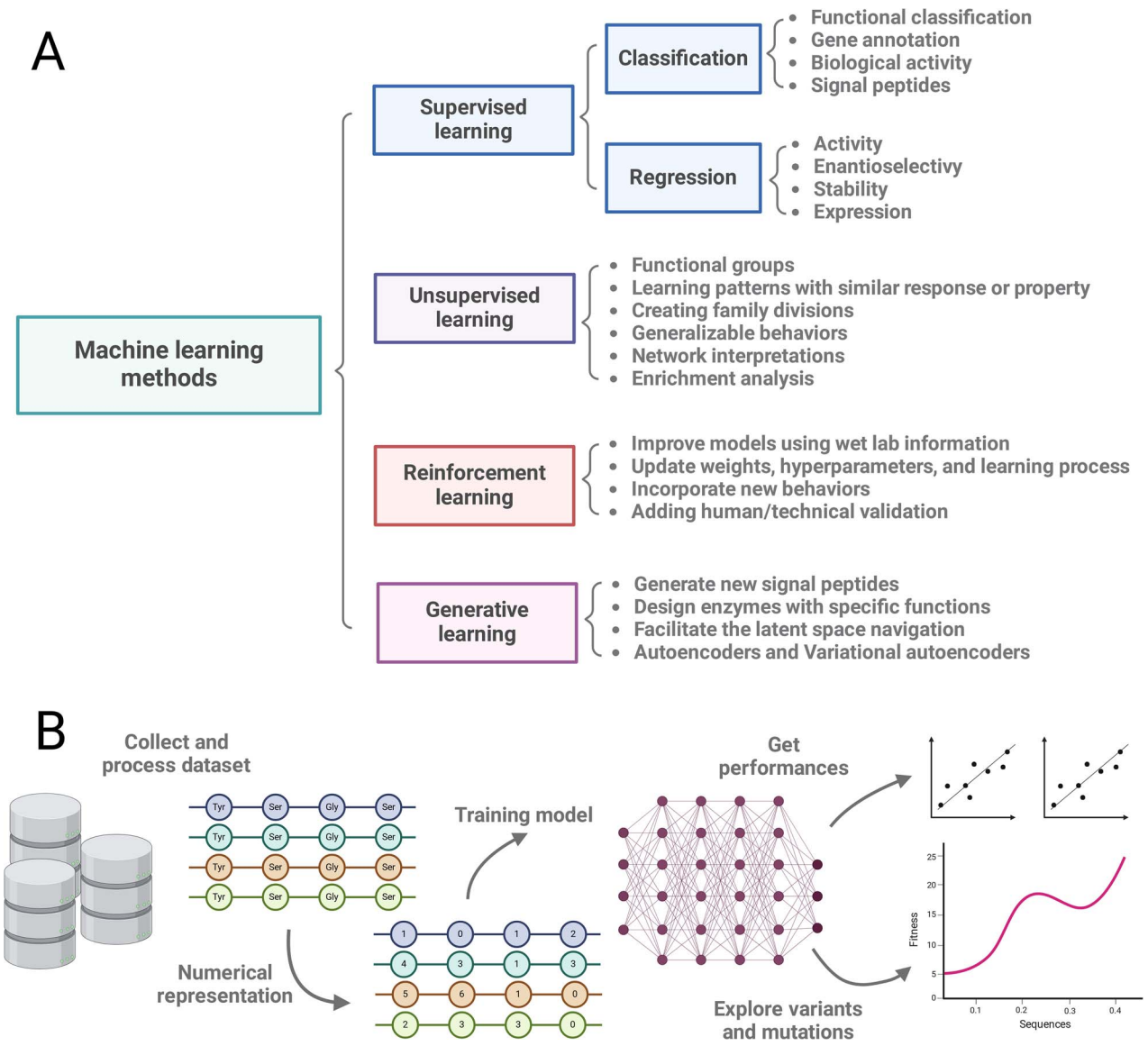

Figure 2. **The main learning type process, the different applications to solve biotechnology and a data-driven pipeline to build predictive models supported by ML strategies**. **A** The four main learning strategies used in biotechnology: (i) Supervised learning is used to build predictive models. (ii) Pattern recognition or unsupervised learning is applied to identify behaviours in an unlabelled dataset. (iii) Reinforcement learning is applied to update predictive models supported by new experimental evidence or knowledge. (iv) Generative learning is applied to create or design new examples and is employed as a landscape navigation strategy. **B** A classic data-driven approach to generate and use predictive models employing protein sequences as input to train the predictive model. First, a dataset is built by collecting information from databases, repositories, public resources or experimental reports. Once the dataset is built, numerical representation strategies are applied to code the protein sequences, generating an interpretable dataset for ML algorithms. The coded dataset is the input used to train a predictive model by applying supervised learning algorithms. Then, the predictive model's performance is evaluated, and the model is validated through new examples or simulated datasets. Finally, the validated model explores the latent space and rebuilds the mutational landscape.

traditional protein engineering strategies generated to assist the peptide design will be analysed and compared with generative approaches. Finally, a proposal pipeline to discover therapeutic peptides will be explored.

Drug discovery supported by ML methods

Drug discovery is the process by which new candidate drugs are developed [82]. Traditional drug discovery involves (i) target identification, (ii) target validation, (iii) lead compound identification and (iv) lead optimization [83]. This process can often take decades and has a cost that can exceed a billion US dollars per target [84], considering that most drugs do not reach the market [85].

In contrast, drug repurposing is the exploration of new therapeutic applications for existing drugs [86]. Drug repurposing has different advantages, including safety, efficacy and accelerating the clinical assessment [87].

The combination of ML approaches with biomedical datasets can facilitate the identification of novel therapeutic targets, studying molecular characteristics, protein interactions, biological activities and adverse effects [4].

Different predictive models have been implemented to assist drug discovery, including methods like BANDIT, a Bayesian ML approach [88], and DeepDTnet, a DL approach that combines phenotypic, genetic, chemical and cellular network profiles to guide the drug discovery [89].

Alternatively, drug–target interaction predictive models have been developed to evaluate drug candidates. These methods have been based on Random Forest algorithms combined with a graph-based features extraction process [90], deep belief network [91] and long short-term memory (LSTM) architectures combined with protein characteristics and drug molecular substructure fingerprints [92].

ML frameworks can also predict drugs' adverse effects at diverse stages of the discovery process. Two relevant approaches have been implemented: a deep neural network to assess the probability of adverse drug reactions in novel pharmaceutical compounds [93], and a geometric self-expressive model to help identify side effects during drug clinical trials [94].

In contrast to the current ML methods implemented to assist drug discovery, generative learning strategies have been developed to design *de novo* molecules or drugs with desirable targets [95]. Methods based on combinations of graph-generative models with Monte Carlo tree search [96] and transformer-decoder architectures to produce *de novo* smiles [97] have demonstrated the usability of the ML approaches to the discovery of new drugs.

While generators generative learning implemented for drug discovery exemplify the potential of AI models for small molecule design [96, 97], their capability can further extend into the domain of peptide design [98].

DGM for the generation of new peptide sequences

DGMs aim to capture the underlying data distribution of a given dataset to generate novel instances that accurately represent the properties of the original data [99, 100]. For peptide design, DGMs can not only generate *de novo* peptide sequences but also perform learning representation tasks and likelihood learning tasks, where the model can learn to assign a greater probability to protein sequences that acquire desired characteristics [12].

There are multiple steps in the pipeline for peptide design using DGMs. First, the model needs to be trained. To this aim, peptide databases, datasets and repositories are collected and processed. Next, the collected peptide sequences require a numerical representation to provide the models with an interpretable input. Peptide representation can rely on feature descriptors, amino acid coding or embedding representations [98]. Then, the model is trained on the represented data; modelling the distribution of the training set allows us to learn the underlying data distribution and generate novel sequences.

In this section, the most common generative strategies are described, then state-of-the-art-related protein design, generation or discovery is analysed. Besides, a comparative analysis of the different generative approaches is addressed, including the advantages and disadvantages of each one.

Variational autoencoders

A VAE is a generative model combining autoencoders and variational inference elements. VAEs are used to learn representations of unsupervised input data and generate new data samples that resemble the training data [12].

Overall, VAEs are powerful generative models capable of learning rich representations of complex data distributions and generating new samples with desirable properties. They have applications in various domains, including image generation, text generation and anomaly detection.

VAE introduces a probabilistic aspect, learning a distribution in the latent space (see Fig. 3A). The most relevant components of a VAE model are (i) an encoder network, (ii) a reparametrization trick, (iii) a latent space, (iv) a decoder network and (v) a loss function [99].

During training, the VAE optimizes the parameters of both the encoder and decoder networks to minimize the combined loss function. This process involves passing input data through the encoder network to obtain latent space representations, sampling from the latent space and then reconstructing the data using the decoder network. The model's parameters are updated using gradient descent techniques.

Once trained, the VAE can generate new data samples by sampling from the learned latent space distribution and passing the samples through the decoder network. By varying the samples in the latent space, the model can generate diverse and realistic data samples that resemble the training data.

Generative adversarial networks

GANs are generative models based on energy-based models (EBMs). Two components are most relevant for a GAN: a generator network and a discriminator network [101] (see Fig. 3B).

The generator is a neural network that takes random noise (often drawn from a simple distribution such as Gaussian) as input and produces data samples as output. Initially, the generator produces random noise, but as training progresses, it learns to generate increasingly realistic samples that resemble the training data [98]. In contrast, the discriminator is another neural network that takes input data samples and predicts whether they are real or fake [98].

The training process of a GAN model is called adversarial training. During the training process, the generator and discriminator are trained simultaneously. The generator aims to produce indistinguishable samples from real data, while the discriminator aims to classify real and generated samples correctly [99]. During training, the generator and discriminator are updated iteratively. The generator creates fake samples and attempts to fool the discriminator, while the discriminator is updated to better distinguish between real and fake samples. This process continues until the generator produces samples that the discriminator is not able to differentiate from real data [101, 102].

Neural language models

Neural language models (NLMs) use neural networks to comprehend and generate human language. Designed to capture intricate patterns and relationships within a given language, these models are proficient in diverse tasks such as language understanding, sentiment analysis, machine translation and text generation (refer to Fig. 3C) [99]. Key components of NLMs are (i) Neural Network Architecture, (ii) Word Embeddings, (iii) Pre-training and Transfer Learning, (iv) Contextual Understanding, (v) Transformer Models and (vi) Generative Language Models.

NLMs predominantly adopt two main frameworks: recurrent neural networks (RNNs) and attention models. In RNNs, LSTM and gated recurrent units are used to construct autoregressive models. These models effectively capture sequential data by retaining historical information in their hidden states [98].

Flows-based and EBMs

Flow-based models are generative models that learn the mapping between data samples and their corresponding probability densities. These methods utilize invertible neural networks to establish a bidirectional mapping between inputs and a latent representation. Unlike some generative models that rely on approximate

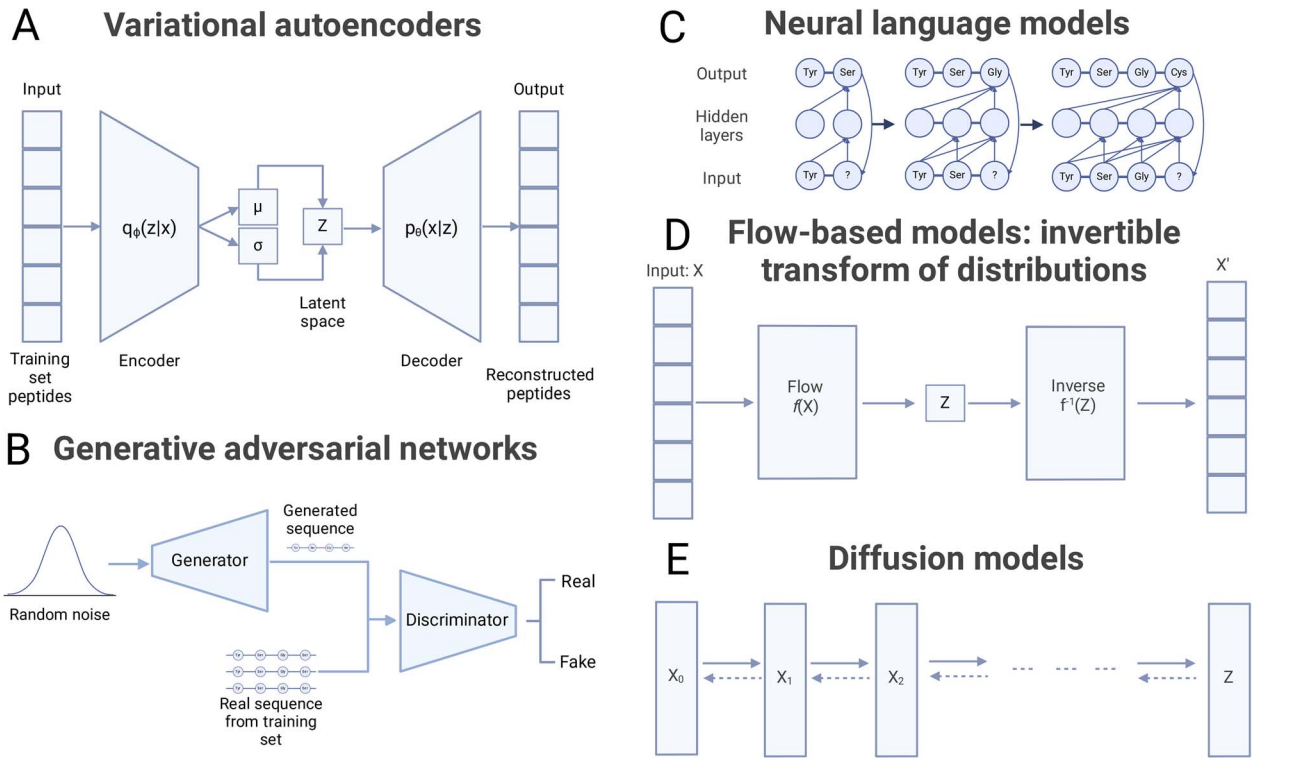

Figure 3. **Classic DL architectures through DGMs employed for peptide design**. **A** An autoencoder is a DL architecture that facilitates encoding the input data into a lower dimensional representation, reconstructing the original input as closely as possible. The autoencoders have two main components: an encoder and a decoder. **B** VAEs are a specific type of autoencoder that introduces probabilistic elements into the model, enabling it to learn a probabilistic distribution of the data in the latent space. VAEs are particularly effective for generating new, similar data points and are often used in generative modelling tasks. **C** NLMs, also known as autoregressive models, are a class of AI models that use neural networks to understand and generate human language. These models are designed to capture the complex patterns and relationships within a given language, enabling them to perform tasks such as language understanding, sentiment analysis, machine translation and text generation. **D** GANs are a class of AI algorithms that generate new, realistic data samples that resemble a given dataset. The main components in a GAN are (i) generator, (ii) discriminator, (iii) adversarial training, (iv) loss function and (v) training process.

sampling, flow-based models often use exact inference and analytical computations [99] (see Fig. 3D).

Different components are necessary to develop a flow-based model, including (i) the base distribution, (ii) the transformation network, (iii) invertible transformation functions and (iv) an output distribution.

EBMs are a diverse category of ML models characterized by their unique approach to modelling data. EBMs are designed to learn an energy function rather than directly learning a probability density function on the input space [103]. According to this energy function, observed or plausible data states are given low values (referred to as 'energies'), whereas unobserved or improbable states are given high values [99]. EBMs differ from conventional generative methods because they do not require normalizing probability distributions.

EBMs are more flexible for handling probability distributions, due to integrating the energy function into the learning model [103].

Diffusion models

Diffusion models are generative models that iteratively apply reversible transformations to a noise process. These models efficiently generate high-quality, diverse samples in various domains, such as images, text and audio [104]. Figure 3E presents a schematic representation of the diffusion model approaches.

The type of model starts with a noise process as the initial input. Then, reversible transformations to the noise process are applied iteratively. Each transformation is designed to gradually change the noise process into a data sample while preserving reversibility [105].

Diffusion models typically incorporate an annealing schedule to control the diffusion rate and ensure stable training. The annealing schedule gradually adjusts the magnitude of the reversible transformations over time, allowing for smoother transitions and better convergence [106].

Diffusion models are trained using maximum likelihood estimation. The training objective is to minimize the negative log-likelihood of the data samples under the diffusion model. This objective encourages the model to learn reversible transformations that can accurately transform the noise process into realistic data samples [105].

Once trained, diffusion models can generate new data samples by running the diffusion process forward from random noise. By varying the random noise inputs, diffusion models can generate diverse, high-quality samples that resemble the training data distribution [104].

Generative approaches applied for peptide design and discovery

Different VAE strategies have been implemented to assist the peptide design or discovery. Methods like PepVAE [107], PepCVAE [108] and GM-Pep [109] have been proposed for AMP design supported by VAE approaches. Two relevant methods for AMP design are proposed in [110] and [111].

[110] used a deep autoencoder to design AMPs, achieving success in 48 days with a 10% hit rate. HydrAMP, a conditional

variational autoencoder (cVAE), uncovers a continuous representation of peptides, leveraging controlled creativity [111]. Using HydrAMP, AMPs, including a superior analogue (Hydraganan-1) of Pexiganan, were generated and experimentally tested against *E. coli* strains [111].

Other examples of peptide design using VAE strategies are target-specific peptide inhibitors [112], peptide inhibitors targeting β-catenin [113], peptides with desired bioactivity and membrane permeability properties [114] and peptide–MHC binding [115].

GANs are among the most widely used generative models for the design of peptides [103]. GANDALF (Generative Adversarial Network Drug-tArget Ligand Fructifier) is a deep convolutional GAN-based method that incorporates information about the active atoms of the protein–ligand interaction domain [116]. GANDALF generates novel peptides and predicts their binding affinity to a specific target. Rather than using protein–protein interaction databases, this model is trained on drug–ligand interactions of protein and FDA-approved peptide drugs based on the THP database [117]. It also uses data from PDB for 3D structures, Uniprot for the names of interacting proteins or peptides and CASTp to calculate protein pockets [118, 119]. This model has generated peptides to target cancer-related proteins, including PD-1, PDL-1 and CTLA-4 [116, 120].

Another GAN-based method for designing peptides, specifically bioactive antiviral peptides, is PandoraGAN [121]. It uses LeakGan, a modified GAN used in text generation, to generate peptides that undergo initial validation based on their amino acid composition, net charge, instability index, repeats and patterns.

Methods like AMPGAN v2 [122], AI4AVP [75] and HelixGAN [123] have also been implemented using GAN strategies to assist therapeutic peptide design.

Applying NLM strategies [124] have implemented AMPTrans-LSTM, a deep generative network-based approach designed to rationalize AMPs. Comprising two interconnected submodels: a long- and short-term memory sampler and a transformer-converter, the model demonstrates a success rate ranging between 30% and 50%. It generates new peptide sequences while preserving essential AMP features.

Flows-based models have been implemented focusing more on drug discovery. An example of these models is TagMol [125]. TagMol is a probabilistic end-to-end EBM for target-specific drug design. This approach uses the EBM to evaluate the binding affinity scores between protein–ligand pairs precisely. TagMol demonstrated that it could generate molecules with binding affinity scores similar to real molecules.

Different approaches have been implemented to assist peptide design, focused on diffusion model strategies [98]. Examples of the implemented strategies are AMP-diffusion. This method integrates latent diffusion with protein language models to generate AMPs [126]. Another relevant method for peptide therapeutic peptide design is the combination of structure prediction networks with diffusion generative approaches [127]. Alternatively, methods like Geometric Latent Diffusion [128] and multi-modal contrastive diffusion models [129] have been designed to support the therapeutic peptide design.

A simple comparative analysis of the different generative strategies employed for peptide design

The different strategies implemented to design therapeutic peptides have proven successful in their application. However, there are differences in their operation that require attention to identify their advantages and disadvantages.

When formulating strategies based on VAEs, leveraging a latent space informed by distributions facilitates the generation of peptide sequences exhibiting an amino acid distribution similar to those in the training data. A more extensive repertoire of sequences enhances the likelihood of producing successful peptides as it expands the scope of potential explorations. However, a higher volume of feasible peptides does not inherently guarantee acquiring the desired biological activity. Given peptides' propensity for moonlighting properties, there exists the possibility of uncovering novel biological activities that may not necessarily be advantageous. Consequently, when employing VAE-based design methodologies, incorporating a validator for the peptide sequences generated by the models becomes imperative.

Alternatively, design methodologies based on GANs require training a discriminator component to distinguish between real and fake sequences. While GAN-based approaches often offer advantages over VAE methods, like a validation process for generated sequences, their implementation requires careful consideration of various factors. First, achieving a balanced dataset is crucial; any class imbalance could impede the model's ability to generalize effectively. Secondly, ensuring accurate differentiation between positive and negative elements, particularly in peptide sequences, demands adopting appropriate representation strategies. Therefore, integrating generative methods with Transformers to leverage pre-trained representation learning models proves beneficial for incorporating decoding techniques and facilitating the discrimination of genuine from spurious sequences. Introducing additional components complicates generative strategies' training, validation and use.

Methods based on NLMs typically incorporate DL architectures such as RNNs or LSTMs. Training these models relies on natural language processing techniques. Achieving robust generalization with these models is challenging, often necessitating a substantial volume of training examples. Their appeal lies in their ability to analyse context and semantics for model development. Nevertheless, akin to VAE-based approaches, evaluating their efficacy requires integrating a classification model, thus complicating the training process. However, recent studies have reported precision rates exceeding 80% in developing AMPs, underscoring the efficacy and utility of these methods.

Recently, advances in peptide design methodologies have showcased remarkable efficacy, particularly in therapeutic peptide development. Among these methods, diffusion-based strategies have emerged as standout performers compared with approaches like VAE or GAN. Leveraging diffusion models enables efficient and high-fidelity reconstruction of new examples. Additionally, unlike flow-based or NLMs, there appears to be no significant correlation between the complexity of the training architectures and model performance. Notably, recent efforts have integrated generative learning with contrastive learning strategies to enhance performance in trained models.

Finally, while various strategies exist to aid peptide design and significant advances have been made in generative AI, several challenges warrant careful consideration before peptides proceed to experimental testing. These challenges encompass the analysis of toxicological and immunogenic properties, which are critical for ensuring safety and non-desirable effects. Additionally, peptides like bacteriocins often feature post-translational modifications (PTMs), a factor not accounted for in standard peptide generation processes. Assessment of pharmacological and pharmacodynamic attributes, such as half-life and quality measured via IC50, is imperative. Moreover, evaluating potential undesirable activities stemming from peptides' moonlighting effects is crucial.

Lastly, given that most therapeutic peptides act as receptor targets, computational affinity assessment against these receptors could enable the development of efficient filters for experimental peptide testing.

Therapeutic peptide discovery through ML and DGMs

Various computational strategies have been used to facilitate the discovery of peptides with therapeutic potential.

These approaches focused on exploring classification systems and reconstructing mutational landscapes to inform rational protein design [130]. Additionally, the advances in structure predictions supported by AlphaFold have generated a new era for peptide-based drug discovery, facilitating the generation or discovery of new therapeutic peptides [131]. Alternatively, methods rooted in DGMs have demonstrated effectiveness in generating or identifying peptides exhibiting antimicrobial and signal peptide activities [110].

However, navigation through a latent space, investigation of diverse conformations and assessment of different functional components tied to the capacity and characteristics of peptides designed using AI remains challenging.

In particular, unique validation techniques and property calculations associated with the activity and quality of the designed peptide are needed.

This section presents a potential pipeline for discovering peptides with therapeutic activity. This pipeline encompasses predictive modelling of peptide properties, tools for predicting interaction affinity and bioinformatics methods that offer *in silico* validation of generated sequences. Experimental validation can be undertaken based on the specific application.

The proposed pipeline is illustrated in Fig. 4, outlining three key components essential for advancing the discovery of peptides with desirable therapeutic attributes.

The initial component (refer to Fig. 4A) is dedicated to designing and implementing models to classify biological activity. These models specialize in binary or multiclass classification of the biological activities of unknown peptides or those generated by generative systems.

For instance, they categorize peptides as antiviral, antihypertensive, antibacterial or peptides with antiviral activity that recognize viruses within the *Retroviridae* family and the HIV species. Beyond models for classifying biological activity, evaluation models may be needed to assess functional mechanisms. For example, evaluating the potential to inhibit the integrase enzyme or assessing peptides' ability to hinder fusion between the GP41 protein and the host in the case of anti-HIV peptides.

Another consideration is a peptide's capacity for multiple activities, such as forming interactions with proteins, membrane permeability and moonlighting capabilities. Introducing new rules or conditions to the peptides explored via classification systems facilitates the integration of regulatory elements into the design process, constrains navigation within the latent space and fosters guided learning for design systems.

The construction of these models follows the strategies outlined earlier in this work, including using numerical representation methods to validate implemented classification models.

The second component focuses on designing and implementing an affinity prediction system between protein and peptide, which only applies to peptides targeting a protein (See Fig. 4B.

The binding affinity predictive component proposes strategies involving matrix or object representations to characterize the interaction complex and train predictive models through DL architectures, focusing on CNN or GCN methods. By applying such architectures, explainable AI methods can be integrated to support predictions from an interpretative and explanatory point of view.

The integration of XAi approaches will allow an understanding of the relevant zones or patterns that dictate the model's response, facilitating the design guidance of peptides exhibiting high affinity for the identified protein zones and key residues that encourage a favourable interaction with the target protein.

Finally, the third component is dedicated to the generative method for exploring novel sequences (See Fig. 4C). Techniques such as VAEs or GANs have been previously utilized to navigate the latent space of peptides awaiting exploration.

Initially, only peptides with a specific biological activity of interest are considered input. Nevertheless, there is also potential to incorporate unknown peptides displaying a specific biological activity, such as adding peptides to the dataset with antiviral properties, even in cases where there is no experimentally validated information, especially concerning anti-HIV activity.

Functional classification models are crucial for assimilating these new sequences in such scenarios. The development of this analytical approach encourages dataset expansion, thereby broadening the latent space to uncover novel peptides.

Once the generative models are trained and validated, the next step is to explore and navigate the latent space. To achieve this, optimization methods that rely on heuristic algorithms can be integrated to explore new sequences.

These identified sequences must be validated using biological activity classification methods and, if necessary, methods predicting affinity to target proteins of interest. In this context, validation serves as a set of rules or checkpoints that must be met to propose a peptide as a potential candidate for experimental validation. Additional validations are also required, such as the analysis of physicochemical properties and stability and the estimation of toxicity.

Reinforcement learning-based methods can be incorporated into the generation and validation stages to update the model responses. Incorporating a layer of active learning enhances the model's performance and improves its generalization. Including experimental validation data further enhances functional classification and prediction models for better guidance when exploring or navigating the latent space.

Challenges and opportunities: perspectives on peptide design in therapeutic applications

The advantages of peptide-based therapeutics include their efficacy, safety, specificity, customizable nature and various synthesis options. Nevertheless, pursuing peptides for medical applications still faces essential challenges despite strides in *in silico* drug discovery and design.

One of the obstacles to IA-assisted peptide design is the lack of a centralized, comprehensive and curated source of peptide information. As described earlier, there are individual efforts such as LAMP [63] for specific biological activity or PepipediaDB [5] for general purpose. However, accelerated advances will require global collaborative efforts.

Individual and fragmented data availability results in inconsistencies and misclassifications; for example, very different biological activities have been reported for the same peptide. However, a significant challenge is the scarcity of systematized experimental

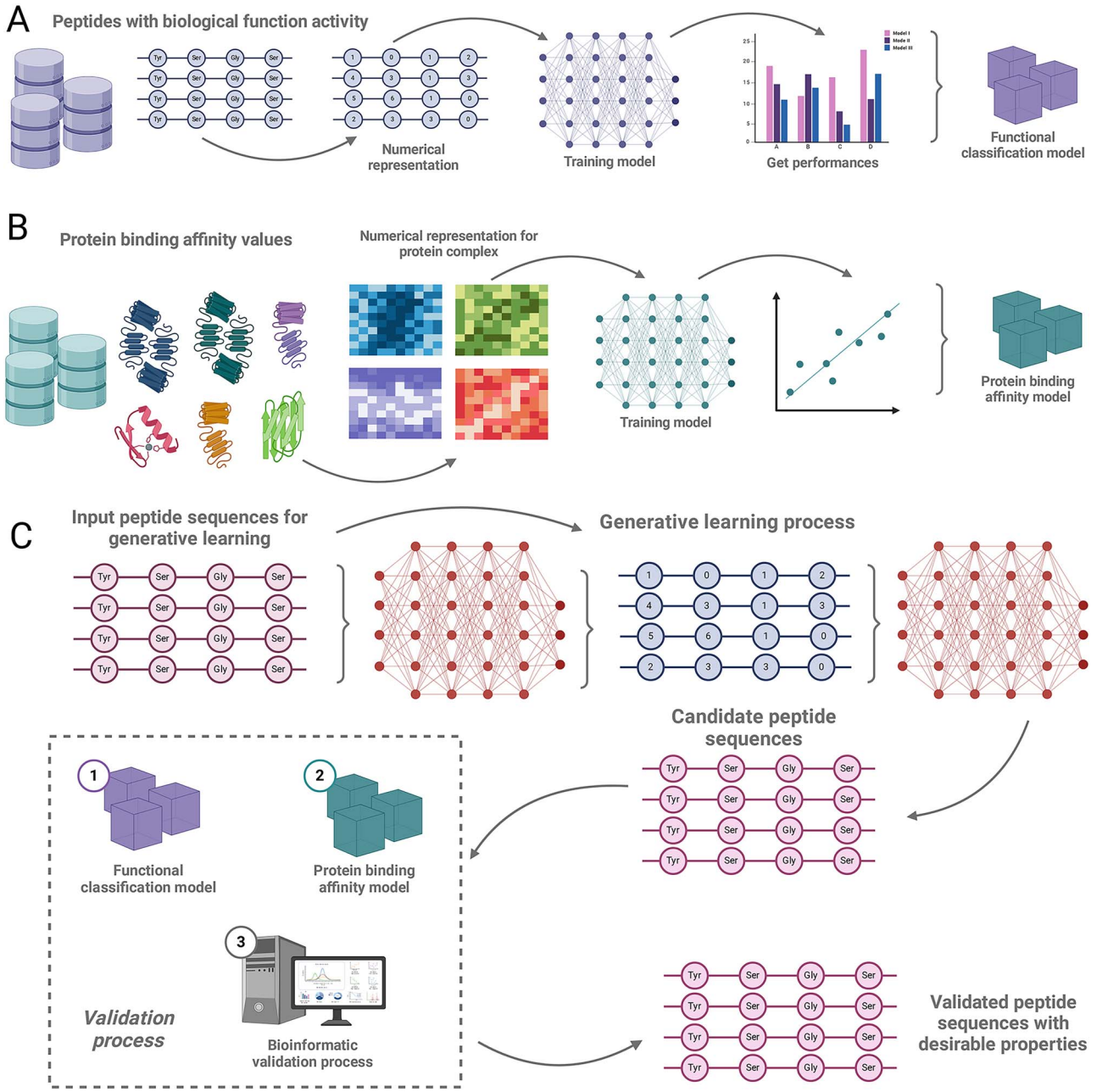

Figure 4. **A generalizable *in silico* pipeline to design therapeutic peptides with desirable properties combining ML, DGMs and biological structural validations**. **A** Pipeline to train functional biological classification models, including the classic steps to develop predictive models using ML approaches. **B** Pipeline to train a binding affinity predictive model using ML algorithms. In this case, it is necessary to represent the protein, the ligand and the complex structure. **C**. A pipeline to build a generative model for designing therapeutic peptides. The pipeline includes the DGM, the functional classification system, the protein-binding affinity predictive model and bioinformatics approaches to validate the designed peptide sequences.

data on the half-life, IC50 and other critical biochemical and biological variables.

A consortium effort to support and maintain a centralized data source and incentives to encourage the deposit of experimental data may overcome this challenge.

Despite the abundance of strategies, techniques and methodologies for developing functional activity classification models and predictive systems for peptides' physicochemical and thermodynamic properties, many of these approaches are challenging to replicate or access. This poses a significant obstacle not only to share results and methodologies but also to compare strategies effectively.

During the training of biological activity classification systems and the development of predictive systems, a crucial aspect involves the techniques used for the representation and coding of the peptide sequence.

Traditional methods like One-Hot encoding or physicochemical property coding often encounter challenges due to variations in sequence lengths. *zero-padding* techniques are typically employed to ensure uniformity during model training, introducing noise that increases with significant sequence length differences.

Solutions focusing on feature engineering-based representation offer an alternative, yet challenges persist in identifying

relevant variables and needs reduction techniques. Moreover, models developed using these strategies often need to improve, rendering them inefficient for therapeutic peptide design. Recently, representation learning methods leveraging pre-trained models have emerged as a promising solution, promoting superior performance in predictive models. However, their computational overhead can be substantial, particularly when handling large datasets.

The issue extends beyond methodological complexity to include the availability and accessibility of benchmark datasets. Thus, there is a pressing need to enhance replicability and ensure persistent access to benchmark datasets. The development of gold-standard datasets capable of validating various training strategies and facilitating performance comparisons is imperative.

On the other hand, peptide design requires a delicate balance between safety, delivery efficiency, stability and preservation of efficacy and specificity. Minor alterations in the peptide sequence can significantly impact binding affinity and susceptibility to degradation.

The moonlighting property of peptides adds another layer of complexity. That can, however, be exploited to achieve a combination of binding properties, transport capacity, mobility and interactions at various molecular levels. Current technical advances, such as implementing deep generative methods to generate new sequences and multi-task modelling, may overcome this challenge.

Alternatively, hierarchical predictive systems or rule-decision classification systems might be implemented for numerical models predicting the desired properties such as IC50, half-life, toxicity, etc.

Peptides, being proteins, are susceptible to PTMs. Unfortunately, these modifications are often overlooked during modelling, creating a gap and introducing uncertainty in peptide design. Notably, bacteriocin peptides, utilized as antimicrobials to combat antibiotic resistance, frequently exhibit such modifications, demonstrating the need to incorporate PTMs predictive models into the pipeline of therapeutic peptide design (See more details in section S3 of Supplementary Material).

Lastly, provided that a comprehensive curated source of peptide information to train potent models is generated, a continuous update and feedback with experimental data will be required. Reinforcement learning will enable the continuous update of the model, which in turn will increase performance and generalization capabilities.

Despite these challenges, automated peptide generation has the potential to produce sequences with desired characteristics, such as enzymatic degradation resistance and specificity. Integrating generative methods, predictive systems and bioinformatics tools provides invaluable support for exploring peptides, uncovering new sequences or peptides with therapeutic activities, and synergizing with traditional methods like directed evolution and rational design. Integrating these tools will accelerate the discovery of novel peptides to complement conventional therapeutic arsenals.

Key Points

- In this work, the properties of different functional peptides, therapeutic and biotechnology applications, relevant repositories, datasets and biological databases for peptide sequences are analysed.
- This work presents the most relevant machine learning strategies applied to develop predictive models using aminoacid sequences or protein structures as input for peptide studies.
- This work describes therapeutic peptide characteristics and strategies for the design and discovery, focusing on generative learning.
- An artificial intelligence pipeline to address the most common problems and challenges related to automated therapeutic peptide design is proposed in this work.

Supplementary data

Supplementary data is available online at *Briefings in Bioinformatics* online.

Funding

D.M.-O. acknowledges ANID for the project 'SUBVENCIÓN A INSTALACIÓN EN LA ACADEMIA CONVOCATORIA AÑO 2022', Folio 85220004. D.M.-O., A.D. and A.O.-N. gratefully acknowledge support from the Centre for Biotechnology and Bioengineering—CeBiB (PIA project FB0001, Conicyt, Chile). A.D. gratefully acknowledges support for the Fondecyt project 11230208. M.A.N. acknowledges ANID for grants Anillo ATE220016, M.A.N and R.U.-P acknowledge support for the Fondecyt 1230298. M.D.D. acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—within the Priority Program Molecular Machine Learning SPP2363 (Project Number 497207454).

Author contributions statement

M.G., M.A.N., A.O.-N. and D.M.-O.: conceptualization. D.M.-O., M.A.N. and A.O.-N.: validation. M.G., M.A.N., A.D., H.A.-K., M.D.D., G.C.-M., L.S.-V. and J.S.-Y.: investigation. M.G., A.D., M.D.D., H.A.-K., M.A.N. and D.M.-O.: writing, review and editing. A.O.-N., M.A.N., R.U.P. and D.M.-O.: supervision and funding resources. A.O.-N., M.A.N. and D.M.-O.: project administration.

Data availability

No new datasets were generated in this work.

References

1. Jakubczyk A, Karaś M, Rybczyńska-Tkaczyk K. *et al*. Current trends of bioactive peptides—new sources and therapeutic effect. *Foods* 2020;**9**:846. https://doi.org/10.3390/foods9070846.
2. Apostolopoulos V, Bojarska J, Chai T-T. *et al*. A global review on short peptides: Frontiers and perspectives. *Molecules* 2021;**26**:430. https://doi.org/10.3390/molecules26020430.
3. Muttenthaler M, King GF, Adams DJ. *et al*. Trends in peptide drug discovery. *Nat Rev Drug Discov* 2021;**20**:309–25. https://doi.org/10.1038/s41573-020-00135-8.
4. Carracedo-Reboredo P, Liñares-Blanco J, Rodríguez-Fernández N. *et al*. A review on machine learning approaches and trends in drug discovery. *Comput Struct Biotechnol J* 2021;**19**:4538–58. https://doi.org/10.1016/j.csbj.2021.08.011.
5. Quiroz C, Saavedra YB, Armijo-Galdames B. *et al*. Peptipedia: a user-friendly web application and a comprehensive

database for peptide research supported by machine learning approach. *Database* 2021;**2021**:baab055. https://doi.org/10.1093/database/baab055.
6. Aronica PGA, Reid LM, Desai N. *et al*. Computational methods and tools in antimicrobial peptide research. *J Chem Inf Model* 2021;**61**:3172–96. https://doi.org/10.1021/acs.jcim.1c00175.
7. Rustagi V, Gupta SRR, Bajaj M. *et al*. Pepanalyzer: predicting peptide properties using its sequence. *Amino Acids* 2023: 1–9.
8. Dara S, Dhamercherla S, Jadav SS. *et al*. Machine learning in drug discovery: a review. *Artif Intell Rev* 2022;**55**:1947–99. https://doi.org/10.1007/s10462-021-10058-4.
9. Szymczak P, MoŻejko M, Grzegorzek T. *et al*. Discovering highly potent antimicrobial peptides with deep generative model hydramp. *Nat Commun* 2023;**14**:1453.
10. Medina-Ortiz D, Salinas P, Cabas-Moras G. *et al*. Exploring machine learning algorithms and numerical representations strategies to develop sequence-based predictive models for protein networks. In: *International Conference on Computational Science and Its Applications*, 2023. p. 231–44. Springer.
11. Wang L, Wang N, Zhang W. *et al*. Therapeutic peptides: current applications and future directions. *Signal Transduct Target Ther* 2022;**7**:48.
12. Zachary W, Johnston KE, Arnold FH. *et al*. Protein sequence design with deep generative models. *Curr Opin Chem Biol* 2021; **65**:18–27.
13. Lau JL, Dunn MK. Therapeutic peptides: historical perspectives, current development trends, and future directions. *Bioorg Med Chem* 2018;**26**:2700–7.
14. Sharma K, Sharma KK, Sharma A. *et al*. Peptide-based drug discovery: current status and recent advances. *Drug Discov Today* 2022:103464.
15. Jinglin F, Nguyen K. Reduction of promiscuous peptides-enzyme inhibition and aggregation by negatively charged biopolymers. *ACS Appl Bio Mater* 2022;**5**:1839–45.
16. Lien S, Lowman HB. Therapeutic peptides. *Trends Biotechnol* 2003;**21**:556–62. https://doi.org/10.1016/j.tibtech.2003.10.005.
17. Taylor SI. Rational design of peptide agonists of cell-surface receptors. *Trends Pharmacol Sci* 2000;**21**:9–10. https://doi.org/10.1016/S0165-6147(99)01414-5.
18. Khan MM, Filipczak N, Torchilin VP. Cell penetrating peptides: a versatile vector for co-delivery of drug and genes in cancer. *J Control Release* 2021;**330**:1220–8. https://doi.org/10.1016/j.jconrel.2020.11.028.
19. Trier N, Hansen P, Houen G. Peptides, antibodies, peptide antibodies and more. *Int J Mol Sci* 2019;**20**:6289. https://doi.org/10.3390/ijms20246289.
20. McTiernan TJ, Diaz DB, Saunders GJ. *et al*. Navigating complex peptide structures using macrocycle conformational maps. *RSC. Chem Biol* 2022;**3**:739–47.
21. Martin-Gómez H, Tulla-Puche J. Lasso peptides: chemical approaches and structural elucidation. *Org Biomol Chem* 2018;**16**:5065–80. https://doi.org/10.1039/C8OB01304G.
22. Ageitos L, Torres MDT, de la Fuente-Nunez C. Biologically active peptides from venoms: applications in antibiotic resistance, cancer, and beyond. *Int J Mol Sci* 2022;**23**:15437.
23. Wang L, Niu D, Wang X. *et al*. A novel machine learning strategy for the prediction of antihypertensive peptides derived from food with high efficiency. *Foods* 2021;**10**:550.
24. Gogineni V, Hamann MT. Marine natural product peptides with therapeutic potential: chemistry, biosynthesis, and pharmacology. *Biochim Biophys Acta Gen Subj* 2018;**1862**:81–196. https://doi.org/10.1016/j.bbagen.2017.08.014.
25. Wegmuller S, Schmid S. Recombinant peptide production in microbial cells. *Curr Org Chem* 2014;**18**:1005–19.
26. Reimer JM, Haque AS, Tarry MJ. *et al*. Piecing together nonribosomal peptide synthesis. *Curr Opin Struct Biol* 2018;**49**:104–13. https://doi.org/10.1016/j.sbi.2018.01.011.
27. Martin V, Egelund PHG, Johansson H. *et al*. Greening the synthesis of peptide therapeutics: an industrial perspective. *RSC Adv* 2020;**10**:42457–92. https://doi.org/10.1039/D0RA07204D.
28. Sharma A, Kumar A, de la Torre BG. *et al*. Liquid-phase peptide synthesis (lpps): a third wave for the preparation of peptides. *Chem Rev* 2022;**122**:13516–46. https://doi.org/10.1021/acs.chemrev.2c00132.
29. Peter J, Burbach H. What are neuropeptides? *Neuropeptides: Methods and protocols, pages* 2011;**1–36**.
30. Yeo XY, Cunliffe G, Ho RC. *et al*. Potentials of neuropeptides as therapeutic agents for neurological diseases. *Biomedicine* 2022;**10**:343.
31. Hancock REW, Haney EF, Gill EE. The immunology of host defence peptides: beyond antimicrobial activity. *Nat Rev Immunol* 2016;**16**:321–34. https://doi.org/10.1038/nri.2016.29.
32. Hemion C, Li J, Kohler C. *et al*. Clearance of neurotoxic peptides and proteins by meningothelial cells. *Exp Cell Res* 2020;**396**:112322. https://doi.org/10.1016/j.yexcr.2020.112322.
33. Meade E, Slattery MA, Garvey M. Bacteriocins, potent antimicrobial peptides and the fight against multi drug resistant species: resistance is futile? *Antibiotics* 2020;**9**:32. https://doi.org/10.3390/antibiotics9010032.
34. Guryanova SV, Ovchinnikova TV. Immunomodulatory and allergenic properties of antimicrobial peptides. *Int J Mol Sci* 2022;**23**:2499.
35. Rodríguez JG, Plaza AV, Rojas SH. *et al*. Moonlighting peptides with emerging function. *PloS One* 2012;**7**:e40125.
36. Jeffery CJ. Protein moonlighting: what is it, and why is it important? *Philos Trans R Soc Lond B Biol Sci* 2018;**373**:20160523.
37. Oliva R, Chino M, Pane K. *et al*. Exploring the role of unnatural amino acids in antimicrobial peptides. *Sci Rep* 2018;**8**:8888. https://doi.org/10.1038/s41598-018-27231-5.
38. Zhang Q, Yanli L, Li S. *et al*. *Peptide-based biosensors*. Elsevier, 2018, 565–601.
39. Vahedifar A, Jianping W. Self-assembling peptides: structure, function, in silico prediction and applications. *Trends Food Sci Technol* 2022;**119**:476–94. https://doi.org/10.1016/j.tifs.2021.11.020.
40. Zhang Y-M, Ye D-X, Liu Y. *et al*. Peptides, new tools for plant protection in eco-agriculture. *Adv Agrochem* 2023;**2**:58–78. https://doi.org/10.1016/j.aac.2023.01.003.
41. Keymanesh K, Soltani S, Sardari S. Application of antimicrobial peptides in agriculture and food industry. *World J Microbiol Biotechnol* 2009;**25**:933–44. https://doi.org/10.1007/s11274-009-9984-7.
42. Schaefer SC, Gasic K, Cammue B. *et al*. Enhanced resistance to early blight in transgenic tomato lines expressing heterologous plant defense genes. *Planta* 2005;**222**:858–66. https://doi.org/10.1007/s00425-005-0026-x.
43. Berillo D, Yeskendir A, Zharkinbekov Z. *et al*. Peptide-based drug delivery systems. *Medicina* 2021;**57**:1209. https://doi.org/10.3390/medicina57111209.
44. Ma Z, Zhang D, Cheng Z. *et al*. Designed symmetrical β-hairpin peptides for treating multidrug-resistant salmonella typhimurium infections. *Eur J Med Chem* 2022;**243**:114769. https://doi.org/10.1016/j.ejmech.2022.114769.
45. Schito GC, Parenti F, Courvalin P. Teicoplanin chemistry and microbiology. *J Chemother* 2000;**12**:5–14.

46. Knudsen LB, Lau J. The discovery and development of liraglutide and semaglutide. *Front Endocrinol* 2019;**10**:155.
47. Shirazi AN, El-Sayed NS, Tiwari RK. *et al*. Cyclic peptide containing hydrophobic and positively charged residues as a drug delivery system for curcumin. *Curr Drug Deliv* 2016;**13**:409–17.
48. Mozaffari S, Bousoik E, Amirrad F. *et al*. Hamidreza Montazeri Aliabadi. Amphiphilic peptides for efficient sirna delivery. *Polymers* 2019;**11**:703.
49. Baig MH, Ahmad K, Saeed M. *et al*. Peptide based therapeutics and their use for the treatment of neurodegenerative and other diseases. *Biomed Pharmacother* 2018;**103**:574–81. https://doi.org/10.1016/j.biopha.2018.04.025.
50. Filichia E, Hoffer B, Qi X. *et al*. Inhibition of drp1 mitochondrial translocation provides neural protection in dopaminergic system in a parkinson's disease model induced by mptp. *Sci Rep* 2016;**6**:32656.
51. Cardoso MH, Orozco RQ, Rezende SB. *et al*. Computer-aided design of antimicrobial peptides: are we generating effective drug candidates? *Front Microbiol* 2020;**10**:3097. https://doi.org/10.3389/fmicb.2019.03097.
52. Sabe VT, Ntombela T, Jhamba LA. *et al*. Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: a review. *Eur J Med Chem* 2021;**224**:113705. https://doi.org/10.1016/j.ejmech.2021.113705.
53. Holzinger A, Keiblinger K, Holub P. *et al*. Ai for life: trends in artificial intelligence for biotechnology. *N Biotechnol* 2023;**74**:16–24. https://doi.org/10.1016/j.nbt.2023.02.001.
54. Jones DT. Setting the standards for machine learning in biology. *Nat Rev Mol Cell Biol* 2019;**20**:659–60. https://doi.org/10.1038/s41580-019-0176-5.
55. Angermueller C, Pärnamaa T, Parts L. *et al*. Deep learning for computational biology. *Mol Syst Biol* 2016;**12**:878. https://doi.org/10.15252/msb.20156651.
56. Xia T, Wei-Shinn K. Geometric graph representation learning on protein structure prediction. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021. p. 1873–1883.
57. Fernández D, Olivera-Nappa Á, Uribe-Paredes R. *et al*. Exploring machine learning algorithms and protein language models strategies to develop enzyme classification systems. In: *International Work-Conference on Bioinformatics and Biomedical Engineering*, 2023. p. 307–19. Springer. https://doi.org/10.1007/978-3-031-34953-9_24.
58. Baek M, DiMaio F, Anishchenko I. *et al*. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021;**373**:871–6. https://doi.org/10.1126/science.abj8754.
59. Jumper J, Evans R, Pritzel A. *et al*. Highly accurate protein structure prediction with alphafold. *Nature* 2021;**596**:583–9. https://doi.org/10.1038/s41586-021-03819-2.
60. Theodoris CV, Xiao L, Chopra A. *et al*. Transfer learning enables predictions in network biology. *Nature* 2023:1–9.
61. Barbero-Aparicio JA, Olivares-Gil A, Rodríguez JJ. *et al*. Addressing data scarcity in protein fitness landscape analysis: a study on semi-supervised and deep transfer learning techniques. *Inf Fusion* 2024;**102**:102035. https://doi.org/10.1016/j.inffus.2023.102035.
62. Singh S, Chaudhary K, Dhanda SK. *et al*. Satpdb: a database of structurally annotated therapeutic peptides. *Nucleic Acids Res* 2016;**44**:D1119–26. https://doi.org/10.1093/nar/gkv1114.
63. Ye G, Wu H, Huang J. *et al*. Lamp2: a major update of the database linking antimicrobial peptides. *Database* 2020;**2020**:baaa061.
64. Wynendaele E, Bronselaer A, Nielandt J. *et al*. Quorumpeps database: chemical space, microbial origin and functionality of quorum sensing peptides. *Nucleic Acids Res* 2013;**41**:D655–9. https://doi.org/10.1093/nar/gks1137.
65. Ramaprasad ASE, Singh S, Venkatesan S. *et al*. Antiangiopred: a server for prediction of anti-angiogenic peptides. *PloS One* 2015;**10**:e0136990. https://doi.org/10.1371/journal.pone.0136990.
66. Hammami R, Zouhir A, Le Lay C. *et al*. Bactibase second release: a database and tool platform for bacteriocin characterization. *BMC Microbiol* 2010;**10**:1–5.
67. Dallago C, Schütze K, Heinzinger M. *et al*. Learned embeddings from deep learning to visualize and predict protein sets. *Curr Protocols* 2021;**1**:e113. https://doi.org/10.1002/cpz1.113.
68. Medina-Ortiz D, Sr, Cabas-Mora G, Sr, Moya-Barria I, Sr. *et al*. Rudeus, a machine learning classification system to study dna-binding proteins. *bioRxiv* 2024:2024–02.
69. Medina-Ortiz D, Contreras S, Quiroz C. *et al*. Development of supervised learning predictive models for highly non-linear biological, biomedical, and general datasets. *Front Mol Biosci* 2020;**7**:13.
70. Lertampaiporn S, Vorapreeda T, Hongsthong A. *et al*. Ensemble-amppred: robust amp prediction and recognition using the ensemble learning method with a new hybrid feature for differentiating amps. *Genes* 2021;**12**:137. https://doi.org/10.3390/genes12020137.
71. Hussain W. Samp-pfpdeep: improving accuracy of short antimicrobial peptides prediction using three different sequence encodings and deep neural networks. *Brief Bioinform* 2022;**23**:bbab487.
72. Lissabet JFB, Belén LH, Farias JG. Antivpp 1.0: a portable tool for prediction of antiviral peptides. *Comput Biol Med* 2019;**107**:127–30. https://doi.org/10.1016/j.compbiomed.2019.02.011.
73. Pang Y, Yao L, Jhong J-H. *et al*. Avpiden: a new scheme for identification and functional prediction of antiviral peptides based on machine learning approaches. *Brief Bioinform* 2021;**22**:bbab263.
74. Timmons PB, Hewage CM. Ennavia is a novel method which employs neural networks for antiviral and anti-coronavirus activity prediction for therapeutic peptides. *Brief Bioinform* 2021;**22**:bbab258.
75. Lin T-T, Sun Y-Y, Wang C-T. *et al*. Ai4avp: an antiviral peptides predictor in deep learning approach with generative adversarial network data augmentation. . *Bioinform Adv* 2022;**2**:vbac080.
76. Khatun MS, Hasan MM, Kurata H. Preaip: computational prediction of anti-inflammatory peptides by integrating multiple complementary features. *Front Genet* 2019;**10**. https://doi.org/10.3389/fgene.2019.00129.
77. Ahmed S, Muhammod R, Khan ZH. *et al*. Acp-mhcnn: an accurate multi-headed deep-convolutional neural network to predict anticancer peptides. *Sci Rep* 2021;**11**:23676. https://doi.org/10.1038/s41598-021-02703-3.
78. Manavalan B, Subramaniyam S, Shin TH. *et al*. Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. *J Proteome Res* 2018;**17**:2715–26. https://doi.org/10.1021/acs.jproteome.8b00148.
79. Qureshi A, Tandon H, Kumar M. Avp-ic50pred: multiple machine learning techniques-based prediction of peptide

antiviral activity in terms of half maximal inhibitory concentration (ic50). *Pept Sci* 2015;**104**:753–63. https://doi.org/10.1002/bip.22703.
80. Olsen TH, Yesiltas B, Marin FI. *et al*. Anoxpepred: using deep learning for the prediction of antioxidative properties of peptides. *Sci Rep* 2020;**10**:21471.
81. Mathur D, Singh S, Mehta A. *et al*. In silico approaches for predicting the half-life of natural and modified peptides in blood. *PloS One* 2018;**13**:e0196829. https://doi.org/10.1371/journal.pone.0196829.
82. Bateman TJ. Drug discovery. In: *Atkinson's Principles of Clinical Pharmacology*, Elsevier, 2022. 563–72. https://doi.org/10.1016/B978-0-12-819869-8.00019-7.
83. Ain Q u, Batool M, Choi S. Tlr4-targeting therapeutics: structural basis and computer-aided drug discovery approaches. *Molecules* 2020;**25**:627.
84. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of r&d costs. *J Health Econ* 2016;**47**:20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012.
85. Vamathevan J, Clark D, Czodrowski P. *et al*. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov* 2019;**18**:463–77.
86. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov* 2004;**3**:673–83. https://doi.org/10.1038/nrd1468.
87. Pushpakom S, Iorio F, Eyers PA. *et al*. Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov* 2019;**18**:41–58. https://doi.org/10.1038/nrd.2018.168.
88. Madhukar NS, Khade PK, Huang L. *et al*. A bayesian machine learning approach for drug target identification using diverse data types. *Nat Commun* 2019;**10**:5221.
89. Zheng S, Li Y, Chen S. *et al*. Predicting drug–protein interaction using quasi-visual question answering system. *Nat Mach Intell* 2020;**2**:134–40. https://doi.org/10.1038/s42256-020-0152-y.
90. Olayan RS, Ashoor H, Bajic VB. Ddr: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches. *Bioinformatics* 2018;**34**:1164–73. https://doi.org/10.1093/bioinformatics/btx731.
91. Wen M, Zhang Z, Niu S. *et al*. Deep-learning-based drug–target interaction prediction. *J Proteome Res* 2017;**16**:1401–9. https://doi.org/10.1021/acs.jproteome.6b00618.
92. Wang Y-B, You Z-H, Yang S. *et al*. A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network. *BMC Med Inform Decis Mak* 2020;**20**:1–9.
93. Mohsen A, Tripathi LP, Mizuguchi K. Deep learning prediction of adverse drug reactions in drug discovery using open tg–gates and faers databases. *Front Drug Discovery* 2021;**1**:10.
94. Galeano D, Paccanaro A. Machine learning prediction of side effects for drugs in clinical trials. *Cell reports Methods* 2022;**2**:100358.
95. Cheng Y, Gong Y, Liu Y. *et al*. Molecular design in drug discovery: a comprehensive review of deep generative models. *Brief Bioinform* 2021;**22**:bbab344.
96. Li Y, Pei J, Lai L. Structure-based de novo drug design using 3d deep generative models. *Chem Sci* 2021;**12**:13664–75. https://doi.org/10.1039/D1SC04444C.
97. Bagal V, Aggarwal R, Vinod PK. *et al*. Molgpt: molecular generation using a transformer-decoder model. *J Chem Inf Model* 2021;**62**:2064–76. https://doi.org/10.1021/acs.jcim.1c00600.
98. Wan F, Kontogiorgos-Heintz D, de la Fuente-Nunez C. Deep generative models for peptide design. *Digital Discovery* 2022;**1**:195–208. https://doi.org/10.1039/D1DD00024A.
99. Strokach A, Kim PM. Deep generative modeling for protein design. *Curr Opin Struct Biol* 2022;**72**:226–36.
100. Ding W, Nakai K, Gong H. Protein design via deep learning. *Brief Bioinform* 2022;**23**.
101. Gui J, Sun Z, Wen Y. *et al*. A review on generative adversarial networks: algorithms, theory, and applications. *IEEE Trans Knowl Data Eng* 2021;**35**:3313–32.
102. Creswell A, White T, Dumoulin V. *et al*. Generative adversarial networks: an overview. *IEEE Signal Process Mag* 2018;**35**:53–65.
103. Lin E, Lin C-H, Lane H-Y. De novo peptide and protein design using generative adversarial networks: an update. *J Chem Inf Model* 2022;**62**:761–74. https://doi.org/10.1021/acs.jcim.1c01361.
104. Cao H, Tan C, Gao Z. *et al*. A survey on generative diffusion models. *IEEE Trans Knowl Data Eng* 2024;
105. Yang L, Zhang Z, Song Y. *et al*. Diffusion models: a comprehensive survey of methods and applications. *ACM Comput Surv* 2023;**56**:1–39.
106. Karras T, Aittala M, Aila T. *et al*. Elucidating the design space of diffusion-based generative models. *Adv Neural Inf Process Syst* 2022;**35**:26565–77.
107. Dean SN, Jerome Anthony E, Alvarez DZ. *et al*. Pepvae: variational autoencoder framework for antimicrobial peptide generation and activity prediction. *Front Microbiol* 2021;**12**:725727. https://doi.org/10.3389/fmicb.2021.725727.
108. Das P, Wadhawan K, Chang O. *et al*. Pepcvae: semi-supervised targeted design of antimicrobial peptide sequences arXiv preprint arXiv:1810.07743. 2018;
109. Chen Q, Yang C, Xie Y. *et al*. Gm-pep: a high efficiency strategy to de novo design functional peptide sequences. *J Chem Inf Model* 2022;**62**:2617–29. https://doi.org/10.1021/acs.jcim.2c00089.
110. Das P, Sercu T, Wadhawan K. *et al*. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations.. *Nat Biomed Eng* 2021;**5**:613–23.
111. Szymczak P, MoŻejko M, Grzegorzek T. *et al*. Discovering highly potent antimicrobial peptides with deep generative model hydramp. *Nat Commun* 2023;**14**:1453. https://doi.org/10.1038/s41467-023-36994-z.
112. Chen S, Lin T, Basu R. *et al*. Design of target specific peptide inhibitors using generative deep learning and molecular dynamics simulations. *Nature. Nat Commun* 2024;**15**:1611.
113. Chen S, Lin T, Basu R. *et al*. Design of peptide inhibitors targeting β-catenin using generative deep learning and molecular dynamics simulations. *Nat Commun* 2024;**15**:1611.
114. Fukunaga I, Matsukiyo Y, Kaitoh K. *et al*. Automatic generation of functional peptides with desired bioactivity and membrane permeability using bayesian optimization. *Mol Inf* 2024;**43**:e202300148. https://doi.org/10.1002/minf.202300148.
115. Bell DR, Domeniconi G, Yang C-C. *et al*. Dynamics-based peptide–mhc binding optimization by a convolutional variational autoencoder: a use-case model for Castelo. *J Chem Theory Comput* 2021;**17**:7962–71. https://doi.org/10.1021/acs.jctc.1c00870.
116. Rossetto A, Zhou W. Gandalf: Peptide generation for drug design using sequential and structural generative adversarial networks. In: *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2020. p. 1–10. ACM.
117. Usmani SS, Bedi G, Samuel JS. *et al*. Thpdb: database of fda-approved peptide and protein therapeutics. *PloS One* 2017;**12**:e0181748. https://doi.org/10.1371/journal.pone.0181748.

118. The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Res* 2023;**51**:D523–31. https://doi.org/10.1093/nar/gkac1052.
119. Dundas J, Ouyang Z, Tseng J. *et al*. Castp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res* 2006;**34**:W116–8.
120. Hoos A. Development of immuno-oncology drugs—from ctla4 to pd1 to the next generations. *Nat Rev Drug Discov* 2016;**15**:235–47. https://doi.org/10.1038/nrd.2015.35.
121. Surana S, Arora P, Singh D. *et al*. Pandoragan: generating antiviral peptides using generative adversarial network. *bioRxiv* page 2021.02.15.431193, 1 2022.
122. Van Oort CM, Ferrell JB, Remington JM. *et al*. Ampgan v2: machine learning-guided design of antimicrobial peptides. *J Chem Inf Model* 2021;**61**:2198–207. https://doi.org/10.1021/acs.jcim.0c01441.
123. Xie X, Valiente PA, Kim PM. Helixgan a deep-learning methodology for conditional de novo design of α-helix structures. *Bioinformatics* 2023;**39**:btad036.
124. Mao J, Guan S, Chen Y. *et al*. Application of a deep generative model produces novel and diverse functional peptides against microbial resistance. *Comput Struct Biotechnol J* 2023;**21**:463–71. https://doi.org/10.1016/j.csbj.2022.12.029.
125. Li J, Beaudoin C, Ghosh S. Energy-based generative models for target-specific drug discovery. *Front Mol Med* 2023;**3**. https://doi.org/10.3389/fmmed.2023.1160877.
126. Chen T, Vure P, Pulugurta R. *et al*. Amp-diffusion: integrating latent diffusion with protein language models for antimicrobial peptide generation. *bioRxiv* 2024:2024–03.
127. Watson JL, Juergens D, Bennett NR. *et al*. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *BioRxiv* 2022:2022–12.
128. Kong X, Huang W, Liu Y. Full-atom peptide design with geometric latent diffusion arXiv preprint arXiv:2402.13555. 2024.
129. Wang Y, Liu X, Huang F. *et al*. A multi-modal contrastive diffusion model for therapeutic peptide generation. *In Proceedings of the AAAI Conference on Artificial Intelligence* 2024;**38**:3–11. https://doi.org/10.1609/aaai.v38i1.27749.
130. Wittmann BJ, Johnston KE, Zachary W. *et al*. Advances in machine learning for directed evolution. *Curr Opin Struct Biol* 2021;**69**:11–8. https://doi.org/10.1016/j.sbi.2021.01.008.
131. Chang L, Mondal A, Singh B. *et al*. Revolutionizing peptide-based drug discovery: advances in the post-alphafold era. *Wiley Interdiscip Rev: Comput Mol Sci* 2024;**14**:e1693.