Chapter 1

Perspectives Chapter: Data-Centric Strategies for Machine Learning-Driven Therapeutic Peptide Design – Challenges and Perspectives

David Medina-Ortiz, Sebastián Escobedo, Norma Murillo-Acevedo, Nicole Soto-García, Diego Fernández-Villegas, Diego Sandoval and Anamaría Daza

Abstract

Machine learning is increasingly applied to the discovery and rational design of therapeutic peptides, offering powerful capabilities for functional prediction, pharmacological profiling, and de novo sequence generation. However, the reliability and generalizability of such approaches critically depend on the quality, completeness, and interoperability of the datasets on which they are built. This chapter provides a comprehensive overview of the computational strategies and best practices for constructing peptide datasets suitable for machine learning applications. We review widely used peptide repositories and examine strategies for data extraction, preprocessing, and standardization, with emphasis on mitigating annotation errors, redundancy, and incomplete metadata. Approaches for redundancy reduction—including sequence homology clustering, structure-based grouping, and embedding-driven similarity analysis—are compared in terms of their impact on dataset diversity and downstream performance. We also discuss challenges posed by scarce and imbalanced datasets, highlighting data-efficient strategies such as transfer learning, positive-unlabeled learning, and contrastive learning. The chapter concludes by outlining future directions, including the development of a peptide-specific markup language, standardized reporting formats, and widespread adoption of FAIR-compliant practices. By combining rigorous preprocessing with interoperable data standards, the field can enable more robust, interpretable, and transferable ML models, accelerating the discovery of safe and effective therapeutic peptides.

Keywords: therapeutic peptides, machine learning, drug discovery, therapeutic peptide, dataset curation, fair principles

1. Introduction

Therapeutic peptides constitute a rapidly expanding class of bioactive molecules distinguished by their high target specificity, low toxicity, and broad applicability in biomedical, pharmaceutical, and industrial domains [1]. Their discovery and optimization have been accelerated by advances in computational biology, the emergence of large-scale curated peptide repositories, and the integration of machine learning (ML) approaches capable of predicting functional, pharmacological, and ADMET-related properties with increasing accuracy [2]. However, the performance, reproducibility, and generalizability of such data-driven strategies are critically dependent on the quality, completeness, and interoperability of the underlying datasets.

This chapter provides a detailed overview of the computational foundations required to build predictive and generative models for therapeutic peptides, with a strong emphasis on rigorous data curation, preprocessing, and compliance with FAIR principles (Findable, Accessible, Interoperable, and Reusable). We examine common issues in dataset construction—such as annotation errors, redundancy, inconsistent metadata, and the scarcity of validated negative examples—and describe strategies to address them through clustering, embedding-based similarity analysis, learning-driven redundancy filters, and standardized data harmonization workflows.

Besides, we discuss the broader challenges of dataset interoperability, the impact of inconsistent reporting practices, and the barriers posed by the lack of unified quality control protocols. We propose future directions that include the establishment of community-endorsed preprocessing guidelines, the development of a machine-readable peptide markup language, and the adoption of standardized reporting formats to facilitate large-scale data integration. By coupling FAIR-compliant practices with robust preprocessing pipelines, the field can unlock the full potential of ML-assisted peptide discovery, enabling more reliable, interpretable, and impactful therapeutic peptide design.

2. Therapeutic peptides: Biological definitions, applications, and translational challenges

Peptides are short chains of amino acids of length between 2 and 100 residues with diverse secondary structure patterns [1]. Peptides participate in different physiological processes, including hormone signaling, immune modulation, and cellular communication [3]. Also, peptides play a significant role in biotechnology, medicine, agriculture, and food preservation [4].

Peptides have demonstrated high selectivity and specificity for various molecular targets [5]. The capability of peptides to bind specific cell surface receptors with high affinity allows precise activation of intracellular pathways, reducing off-target effects and adverse reactions compared to conventional small-molecule drugs [6]. Additional properties of these molecules, like enhanced tissue penetration, low immunogenicity, and biodegradability, positioned peptides as promising therapeutic agents [7].

Despite their numerous advantages, therapeutic peptides face several limitations that hinder their clinical translation [8]. These molecules typically suffer from poor oral bioavailability, rapid enzymatic degradation, limited membrane permeability, and short systemic half-lives, often requiring parenteral administration [9]. Such unfavorable pharmacokinetic properties can compromise both patient adherence

DOI: http://dx.doi.org/10.5772/intechopen.1013230

and therapeutic efficacy [7]. To address these challenges, a range of strategies such as peptide cyclization, PEGylation, lipidation, nanoparticle-based delivery systems, and the integration of cell-penetrating motifs are being explored to improve stability and prolong biological activity [10].

Advances in computational biology, high-throughput screening, and peptide engineering have substantially accelerated both drug discovery and repurposing efforts [11]. In recent years, ML-based strategies have been increasingly integrated into conventional pipelines, enabling more efficient identification and optimization of therapeutic peptides [12]. The following section provides an overview of key ML-based strategies used to accelerate the discovery process and support the rational design of peptides with desirable pharmacological properties.

3. Machine learning strategies for drug discovery and drug repurposing

Machine learning has emerged as a key component in the discovery, design, and repurposing of therapeutic peptides, offering powerful tools to facilitate early-stage screening, predict pharmacokinetic behavior, and generate novel bioactive sequences. Using curated datasets and high-dimensional sequence embeddings, ML strategies have significantly improved the efficiency, scalability, and precision of traditional pipelines. This section provides an integrative overview of three key ML applications, including functional and toxicity classification, prediction of pharmacological and ADMET-related properties, and *de novo* peptide generation. **Figure 1** further summarizes the main ML-based strategies employed in the study of therapeutic peptides.

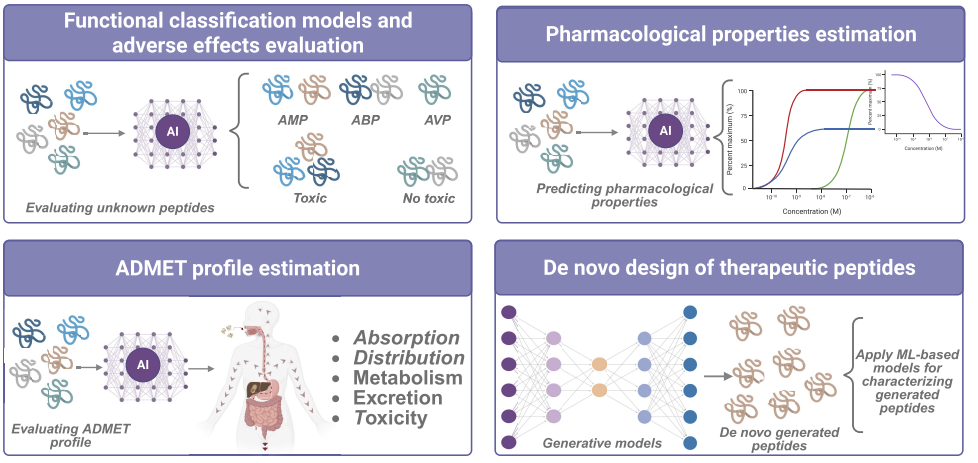

Figure 1.
Applications of machine learning in therapeutic peptide discovery and optimization. Machine learning has been applied to four complementary strategies in the context of accelerating discovery and assisting rational design of therapeutic peptides. First, functional classification models and adverse effect evaluation enable the discrimination of bioactive peptides from inactive or toxic sequences. Second, pharmacological property prediction supports the assessment of potency, specificity, and therapeutic relevance. Third, ADMET profile estimation—covering absorption, distribution, metabolism, excretion, and toxicity—facilitates early identification of candidates with favorable pharmacokinetics and safety profiles. Finally, de novo design approaches using generative algorithms to create novel peptide sequences, which can be iteratively refined using the preceding predictive models.

3.1 Functional classification and toxicity prediction

One of the most relevant applications of ML in peptide discovery involves the classification of sequences based on their biological activity and safety profiles. Functional classification models are trained to identify peptides with specific therapeutic functions such as antimicrobial, anticancer, antidiabetic, or antihypertensive activity by learning sequence-to-label relationships from experimentally validated datasets [13]. These models not only accelerate candidate prioritization but also reduce the dependence on time-consuming and resource-intensive experimental assays [14].

Beyond bioactivity, the safety of candidate peptides is equally critical. Toxicity prediction models are designed to detect adverse properties, such as hemolysis, cytotoxicity, or allergenicity, which are the leading causes of late-stage weakening in peptide-based therapeutics. Several computational tools, including ToxinPred3.0, tAMPer, ToxIBTL, and ATSE [15, 16, 17, 18], have been developed using a range of ML strategies, from support vector machines and ensemble models to deep learning architectures like CNNs, BiLSTMs, and transformers. These models increasingly incorporate both sequence and structural features and often benefit from ensemble or multimodal integration to improve generalizability across peptide classes [19].

Advanced frameworks also address more specific safety criteria. For instance, hemolytic activity and immunogenicity are now routinely predicted using classifiers that combine traditional feature engineering with embeddings derived from pretrained protein language models. These approaches facilitate early toxicity filtering, supporting safer and more efficient peptide development pipelines [20].

3.2 Predicting pharmacological properties and ADMET profiles

Following functional and safety evaluation, the next step in peptide development involves assessing their pharmacokinetic and pharmacodynamic properties. Pharmacokinetics and toxicity are typically summarized by the ADMET profile—Absorption, Distribution, Metabolism, Excretion, and Toxicity [21]. These parameters determine the systemic behavior of peptides, guiding formulation, dosage, and route-of-administration decisions [22].

ML methods have proven effective in modeling physicochemical and pharmacokinetic variables related to ADMET, including solubility, stability, permeability, and half-life [23]. While early models have been developed based on traditional classifiers such as random forests and SVMs, recent strategies increasingly adopt deep learning architectures capable of multitask prediction [22]. Alternatively, attention-based models and transfer learning strategies have enabled simultaneous prediction of multiple pharmacokinetic properties, improving both accuracy and computational efficiency [14].

A major challenge in this field is the limited availability of peptide-specific ADMET datasets [13]. Most predictive frameworks described in the literature—such as PharmaBench [21]—are primarily curated for small molecules and exclude peptides, largely due to fundamental structural and chemical differences [16]. To bridge this gap, recent efforts have introduced peptide-oriented ADMET predictors that combine physicochemical descriptors with sequence embeddings derived from pre-trained protein language models, including ESM-2 and ProtT5 [14].

DOI: http://dx.doi.org/10.5772/intechopen.1013230

Furthermore, emerging frameworks increasingly incorporate ADMET predictors into multi-objective optimization workflows. In these scenarios, reinforcement learning approaches are used to guide peptide design toward optimal pharmacological profiles [24]. For example, tools now exist to predict the half-life of peptides in digestive environments, model the stability of the blood, or estimate the minimum inhibitory concentration in antimicrobial peptides. Nevertheless, the success of these methods remains tightly linked to the availability of well-annotated experimental data, which is still limited due to inconsistent assay formats and incomplete metadata reporting.

3.3 *De novo* design and optimization via generative models

One of the most transformative developments in ML-assisted peptide research is the application of generative models for the *de novo* design of therapeutic sequences [1]. Unlike conventional strategies based on motif extraction or sequence alignment, generative models enable exploration of vast, uncharted regions of sequence space, facilitating the discovery of novel peptides with tailored biological and pharmacokinetic properties [13].

Several architectures have been adapted for *de novo* design of therapeutic peptides, including variational autoencoders (VAEs), generative adversarial networks (GANs), and, more recently, diffusion models [25].

A key advancement in this area is the integration of protein language models such as ProtBERT, ProtT5, and ESM-2 into the generative workflow [26]. These models generate dense embeddings that capture both local sequence motifs and global structural patterns, improving the biological plausibility of generated peptides [16]. ESM3, in particular, represents a new frontier in generative modeling. As a multimodal model trained on billions of sequences and structures, ESM3 encodes sequence, structure, and function in a shared latent space, enabling conditional generation that responds to complex prompts and design constraints [27].

After generation, peptides typically experience a post-processing pipeline that includes *in silico* screening for undesirable properties such as ambiguity, instability, or predicted toxicity [22]. Filtering steps may involve rule-based exclusion, toxicity classifiers, or property-based ranking functions. In more advanced workflows, generated sequences are iteratively optimized using reinforcement learning or genetic algorithms to refine functionality, enhance stability, or reduce immunogenicity [13].

Despite the promise of the ML-based techniques, current generative frameworks still face important limitations. Many available training datasets lack detailed functional annotations, structural metadata, or pharmacokinetic labels, factors essential for robust model training and reliable prediction[14]. As a result, the more general applicability of generative peptide design will depend on continued efforts to develop standardized, high-quality, and openly accessible peptide datasets.

4. Curated databases and specialized repositories for therapeutic peptide research

Computational methods and the development of curated biological datasets have supported the increase of peptide-based therapeutic compounds [28]. In this

scenario, well-structured peptide databases are relevant for generating centralized resources. These resources can include various peptide information, such as functional annotation, pharmacological and physicochemical properties, ADMET profiles, and sequence and structural properties of peptides [29]. Besides, these resources contribute to multiple stages of the modern drug discovery pipelines, including peptide screening, optimization, repurposing, and validation [30].

Traditional resources such as UniProt and the Protein Data Bank (PDB) are critical for peptide research by providing standardized data [31, 32]. However, the growing demand for peptide-centered data has led to the multiplication of specialized repositories, many of which focus on particular biological functions, therapeutic applications, or biochemical properties [33]. General-purpose databases like PepBank and StraPep aggregate diverse peptide sequences, often derived from text mining or literature curation, with limited therapeutic contextualization [34, 35]. In contrast, domain-specific repositories such as dbAMP, DBAASP, DRAMP, CAMP, and LAMP contain specific information on antimicrobial peptides (AMPs), including sources, mechanisms of action, resistance profiles, and optimization strategies [36–40]. Additional resources such as InverPep, SATPdb, and APD3 provide structural annotations and experimentally validated AMP data [41–43].

Databases dedicated to functional classes of peptides have also emerged. AVPdb and CancerPPD, for example, focus on antiviral and anticancer peptides, respectively, usually providing structural motifs, activity values (e.g., IC_{50}), and target pathways [44, 45]. Other examples include Hemolytik (hemolytic activity), B3Pdb (blood–brain barrier penetration), Quorumpeps (quorum-sensing peptides), and AntiAngioPred (antiangiogenic activity) [46–49]. These resources support specific areas of therapeutic interest and are increasingly used to guide design strategies in limited biomedical domains [33].

Furthermore, BioDADPep, AHTPDB, and related datasets report potency values for peptides with antihypertensive, antidiabetic, and metabolic effects, with implications for both clinical therapeutics and functional food development [50, 51]. In addition, pharmacokinetic properties are another area of focus. Databases and tools such as PEPlife and PLifePred provide experimentally determined or computationally predicted half-lives [52, 53].

Regulatory-oriented repositories like THPdb and PepTherDia catalogue peptide-based drugs approved by agencies such as the FDA [54, 55]. These databases include information on pharmacokinetics, mechanisms of action, clinical indications, and administration routes, serving as reference points for drug repurposing, benchmarking, and mechanistic studies [56].

Peptipedia v2.0 exemplifies a more recent generation of peptide knowledge bases designed to aggregate, annotate, and organize peptide-related information from multiple sources [28]. While extensive in scope—reporting over 3.9 million peptides from more than 70 public databases and publications—its strength lies in the integration of curated annotations covering sequence, structure (e.g., linear, cyclic, disulfide-rich), biological function (e.g., antimicrobial, anticancer), and physicochemical properties [57]. While Peptipedia represents a comprehensive tool for peptide exploration, its performance, like that of other databases, depends on the quality and consistency of its integrated sources [58].

The diversity of these resources reflects the increasing complexity of peptide-based research and its applications in oncology, infectious disease, neurology, and metabolic disorders [2]. As the field matures, the adoption of FAIR principles

DOI: http://dx.doi.org/10.5772/intechopen.1013230

(Findable, Accessible, Interoperable, Reusable) becomes essential for ensuring data quality, transparency, and reuse [59]. Many repositories now support programmatic access, cross-database referencing, and machine-readable formats that enable their integration into AI-based frameworks for predictive modeling, rational design, and *de novo* peptide generation [60].

5. Strategies for data extraction and standardization

Building datasets for computational biology applications studying therapeutic peptides requires access to trustworthy data repositories, the implementation of well-documented data-extraction strategies, and the development of standardized approaches [61]. In the context of therapeutic peptides, where data sources are diverse, encompassing various data formats, different knowledge, and, in some cases, unstructured scientific texts, the development of data extraction, preprocessing, and data standardization is crucial for ensuring data quality, reproducibility, and reusability [62]. An overview of these processes is illustrated in **Figure 2**.

This section describes the most common strategies employed for data extraction and dataset construction in the context of therapeutic peptides, with a primary focus on data-driven applications based on ML-based approaches.

5.1 Extraction of peptide data from heterogeneous sources

Traditionally, extracting data has involved manually filtering through peptide sequences and annotations from various public databases, such as UniProt and PDB, as well as specialized peptide libraries [63]. These platforms typically allow users to download data in structured formats, such as FASTA, XML, or CSV, which serve as the building blocks for creating datasets [64]. However, while this manual approach is common, it has its limitations in scalability, traceability, and error management.

To overcome these challenges, a more effective and reproducible method involves using Application Programming Interfaces (APIs) [65]. APIs enable users to access database content in a structured and programmable way, facilitating the automation of complex queries, the retrieval of large-scale datasets, and the integration of data collection into computational workflows [66].

When APIs are unavailable, web scraping offers a practical but limited alternative [67]. Web scraping can be implemented by employing different tools, including BeautifulSoup, Selenium, and Scrapy [68, 69, 70]. Web scraping proves helpful in accessing older databases or pulling data from supplementary sources found in online articles, showcasing the adaptability of researchers in handling various challenges [71].

Another powerful yet often overlooked strategy is using natural language processing and text mining to extract valuable data from scientific literature, patents, or experimental reports [72]. Many relevant properties related to peptides, such as activity measurements, pharmacokinetics, or structural changes, are often presented in text rather than being available in databases [73]. Tools like PubTator, BioBERT, and ScispaCy can automatically identify named entities, specific terms found in the text, like peptide names or organisms, and extract relevant numerical values (like IC_{50} or half-life) alongside contextual information [74–76].

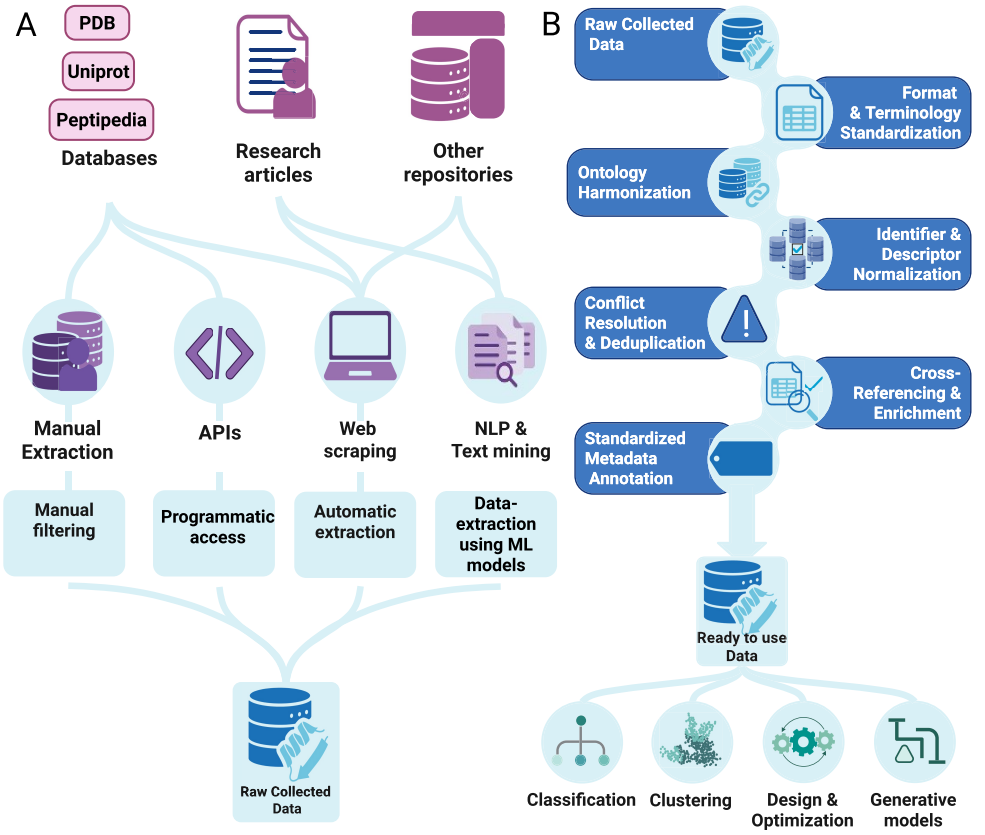

Figure 2.
Integrated workflow for therapeutic peptide dataset construction and application. (A) illustrates various data extraction strategies used to collect peptide-related data from structured databases, research articles, and other repositories. The extraction methods include manual curation, APIs, web scraping, and natural language processing (NLP) and text mining techniques. (B) shows the subsequent standardization and harmonization process, which ensures consistent formatting, ontology alignment, conflict resolution, and metadata annotation. The resulting ready-to-use data supports downstream applications such as peptide classification, clustering, design and optimization, and generative modeling.

The NLP-based method serves as a complement to common data mining-based strategies, enabling researchers to uncover novel peptides that appear only in original research articles as text or in tables and are not directly represented in traditional databases [73].

Alternatively, hybrid approaches that combine automated data extraction with expert curation are also relatively common, especially when dealing with noisy or inconsistent data sources [77]. In this approach, computational pipelines handle the initial data parsing, annotation, and filtering [78]. After this automated phase, domain experts review the results and make any necessary corrections or additions as needed [77]. These semi-automated curation processes are particularly valuable for creating high-confidence datasets with comprehensive and accurate coverage [79, 80].

Each data extraction strategy discussed presents a distinct trade-off between scalability, precision, and accessibility. Manual extraction from public databases like UniProt or PDB offers high interpretability and user control but suffers from limited

DOI: http://dx.doi.org/10.5772/intechopen.1013230

scalability, susceptibility to human error, and challenges in reproducibility. In contrast, APIs provide a structured and reproducible method for accessing large-scale, up–to-date datasets programmatically, making them ideal for integrative pipelines [65]. However, their availability is limited to well-maintained platforms, and they require a certain level of technical expertise [81]. Web scraping emerges as an alternative when APIs are missing, but it is inherently breakable, often sensitive to website updates, and raises ethical and legal concerns regarding data use [68]. NLP-based methods unlock valuable information embedded in scientific literature and patents—often absent from structured repositories—by enabling the automated extraction of entities and experimental data [82]. However, the performance of NLP-based approaches is highly dependent on model quality, training data, and the complexity of natural language contexts [83]. Hybrid approaches that combine automated parsing with expert curation provide a balance between efficiency and accuracy [84]. Selecting the appropriate strategy depends on the nature of the source, the scale of extraction, and the required level of data fidelity and interoperability [85].

5.2 Standardization and harmonization of extracted data

Once the data has been collected and extracted, the next step is standardization. This process aims to harmonize formats, terminologies, and data structures to generate coherent, interoperable datasets compatible with downstream statistical analyses and computational tools.

Standardization typically begins by aligning dataset fields with established ontologies and controlled vocabularies, such as the Gene Ontology (GO), ChEBI, and EDAM [86, 87]. Ontology mapping facilitates consistent functional annotation and categorical characterization of biological entities using widely recognized semantic frameworks [63].

Following ontology alignment, harmonization ensures coherence across multiple ontologies and data types [88]. This step often involves converting entries into standardized formats—such as CSV, TSV, JSON-LD, FASTA, or RDF—to support interoperability and facilitate the integration with bioinformatics pipelines and machine learning workflows [89]. These formats are widely supported by tools such as OpenRefine, Biopython, and pandas, enabling efficient parsing, transformation, and analysis of structured biological data [90, 91, 92].

Once the dataset is harmonized, further normalization steps are applied to unify identifiers, descriptors, and sequence formats. Amino acid sequences are typically encoded using IUPAC standards [93], while unique identifiers are cross-referenced against curated databases [32, 94, 95]. However, due to inconsistencies and incomplete coverage in existing databases, many sequences cannot be directly mapped. In such cases, alignment tools like BLAST are used to identify corresponding entries [96]. For datasets reporting physicochemical properties, normalization also involves expressing values in consistent units and formats [97]. Properties such as pKa, isoelectric point, and hydrophobicity index should be calculated using uniform algorithms to avoid discrepancies across datasets or experimental conditions.

Integrating heterogeneous data sources frequently reveals redundant or conflicting entries [98]. Addressing these inconsistencies requires deduplication protocols based on sequence hashing, similarity thresholds, or prioritization rules

informed by data provenance and validation status [99]. Curated records from peer-reviewed literature or regulatory repositories are generally preferred over predictive or inferred annotations [79].

Cross-referencing with additional resources is strongly recommended to enhance the interpretability and biological value of peptide datasets. For example, antimicrobial peptides can be linked to resistance mechanisms, pathogenic targets, or clinical trial information [100]. Supplementary data from resources such as PDB (structure), Reactome (biological pathways), or DrugBank (pharmacological data) enrich the dataset and expand its applicability in systems biology and translational research [101, 102].

The systematic annotation of metadata using standardized schemas such as MIAPE, MIABE, or EnzymeML is also a relevant point on the annotation and standardization of collected data [103, 104, 105]. Metadata should include details on peptide synthesis protocols, assay conditions, the organism of origin, and the therapeutic context. Such information is essential for ensuring reproducibility, enabling comparative analyses, and supporting the generalizability of predictive models [106].

In the context of therapeutic peptide discovery––where machine learning, structural modeling, and rational design approaches depend critically on data granularity and reliability––rigorous curation is not merely a best practice; it is a foundational requirement [78, 79, 106]. As datasets continue to expand in volume and complexity, and as FAIR-compliance becomes a prerequisite for reproducible science, these strategies will remain central to the development of interoperable and impactful peptide-based solutions in biotechnology and medicine [97].

5.3 Implementing FAIR principles in peptide data integration: Opportunities and challenges

The FAIR principles (Findability, Accessibility, Interoperability, and Reusability) are a set of guidelines to ensure the transparency and long-term usability of scientific data [107] (see **Figure 3**). FAIR principles support the development of structured and well-documented datasets. The FAIR principles are particularly relevant to therapeutic peptide applications, given the high heterogeneity of the data and the need for reproducibility [97].

Applying the FAIR principles to peptide datasets presents different challenges and complexities [108]. For findability, the data must be richly annotated with standardized metadata and indexed in searchable repositories using persistent identifiers [107]. Resources such as Peptipedia or DRAMP, which expose metadata and sequence data through indexed platforms, partially fulfill this requirement [38, 57]. Nevertheless, many peptide datasets remain isolated in supplementary files or poorly annotated repositories [79].

Accessibility requires that data be made available through standardized protocols, such as HTTPS or FTP, and ideally through APIs or SPARQL endpoints that enable automated querying [109]. While many modern databases offer this level of access, legacy datasets—especially those found in older literature—still pose significant barriers [28]. Additionally, licensing restrictions and the absence of clear data usage policies can limit the accessibility of curated peptide datasets, particularly those generated in industrial or clinical environments [110].

DOI: http://dx.doi.org/10.5772/intechopen.1013230

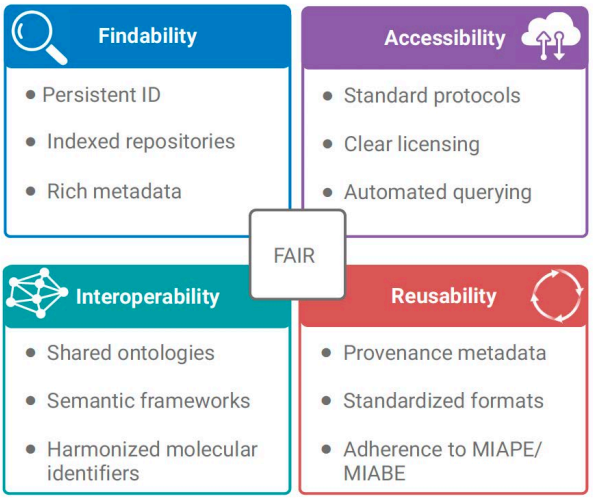

Figure 3.
Implementation of FAIR principles in peptide data integration. The four core components of the FAIR framework—Findability, Accessibility, Interoperability, and Reusability—are discussed in the context of peptide-related datasets. Findability emphasizes the use of persistent identifiers, indexed repositories, and rich metadata to ensure datasets can be efficiently located and referenced. Accessibility highlights the importance of standardized protocols, clear licensing, and automated querying to promote controlled yet seamless data access. Interoperability focuses on aligning datasets through shared ontologies, semantic frameworks, and harmonized molecular identifiers, enabling their integration across heterogeneous resources. Finally, Reusability underscores the role of provenance metadata, standardized formats, and compliance with community standards such as MIAPE or MIABE in facilitating reliable reuse across diverse computational pipelines and research contexts.

Ensuring interoperability requires aligning datasets with shared ontologies and semantic frameworks to enable consistent interpretation and integration across resources [111]. In the peptide domain, this may involve mapping sequences to UniProt identifiers, referencing biochemical terms through ChEBI or EDAM ontologies, and linking functional annotations to GO terms [107]. For example, integrating a peptide dataset from a pharmacological repository with a structural database such as the Protein Data Bank would require harmonizing molecular identifiers and functional annotations to ensure that entries referring to the same peptide are consistently recognized across both systems. Nevertheless, the domain-specific nature of peptide bioactivity and structural diversity often complicates ontology mapping, as many descriptors relevant to peptide engineering lack standardized terms across biomedical vocabularies [112].

Reusability is maybe the most impactful yet challenging aspect to achieve [113]. It involves providing well-described metadata, clear provenance, and standardized file formats to allow seamless integration into different computational pipelines [114]. For example, peptide datasets expressed in TSV or JSON-LD formats and annotated with MIAPE or MIABE descriptors are more likely to be adopted in external projects [115].

Despite the promise of FAIR integration, several challenges remain. First, achieving full compliance requires coordinated efforts between data generators, curators, and platform developers [108]. Manual curation remains critical for ensuring data quality, especially when integrating bioactivity data from literature, patents, or internal assays [116]. Also, standardization can introduce rigid constraints that may obscure context-specific annotations.

However, the benefits of applying FAIR principles to peptide datasets are significant. They promote data transparency, enabling researchers to trace annotations and decisions made during the curation process [117]. FAIR principles also facilitate collaboration and data sharing across institutions and disciplines, accelerating the discovery of peptides [118]. Besides, these principles enable the integration of peptide data into machine learning models, which rely heavily on standardized formats, annotated metadata, and consistent labeling to deliver meaningful predictions [119].

Several initiatives are advancing the *FAIRification* of peptide-related data. For example, integrating peptide datasets into BioSchemas or FAIRsharing registries enables their metadata to be discoverable by search engines and interoperable with other omics data sources [120, 121]. Tools like FAIR-Checker and OpenRefine extensions have also been adapted to assist in validating and converting peptide metadata into machine-readable formats [122].

6. Data quality and confidence in therapeutic peptide datasets: Addressing annotation errors, redundancy, and reliability

The quality and reliability of peptide datasets play a central role in the development of predictive models and computational strategies for therapeutic peptide discovery. Several challenges can compromise model performance and reproducibility. This section discusses key factors that affect data confidence. First, we discuss common issues in dataset generation, including misannotations and a lack of standard validation protocols. Next, we explore redundancy reduction strategies, from sequence-based clustering to advanced learning-based filters. Finally, we highlight the importance of building reliable negative datasets, a critical but often overlooked component in supervised and semi-supervised machine learning tasks. These subsections address essential aspects of data curation needed to ensure trustworthy and high-quality peptide datasets.

6.1 Quality of the therapeutic peptide datasets and common challenges in the dataset generation

The usability and success of computational models in therapeutic peptide research are directly linked to the quality and reliability of the employed data [123]. High-quality peptide datasets are critical for model training, benchmarking, and reproducibility [124]. However, ensuring trust and confidence in peptide data remains a considerable challenge due to several factors, including annotation errors, redundancy, missing or inconsistent metadata, and the variable provenance of source information [125, 126].

A main problem in peptide datasets is annotation error, which can occur at multiple levels [127]. For example, incorrect labeling of peptide activity can result from manual curation mistakes, outdated classification criteria, or inconsistent experimental conditions [128]. Similarly, structural annotations may differ depending on the computational tool or experimental method employed [129]. These inconsistencies can propagate through databases, especially in aggregated resources that pull data from multiple sources without strict harmonization protocols [130].

DOI: http://dx.doi.org/10.5772/intechopen.1013230

Redundancy is another frequent issue, particularly in datasets derived from public repositories or those obtained through literature mining [125]. Identical or highly similar peptide sequences may be present multiple times under different identifiers or experimental contexts [131]. While some redundancy can be beneficial—highlighting reproducibility or context-specific effects—uncontrolled duplication inflates dataset size, introduces sampling biases, and may lead to misleading performance evaluations in predictive modeling [132].

To assess and improve dataset reliability, several data quality indicators have been proposed and can be applied to therapeutic peptide datasets, including i) Completeness Score, ii) Annotation Consistency Index, iii) Redundancy Ratio, iv) Provenance Traceability Score, and v) Reproducibility Potential [133].

Despite the relevance of these metrics, few peptide databases provide them explicitly. As a result, users are often left to estimate data reliability based on indirect signals, such as citation counts, the reputation of the source database, or coverage in benchmark datasets [61, 134]. This lack of transparency can compromise confidence in datasets, particularly in high-stakes applications such as therapeutic design or toxicity prediction [23].

Another critical limitation is the lack of standardized validation protocols across datasets. Peptide annotations, particularly those related to functional properties, are often derived from heterogeneous experimental configurations [83]. The absence of standardized thresholds or reference conditions can limit comparative analyses and introduce noise into supervised learning tasks.

To mitigate these challenges and enhance trust in peptide data, several strategies have been proposed. First, implementing multi-layered validation—where data entries are corroborated through orthogonal sources (e.g., literature, experimental evidence, structural models)—can reduce reliance on single-point annotations [135]. Second, transparent versioning and change logs in peptide databases allow users to track updates and revisions, a critical factor when building reproducible pipelines [136]. Finally, community curation initiatives, where domain experts collaboratively validate and annotate data, offer a promising approach for improving annotation quality and confidence [137]. An integrated summary of the most frequent data quality issues, their corresponding indicators, and practical strategies to address them is presented in **Figure 4**.

6.2 Redundancy reduction: From sequence clustering to learning-based filters

As peptide datasets grow in size and complexity, primarily through the integration of data from heterogeneous sources, redundancy becomes a growing challenge [132]. Redundant entries, typically duplicated or highly similar sequences, can distort statistical distributions, introduce sampling bias, and compromise the generalizability of machine learning models [99]. The redundancy problem in supervised learning tasks is associated with repeated or near-identical entries. This input data may inflate performance metrics due to unintentional data leakage between training and test sets [138]. For these reasons, reducing redundancy is a critical preprocessing step in building reliable peptide datasets [139]. **Figure 5A** summarizes the current strategies to address redundancy reduction challenges.

One of the most common approaches to address this issue is sequence clustering based on homology thresholds [99]. Tools like CD-HIT and MMseqs2 group sequences based on predefined percentage identity (e.g., 90%, 80%, 70%), retaining

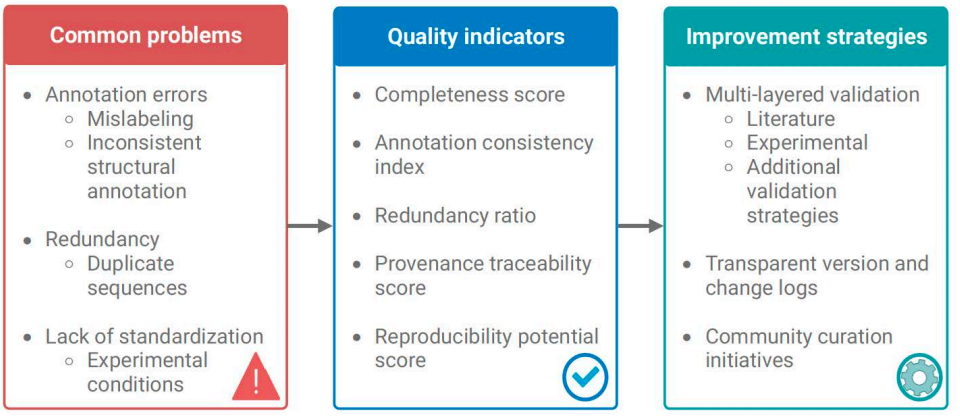

Figure 4.
Factors affecting data confidence and strategies for improvement in therapeutic peptide datasets. Three interconnected dimensions influencing the reliability of peptide datasets. Common problems include annotation errors—such as mislabeling of activity or inconsistent structural assignments—redundancy from duplicate or near-identical sequences, and lack of standardization in experimental conditions or metadata formats. Quality indicators, such as completeness score, annotation consistency index, redundancy ratio, provenance traceability score, and reproducibility potential, provide measurable criteria for assessing dataset reliability. Improvement strategies include multi-layered validation using orthogonal evidence (e.g., literature, experimental results, and structural corroboration), transparent versioning and change logs, and community-driven curation efforts, all aimed at enhancing trust and usability of therapeutic peptide data for computational and experimental applications.

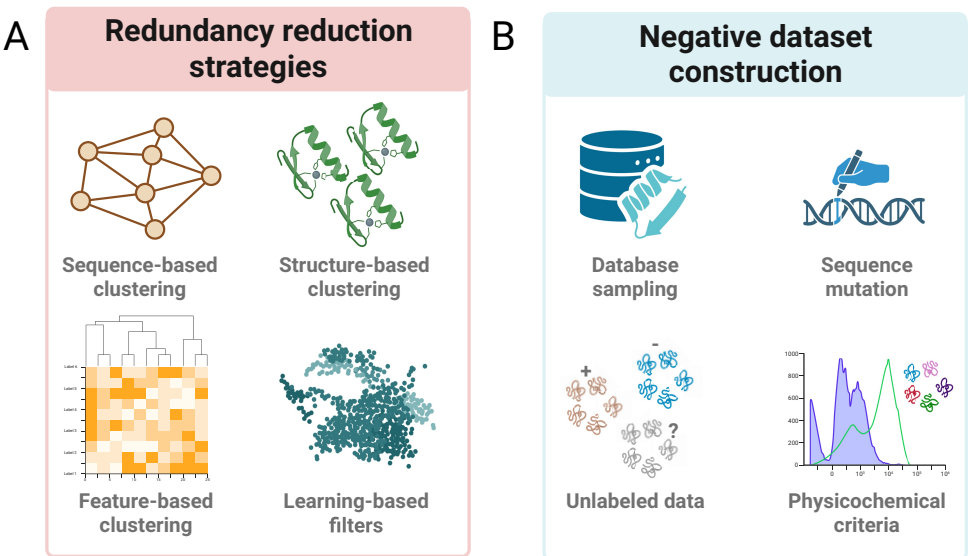

Figure 5.
Overview of redundancy reduction approaches and negative dataset construction strategies for therapeutic peptide machine learning workflows. (A) A summary of redundancy reduction methods, including sequence-based clustering, structure-based clustering, feature/embedding-based clustering, and learning-based filtering. (B) Strategies for generating negative datasets, ranging from heuristic and similarity filtering to the use of experimentally validated non-active sequences, synthetic negatives, and physicochemical or structural dissimilarity filters. A bidirectional arrow connecting both panels highlights their integration as a critical step for building robust and generalizable ML models in therapeutic peptide discovery.

DOI: http://dx.doi.org/10.5772/intechopen.1013230

a single representative from each cluster [140, 141]. These methods use heuristics or greedy algorithms and are valued for their speed and ease of use, making them a staple in large-scale curation pipelines [142]. However, this type of clustering has notable limitations: it relies solely on primary sequence similarity and may overlook important functional or structural relationships [143]. Moreover, overly strict identity thresholds can remove biologically meaningful variants, reducing the overall diversity of the dataset and potentially weakening downstream analyses.

To address these limitations, structure-based clustering approaches have been proposed [144]. These methods take into account higher-order features such as predicted secondary structures, solvent accessibility, or 3D conformations, often using tools like AlphaFold2 or ESMFold [145, 146]. Clustering based on structural fingerprints or spatial embeddings can provide a more biologically meaningful grouping, particularly when a peptide's activity depends more on its tridimensional configuration than on its sequence [147]. Nevertheless, this approach is computationally expensive and heavily dependent on the accuracy of structural predictions [148]. For short or disordered peptides, the confidence of these predictions tends to be low, which introduces further uncertainty [149].

A more flexible approach involves feature-based clustering, where peptides are encoded using physicochemical descriptors, sequence-derived features (e.g., hydrophobicity, charge, or molecular weight), or learned representations (embeddings) from protein language models such as ProtT5 or ESM [150, 151]. Embeddings offer a richer, multidimensional view of the data and can be clustered using unsupervised learning algorithms like k-means, DBSCAN, or hierarchical clustering [152, 153, 154]. This approach allows the grouping of peptides based on latent similarity patterns that may not be apparent at the sequence level [144]. Moreover, it supports modular integration of multiple data types—such as activity labels, structural motifs, and functional annotations—into the clustering process [155]. However, feature-based clustering is sensitive to the choice of representation, distance metric, and clustering algorithm, which can lead to variable outcomes and requires careful validation [156].

More recently, learning-based filtering strategies have emerged. Instead of treating redundancy as a static similarity problem, these approaches consider it a dynamic decision task [157]. In this approach, deep learning models or contrastive learning methods are used to detect and filter out redundant or low-information sequences based on their contribution to model generalization or prediction confidence. For instance, peptides that produce indistinguishable latent embeddings or fail to alter decision boundaries in classifier models may be flagged as redundant [158]. These strategies are especially useful in high-dimensional and multi-label contexts, where redundancy may not be obvious at the raw sequence level [159]. However, the main disadvantage is that they can be computationally intensive and may lack interpretability unless supported by robust exploratory tools [160].

Across all these methods, there is a clear trade-off between interpretability, scalability, and representational depth. Sequence identity clustering is fast and intuitive but exhibits problems in its approximation of functional similarity [143, 148]. Structure-based methods provide biological insights but are limited by their prediction accuracy and resource requirements [144]. Feature and embedding-based clustering strike a balance between complexity and flexibility, but require rigorous validation [161]. Learning-based filters introduce adaptive, task-specific redundancy

reduction, but may lack reproducibility and interpretability if not supported by robust explanatory mechanisms [162].

In practice, hybrid strategies often yield the most reliable results. For instance, an initial round of homology clustering can be used to reduce obvious duplication, followed by feature-based or learning-guided refinement to address latent redundancies. Additionally, filtering steps can incorporate domain-specific constraints, such as preserving representatives across taxonomic groups or therapeutic categories, that can serve to maintain functional variability while still ensuring data quality [163].

6.3 An additional task toward building datasets for machine learning applications: working with negative examples

High-quality negative datasets represent a critical, yet often overlooked, component in the development of robust data-driven models for bioactive peptide prediction. While the importance of well-annotated positive examples has been widely acknowledged, the scarcity of reliable negative counterparts—particularly for non-antimicrobial peptides—remains a significant limitation [164]. This challenge is compounded by the fact that few efforts have focused on the systematic collection and annotation of peptides lacking antimicrobial activity [37, 61] (see **Figure 5B**).

In other domains of molecular biology, such as microRNA classification or protein–protein interaction prediction, strategies like one-class classification and positive-unlabeled (PU) learning have been employed to address similar challenges [61, 165, 166]. The former relies exclusively on positive data for training, while the latter utilizes unlabeled data in place of explicitly negative examples. However, until recently, such approaches were rarely applied to the classification of antimicrobial peptides (AMPs), where the most prevalent strategy for assembling negative datasets has involved sampling from general-purpose peptide databases like UniProt [167, 168]. At its core, this sampling-based approach typically involves two main steps. First, extracting all peptide sequences within a predefined length range, then removing sequences annotated as antimicrobial or with related functional terms.

To ensure the resulting dataset's validity, further refinement is essential. First, duplicate sequences must be removed by comparing the negative dataset to the positive one. Second, redundancy should be reduced, most commonly using CD-HIT at a 40% sequence identity threshold [169, 170]. Third, class balance must be maintained, generally by downsampling the larger class—typically the negative set—to match the size of the smaller one. In some cases, sequences containing non-canonical or chemically modified residues are excluded, depending on the research objective.

While this heuristic method yields workable negative datasets, it is not without important challenges. The absence of standardized, universally accepted exclusion criteria introduces uncertainty into the model's generalizability, especially when applied to novel datasets [134, 164]. Moreover, the lack of benchmark AMP datasets further complicates the assessment of model performance and reproducibility.

The reliability of annotation-based filtering is also limited by the incompleteness and potential ambiguity of database annotations, which may lead to false negatives or introduce bias against well-annotated taxonomic or functional groups [171].

Alternative strategies have emerged to address these limitations. Functional descriptors such as minimum inhibitory concentration or half-maximal inhibitory

DOI: http://dx.doi.org/10.5772/intechopen.1013230

concentration can be used to more precisely identify inactive peptides [172, 173], particularly when species- or strain-specific, experimentally validated values are available. Such information can be found in domain-specific databases like APD, DBAASP, DRAMP, CAMP, dbAMP, AmPep, AVPdb, BioPep-UWM, ParaPep, and PlantPepDB [62, 64, 174]. When the objective is to predict a specific bioactivity (e.g., antiparasitic), negative datasets may be constructed using peptides with unrelated but known activities (e.g., antiviral or anticancer), assuming functional exclusivity in the learning task [175, 176].

Another refinement strategy involves sequence similarity filtering using tools such as BLAST, where sequences closely related to known AMPs (based on an e-value threshold) are removed from the negative dataset [177]. Specific databases already include experimentally validated non-AMP sequences, which can serve as a valuable source of high-confidence negative examples [37].

More recent strategies have also been explored. These include training on unlabeled data through contrastive learning or other semi-supervised deep learning techniques [165, 166, 178, 179], as well as generating synthetic negative samples either by random sequence generation or by introducing mutations into known AMPs [178, 180, 181]. Additionally, some studies propose selecting negative samples based on physicochemical or structural dissimilarity to known AMPs, using features such as charge, hydrophobicity, or secondary structure propensity [182].

To mitigate biases and improve robustness, one promising approach involves building multiple negative datasets using different strategies and training independent models on each. These models can then be evaluated on an external benchmark dataset to assess generalizability and consensus predictions [61, 183].

Given the absence of a universally accepted standard for negative dataset construction, researchers must make careful decisions based on the quality of available data, the computational resources at hand, and the specific goals of the predictive model.

7. Data integration for machine learning applications: Traditional data-driven pipelines and demonstrative examples

The integration of therapeutic peptide data into machine learning frameworks depends on well-structured pipelines that ensure both methodological rigor and biological relevance. This section presents a comprehensive overview of traditional data-driven strategies commonly employed to develop predictive models, outlining each stage—from data collection and preprocessing to model training, evaluation, and deployment. We also discuss strategies tailored to low-data (Low-N) scenarios, highlighting the role of transfer learning in enhancing model performance when data is scarce.

7.1 Traditional strategies to develop predictive models using machine learning strategies

Traditional data-driven machine learning approaches applied to the study of therapeutic peptides can be grouped into four main categories: classification, regression, clustering, and generative modeling [1]. Regardless of the specific task, these workflows share a series of core stages, including dataset collection and

curation, numerical encoding of peptide sequences, model training using suitable algorithms, performance evaluation with task-appropriate metrics, and the application of the trained model to predict, cluster, or generate peptide sequences with desired properties [124].

The process begins with the collection and cleaning of peptide sequences [184]. As discussed previously, this involves removing duplicates, incomplete entries, or inconsistencies to ensure the quality and reliability of the data [185]. For supervised learning, as in classification and regression, the dataset must include the relevant labels or target values, while in unsupervised clustering and generative modeling, the emphasis lies in capturing the diversity and representativeness of the peptide space [186].

Once curated, peptide sequences are transformed into numerical representations suitable for computational processing. Encoding strategies range from traditional approaches, such as one-hot encoding and physicochemical descriptors, to more advanced embeddings derived from pre-trained protein language models. In clustering and generative modeling tasks, embeddings or latent representations are especially important, as they capture structural, functional, or compositional patterns within the sequence space [187].

The encoded dataset is typically split into training, validation, and testing subsets, often in ratios such as 70:20:10 or 80:20 [184]. In low-data regimes, cross-validation techniques are employed to mitigate overfitting and enhance generalizability. While supervised tasks use these partitions directly for model assessment, clustering models rely on internal or external validation metrics, and generative models are frequently trained on the complete dataset, reserving a portion for evaluating novelty, diversity, and validity of the generated sequences [188].

Model training depends on the learning paradigm. Classification models, used for tasks such as activity prediction or functional annotation, may rely on ensemble algorithms, kernel-based methods, or deep learning architectures, while regression models, designed to predict continuous properties such as solubility or binding affinity, often employ linear regression variants, boosting-based regressors, or deep neural networks adapted for regression tasks [189]. Clustering approaches group peptides according to similarity in their feature space, applying algorithms such as centroid-based, density-based, or hierarchical methods, often in combination with dimensionality reduction techniques to improve interpretability. Generative modeling focuses on learning the distribution of peptide sequences to produce new candidates, employing architectures such as variational autoencoders, generative adversarial networks, or autoregressive models [190].

Performance evaluation also varies with the task. Classification models are typically assessed through accuracy, precision, recall, F1-score, Matthews correlation coefficient, and curve-based metrics such as ROC-AUC or PR-AUC [184]. Regression models are evaluated using correlation measures like Pearson or Spearman coefficients, error-based metrics such as mean absolute error and root mean squared error, and determination coefficients [191]. Clustering evaluation relies on internal metrics such as the Silhouette score, Davies–Bouldin index, or Calinski–Harabasz index, and when reference labels are available, on external metrics such as the Adjusted Rand Index or Normalized Mutual Information [192]. Generative models are judged by their validity, uniqueness, and novelty, alongside domain-specific evaluations, which may include the predicted biological activity of generated sequences or the similarity of their property distributions to reference datasets.

DOI: http://dx.doi.org/10.5772/intechopen.1013230

Hyperparameter optimization is common across all paradigms and can be performed through exhaustive search, evolutionary strategies, or Bayesian optimization [124]. In generative modeling, additional tuning often balances the exploration of sequence space with the targeted optimization of specific properties, sometimes integrating reinforcement learning strategies. The independent test set, or an equivalent held-out evaluation, provides the final performance benchmark. In clustering and generative tasks, this may be complemented by downstream task performance or case studies to illustrate the model's practical relevance. Once validated, models can be deployed in a wide range of applications, including peptide screening, drug discovery, rational design of therapeutic candidates, unsupervised exploration of peptide diversity, and *de novo* generation of optimized sequences [184].

7.2 Working under low-N regimes, transfer learning strategies to develop functional classification models

Functional classification of therapeutic peptides presents significant challenges due to the need for high-quality annotations, diverse sequence representations, and robust machine learning frameworks. However, most available datasets are small, noisy, and imbalanced. These are classic examples of low-N scenarios, where only a limited number of labeled examples are available for training [193]. This scarcity of data undermines the assumptions of traditional supervised learning, which typically relies on large, well-balanced datasets to achieve reliable generalization [194].

In this context, transfer learning and other data-efficient strategies have emerged as powerful solutions. Among the most promising tools are protein language models (PLMs), which are trained on massive collections of unlabeled protein sequences. These models generate embeddings that capture biochemical, structural, and evolutionary information [195], providing a strong foundation for downstream classification tasks when labeled data is limited [196].

PLMs can be used as fixed feature extractors or fine-tuned on domain-specific datasets. In both cases, they offer substantial improvements in predictive performance under data-constrained settings [197]. Their ability to transfer learned biological knowledge from large corpora to specific peptide tasks makes them particularly valuable in therapeutic discovery pipelines.

Positive-unlabeled (PU) learning is another critical approach, especially relevant when only positive samples are available and the negative class is undefined or unlabeled [198]. PU learning allows classifiers to be trained in the presence of incomplete labels while maintaining strong generalization capabilities [199].

Complementing these methods, contrastive learning techniques help refine the quality of representations. These approaches train models to maximize the similarity between related peptides and distinguish them from unrelated ones. As a result, contrastive learning improves the discriminative power of the model even when supervision is weak or noisy [200, 201].

Together, these strategies form a flexible and effective framework for functional peptide classification in settings where labeled data is scarce. By combining pretrained embeddings from PLMs with weak supervision and representation learning objectives, researchers can build predictive models that are accurate, interpretable, and scalable. These advances not only address the limitations of low-N datasets but also accelerate the identification and prioritization of peptides with therapeutic potential.

8. Open problems, limitations, and perspectives for peptide data integration

Despite the increasing availability of peptide-related data and the expansion of computational tools for their analysis, significant challenges persist that limit the full potential of data-driven strategies in therapeutic peptide research. One of the most critical issues is the inconsistency and heterogeneity in how peptide data are reported, annotated, and structured across databases. These discrepancies emerge from the diverse origins of the data and manifest as variable nomenclature, inconsistent functional labels, and missing metadata. As a result, dataset integration becomes a non-trivial task, often requiring extensive manual curation and harmonization efforts that can introduce subjectivity and further inconsistencies [202].

Annotation errors and uncontrolled redundancy are additional barriers that undermine confidence in peptide datasets. Mislabeling of functional activities, discrepancies in structural annotations due to conflicting experimental evidence or prediction tools, and duplication of sequences across repositories compromise data quality. These issues can propagate undetected through aggregated databases, inflating dataset size and introducing hidden biases. Particularly in supervised learning contexts, redundancy can lead to data leakage between training and testing partitions, producing overoptimistic performance estimates and models that fail to generalize. The lack of standardized criteria for quality control and validation exacerbates this situation, leaving researchers to rely on ad hoc heuristics or database reputation rather than objective, traceable quality indicators [203].

These data quality problems directly impact the development of predictive models. Many of the challenges encountered during training, including model overfitting, poor generalizability, and lack of reproducibility, can often be traced back to issues in dataset construction. Inaccurate or noisy annotations weaken the signal that machine learning models are designed to capture, while poor class balance and ambiguous negative sets limit the interpretability of evaluation metrics. Although advanced strategies such as transfer learning and contrastive learning offer partial relief in low-data regimes, their effectiveness still hinges on the availability of reliable and representative training data [19].

To address these problems and strengthen the foundations of data-driven peptide research, several directions for future work are urgently needed. First, the development of a standardized markup language for peptides that enables structured, interoperable, and machine-readable annotation of peptide data, including sequence, structure, activity, context, and experimental provenance. Such a format would greatly facilitate integration across databases, promote transparency, and streamline the creation of curated datasets for machine learning applications.

Second, aligning peptide datasets with FAIR principles (Findable, Accessible, Interoperable, Reusable) should become a priority across the community. This includes not only technical implementations (e.g., metadata schemas, persistent identifiers, APIs) but also community-endorsed standards for data submission, versioning, and validation. Building centralized peptide registries with traceable annotations and transparent update logs would further enhance reproducibility and trust [204].

DOI: http://dx.doi.org/10.5772/intechopen.1013230

Third, the development of a formalized guideline of best practices for dataset construction and integration in peptide-focused machine learning studies. This guideline should encompass recommendations for data preprocessing, annotation validation, redundancy reduction, negative dataset generation, and evaluation protocols. Ideally, it would be supported by a consortium of domain experts, database curators, and computational method developers, ensuring that it reflects both practical needs and technical feasibility [205].

Promoting interdisciplinary collaborations will be key to addressing these challenges. The construction of high-quality datasets requires joint efforts from experimentalists, computational biologists, and data scientists. Initiatives that promote collaborative data curation, community-driven benchmarking, and open sharing of intermediate datasets will accelerate progress and contribute to the development of more robust, generalizable, and interpretable predictive models for therapeutic peptides [206].

9. Conclusions

The rapid integration of machine learning into therapeutic peptide research has emphasized the central importance of high-quality, well-preprocessed, and interoperable datasets. As demonstrated throughout this chapter, the reliability, reproducibility, and generalizability of predictive and generative models are determined not only by algorithmic innovation but also by the integrity and representational depth of the data that form their foundation. Rigorous preprocessing, including systematic extraction, standardization, and redundancy reduction, is therefore essential for building trustworthy computational pipelines.

Adopting the FAIR principles offers a practical and widely applicable framework for addressing persistent challenges such as inconsistent annotation, heterogeneous formats, and incomplete metadata. FAIR-compliant datasets facilitate integration across diverse sources, enable transparent benchmarking, and support reproducible workflows. In the context of therapeutic peptides, where data heterogeneity is the norm, such practices are indispensable for scaling ML applications with confidence.

Realizing this vision will require community-wide commitment to shared standards. Establishing unified protocols for preprocessing, annotation validation, and redundancy reduction—together with standardized reporting formats and a machine-readable peptide markup language—will ensure interoperability, transparency, and long-term usability. Centralized, versioned repositories with traceable provenance and open programmatic access will further enable collaborative curation and integrative modeling.

Investing in these infrastructures today will define the reliability, interpretability, and translational impact of tomorrow's ML-driven peptide discovery. By combining FAIR-compliant data practices with rigorous preprocessing pipelines, the field can advance toward robust, transferable, and transparent computational models, accelerating the design of safe and effective therapeutic peptides.

Acknowledgments

DM-O, DF, and NS-G acknowledge ANID for the project "SUBVENCIÒN A INSTALACIÒN EN LA ACADEMIA CONVOCATORIA Año 2022", Folio 85220004.

DM-O, NM-A, and SE acknowledge ANID for funding through the "FONDECYT Iniciación project 11250295". DM-O and AD gratefully acknowledge support from the Centre for Biotechnology and Bioengineering - CeBiB (PIA project FB0001, AFB240001, ANID, Chile). AD acknowledges ANID for funding through the "FONDECYT Iniciación project 11230208".

Conflict of interest

The authors declare no conflict of interest.

Author details

David Medina-Ortiz[1,2*], Sebastián Escobedo[1], Norma Murillo-Acevedo[1], Nicole Soto-García[1], Diego Fernández-Villegas[1], Diego Sandoval[1] and Anamaría Daza[2*]

1 Departamento de Ingeniería en Computación, Universidad de Magallanes, Punta Arenas, Chile

2 Centre for Biotechnology and Bioengineering, CeBiB, Universidad de Chile, Santiago, Chile

*Address all correspondence to: david.medina@umag.cl; ana.sanchez@ing.uchile.cl

References

[1] Goles M, Daza A, Cabas-Mora G, Sarmiento-Varón L, Sepúlveda-Yañez J, Anvari-Kazemabad H, Davari MD, Uribe-Paredes R, Olivera-Nappa Á, Navarrete MA, et al. Peptide-based drug discovery through artificial intelligence: Towards an autonomous design of therapeutic peptides. Briefings in Bioinformatics. 2024;**25**(4): DOI: 10.1093/bib/bbae275

[2] Wang L, Wang N, Zhang W, Cheng X, Yan Z, Shao G, Wang X, Wang R, Caiyun F. Therapeutic peptides: Current applications and future directions. Signal Transduction and Targeted Therapy. 2022;**7**(1):48

[3] Jimeng S, Song Y, Zhu Z, Huang X, Fan J, Qiao J, Mao F. Cell–cell communication: New insights and clinical implications. Signal Transduction and Targeted Therapy. 2024;**9**(1):1–52

[4] Rossino G, Marchese E, Galli G, Verde F, Finizio M, Serra M, Linciano P, Collina S. Peptides as therapeutic agents: Challenges and opportunities in the green transition era. Molecules. 2023;**28**(20):7165

[5] Wang L, Wang N, Zhang W, Cheng X, Yan Z, Shao G, Wang X, Wang R, Caiyun F. Therapeutic peptides: Current applications and future directions. Signal Transduction and Targeted Therapy. 2022;**7**(2):1–27

[6] Xiao W, Jiang W, Chen Z, Huang Y, Mao J, Zheng W, Yonghe H, Shi J. Advance in peptide-based drug development: Delivery platforms, therapeutics and vaccines. Signal Transduction and Targeted Therapy. 2025;**10**(1):1–56

[7] Rosson E, Lux F, David L, Godfrin Y, Tillement O, Thomas E. Focus on therapeutic peptides and their delivery. International Journal of Pharmaceutics. 2025;**675**(4):125555

[8] Zheng B, Wang X, Guo M, Tzeng CM. Therapeutic peptides: Recent advances in discovery, synthesis, and clinical translation. International Journal of Molecular Sciences. 2025;**26**(11):5131

[9] Öngoren B, Kara A, Serrano DR, Lalatsa A. Novel enabling strategies for oral peptide delivery. International Journal of Pharmaceutics. 2025;**681**(8):125888

[10] Baral KC, Choi KY. Barriers and strategies for oral peptide and protein therapeutics delivery: Update on clinical advances. Pharmaceutics. 2025;**17**(4):397

[11] Singh S, Gupta H, Sharma P, Sahi S. Advances in artificial intelligence (ai)-assisted approaches in drug screening. Artificial Intelligence Chemistry. 2024;**2**(1):100039

[12] Hashemi S, Vosough P, Taghizadeh S, Savardashtaki A. Therapeutic peptide development revolutionized: Harnessing the power of artificial intelligence for drug discovery. Heliyon. 2024;**10**(22):e40265

[13] Rezaee K, Eslami H. Bridging machine learning and peptide design for cancer treatment: A comprehensive review. Artificial Intelligence Review. 2025;**58**(5). DOI: 10.1007/s10462-025-11148-3

[14] Guan J, Xie P, Meng D, Yao L, Dan Y, Chiang YC, Lee TY, Wang J.

Toxipep: Peptide toxicity prediction via fusion of context-aware representation and atomic-level graph. Computational and Structural Biotechnology Journal. 2025;**27**(1):2347–2358

[15] Gupta S, Kapoor P, Chaudhary K, Gautam A, Kumar R. Open Source Drug Discovery Consortium, and Gajendra PS Raghava. In silico approach for predicting toxicity of peptides and proteins. PloS One. 2013;**8**(9):e73957

[16] Ebrahimikondori H, Sutherland D, Yanai A, Richter A, Salehi A, Chenkai L, Coombe L, Kotkoff M, Warren RL, Birol I. Structure-aware deep learning model for peptide toxicity prediction. Protein Science. 2024;**33**(7):e5076

[17] Wei L, Xiucai Y, Sakurai T, Zengchao M, Wei L. Toxibtl: Prediction of peptide toxicity based on information bottleneck and transfer learning. Bioinformatics. 2022;**38**(6):1514–1524

[18] Wei L, Xiucai Y, Xue Y, Sakurai T, Wei L. Atse: A peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. Briefings in Bioinformatics. 2021;**22**(5):bbab041

[19] Badrinarayanan S, Chakradhar Guntuboina PM, Farimani A. Multi-peptide: Multimodality leveraged language-graph learning of peptide properties. arXiv, 2024

[20] Robles-Loaiza AA, Pinos-Tamayo EA, Mendes B, Ortega-Pila JA, Proaño-Bolaños C, Plisson F, Teixeira C, Gomes P, Almeida JR. Traditional and computational screening of non-toxic peptides and approaches to improving selectivity. Pharmaceuticals. 2022;**15**(3):323

[21] Niu Z, Xiao X, Wenfan W, Cai Q, Jiang Y, Jin W, Wang M, Yang G, Kong L, Jin X, Yang G, Chen H. Pharmabench: Enhancing admet benchmarks with large language models. Scientific Data. 2024;**11**(1):1–15

[22] Venkataraman M, Gopi C, Rao JK, Madavareddi S, Rao M. Leveraging machine learning models in evaluating admet properties for drug discovery and development. ADMET & DMPK. 2025;**13**(3):2772

[23] Wang JH, Sung TY. Toxteller: Predicting peptide toxicity using four different machine learning approaches. ACS Omega. 2024;**9**(29):32116–32123

[24] Hesamzadeh P, Seif A, Mahmoudzadeh K, Koli MG, Mostafazadeh A, Nayeri K, Mirjafary Z, Saeidian H. De novo antioxidant peptide design via machine learning and dft studies. Scientific Reports. 2024;**14**(3):1–18

[25] Jin S, Zeng Z, Xiong X, Huang B, Tang L, Wang H, Ma X, Xiaochun T, Shao G, Huang X, Lin F. Ampgen: An evolutionary information-reserved and diffusion-driven generative model for de novo design of antimicrobial peptides. Communications Biology. 2025;**8**(1):1–14

[26] Zhang M, Zhou J, Wang X, Wang X, Fang G. Deepbp: Ensemble deep learning strategy for bioactive peptide prediction. BMC Bioinformatics. 2024;**25**(1). DOI: 10.1186/s12859-024-05974-5

[27] Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, Lin Z, Verkuil R, Tran VQ, Deaton J, Wiggert M, Badkundri R, Shafkat I, Gong J, Derry A, Molina RS, Thomas N, Khan Y, Mishra C, Kim C, Bartie LJ, Nemeth M, Hsu PD, Sercu T,

DOI: http://dx.doi.org/10.5772/intechopen.1013230

Candido S, Rives A. Simulating 500 million years of evolution with a language model. bioRxiv.600583, **7** 2024

[28] Quiroz C, Saavedra YB, Armijo-Galdames B, Amado-Hinojosa J, Olivera-Nappa Á, Sanchez-Daza A, Medina-Ortiz D. Peptipedia: A user-friendly web application and a comprehensive database for peptide research supported by machine learning approach. Database. 2021;**2021**:baab055

[29] Zhou X, Liu G, Cao S, Lv J.Deep learning for antimicrobial peptides: Computational models and databases. Journal of Chemical Information and Modeling. 2025;**65**(4):1708–1717

[30] Hashemi S, Vosough P, Taghizadeh S, Savardashtaki A. Therapeutic peptide development revolutionized: Harnessing the power of artificial intelligence for drug discovery. Heliyon. 2024;**10**(22):e40265

[31] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Research. 2000;**28**(1):235–242

[32] The UniProt Consortium. Uniprot: The universal protein knowledgebase in 2025. Nucleic Acids Research. 2025, **53** (D1), D609–D617

[33] Porto WF, Pires AS, Franco OL. Computational tools for exploring sequence databases as a resource for antimicrobial peptides. Biotechnology Advances. 2017;**35**(3):337–349

[34] Shtatland T, Guettler D, Kossodo M, Pivovarov M, Weissleder R. Pepbank-a database of peptides based on sequence text mining and public peptide data sources. BMC Bioinformatics. 2007;**8**:1–10

[35] Wang J, Yin T, Xiao X, Dan H, Xue Z, Jiang X, Wang Y. Strapep: A structure database of bioactive peptides. Database. 2018;**2018**:bay038

[36] Jhong J-H, Yao L, Pang Y, Zhongyan L, Chung C-R, Wang R, Shangfu L, Wenshuo L, Luo M, Renfei M, et al. dbamp 2.0: Updated resource for antimicrobial peptides with an enhanced scanning method for genomic and proteomic data. Nucleic Acids Research 2022;**50**:D460–D470

[37] Pirtskhalava M, Amstrong AA, Grigolava M, Chubinidze M, Alimbarashvili E, Vishnepolsky B, Gabrielian A, Rosenthal A, Hurt DE, Tartakovsky M. Dbaasp v3: Database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. Nucleic Acids Research. 2021;**49**(D1): D288–D297

[38] Shi G, Kang X, Dong F, Liu Y, Zhu N, Yuxuan H, Hanmei X, Lao X, Zheng H. Dramp 3.0: An enhanced comprehensive data repository of antimicrobial peptides. Nucleic Acids Research. 2022;**50**(D1):D488–D496

[39] Waghu FH, Barai RS, Gurung P, Idicula-Thomas S. Campr3: A database on sequences, structures and signatures of antimicrobial peptides. Nucleic Acids Research. 2016;**44**(D1):D1094–D1097

[40] Guizi Y, Hongyu W, Huang J, Wang W, Kuikui G, Guodong L, Zhong J, Huang Q. Lamp2: A major update of the database linking antimicrobial peptides. Database: The Journal of Biological Databases and Curation. 2020;**2020**:baaa061

[41] Gómez EA, Giraldo P, Orduz S. Inverpep: A database of invertebrate antimicrobial peptides. Journal of

Global Antimicrobial Resistance. 2017;**8**:13–17

[42] Singh S, Chaudhary K, Dhanda SK, Bhalla S, Usmani SS, Gautam A, Tuknait A, Agrawal P, Mathur D, Raghava GPS. Satpdb: A database of structurally annotated therapeutic peptides. Nucleic Acids Research. 2016;**44**(D1): D1119–D1126

[43] Wang G, Xia L, Wang Z. Apd3: The antimicrobial peptide database as a tool for research and education. Nucleic Acids Research. 2016;**44**(D1):D1087–D1093

[44] Qureshi A, Thakur N, Tandon H, Kumar M. Avpdb: A database of experimentally validated antiviral peptides targeting medically important viruses. Nucleic Acids Research. 2014;**42**(D1):D1147–D1153

[45] Tyagi A, Tuknait A, Anand P, Gupta S, Sharma M, Mathur D, Joshi A, Singh S, Gautam A, Raghava GPS. Cancerppd: A database of anticancer peptides and proteins. Nucleic Acids Research. 2015;**43**(D1):D837–D843

[46] Gautam A, Chaudhary K, Singh S, Joshi A, Anand P, Tuknait A, Mathur D, Varshney GC, Raghava GPS. Hemolytik: A database of experimentally determined hemolytic and non-hemolytic peptides. Nucleic Acids Research. 2014;**42**(D1):D444–D449

[47] Kumar V, Patiyal S, Kumar R, Sahai S, Kaur D, Lathwal A, Raghava GPS. B3pdb: An archive of blood–brain barrier-penetrating peptides. Brain Structure and Function. 2021;**226**: 2489–2495

[48] Wynendaele E, Bronselaer A, Nielandt J, D'Hondt M, Stalmans S, Bracke N, Verbeke F, De Wiele CV, De Tré G, De Spiegeleer B. Quorumpeps database: Chemical space, microbial origin and functionality of quorum sensing peptides. Nucleic Acids Research. 2013;**41**(D1):D655–D659

[49] Ramaprasad ASE, Singh S, S RGP, Venkatesan S. Antiangiopred: A server for prediction of anti-angiogenic peptides. PloS One. 2015;**10**(9): e0136990

[50] Roy S, Teron R. Biodadpep: A bioinformatics database for anti diabetic peptides. Bioinformation. 2019;**15** (11):780

[51] Kumar R, Chaudhary K, Sharma M, Nagpal G, Chauhan JS, Singh S, Gautam A, Raghava GPS. Ahtpdb: A comprehensive platform for analysis and presentation of antihypertensive peptides. Nucleic Acids Research. 2015;**43**(D1):D956–D962

[52] Mathur D, Prakash S, Anand P, Kaur H, Agrawal P, Mehta A, Kumar R, Singh S, Raghava GPS. Peplife: A repository of the half-life of peptides. Scientific Reports. 2016;**6**(1):36617

[53] Mathur D, Singh S, Mehta A, Agrawal P, Raghava GPS. In silico approaches for predicting the half-life of natural and modified peptides in blood. PloS One. 2018;**13**(6):e0196829

[54] D'Aloisio V, Dognini P, Hutcheon GA, Coxon CR. Peptherdia: Database and structural composition analysis of approved peptide therapeutics and diagnostics. Drug Discovery Today. 2021;**26**(6):1409–1419

[55] Usmani SS, Bedi G, Samuel JS, Singh S, Kalra S, Kumar P, Ahuja AA, Sharma M, Gautam A, Raghava GPS. Thpdb: Database of fda-approved peptide and protein therapeutics. PloS One. 2017;**12**(7):e0181748

[56] Barman P, Joshi S, Sharma S, Preet S, Sharma S, Saini A. Strategic approaches to improvise peptide drugs as next generation therapeutics. International Journal of Peptide Research and Therapeutics. 2023;**29**(4):61

[57] Cabas-Mora G, Daza A, Soto-Garca N, Garrido V, Alvarez D, Navarrete M, Sarmiento-Varón L, Sepúlveda Yañez JH, Davari MD, Cadet F, et al. Peptipedia v2. 0: A peptide sequence database and user-friendly web platform. a major update. Database. 2024;**2024**:baae113

[58] Zhang D-E, Tong H, Shi T, Huang K, Peng A. Trends in the research and development of peptide drug conjugates: Artificial intelligence aided design. Frontiers in Pharmacology. 2025;**16**:1553853

[59] Perez-Riverol Y, Bittremieux W, Noble WS, Martens L, Bilbao A, Lazear MR, Grüning B, Katz DS, MacCoss MJ, Dai C, Eng JK, Bouwmeester R, Shortreed MR, Audain E, Sachsenberg T, Van Goey J, Wallmann G, Wen B, Käll L, Fondrie WE. Open-source and fair research software for proteomics. Journal of Proteome Research. 2025;**24**(5):2222–2234

[60] Lauterbach S, Dienhart H, Range J, Malzacher S, Spoering J-D, Rother D, Pinto MF, Martins P, Lagerman CE, Bommarius AS, et al. Enzymeml: Seamless data flow and modeling of enzymatic data. Nature Methods 2023;**20**:400–402

[61] Sidorczuk K, Gagat P, Pietluch F, Kała J, Rafacz D, Bakała L, Słowik J, Kolenda R, Rödiger S, Legana CHWF, Cooke IR, Mackiewicz P, Burdukiewicz M. Benchmarks in antimicrobial peptide prediction are biased due to the selection of negative data. Briefings in Bioinformatics. 2022;**23**(5):bbac343

[62] Ramazi S, Mohammadi N, Allahverdi A, Khalili E, Abdolmaleki P. A review on antimicrobial peptides databases and the computational tools. Database. 2022;**2022**:baac011

[63] Chen Z, Wang R, Guo J, Wang X. The role and future prospects of artificial intelligence algorithms in peptide drug development. Biomedicine & Pharmacotherapy. 2024;**175**(116709):116709

[64] Marczak B, Bocian A, Lyskowski A. Antimicrobial peptide databases as the guiding resource in new antimicrobial agent identification via computational methods. Molecules. 2025;**30**(6):1318

[65] Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, Niculae V, Prettenhofer P, Gramfort A, Grobler J, et al. Api design for machine learning software: Experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*, 2013

[66] Nightingale A, Antunes R, Alpi E, Bursteinas B, Gonzales L, Liu W, Luo J, Guoying Q, Turner E, Martin M. The proteins api: Accessing key integrated protein and genome information. Nucleic Acids Research. 2017;**45**(W1):W539–W544

[67] Bale AS, Naveen Ghorpade SK, Rohan BS, et al. Web scraping approaches and their performance on modern websites. In: *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 956–959. IEEE, 2022

[68] Abodayeh A, Hejazi R, Najjar W, Shihadeh L, Latif R. Web scraping for

data analytics: A beautifulsoup implementation. In: *2023 sixth international conference of women in data science at prince Sultan University (WiDS PSU)*, 65–69. IEEE, 2023

[69] Manjari KU, Rousha S, Sumanth D, Devi JS. Extractive text summarization from web pages using selenium and tf-idf algorithm. In: *2020 4th international conference on trends in electronics and informatics (ICOEI)(48184)*, 648–652. IEEE, 2020

[70] Myers D, McGuffee JW. Choosing scrapy. Journal of Computing Sciences in Colleges. 2015;**31**(1):83–89

[71] Glez-Peña D, Lourenço A, López-Fernández H, Reboiro-Jato M, Fdez-Riverola F. Web scraping technologies in an api world. Briefings in Bioinformatics. 2014;**15**(5):788–797

[72] Kwabena AE, Wiafe O-B, John B-D, Bernard A, Boateng FAF. An automated method for developing search strategies for systematic review using natural language processing (nlp). MethodsX. 2023;**10**:101935

[73] Jinchan Q, Steppi A, Zhong D, Hao J, Wang J, Lung P-Y, Zhao T, Zhe H, Zhang J. Triage of documents containing protein interactions affected by mutations using an nlp based machine learning approach. BMC Genomics. 2020;**21**(1):773

[74] Neumann M, King D, Beltagy I, Ammar W. Scispacy: Fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*, 2019

[75] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. Biobert: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;**36**(4):1234–1240

[76] Wei C-H, Allot A, Lai P-T, Leaman R, Tian S, Luo L, Jin Q, Wang Z, Chen Q, Zhiyong L. Pubtator 3.0: An ai-powered literature resource for unlocking biomedical knowledge. Nucleic Acids Research. 2024;**52**(W1): W540–W546

[77] Vollmar M, Tirunagari S, Harrus D, Armstrong D, Gaborova R, Gupta D, Afonso MQL, Evans G, Velankar S. Dataset from a human-in-the-loop approach to identify functionally important protein residues from literature. Scientific Data. 2024;**11** (1):1032

[78] Riquelme-Garca A, Mulero-Hernández J, Tomás Fernández-Breis J. Annotation of biological samples data to standard ontologies with support from large language models. Computational and Structural Biotechnology Journal. 2025;**27**:2155–2167

[79] Chen Q, Britto R, Erill I, Jeffery CJ, Liberzon A, Magrane M, Onami J-I, Robinson-Rechavi M, Sponarova J, Zobel J, et al. Quality matters: Biocuration experts on the impact of duplication and other data quality issues in biological databases. Genomics, Proteomics & Bioinformatics 2020;**18**:91–103

[80] Kumawat H, Kaalia R, Ghosh I. Role of ontology in biomedical text mining. In Sharan A, Malik N, Imran H, Ghosh I, editors. Text Mining Approaches for Biomedical Data. Singapore: Springer; 2024:117–125

[81] Huang Q, Sun Y, Xing Z, Cao Y, Chen J, Xiwei X, Jin H, Jiaxing L. Let's discover more api relations: A large

language model-based ai chain for unsupervised api relation inference. ACM Transactions on Software Engineering and Methodology. 2024;**33**:1–34. To appear

[82] Laurent JM, Janizek JD, Ruzo M, Hinks MM, Hammerling MJ, Narayanan S, Ponnapati M, White AD, Rodriques SG. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024

[83] Ofer D, Brandes N, Linial M. The language of proteins: Nlp, machine learning & protein sequences. Computational and Structural Biotechnology Journal. 2021;**19**:1750–1758

[84] Burger JD, Doughty E, Khare R, Wei C-H, Mishra R, Aberdeen J, Tresner-Kirsch D, Wellner B, Kann MG, Zhiyong L, et al. Hybrid curation of gene–mutation relations combining automated extraction and crowdsourcing. Database. 2014;**2014**:bau094

[85] Waskom ML, Tan KT, Wiberg H, Cohen AB, Wittmershaus B, Shapiro W. A hybrid approach to scalable real-world data curation by machine learning and human experts. medRxiv, 2023. preprint.

[86] Aleksander SA, Balhoff J, Seth Carbon JMC, Drabkin HJ, Ebert D, Feuermann M, Gaudet P, Harris NL, et al. The gene ontology knowledgebase in 2023. Genetics 2023;**224**:iyad031

[87] Strömert P, Hunold J, Castro A, Neumann S, Koepler O. Ontologies4chem: The landscape of ontologies in chemistry. Pure and Applied Chemistry. 2022;**94**(6):605–622

[88] Verbitsky A, Boutet P, Eslami M. Metadata harmonization from biological datasets with language models. bioRxiv, 2025–01, 2025

[89] Deutsch EW. File formats commonly used in mass spectrometry proteomics. Molecular & Cellular Proteomics. 2012;**11**(12): 1612–1621

[90] Chapman B, Chang J. Biopython: Python tools for computational biology. ACM Sigbio Newsletter. 2000;**20**(2): 15–19

[91] Miller M, Vielfaure N. Openrefine: An approachable open tool to clean research data. Bulletin - Association of Canadian Map Libraries and Archives (ACMLA). 2022;(170). Available from: https://openjournals.uwaterloo.ca/index.php/acmla/article/view/4873

[92] Reback J, McKinney W, Bossche JVD, Augspurger T, Cloud P, Klein A, Hawkins S, Roeschke M, Tratner J, She C, et al. pandas-dev/pandas: Pandas 1.0. 5. Zenodo. 2020. Available from: https://ui.adsabs.harvard.edu/abs/2020zndo...3898987R/

[93] McNaught A. The iupac international chemical identifier. Chemistry International. 2006;**28**(6):12–15

[94] Gaulton A, Hersey A, Michałnowotka APB, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, et al. The chembl database in 2017. Nucleic Acids Research 2017;**45**:D945–D954

[95] Kim S, Chen J, Cheng T, Gindulyte A, Jia H, He S, Qingliang L, Shoemaker BA, Thiessen PA, Yu B, et al. Pubchem 2023 update. Nucleic Acids Research 2023;**51**:D1373–D1380

[96] Pertsemlidis A, Fondon III. JW. Having a blast with bioinformatics (and avoiding blastphemy). Genome Biology. 2001;**2**(10):reviews2002–1

[97] Wise J, Grebe de Barron A, Splendiani A, Balali-Mood B, Vasant D, Little E, Mellino G, Harrow I, Smith I, Taubert J, et al. Implementation and relevance of fair data principles in biopharmaceutical r&d. Drug Discovery Today 2019;**24**:933–938

[98] Ebrahimikondori H, Sutherland D, Yanai A, Richter A, Salehi A, Chenkai L, Coombe L, Kotkoff M, Warren RL, Birol I. Structure-aware deep learning model for peptide toxicity prediction. Protein Science. 2024;**33**(7):e5076

[99] Chen Q, Wan Y, Zhang X, Lei Y, Zobel J, Verspoor K. Comparative analysis of sequence clustering methods for deduplication of biological databases. Journal of Data and Information Quality (JDIQ). 2018;**9**(3):1–27

[100] Bucataru C, Ciobanasu C. Antimicrobial peptides: Opportunities and challenges in overcoming resistance. Microbiological Research. 2024;**286**(127822). DOI:10.1016/j.micres.2024.127822

[101] Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J, Gillespie M, Haw R, et al. The reactome pathway knowledgebase. Nucleic Acids Research 2020;**48**:D498–D503

[102] Knox C, Wilson M, Klinger CM, Franklin M, Oler E, Wilson A, Pon A, Cox J, Chin NE, Strawbridge SA, et al. Drugbank 6.0: The drugbank knowledgebase for 2024. Nucleic Acids Research 2024;**52**:D1265–D1275

[103] Orchard S, Al-Lazikani B, Bryant S, Clark D, Calder E, Dix I, Engkvist O, Forster M, Gaulton A, Gilson M, et al. Minimum information about a bioactive entity (miabe). Nature Reviews Drug Discovery 2011;**10**:661–669

[104] Vempati UD, Chung C, Mader C, Koleti A, Datar N, Vidovic D, Wrobel D, Erickson S, Muhlich JL, Berriz G, et al. Metadata standard and data exchange specifications to describe, model, and integrate complex and diverse high-throughput screening data from the library of integrated network-based cellular signatures (lincs). Journal of Biomolecular Screening 2014;**19**:803–816

[105] Range J, Halupczok C, Lohmann J, Swainston N, Kettner C, Bergmann FT, Weidemann A, Wittig U, Schnell S, Pleiss J. Enzymeml—a data exchange format for biocatalysis and enzymology. The FEBS Journal. 2022;**289**(19):5864–5874

[106] Yasset Perez-Riverol and European Bioinformatics Community for Mass Spectrometry. Toward a sample metadata standard in public proteomics repositories. Journal of Proteome Research. 2020;**19**(10):3906–3909

[107] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, et al. The fair guiding principles for scientific data management and stewardship. Scientific Data 2016;**3**:1–9

[108] Garcia L, Bolleman J, Gehant S, Redaschi N, Martin M. Fair adoption, assessment and challenges at uniprot. Scientific Data. 2019;**6**(1):175

[109] Lautenbacher L, Samaras P, Muller J, Grafberger A, Shraideh M, Rank J, Fuchs ST, Schmidt TK, The M, Dallago C, et al. Proteomicsdb: Toward a fair open-source resource for life-science research. Nucleic Acids Research 2022;**50**:D1541–D1552

[110] Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, Haak LL, Haendel M, Herman I, Hodson S, et al. Achieving human and machine accessibility of cited data in scholarly publications. PeerJournal of Computer Science. 2015;**1**:e1

[111] LeRoy NJ, Khoroshevskyi O, O'Brien A, Stepien R, Arslan A, Sheffield NC. Pephub: A database, web interface, and api for editing, sharing, and validating biological sample metadata. GigaScience. 2024;**13**:giae033

[112] Breiman L, Jacquemard C, Giansanti V, Tinti M, Fraternali F, Raimondo D. Aaontology: An ontology of amino acid scales for peptide and protein characterization. Journal of Molecular Biology. 2024;**436**(15):168325

[113] Thanos C. Research data reusability: Conceptual foundations, barriers and enabling technologies. Publications. 2017;**5**(1):2

[114] Perez-Riverol Y, Alpi E, Wang R, Hermjakob H, Vizcano JA. Making proteomics data accessible and reusable: Current state of proteomics databases and repositories. Proteomics. 2015; **15**(5-6):930–950

[115] Klein JA, Lam H, Mak TD, Bittremieux W, Pérez-Riverol Y, Gabriels R, Shofstahl J, Hecht H, Binz P-A, Kawano S, Bossche TVD, Carver JJ, Neely BA, Mendoza L, Suomi T, Claeys T, Payne T, Schulte D, Sun Z, Hoffmann N, Zhu Y, Neumann S, Jones AR, Bandeira N, Vizcano JA, Deutsch EW. The proteomics standards initiative standardized formats for spectral libraries and fragment ion peak annotations: Mzspeclib and mzpaf. Analytical Chemistry. 2024;**96** (46):18491–18501

[116] Papadatos G, Gaulton A, Hersey A, Overington JP. Activity, assay and target data curation and quality in the chembl database. Journal of Computer-aided Molecular Design. 2015;**29** (9):885–896

[117] Mitchell SN, Lahiff A, Cummings N, Hollocombe J, Boskamp B, Field R, Reddyhoff D, Zarebski K, Wilson A, Viola B, et al. Fair data pipeline: Provenance-driven data management for traceable scientific workflows. Philosophical Transactions of the Royal Society A 2022;**380**:20210300

[118] Gadiya Y, Ioannidis V, Henderson D, Gribbon P, Rocca-Serra P, Satagopam V, Sansone S-A, Wei G. Fair data management: What does it mean for drug discovery? Frontiers in Drug Discovery. 2023;**3**:1226727

[119] Leipzig J, Nüst D, Hoyt CT, Ram K, Greenberg J. The role of metadata in reproducible computational research. Patterns. 2021;**2**(9):100322

[120] Gray AJG, Goble C, Jimenez RC. Bioschemas: From potato salad to protein annotation. In *The 16th International Semantic Web Conference 2017*. RWTH Aachen University, 2017

[121] Sansone S-A, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL. Milo Thurston, and FAIRsharing Community. Fairsharing as a community approach to standards, repositories and policies. Nature Biotechnology. 2019;**37**(4):358–367

[122] Gaignard A, Rosnet T, De Lamotte F, Lefort V, Devignes M-D. Fair-checker: Supporting digital resource findability and reuse with knowledge graphs and semantic web standards. Journal of Biomedical Semantics. 2023;**14**(1):7

[123] Aronica PGA, Reid LM, Desai N, Jianguo L, Fox SJ, Yadahalli S, Essex JW, Verma CS. Computational methods and tools in antimicrobial peptide research. Journal of Chemical Information and Modeling. 2021;**61**(7):3172–3196

[124] Asim MN, Asif T, Mehmood F, Dengel A. Peptide classification landscape: An in-depth systematic literature review on peptide types, databases, datasets, predictors architectures and performance. Computers in Biology and Medicine. 2025;**188**:109821. DOI: 10.1016/j.compbiomed.2025.109821

[125] Peng S, Rajjou L. Unifying antimicrobial peptide datasets for robust deep learning-based classification. Data in Brief. 2024;**56**:110822. DOI: 10.1016/j.dib.2024.110822

[126] Ren F, Wei J, Chen Q, Mengling H, Yu L, Jianing M, Zhou X, Qin D, Jianming W, Anguo W. Artificial intelligence-driven multi-omics approaches in Alzheimer's disease: Progress, challenges, and future directions. Acta Pharmaceutica Sinica B. 2025;**15**:4327–4385

[127] Jones CE, Brown AL, Baumann U. Estimating the annotation error rate of curated go database sequence annotations. BMC Bioinformatics. 2007;**8**(1):170

[128] Gawor A, Bulska E. A standardized protocol for assuring the validity of proteomics results from liquid chromatography–high-resolution mass spectrometry. International Journal of Molecular Sciences. 2023;**24**(7):6129

[129] Yang Y, Sun S, Yang S, Yang Q, Xinqiong L, Wang X, Quan Y, Huo X, Qian X. Structural annotation of unknown molecules in a miniaturized mass spectrometer based on a transformer enabled fragment tree method. Communications Chemistry. 2024;**7**(1):109

[130] Vahabi N, Michailidis G. Unsupervised multi-omics data integration methods: A comprehensive review. Frontiers in Genetics. 2022;**13**:(854752). DOI: 10.3389/fgene.2022.854752

[131] Blakeley-Ruiz JA, Kleiner M. Considerations for constructing a protein sequence database for metaproteomics. Computational and Structural Biotechnology Journal. 2022;**20**:937–952

[132] Aguilera-Mendoza L, Marrero-Ponce Y, Tellez-Ibarra R, Llorente-Quesada MT, Salgado J, Barigye SJ, Liu J. Overlap and diversity in antimicrobial peptide databases: Compiling a non-redundant set of sequences. Bioinformatics. 2015;**31**(15):2553–2559

[133] Gong Y, Liu G, Xue Y, Rui L, Meng L. A survey on dataset quality in machine learning. Information and Software Technology. 2023;**162**:(107268)107268 DOI: 10.1016/j.infsof.2023.107268

[134] Garca-Jacas CR, Pinacho-Castellanos SA, Garca-González LA, Brizuela CA. Do deep learning models make a difference in the identification of antimicrobial peptides? Briefings in Bioinformatics. 2022;**23**(3):bbac094

[135] Zhou W-J, Yang H, Zeng W-F, Zhang K, Chi H, Si-Min H. Validation beyond the target-decoy approach for peptide identification in shotgun proteomics. Journal of Proteome Research. 2019; **18**: 2747–2758. pvalid

[136] Dooley DM, Petkau AJ, Domselaar G, Hsiao WWL. Sequence database versioning for command line and galaxy bioinformatics servers. Bioinformatics. 2015;**32**:1275–1277

[137] Ramsey J, McIntosh B, Renfro D, Aleksander SA, LaBonte SA, Ross C, Zweifel A, Liles N, Farrar S, Gill J, Erill I, Ades S, Berardini T, Bennett JA, Brady S, Britton R, Carbon S, Caruso SM, Clements D, Dalia RR, Defelice M, Doyle EL, Friedberg I, Gurney S, Hughes L, Johnson AA, Kowalski JM, Li D, Lovering R, Mans TL, McCarthy F, Moore S, Murphy R, Paustian T, Perdue S, Peterson CN, Prüß B, Saha M, Sheehy RR, Tansey J, Temple L, Thorman AW, Trevino SR, Vollmer A, Walbot V, Willey JM, Siegele D, Hu JC. Crowdsourcing biocuration: The community assessment of community annotation with ontologies (cacao). PLOS Computational Biology. 2021;**17** (10):e1009463

[138] Bernett J, Blumenthal DB, Grimm DG, Haselbeck F, Joeres R, Kalinina OV, List M. Guiding questions to avoid data leakage in biological machine learning applications. Nature Methods. 2024;**21** (8):1444–1453

[139] Maharana K, Mondal S, Nemade B. A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings. 2022;**3**(1):91–99

[140] Steinegger M, Söding J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology. 2017;**35** (11):1026–1028

[141] Limin F, Niu B, Zhu Z, Sitao W, Weizhong L. Cd-hit: Accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;**28**(23):3150–3152

[142] Chen Q, Wan Y, Lei Y, Zobel J, Verspoor K. Evaluation of cd-hit for constructing non-redundant databases. In: *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 703–706. IEEE, 2016

[143] Liu J, Xiao J, Xunwen S, Wang Y. Deep learning–enhanced clustering and classification of protein molecule tertiary structures using weighted distance matrices. Briefings in Bioinformatics. 2025;**26**(4):bbaf331

[144] Pereira J, Pantolini L, Durairaj J, Schwede T. Large-scale protein clustering in the age of deep learning. Current Opinion in Structural Biology. 2025;**94**:103078

[145] Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, Tsenkov M, Nair S, Mirdita M, Yeo J, et al. Alphafold protein structure database in 2024: Providing structure coverage for over 214 million protein sequences. Nucleic Acids Research 2024;**52**:D368–D375

[146] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. Highly accurate protein structure prediction with alphafold. nature 2021;**596**:583–589

[147] Hartman E, Forsberg F, Kjellström S, Petrlova J, Luo C, Scott A, Puthia M, Malmström J, Schmidtchen A.

Peptide clustering enhances large-scale analyses and reveals proteolytic signatures in mass spectrometry data. Nature Communications. 2024;**15**(1):7128

[148] Barrio-Hernandez I, Yeo J, Jänes J, Mirdita M, Gilchrist CLM, Wein T, Varadi M, Velankar S, Beltrao P, Steinegger M. Clustering predicted structures at the scale of the known protein universe. Nature. 2023;**622**(7983):637–645

[149] Carugo O. Accuracy of alphafold models: Comparison with short n.. o contacts in atomic resolution protein crystal structures. Computational Biology and Chemistry. 2024;**110**:108069

[150] Schwämmle V, Jensen ON. Vsclust: Feature-based variance-sensitive clustering of omics data. Bioinformatics. 2018;**34**(17):2965–2972

[151] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Wenting L, Smetanin N, Dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. bioRxiv, 2022

[152] Ahmed M, Seraj R, Islam SMS. The k-means algorithm: A comprehensive survey and performance evaluation. Electronics. 2020;**9**(8):1295

[153] Schubert E, Sander J, Ester M, Kriegel HP, Xiaowei X. Dbscan revisited, revisited: Why and how you should (still) use dbscan. ACM Transactions on Database Systems (TODS). 2017;**42**(3):1–21

[154] Murtagh F, Contreras P. Algorithms for hierarchical clustering: An overview, ii. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2017;**7**(6):e1219

[155] Zhang X, Zhou Z, Hanfei X, Liu C-T. Integrative clustering methods for multi-omics data. Wiley Interdisciplinary Reviews: Computational Statistics. 2022;**14**(3):e1553

[156] McCrory MNF, Thomas SA. Cluster metric sensitivity to irrelevant features. In: Computational Problems in Science and Engineering II. Springer; 2025. p 85–95

[157] Alam FF, Rahman T, Shehu A. Learning reduced latent representations of protein structure data. In: *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, p. 592–597, 2019

[158] Feng S, Sterzenbach R, Guo X. Deep learning for peptide identification from metaproteomics datasets. Journal of Proteomics. 2021;**247**:104316. 104316

[159] Dai J, Chen J, Liu Y, Hu H. Novel multi-label feature selection via label symmetric uncertainty correlation learning and feature redundancy evaluation. Knowledge-Based Systems. 2020;**207**(106342):106342

[160] Medina-Ortiz D, Khalifeh A, Anvari-Kazemabad H, Davari MD. Interpretable and explainable predictive machine learning models for data-driven protein engineering. Biotechnology Advances. 2025;**79**:108495

[161] Guo G, Chen L, Yanfang Y, Jiang Q. Cluster validation method for determining the number of clusters in categorical sequences. IEEE Transactions on Neural Networks and Learning Systems. 2017;**28**:2936–2948

[162] Matthew P, Irizarry R. Interpretable convolution methods for learning genomic sequence motifs. bioRxiv, 2018. preprint

[163] Wang H, Zhang Y, Wen L, Wei Z, Wang Z, and Yang M. Clcluster: A redundancy-reduction contrastive learning-based clustering method of cancer subtype based on multi-omics data. Molecular Therapy - Nucleic Acids. 2025;**36**:(2). DOI:10.1016/j.omtn.2025.102534

[164] Sidorczuk K, Gagat P, Pietluch F, Kała J, Rafacz D, Bakała L, Słowik J, Kolenda R, Roediger S, Fingerhut LCHW, et al. The impact of negative data sampling on antimicrobial peptide prediction. bioRxiv, 2022–2025, 2022

[165] Xue Y, Dian L, Liu G. Improve multi-modal embedding learning via explicit hard negative gradient amplifying. arXiv preprint arXiv:2506.02020, 2025

[166] Yuan B, Gong C, Tao D, Yang J. Weighted contrastive learning with hard negative mining for positive and unlabeled learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2025

[167] Bhadra P, Yan J, Jinyan L, Fong S, Siu SWI.Ampep: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. Scientific Reports. 2018;**8**(1):1697

[168] Veltri D, Kamath U, Shehu A. Deep learning improves antimicrobial peptide recognition. Bioinformatics. 2018;**34**(16):2740–2747

[169] Szymczak P, Możejko M, Grzegorzek T, Jurczak R, Bauer M, Neubauer D, Sikora K, Michalski M, Sroka J, Setny P, et al. Discovering highly potent antimicrobial peptides with deep generative model hydramp. Nature Communications 2023;**14**:1453

[170] Xing W, Zhang J, Chen L, Huo Y, Dong G. iamp-attenpred: A novel antimicrobial peptide predictor based on bert feature extraction method and cnn-bilstm-attention combination model. Briefings in Bioinformatics. 2023;**25**:(1):bbad443 DOI: 10.1093/bib/bbad443

[171] Rádai Z, Kiss J, Nagy NA. Taxonomic bias in amp prediction of invertebrate peptides. Scientific Reports. 2021;**11**(1):17924

[172] Aguilera-Puga MDC, Plisson F. Structure-aware machine learning strategies for antimicrobial peptide discovery. Scientific Reports. 2024;**14**(1):11995

[173] Bournez C, Riool M, de Boer L, Cordfunke RA, de Best L, van Leeuwen R, Drijfhout JW, Zaat SAJ, van Westen GJP. Calcamp: A new machine learning model for the accurate prediction of antimicrobial activity of peptides. Antibiotics. 2023;**12**(4):725

[174] Jaiswal M, Singh A, Kumar S. Ptpamp: Prediction tool for plant-derived antimicrobial peptides. Amino Acids. 2023;**55**(1):1–17

[175] Pang Y, Yao L, Jhong J-H, Wang Z, Lee T-Y. Avpiden: A new scheme for identification and functional prediction of antiviral peptides based on machine learning approaches. Briefings in Bioinformatics. 2021;**22**(6):bbab263

[176] Periwal N, Arora P, Thakur A, Agrawal L, Goyal Y, Rathore AS, Anand HS, Kaur B, Sood V. Antiprotozoal

peptide prediction using machine learning with effective feature selection techniques. Heliyon. 2024;**10**(16): e36163

[177] Shao J, Zhao Y, Wei W, Vaisman II. Agramp: Machine learning models for predicting antimicrobial peptides against phytopathogenic bacteria. Frontiers in Microbiology. 2024;**15** (1304044). DOI: 10.3389/fmicb. 2024.1304044

[178] Dong H, Long X, Yun L. Synthetic hard negative samples for contrastive learning. Neural Processing Letters. 2024;**56**(1):33

[179] Ansari M, White AD. Learning peptide properties with positive examples only. Digital Discovery. 2024;**3**(5):977–986

[180] Lin T-T, Li-Yen Yang I-HL, Cheng W-C, Hsu Z-R, Chen S-H, Lin C-Y. Ai4amp: An antimicrobial peptide predictor using physicochemical property-based encoding method and deep learning. Msystems. 2021;**6**(6): e00299–21

[181] Maasch JRMA, Torres MDT, Melo MCR, de la Fuente-nunez C. Molecular de-extinction of ancient antimicrobial peptides enabled by machine learning. Cell Host & Microbe. 2023;**31**(8): 1260–1274

[182] Fu L, Zheng X, Luo J, Zhang Y, Gao X, Jin L, Liu W, Zhang C, Gao D, Bocheng X, et al. Machine learning accelerates the discovery of epitope-based dual-bioactive peptides against skin infections. International Journal of Antimicrobial Agents 2024;**64**:107371

[183] Pang Y, Wang Z, Jhong J-H, Lee T-Y. Identifying anti-coronavirus peptides by incorporating different negative datasets and imbalanced learning strategies. Briefings in Bioinformatics. 2021;**22**(2):1085–1095

[184] Lin C, Xiong S, Cui F, Zhang Z, Shi H, Wei L. Deep learning in antimicrobial peptide prediction. Journal of Chemical Information and Modeling. 2025;**65**:7373–7392

[185] Medina-Ortiz D, Contreras S, Fernández D, Soto-Garca N, Moya I, Cabas-Mora G, Olivera-Nappa Á. Protein language models and machine learning facilitate the identification of antimicrobial peptides. International Journal of Molecular Sciences. 2024;**25** (16):8851

[186] Kouba P, Kohout P, Haddadi F, Bushuiev A, Samusevich R, Sedlar J, Damborsky J, Pluskal T, Sivic J, Mazurenko S. Machine learning-guided protein engineering. ACS Catalysis. 2023;**13**(21):13863–13895

[187] Wan X. Protein sequence feature extraction techniques: Research advances, progress, and applications. 2025

[188] Mardikoraem M, Wang Z, Pascual N, Woldring D. Generative models for protein sequence modeling: Recent advances and future directions. Briefings in Bioinformatics. 2023;**24**(6): bbad358

[189] Attique M, Farooq MS, Khelifi A, Abid A. Prediction of therapeutic peptides using machine learning: Computational models, datasets, and feature encodings. Ieee Access. 2020;**8**:148570–148594

[190] Lin E, Lin C-H, Lane H-Y. De novo peptide and protein design using generative adversarial networks: An update. Journal of Chemical

Information and Modeling. 2022;**62**(4):761–774

[191] Terven J, Cordova-Esparza D-M, Romero-González J-A, Ramrez-Pedraza A, Chávez-Urbiola EA. A comprehensive survey of loss functions and metrics in deep learning. Artificial Intelligence Review. 2025;**58**(7):195

[192] Shahbazy M, Ramarathinam SH, Chen L, Illing PT, Faridi P, Croft NP, and Purcell AW. Mhcplogics: An interactive machine learning-based tool for unsupervised data visualization and cluster analysis of immunopeptidomes. Briefings in Bioinformatics. 2024;**25**(2): DOI: 10.1093/bib/bbae087

[193] Biswas S, Khimulya G, Alley EC, Esvelt KM, Church GM. Low-n protein engineering with data-efficient deep learning. Nature Methods. 2021;**18**(4):389–396

[194] Yang J, Francesca-Zhoufan L, Arnold FH.Opportunities and challenges for machine learning-assisted enzyme engineering. ACS Central Science. 2024;**10**(2):226–241

[195] Zhou Z, Zhang L, Yuanxi Y, Mingchen L, Hong L, Tan P. Enhancing the efficiency of protein language models with minimal wet-lab data through few-shot learning, 2024

[196] José -AB-A, Olivares-Gil A, Rodrguez JJ, Garca-Osorio C, Dez-Pastor JF. Addressing data scarcity in protein fitness landscape analysis: A study on semi-supervised and deep transfer learning techniques. Information Fusion. 2024;**102**(102035): DOI: 10.1016/j.inffus.2023.102035

[197] Schmirler R, Heinzinger M, Rost B. Fine-tuning protein language models boosts predictions across diverse tasks. Nature Communications. 2024;**15**(1):7407

[198] Hao Y, Li B, Huang D, Sijin W, Wang T, Lei F, Liu X. Developing a semi-supervised approach using a pu-learning-based data augmentation strategy for multitarget drug discovery. International Journal of Molecular Sciences. 2024;**25**(15):8239

[199] Jaskie K, Spanias A. Positive and unlabeled learning algorithms and applications: A survey. In: *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, p. 1–8. IEEE, 2019

[200] Wang J, Liu Y, Tian B. Protein-small molecule binding site prediction based on a pre-trained protein language model with contrastive learning. Journal of Cheminformatics. 2024;**16**(1):125

[201] Heinzinger M, Littmann M, Sillitoe I, Bordin N, Orengo C, Rost B. Contrastive learning on protein embeddings enlightens midnight zone. NAR Genomics and Bioinformatics. 2022;**4**(2):lqac043

[202] Flissi A, Ricart E, Campart C, Chevalier M, Dufresne Y, Michalik J, Jacques P, Flahaut C, Lisacek F, Leclère V, Pupin M. Norine: Update of the nonribosomal peptide resource. Nucleic Acids Research. 2019;**48**:D465–D469

[203] Upadhyay U, Pucci F, Herold J, Schug A. Nucleoseeker: Precision filtering of rna databases to curate high-quality datasets. bioRxiv, 2024

[204] Ammar A, Bonaretti S, Winckers LA, Quik J, Bakker M, Maier D, Lynch I, van Rijn J, Willighagen E. A semi-automated workflow for fair maturity indicators in the life sciences. Nanomaterials. 2020;**10**(10):2068

[205] Attafi OA, Clementel D, Kyritsis K, Capriotti E, Farrell G, Fragkouli S-C, Castro LJ, Hatos A, Lenaerts T, Mazurenko S, Mozaffari S, Pradelli F, Ruch P, Savojardo C, Turina P, Piovesan D, Monzon AM, Psomopoulos F, Tosatto SCE. Dome registry: Implementing community-wide recommendations for reporting supervised machine learning in biology. arXiv. 2024;**13** DOI: 10.1093/gigascience/giae094

[206] Kouper I. Data curation in interdisciplinary and highly collaborative research. International Journal of Digital Curation. 2023;**17** (1):20