Chemical Science

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: E. LOPEZ LOPEZ, J. P. Sánchez Castañeda, M. S. Martinez-Cortés, C. de la Fuente-Nunez and J. L. Medina-Franco, *Chem. Sci.*, 2026, DOI: 10.1039/D5SC04465K.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/chemical-science

View Article Online
DOI: 10.1039/D5SC04465K

Review

Exploring and Expanding the Chemical Multiverse of Peptides

Edgar López-López,[1,2] Jean Paul Sánchez,[1] Massyel S. Martinez-Cortés,[1] Cesar de la Fuente-Nunez,[3-6,*] José L. Medina-Franco,[1,*]

[1] *DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico*

[2] *Department of Chemistry and Graduate Program in Pharmacology, Center for Research and Advanced Studies of the National Polytechnic Institute, Section 14-740, 07000 Mexico City, Mexico*

[3] *Machine Biology Group, Departments of Psychiatry and Microbiology, Institute for Biomedical Informatics, Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America*

[4] *Departments of Bioengineering and Chemical and Biomolecular Engineering, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America*

[5] *Department of Chemistry, School of Arts and Sciences, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America*

[6] *Penn Institute for Computational Science, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America*

**Corresponding authors: medinajl@unam.mx, cfuente@upenn.edu*

Abstract

Peptides occupy a unique and rapidly expanding domain within the broader chemical space, offering exciting opportunities for therapeutic, nutritional, cosmetic, and materials applications. While efforts to characterize chemical space have traditionally focused on small molecules, growing evidence underscores the value of extending this concept to include peptide-based compounds. In this review, we survey recent advances in the exploration and expansion of the peptide chemical space, focusing on short peptides that bridge chemoinformatics and bioinformatics perspectives. We begin by briefly discussing the impact and applications of peptides in various research and industry areas and then examining the theoretical and practical size of the peptide chemical space, emphasizing how naturally occurring and synthetic peptides vastly increase its diversity. We then discuss molecular representations—from conventional notations to specialized peptide descriptors—highlighting their impact on library design, structural analyses, and activity predictions. Visualization methods and machine learning models are presented as tools for mapping structure–property and structure–function relationships. Next, we explore computational strategies for *de novo* peptide generation, driven by advances in generative modeling and high-throughput

View Article Online
DOI: 10.1039/D5SC04465K

screening. Throughout, we emphasize the role of open-source resources and integrated computational pipelines that combine chemo- and bioinformatics approaches to enhance data quality and predictive performance. We conclude by identifying major challenges—such as the complex structural landscape of peptides, data curation, and the need for consensus screening methods—and outline emerging opportunities for further expanding and refining the peptide chemical space.

Keywords: bioinformatics; chemical space; chemoinformatics; drug discovery; machine learning; molecular design; natural products; peptide design; representation.

Abbreviations: 3D-VAIF, three-dimensional vector of atomic interaction field; AMPs, antimicrobial peptides; anti-MRSA, Methicilin-resistant *Staphylococcus aureus;* BAPs, bioactive peptides; cAAs, canonical amino acids; ECFP, extended connectivity fingerprints; FDA, Food and Drug Administration; GalNAc, N-acetilgalactosamine; GLP-1 RA, glucagon-like peptide-1 receptor agonist; HELM, Hierarchical Editing Language for macromolecules; HNP-1, human neutrophil peptide 1; MAP, MinHashed atom-pair fingerprint; MXFP, macromolecule extended atom-pair fingerprint; ncAAs, non-canonical amino acids; npAAs, non-proteinogenic amino acids; nsAAs, nos-standard amino acids; PDB, Protein Data Bank; PLN, Protein Line Notation; PTMs, post-translational modification; QED, quantitative estimate of drug-likeness; SCSR, Self-Contained Sequence Representation; siRNA, small interfering RNA; SPPS, solid-phase peptide synthesis; unAAs unnatural amino acids.

1 Introduction

Chemical space has various definitions that can be classified into two general perspectives that have been recently reviewed and collected.[1] One set of definitions is focused on the “total number of chemical compounds” that can exist. Other definitions, in addition to the N-number of compounds, include a set of M-descriptors that generate a multi-dimensional space in which the compounds are located, like the concept of chemical space proposed by Virshup et al.[2] The latter view of the chemical space gives rise to the high-dependence of chemical space on the number and type of descriptors that are used to construct or define the space. Such variability of dependence has led to the concept of “chemical multiverse” which

View Article Online
DOI: 10.1039/D5SC04465K

has been defined as the group or collection of chemical spaces of a given set of compounds, each space defined by a specific set of descriptors.[1]

Since, in practical applications, the chemical space depends on the N-number and type of chemical compounds that are being studied, it would be appropriate to refer to “chemical subspaces”, for example, the subspace of drug-like organic compounds, metal-containing drugs, food chemicals, materials, and peptides.

Chemical space analysis, in particular, visual and qualitative/quantitative analysis, has multiple applications such as library design, compound selection, structure-property relationships, and chemical diversity analysis. To date, most common applications of chemical space are focused on small molecules (organic compounds) and peptides, with emerging and growing applications in other areas (such as food chemicals, materials, and organometallic molecules).[3]

Peptides as bioactive compounds are attractive in research areas like drug discovery, owing to their unique properties. Their high specificity and affinity to interact with targets make them ideal for certain applications in multiple fields, including therapeutics, material science, cosmetics, food chemistry, and biotechnology.[4-9] The first proof of peptides’ potential is given by the medical use of insulin in 1922 for type 1 diabetes, starting their increasing relevance in modern biomedicine.[6] The architectural diversity and operationally versatile characteristics of peptides further contribute to their constantly increasing importance in these areas.[10,11]

These extensive advantageous characteristics have led the researcher to explore the combinatorial potential of peptides, as their sequence variability generates a virtually limitless number of possible structures.[12] The peptide’s features stem from the amino acid chain’s length and conformation, which can be composed of canonical amino acids (cAAs) or can be further expanded by incorporating non-canonical amino acids (ncAAs), synthetic modifications, and post-translational modifications (PTMs).[13] Various elements serve to create a vast and versatile chemical domain that is under continuous exploration and expansion.[14]

Considering that a peptide comprises a series of amino acid monomers linked in a linear sequence, the range of potential peptides increases dramatically as the length extends.[10,11] This theoretical diversity establishes the basis of peptide chemical space for discovering new bioactive molecules and refining

View Article Online
DOI: 10.1039/D5SC04465K

potential therapeutic options.[12] The extensive nature of this space encounters several difficulties regarding synthesis, stability, and the feasibility of practical application. By mapping peptide chemical space, it is possible to seize its potential for developing new drugs, biomaterials, and functional molecules tailored to specific applications.

In the past few years, the increased interest in studying peptides from various applications (including the deep knowledge and characterization of bioactive peptides that are toxic compounds - biotoxins) has attracted the attention of the community to chart the chemical space of peptides.[12,15,16] Similar to the exploration of the chemical space of small organic compounds, a key aspect in the charting of the chemical space of peptides is the nature/type and number of descriptors used to define the space, along with the number and type of specific peptides under study.

The goal of this review is to survey the progress on the exploration and expansion of peptide chemical space. The review is organized into five main sections. After this Introduction, we discuss practical applications of peptides. Section 3 presents an analysis of the size of the chemical space of peptides. The next section discusses approaches to systematically explore - quantitatively and visually - and expand the chemical space of peptides. The last section presents concluding remarks. The review focuses on short and long peptides (with < 50 amino acids).

2 Practical Applications of Peptides

Due to their structural versatility, biocompatibility, and ease of synthesis, small peptides have gained increasing interest across various fields ranging from biomedical sciences to food technology and materials engineering. Hereunder, we discuss exemplary practical applications highlighting their therapeutic and industrial potential.

2.1 Drug Discovery: Novel Peptide Therapeutics

In drug discovery, small peptides serve as promising candidates for novel therapeutics due to their high specificity and low toxicity.[17,18] Since the discovery of insulin in 1922, peptide drugs have been developed to treat a wide range of diseases, including cancer, immunological diseases, metabolic disorders, viral

View Article Online
DOI: 10.1039/D5SC04465K

infections, cardiovascular diseases, and other chronic diseases.[19,20] The advancement in peptide technology over the past decades is changing the drug discovery landscape.

Insulin was isolated by Banting and Best from dog pancreas and later from bovine sources, after which it was further purified and its amino acid sequence determined. In 1982, human recombinant insulin was produced for the first time in *E. coli* and yeast. Today, rapid-acting insulin analogs are being developed to optimize glycemic control.[21]

Pharmacologically active peptides are hard to formulate as drug products, as compared to small molecules, due to the various challenges in administration and delivery of therapeutic peptides into cancer cells and tumor sites. Typically, peptide drugs (with cAAs) exhibit shorter circulation half-lives, lower cell permeability, and typically higher rates of enzymatic degradation. Nevertheless, therapeutic peptides with cAAs and ncAAs have the advantage of high target specificity and low toxicity. Overcoming their current limitations will lead to safer and more effective drugs.[19, 22]

The development of therapeutic peptides has followed diverse paths, illustrating the main challenges and solutions in the field. An emblematic case is that of the incretin peptide GLP-1, initially limited by its rapid degradation in blood, which led to the design of analogs resistant to the DPP-4 enzyme and with structural modifications that prolonged their half-life, giving rise to successful drugs such as liraglutide and semaglutide (Figure 1). Glucagon was first isolated in 1923 and approved by the FDA in 1960 for treating severe hypoglycemia. In 1982, the glucagon gene was identified in the Atlantic anglerfish, enabling the discovery of mammalian glucagon genes and the production of recombinant glucagon in bacteria. By the late 1990s, recombinant glucagon became commercially available, yet it retained the same stability issues.[23, 24]

The use of partially or fully substituted *L*-amino acids with *D*-amino acids is a strategy to decrease proteolytic cleavage and lower immunogenicity. An example is octreotide, an FDA-approved octapeptide (Figure 1), that is an unnatural *D*-enantiomer modification, which is used in the treatment of gastrointestinal tumors. The development of octreotide traces back to the discovery of somatostatin and the desire to harness its inhibitory effects on hormone secretion, while overcoming somatostatin´s extremely short half-life. Researchers adopted a strategy of peptide analog design, selecting shorter cyclic peptides that could maintain receptor binding yet resist proteolytic degradation.[25] In particular, they introduced *D*-amino acids

View Article Online
DOI: 10.1039/D5SC04465K

and chemically constrained the peptide by cyclization (disulfide bridge) to increase metabolic stability and receptor affinity. Lead optimization employed structure-activity relationship (SAR) studies to refine receptor subtype selectivity and pharmacokinetic properties. The resulting octreotide with an enhanced half-life, high affinity for somatostatin receptor subtypes 2 and 5, and improved bioavailability.[18,20,26]

Significant achievements have been made in the efficacy and selectivity of therapeutic peptide delivery. The bioavailability and stability of therapeutic peptides have been increased due to the development of several formulation and delivery methods, including prodrug approaches, direct chemical modifications, applying special drug delivery systems, co-administration of enzyme inhibitors and absorption enhancers, because free peptides are not systematically stable without modifications.[27,28] For example, peptide cyclization is a structural manipulation where the constrained geometries result in dramatically reduced proteolytic degradation by amino and carboxypeptidases. Octreotide is the stable analogue of the parent peptide, somatostatin. Similarly, rational structural optimization played a central role in the design of bivalirudin (Figure 1), which was developed through a structure-based approach aimed at creating a safer and more controllable anticoagulant than heparin. Inspired by the natural thrombin inhibitor hirudin, researchers used the crystal structure of the thrombin–hirudin complex to engineer shorter synthetic analogs. Peptide synthesis and biochemical assays led to the identification of bivalirudin, a 20-amino-acid peptide that binds reversibly to thrombin's active site and exosite I, providing potent yet transient anticoagulation.[29,30] While eptifibatide (Figure 1) was engineered from the snake venom peptide barbourin, using SAR-guided optimization and cyclic peptide synthesis to enhance receptor selectivity, stability, and pharmacokinetic properties.[31]

Setmelanotide (Figure 1) and zilucoplan (chemical structure not shown) exemplify the application of rational peptide design and optimization in modern drug discovery. Setmelanotide, an eight–amino acid agonist of the melanocortin-4 receptor (MC4R), was developed through SAR studies and receptor binding assays to enhance potency, selectivity, and signaling bias toward Gs-mediated pathways involved in appetite regulation. Similarly, zilucoplan, a synthetic 15-residue macrocyclic peptide, emerged from an mRNA display screening platform that identified high-affinity binders to complement component C5. The lead sequence was optimized through solid-phase peptide synthesis, structural cyclization, and lipophilic

View Article Online
DOI: 10.1039/D5SC04465K

modification to improve stability and pharmacokinetic properties, demonstrating how diverse engineering strategies contribute to the development of potent and selective therapeutic peptides.[32-34]

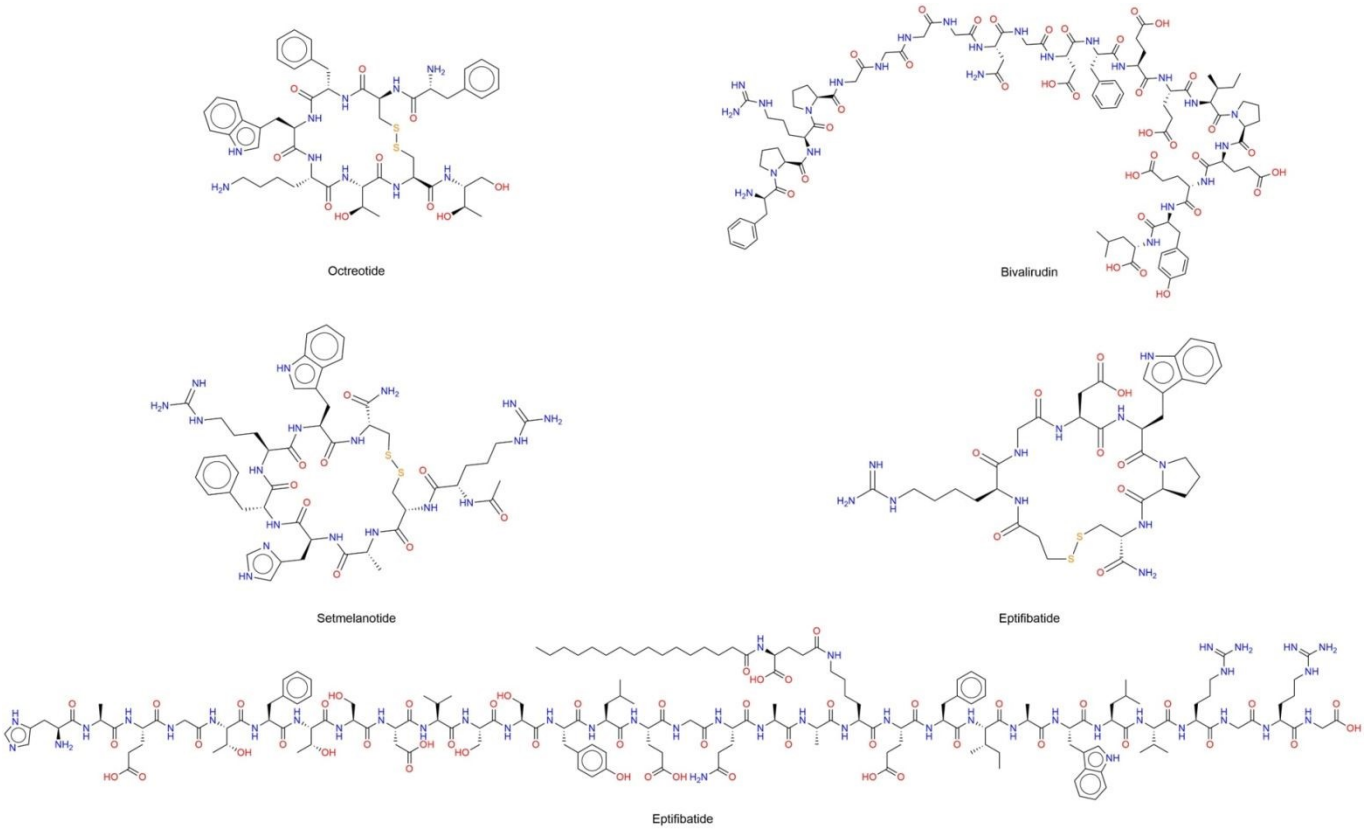

Figure 1. Chemical structure of selected peptides approved by the Food and Drug Administration (FDA) of the United States for clinical use. Linear amino acid sequences (one-letter code) of selected therapeutic peptides: Octreotide (FCFWKTCT), Bivalirudin (FPRPGGGGNGDFEEIPEEYL), Setmelanotide (RCAHFRWC), Etifibatide (CXGDWPC), and Liraglutide (HAEGTFTSDVSSYLEGQAAK(¹)EFIAWLVRGRG), where K(¹) denotes a lysine residue acylated with palmitoyl-γ-Glu (palmitic acid attached via a glutamic acid linker).

Antibiotics invented or discovered like penicillin by Alexander Fleming, have been used as a wonder drug for almost a century. However, those antibiotics are becoming failures due to their extensive overuse in recent decades, resulting in antimicrobial resistance. This prompted scientists to emphasize other alternatives like ocellatin. Ocellatin is a peptide derived from the skin secretions of *Leptodactylus* genus frogs and has a broad spectrum of antibacterial activities, specifically in gram-negative bacteria.[35]

Notable progress in the development of vaccines and small molecules with antiviral therapies, the continued emergence and re-emergence of viral outbreaks, along with rising antiviral resistance, have driven researchers to constantly seek new antiviral candidates.[36] Hence, antiviral peptides that mainly originate from antimicrobial peptides with antiviral activities can be prospective antiviral agents to fight viral

View Article Online
DOI: 10.1039/D5SC04465K

infections. Antiviral peptides typically are short (12-50 amino acid residues), and hydrophobicity is likely to be a key characteristic for antiviral peptides to target enveloped viruses. These antiviral peptides act against enveloped viruses by interrupting the fundamental stages of their life cycle of entry, synthesis, or assembly. Naturally, antimicrobial peptides with antiviral properties have been found in almost all multicellular organisms, like plants, animals, mammals, and microbes. It is important to mention that marine organisms are highly regarded reservoirs of pharmacologically active molecules, including peptides. Marine organisms biosynthesize structurally unique and bioactive compounds as an adaptive response to the harsh, highly competitive, and physiologically demanding conditions of the environment; conditions that markedly contrast with those of terrestrial ecosystems. These extreme ecological pressures drive the evolution of potent and functionally diverse molecular architectures.[37, 38]

Marine-derived cyclic and linear peptides have significantly advanced our understanding of ion channel modulation, antimicrobial activity, cytotoxic mechanisms, and other pharmacologically relevant properties, thereby positioning marine peptides as promising candidates for innovative therapeutic development.[37]

An example of the antiviral peptide is the HIV-1 targeting human neutrophil peptide HNP-1 exhibits an indirect mechanism of action by binding both the viral envelope glycoprotein Env and host cell surface molecules, including CD4 and co-receptors, in a manner that is independent of glycan interaction and serum components.[36] Additionally, its capacity for oligomerization or conformational rearrangement may sterically hinder the fusion process.[27, 39, 40]

2.2 Food and Nutrition: Functional Peptides with Specific Health Benefits

In recent years, food has been increasingly recognized not only as a source of essential nutrients, but also as a reservoir of biologically active compounds capable of promoting human health and enhancing physiological functions. Among these compounds, bioactive peptides (BAPSs) (short sequences of 2 to 20 amino acid residues with molecular weights ranging from 0.4 to 2 kDa) have attracted considerable attention due to their diverse health-promoting properties. Although less common, longer peptides such as lunasin (a 43-residue peptide from soy) have also been identified, exhibiting anticancer and hypocholesterolemic effects.[41, 42]

View Article Online
DOI: 10.1039/D5SC04465K

BAPS are typically released from parent proteins during enzymatic hydrolysis (*e.g.*, using trypsin, pepsin, alcalase) or through microbial fermentation. In addition to their role in basic nutrition, food-derived protein hydrolysates can exert immunomodulatory, anticancer, antihypertensive, antioxidant, antimicrobial, antidiabetic and anti-inflammatory effects. BAPs have been isolated from a wide range of sources including milk, egg, fish, soybean, rice, pea, oyster, mussel, chlorella and spirulina.[43-45]

Microalgae have emerged as sustainable, protein-rich organisms with the ability to synthesize a wide array of primary and secondary metabolites. Their high content of essential amino acids, combined with a rich profile of bioactive compounds, positions microalgae as promising therapeutic agents and valuable sources of functional food ingredients. For instance, *Arthrospira platensis* (Spirulina), a blue-green microalga consumed globally as a nutraceutical supplement, displays notable anti-inflammatory activity through the suppression of pro-inflammatory cytokines and gene expression.[46] Microalgae- derived peptides have demonstrated a wide range of bioactivities, including antihypertensive, antioxidant, anti-inflammatory, anticancer, antibacterial, antiallergic and antidiabetic effects. These peptides can be incorporated into various functional food products such as beverages, baked goods, pasta, yogurts and sports supplements, offering enhanced nutritional profiles without compromising sensory quality.[47]

A particularly relevant class of bioactive peptides is antimicrobial peptides (AMPs). AMPs display broad-spectrum antimicrobial activity, unique structural features and mechanisms of action that reduce the risk of drug resistance. Beyond disrupting bacterial membranes, some AMPs can penetrate cells and inhibit nucleic acid or protein synthesis, showing potential against multidrug-resistant strains.[48]

Fermented foods, especially in Asian countries, are another important source of bioactive peptides. Traditional fermented products like soybean (*e.g.*, sufu), fish, and milk derivatives are rich in peptides with antioxidant, antihypertensive, antimicrobial, antidiabetic, and anticancer properties.[49] Fermentations enhance not only shelf life but also flavor, texture, and nutritional value, due to proteolytic activity that releases bioactive peptides during ripening. Despite their benefits, some protein hydrolysates, particularly from soy, may generate bitter-tasting peptides during enzymatic hydrolysis, affecting palatability. Interestingly, the protease produced by *Mucor* species, used in fermented products like sufu, can degrade soybean protein without producing bitter peptides, while still generating bioactive polypeptides.

View Article Online
DOI: 10.1039/D5SC04465K

Fish proteins also serve as a high-quality source of peptides, particularly due to their content of essential amino acids and polyunsaturated fatty acids. Fermented fish products have been reported to exhibit antioxidant and ACE-inhibitory activity, largely due to the presence of low-molecular-weight peptides formed during processing.[48, 49]

Peptides obtained from various dietary proteins have been shown to exhibit diverse biological activities, including immunomodulatory, anticancer, antihypertensive, antioxidant, anti-inflammatory, mineral-chelating, lipid-lowering, bone-protective, and antimicrobial properties.[8, 50]

Modulating immune function through dietary components has proven to be a practical and effective approach; additionally, the identification of new immune-modulating peptides derived from food proteins may offer added benefits in dietary-based therapies.[50]

2.3 Cosmetics and Materials: Peptides for Biomaterials, Nanotechnology, and Skin Penetration Enhancers

The field of medical aesthetic skin care includes a vast array of ingredients and topical formulations, emphasizing the importance of a careful and evidence-based selection process. Evaluating the quality of scientific support for manufacturer claims is essential, including *in vivo* and *in vitro* studies that validate ingredient efficacy and practitioner preferences; these factors are key in determining product use, although not exhaustive. Peptides have advantages over small chemical molecules in specificity and selectivity, but they often have poor ability to penetrate skin.[51] Due to their multifunctional and regenerative capabilities, peptides have become a topic of growing scientific interest. Their biological activity depends largely on their structure and includes antioxidant, anti-aging, moisturizing, promoting collagen production, and wound-healing effects.[52]

Peptides can be classified by their mechanism of action into several functional categories: signal, carrier, neurotransmitter-inhibiting, enzyme-inhibiting, and antimicrobial peptides. The signal peptides stimulate the synthesis of collagen and elastin, one of the first cosmetic signal peptides to demonstrate this effect was palmitoyl peptide (Figure 2).[4,53]

Carrier peptides facilitate the delivery of essential trace elements involved in enzymatic activity and tissue repair, contributing to improve skin elasticity, the first commercialized carrier peptide was formulated

View Article Online
DOI: 10.1039/D5SC04465K

to deliver copper, which is a trace element necessary for wound healing.[38] Neurotransmitter-inhibiting peptides act by reducing neurotransmitter release, a process responsible for muscle contraction. By modulating this mechanism, they help diminish the appearance of fine lines and wrinkles. Acetyl hexapeptide-3, pentapeptide-3, pentapeptide-18, and tripeptide-3 (Figure 2) exhibit neuro-suppressive abilities and are referred to as neurotransmitter peptides.[14,54] Enzyme-inhibiting peptides prevent collagen degradation by inhibiting specific enzymes, thereby maintaining the integrity of skin structure. An example that is used in skincare is Oligopeptide-68 (Figure 2). Antimicrobial peptides defend against pathogens, including bacteria, fungi, and viruses, by compromising microbial membrane integrity. Myristoyl tetrapeptide-13 (Figure 2) is an example of a synthetic lipopeptide with potent antimicrobial activity.[54, 55]

The amino acid sequence of a cosmetic peptide plays a crucial role in determining its effects on the skin. Each amino acid in a peptide sequence contributes to the shape and charge of the molecule, therefore determining how the peptide interacts with the receptors and enzymes and how it diffuses through the lipid layer. Considering this, peptides containing amino acids with a positive charge, such as lysine, bind with a higher frequency to the membrane if they are located at the extremities of the sequence. Nevertheless, peptides composed of less hydrophobic, polar residues are much less likely to adsorb to membranes than phenylalanine-based peptides.[53]

Peptides used for cosmetic applications can be combined with zinc sulfate to enhance their antimicrobial effect and lesion-healing. Also, peptides can be formulated with vitamin E in antioxidant formulations, addressing structural and oxidative damage. Peptides contribute to the reduction in oxidative stress in the skin by scavenging free radicals through different pathways, resulting in delaying the skin's aging process.[51,52]

Recent research has shifted toward evaluating not only the biological activity of peptides, but also their bioavailability and formulation stability. While peptides offer multiple advantages as active ingredients in cosmetic applications, the development of new formulations is often constrained by issues related to stability, solubility, and skin permeability. One of the main challenges in the manufacturing process is preserving the structural integrity and bioactivity of peptides, which may be compromised by factors such as interactions with other formulation components, pH changes, temperature variations, and processing

View Article Online
DOI: 10.1039/D5SC04465K

methods. To address these challenges, it is crucial to select excipients that are chemically inert or minimally reactive to reduce the risk of degradation.[51, 52]

Development of bioactive peptides as safe and effective skin-care products, including dermatological applications such as wound healing, requires an understanding of their interaction with the various components present in the skin. Preclinical formulation of cosmetic and dermatological creams by observing the epidermal properties after application allows for a more complete understanding of the safety and efficacy of the product.[53,55]

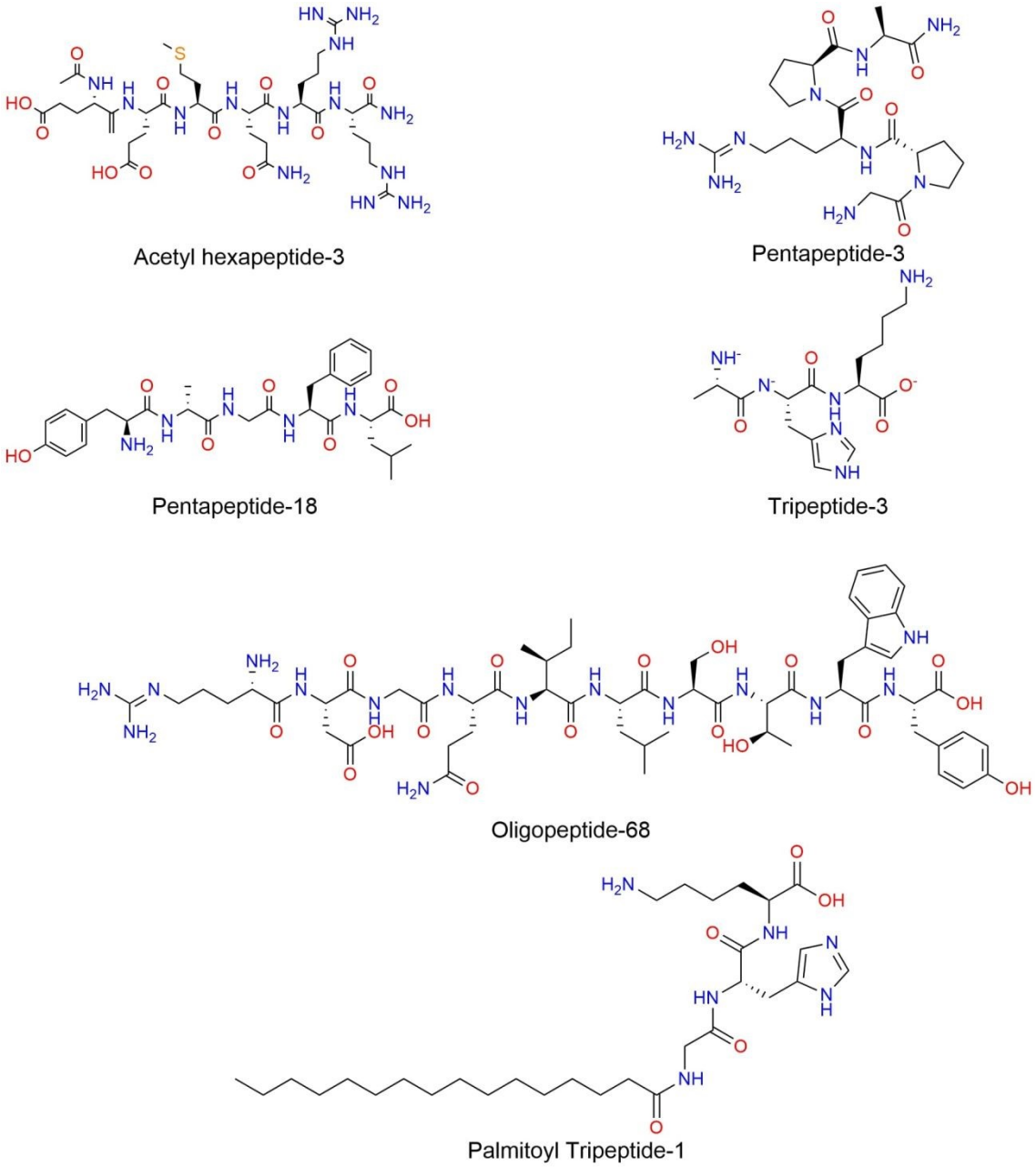

Figure 2. Examples of therapeutic and functional peptides commonly used in cosmetic formulations, discussed in the manuscript.

View Article Online
DOI: 10.1039/D5SC04465K

3 The Size of the Chemical Space of Peptides

Currently, peptides offer many opportunities in therapeutic and nontherapeutic areas, such as drug discovery, materials science, cosmetics, nutrition, and synthetic biology.[5,53] Versatility stems from the colossal combinatorial diversity of amino acids in peptides, highlighting the theoretically limitless peptide sequences conceivable.[54] The variety of potentially viable peptides has escalated by accounting for ncAAs, synthetic amino acids, and PTMs.

A peptide canonically consists of a linear chain with an “n” number of amino acids, where each link in the chain comes from a pool of 20 possible cAAs or proteinogenic amino acids.[56] The length of a peptide starts with two amino acids and has an arbitrary cut-off, typically set below 50 or 100 residues.[57-59] The number of possible sequences for a given peptide length is obtained by the formula:[15]

$$P(n) = A^n$$

where P(n) defines the total peptide sequences, A represents the count of selected amino acids as building blocks (like 20 proteinogenic forms), and “n” designates the peptide’s length.

The count of theoretical chains escalates exponentially as the target protein length augments, and with it the possible peptides for exploration. This group of peptides displays all the peptide combinations up to a certain length limit (usually between 50 – 100 residues). This chemical space provides a multitude of potential therapeutic peptides, crucial for drug discovery.[8] However, the number of peptides expands indefinitely when we add the non-canonical, synthetic amino acids and PTMs.[60] This trend can be appreciated, for instance, in the work of Orsi and Reymond,[11] who reported a virtual library comprising approximately 1×10^{60} peptide-like molecules, which were generated from the assembly of 100 commercially available peptide and peptoid building blocks into linear and cyclic oligomers of up to 30 units. This represents nearly 21 orders of magnitude more than the number of theoretical peptides of equivalent length derived from the 20 cAAs. The diversity increases even further when considering more building blocks like those 545 catalogued in the NORINE database,[61] which also currently contains 1,744 unique entries of nonribosomal peptides. As new peptides and monomers continue to be discovered and incorporated into such databases, the accessible peptide chemical space expands.[59]

View Article Online
DOI: 10.1039/D5SC04465K

3.1 Peptide Databases

Table 1. Examples of peptide-related databases and computational resources for peptide research.

Database or resource	Description	Number elements	Website	Ref.
PDB (Protein Data Bank)	An open-access repository that archives three-dimensional structural data of biological macromolecules determined mainly by X-ray crystallography, NMR spectroscopy, and cryo-EM.	> 100,000 entries	https://www.rcsb.org/	62
PeptideAtlas	Repository of experimental proteomics data (mass spectrometry)	6,636,295,537 peptide spectrum matches, 114 builds	https://peptideatlas.org/	63
UniProt	Comprehensive protein/peptide resource	253,635,358 entries	https://www.uniprot.org/	64
Peptipedia	Integrates >30 peptide databases (antimicrobial, anticancer, etc.) in a unified resource	3,983,654 peptides	https://peptipedia.cl/	65
NIST	Peptide mass spectral libraries	>4,300,000 spectra; 1,260,000 entities	https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:start	66
IEDB	The Immune Epitope Database	>1,600,000 Peptidic Epitopes	https://www.iedb.org/	67
ProteomicsDB	Human protein expression and PTM data	684,691 human peptides	https://www.proteomicsdb.org/	68
PRIDE	The PRoteomics IDEntifications database	>500,000 peptides (PRIDE Crosslinking)	https://www.ebi.ac.uk/pride/	69
SATPdb	Structurally Annotated Therapeutic Peptide Database	37,100 entries	http://crdd.osdd.net/raghava/satpdb/	70
DRAMP 4.0	Curated Data Repository of Antimicrobial Peptides	30,260 entries	http://dramp.cpu-bioinfor.org/	71

View Article Online
DOI: 10.1039/D5SC04465K

DBAASP v3	Antimicrobial peptides: activity and structure	>23,000 Monomer, Multimer 420, Multi Peptide 236	https://dbaasp.org/home	72
$CAMP_{R4}$	Collection of Antimicrobial Peptides	>24,000 sequences	https://camp.bicnirrh.res.in/	73
Propedia	Database of peptide–protein interactions	>19,000 peptide-protein complexes	https://bioinfo.dcc.ufmg.br/propedia/	74
CancerPDF	Cancer Peptidome Database of bioFluids	14,367 experimentally validated peptides	https://webs.iiitd.edu.in/raghava/cancerpdf/	75
PepBDB	Peptide Binding DataBase	13,299 structures of peptide-mediated protein interactions	http://huanglab.phys.hust.edu.cn/pepbdb/	76
HORDB	Hormone Peptide Database	7,390 peptide hormones. Includes structural, functional, and bioactivity data	http://hordb.cpu-bioinfor.org/	77
CancerPPD2	A repository of experimentally verified anticancer peptides and anticancer proteins	6,521 entries	https://webs.iiitd.edu.in/raghava/cancerppd2/	78
CAD v1.0	Cancer Antigenic Peptide Database	6,000 simulated neopeptides; 800 cancer antigens	http://cad.bio-it.cn/	79
APD6 (Antimicrobial Peptide Database)	Comprehensive antimicrobial peptide database (original version)	5,680 peptides	https://aps.unmc.edu/	80
NeuroPep 2.0	Neuropeptides	11,417 unique neuropeptide entries	http://isyslab.info/NeuroPep/	81
DADP	Database of Anuran Defense Peptides	2571 entries	http://split4.pmfst.hr/dadp/	82
PepLife	Half-life information for therapeutic peptides (experimental data)	2,172 entries	http://crdd.osdd.net/raghava/peplife/	83
Norine	Database of non-ribosomal peptides	1,744 peptides	https://norine.univ-lille.fr/norine/	84
MBPDB	Bioactive peptides derived from milk proteins	691 bioactive peptides sequences	https://mbpdb.nws.oregonstate.edu/	85

View Article Online
DOI: 10.1039/D5SC04465K

THPdb	Therapeutic Peptide Database	852 entries	http://crdd.osdd.net/raghava/thpdb/	86
BIOPEP-UWM	Database of bioactive peptides	7,857 entries	https://biochemia.uwm.edu.pl/en/biopep-uwm-2/	87

3.2 Comparison of Theoretical vs. Practically Synthesizable Peptides

While the theoretical total number of peptides (peptide chemical space) is astronomically large, we can distinguish between unexplored and explored chemical spaces, each containing bioactive and non-bioactive compounds.[59,88] Comparing the vast unexplored chemical space with explored or known space, represented by accessible peptides in libraries;[8] uncovers that the most promising active peptides are still undiscovered, and eligible for synthesis and experimentation.[89] Instead, the variety of peptides that can be feasibly created is constrained by numerous variables, which fluctuate based on the method of production and the attributes of the amino acid residues. These limitations are present in synthesis methods, folding stability, cost, and efficiency.

Chemical and enzymatic approaches to synthesizing longer peptides face limitations related to peptide length, purity, and complexity. As peptide length increases, yields tend to decrease, and synthesizing complex sequences often requires advanced techniques. Additionally, synthesizing peptides with PTMs is challenging, for instance, in addressing aggregation issues. These difficulties can result in low yields, reduced purity, and, in some cases, failure to obtain the desired peptide.[90]

Some peptide sequences tend to aggregate or misfold, making them challenging to synthesize, isolate, and study. The difficulties associated with synthesizing certain peptides are encapsulated in the term “difficult peptides”, introduced in the 1980s. These peptides share a common characteristic which is a high propensity for aggregation. This phenomenon arises from significant inter- or intra-molecular β-sheet interactions, which promote aggregation during synthesis. These structural interactions are stabilized by hydrogen bonds along the peptide backbone, making certain sequences prone to aggregation.[91]

The synthesis of longer and more complex peptides gets costly and time-consuming, thus restricting the feasible peptide chemical space. Improvements in peptide synthesis methods improve their scalability and yield. Advancements in reactive materials like resins, amino acid derivatives, and coupled reagents; along with better purification methods, have been vital for cost-cutting, boosting yield and purity, and

View Article Online
DOI: 10.1039/D5SC04465K

facilitating complex modifications. As a result, peptide companies offer peptides of 15 to 20 amino acids, including those with natural and non-natural or modified residues, at relatively low costs.[92] Thus, while the theoretical number of peptides is nearly limitless, practical synthesis and study are constrained by technical and economic factors.

Moreover, enzymatic synthesis, while potentially more efficient, but costly cofactors limit large-scale production. These reagents limit the variety of studyable peptides, showing a significant gap between the theoretical peptide chemical space and the subset accessible for experimental exploration.[93]

Additionally, availability of building blocks, especially non-proteinogenic amino acids (npAAs), limit the space for practically made peptides. Not all npAAs have been made yet,[94] and their chemical synthesis have their own challenges due to several issues such as stereoselectivity and low production yields. Many amino acids are chiral, with 19 of the 20 canonical ones containing at least one chiral center, excepting glycine, which is the only achiral cAAs. Among them, Isoleucine and threonine each have two chiral centers at α- and β-carbons.[95] Stereochemistry is key to defining the structural and functional properties of peptides.[96] Variations in the spatial arrangement of atoms influence backbone conformation, side-chain orientation, and overall molecular dynamics, which impact biological recognition, binding affinity, and functional selectivity.[96-98] From a broader perspective, stereochemical diversity represents an additional dimension of the peptide chemical space (stereochemical space),[99] enabling the exploration of novel structural motifs and conformational states beyond those encoded by natural amino acids. Expanding this stereochemical landscape not only enhances the potential for discovering peptides with improved stability, bioavailability, or activity but also provides a richer framework for *in silico* modeling and rational design of bioactive peptide scaffolds.[100,101]

Beyond stereochemical variation, molecular diversity also arises from the incorporation of non-canonical and chemically modified amino acids. While stereochemical changes influence the conformational landscapes of amino acids, the introduction of new monomers expands peptide chemical space by adding novel functional groups, backbone structures, and reactivities. Together, these complementary strategies—stereochemical diversification and amino acid development—offer enhanced opportunities for designing peptide structure, dynamics, and function.

View Article Online
DOI: 10.1039/D5SC04465K

3.3 Incorporation of Non-Canonical and Synthetic Amino Acids, Post-Translational Modifications

The chemical space of peptides has greatly expanded through the incorporation of ncAAs and PTMs as building blocks.[92,102] ncAAs, also known as unnatural or npAAs, non-standard amino acids (nsAAs), or unnatural amino acids (unAAs),[14] come from different types of amino acids not present genetic code.[103] Most ncAAs are synthesized chemically or semi-synthetically, while only a few can be produced through natural *in vivo* pathways.[103] Their inclusion dramatically expands the chemical space of peptides, offering unprecedented opportunities to fine-tune peptide structure-related functions. For example, amino acids such as selenocysteine and pyrrolysine enhance peptide reactivity and structural versatility.[92, 104, 105] Moreover, modified amino acids can improve peptide pharmacokinetic properties and increase thermal stability or resistance to enzymatic degradation.[106]

PTMs, like phosphorylation, acetylation, methylation, and glycosylation, contribute to the expansion of peptidic chemical space.[107] Those modifications extend their potential application by expanding their functions, stability, and interactions.[108-112] Thus, the introduction of PTMs is relevant for therapeutic purposes, mimicking natural alterations by their chemical or enzymatic addition.[110,113] On the one hand, chemical synthesis provides a uniform introduction of particular modifications at specified sites within a protein or peptide of interest. Advancements in peptide ligation methods and efficient coupling agents now enable the synthesis of long peptides with tailored modification and PTMs, therefore influencing their structural integrity, functional properties, and overall stability.[114-117]

On the other hand, the *in vitro* peptide enzymatic modifications often lead to heterogeneous products due to limited specificity and insufficient control over the extent of modification. Besides, complementary biological strategies, such as genetic code expansion technology, allow precise incorporation unAAs into target proteins by utilizing engineered orthogonal aminoacyl-tRNA synthetase/tRNA pairs.[116-119] This technique permits including diverse site-specific PTMs into recombinant proteins, including modifications such as acetylation, methylation, phosphorylation, and nitration. This method relies on the accessibility of an orthogonal tRNA synthetase specifically designed for the intended modification. However, the efficiency of this approach decreases when incorporating multiple PTMs within a single peptide or protein, presenting significant challenges for achieving large-scale combinatorial modifications.[116-119]

View Article Online
DOI: 10.1039/D5SC04465K

Head-to-tail macrocyclization is a naturally present PTM that stabilizes the protein and peptide fold, enhancing thermal stability and resistance to exoprotease proteolytic degradation.[120] In nature, cyclic peptides are present in bacteria, fungi, plants, and marine species, displaying remarkable diversity in shape, size, and chemical composition.[120,121] Their therapeutic potential arises from their capacity to inhibit enzymes, interfere with protein-protein interaction, modulate cell signaling, and regulate immune responses.[121] Their exceptional stability and selectivity make them ideal candidates for drug design. Additionally, cyclic peptides serve as crucial tools for drug discovery, functioning as molecular probes for detecting protein function, disease mechanisms, or therapeutic targets. Novel developments in methodologies including solid-phase peptide synthesis (SPPS), chemoenzymatic synthesis, and orthogonal protection strategies, have enhanced the specificity and complexity in the fabrication of cyclic peptides.[121]

Integration of ncAAs, cyclic peptides, and PTMs notably widens the peptide chemical multiverse, facilitating the creation of functional peptides beyond the capabilities of standard amino acids.

4. Exploration of the Known Chemical Space of Peptides

The exploration and systematic representation of peptide chemical space have become essential in both bioinformatics and chemoinformatics. Peptides occupy a unique position between small molecules and proteins, exhibiting complexity in their sequences, conformations, and physicochemical properties, which makes them highly diverse. To enable the rational exploitation of this diversity, a variety of computational frameworks, databases, and molecular representations have been developed. These resources offer standardized notations, molecular fingerprints, and structural encodings, facilitating the study and comparison of peptide structures, analog identification, property prediction, and machine learning–based modeling. Table 2 summarizes representative tools and resources, highlighting their main functions and the types of chemical or structural information they provide. The subsequent sections further discuss the principles and applications of selected methods in the context of peptide informatics and chemical space exploration.

View Article Online
DOI: 10.1039/D5SC04465K

Table 2. Selected bioinformatic and chemoinformatic resources for exploring peptide chemical space.

Tool Name	Website (when available) / Short description	Ref.
CHUCKLES (Chirality-Oriented Chemical Representation)	Method that interconverts peptide or peptoid sequences with SMILES, enabling both sequence- and structure-based searches, including branching and cyclic structures.	122
SCSR (Self-Contained Sequence Representation)	Method that encodes amino acid sequences with side-chain chemical info for modeling.	123
MAP4 (MinHashed atom-pair fingerprint up to a diameter of four bonds)	A molecular fingerprint that combines atom-pair concepts with MinHashing to efficiently represent both small molecules and large biomolecules. It captures structural and topological features up to four bonds apart, enabling fast and scalable molecular similarity searches across diverse chemical spaces. https://github.com/reymond-group/map4	124
MAP (Modification and Annotation in Proteins)	Format extends the traditional FASTA format by enabling annotation of modified residues, post-translational modifications, binding sites, mutations, and protein metadata. https://webs.iiitd.edu.in/raghava/maprepo/	125
ECFP (Extended Connectivity Fingerprint)	A circular molecular fingerprint that encodes atomic neighborhoods based on connectivity patterns. Widely used in cheminformatics for similarity searching, clustering, and QSAR modeling, ECFP captures local structural features around each atom to represent molecular topology in a compact, numerical form. https://docs.chemaxon.com/display/docs/fingerprints_extended-connectivity-fingerprint-ecfp.md	126
HELM (Hierarchical Editing Language for Macromolecules)	Enables standardized representation of complex biomolecules, including proteins, nucleotides, and antibody–drug conjugates.	127
PLN (Protein Line Notation)	A linear, text-based representation for describing protein and peptide sequences in a compact, machine-readable format. PLN encodes sequence information, modifications, and structural annotations, facilitating data exchange, database indexing, and computational analysis of proteins and peptides. http://www.biochemfusion.com/doc/PLN_Guide/PLN_Guide.html	128
3D-VAIF (three-dimensional vector of atomic interaction field)	Structural descriptor method encoding electrostatic and steric atomic interactions for 3D peptide representation.	129

View Article Online
DOI: 10.1039/D5SC04465K

MXFP (Macromolecule Extended Atom-Pair Fingerprint)	A 217-dimensional atom-pair fingerprint designed to encode large molecules (e.g., peptides, macrocycles, natural-products) by representing pharmacophore-group atom-pairs and their topological distances; useful for similarity searching and chemical-space mapping of non-Lipinski or biomolecular compounds. https://github.com/reymond-group/mxfp_python	130
KNIME (The Konstanz Information Miner)	An open-source, modular platform for end-to-end data analytics that enables users to visually build, execute and monitor data workflows—covering extraction, transformation, modelling and visualization—without needing extensive coding. https://www.knime.com/	131
Datawarrior	Software for data analysis and visualization. https://openmolecules.org/datawarrior/	132, 133
PepINVENT	Peptide design tool extending the REINVENT platform https://github.com/MolecularAI/PepINVENT/	134
PepSMI	A web-tool provided by NovoPro Bioscience Inc. that converts a peptide amino-acid sequence (using one-letter codes) into a SMILES (Simplified Molecular Input Line Entry System) string, enabling computational representation of the peptide's molecular structure. https://www.novoprolabs.com/tools/prot-sol	135
SignalP	Signal peptide prediction tool. https://services.healthtech.dtu.dk/services/SignalP-6.0/	136
Unipept	Peptide-based metaproteomics tool (biodiversity, biomarker discovery). https://unipept.ugent.be/	137

[a] Websites are provided when available; otherwise, a short description is included, along with the reference.

4.1 Molecular Representation

One of the most important considerations for generating representative chemical spaces that serve the purpose of the project's goals (*e.g.*, meaningful chemical spaces) is the appropriate use of the molecular representation and the descriptors that will be the basis to define the (multi) dimensional space. Towards this end, novel representations have been developed to condense structural information of complex molecules, like peptides.[122-130] For example, hashed fingerprints (*e.g.*, extended connectivity fingerprints (ECFP) and MinHashed atom-pair fingerprint (MAP)) have been distinguished from other unidimensional representations because they can codify the atom connectivity and neighborhoods of complex molecules.[124,126] However, these kinds of representations could be redundant for polymeric compounds, like large peptides, antibodies, or other kinds of proteins. For polymeric compounds, there have been

View Article Online
DOI: 10.1039/D5SC04465K

developed sequence-based representations that take advantage of the molecular redundancy of each compound to simplify their representations. For example, the CHUCKLES notation compacts the structural data of each amino acid into a unique letter, which can codify simple post-structural modifications like cysteine bridges.[122] Other notations, such as the Hierarchical Editing Language for macromolecules (HELM) and the Self-Contained Sequence Representation (SCSR), can represent more complex amino acid-based compounds like large canonical and non-canonical peptides and antibodies.[123,127] Interestingly, for polypeptides or proteins with more than 50 amino acids, the most common representations are the PDB entries that contain the coordinates of each atom of the polypeptide/protein and other non-covalent bound atoms (*i.e.*, solvent and ligands atoms), considering the presence of multiple chains, subunits, or multi-domain complexes. However, there exist unidimensional representations to codify protein connectivity information like Protein Line Notation (PLN) and Boehringer Ingelheim Line Notation (BILN) which can convert sequence representation to atom connectivity data without considering three-dimensionality features.[128,138]

Similarly, chemo-bioinformatics representations are innovative strategies to codify the atom connectivity of peptides at the same time as other important features, which has accelerated the development of novel peptide-based molecules with specific features. For example, the three-dimensional vector of atomic interaction field (3D-VAIF) approach captures the information of electrostatic and steric interaction between different types of atoms in peptides.[129] The macromolecule extended atom-pair fingerprint (MXFP) can describe molecular shapes and pharmacophores.[17,130] However, one of the major limitations of these representations is their low interpretability for the user, as they consist of alphanumeric codes that can only be understood through mathematical and computational processes (Figure 3D).

In addition, hybrid fingerprints inspired by different kinds of data, like amino acid sequence, atom-connectivity, and physicochemical properties, have been developed to codify most strictly the chemo-biological features of peptides.[139,140] Consequently, after the development of more informative molecular representations for peptides, the thin line between traditional bioinformatics and cheminformatics approaches to represent peptides is becoming increasingly blurred. However, there is an unwritten rule of thumb about the use of molecular fingerprints to codify the chemical structure of small peptides (<50

View Article Online
DOI: 10.1039/D5SC04465K

residues) and atom coordinates to represent large peptidic structures (>50 residues).[12] A few representative representations are illustrated in Figure 3 for a representative (given) peptide.

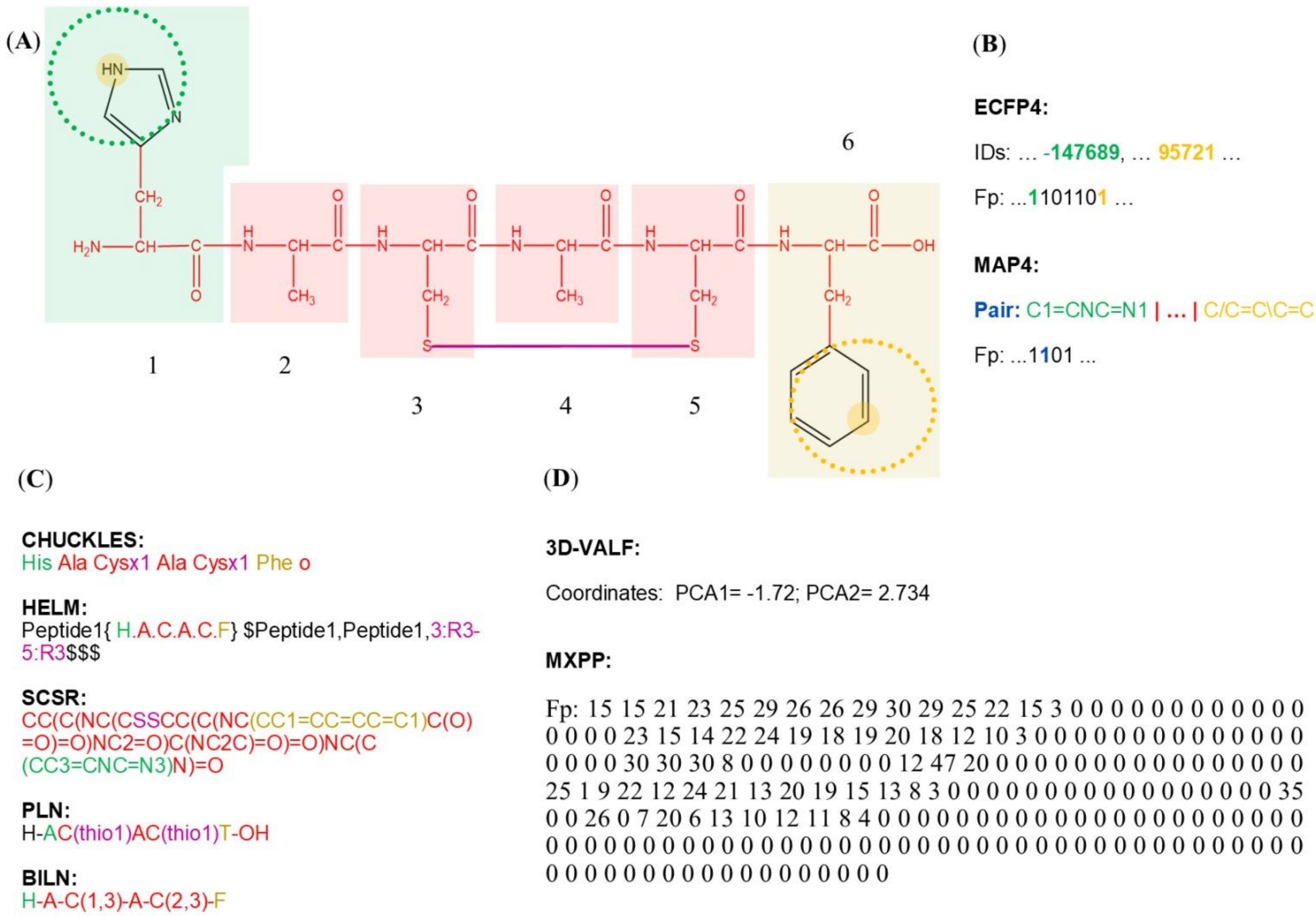

Figure 3. Molecular representations commonly used for peptides. (**A**) Peptide example. Their chemical structure contains amino acids in a specific order (alanine, cysteine, alanine, cysteine, and threonine), in which their cysteines form a disulfide bond; (**B**) Examples of hashed fingerprints. ECFP4 represents data of each atom in the structure and their connectivity with 2 atoms of distance from these, which is condensed bits after collision methods to avoid redundant connectivity information. A similar example is the MAP4 fingerprint, which uses this same strategy but captures the information on the connectivity of paired atoms; (**C**) Amino acid-based notations. HELM and PLM use the conventional one-letter notation to represent implicitly the amino acid connectivity and use specific codifications to represent out-peptide bonds, like disulfide bonds. However, examples like SCSR, PLN, and BILN representations offer an alternative to represent explicitly the atom connectivity of peptides; and (**D**) Three dimensional-based representations. For example, 3D-VALF uses conformational data to create new vectors from dimensional reduction methods capable of condensing the atom connectivity and property data of whole peptides; MXFP uses fingerprint-based representation to codify the presence of pharmacophoric features. The peptide representations shown in B and D panels are illustrative of the type of data generated by each representation.

View Article Online
DOI: 10.1039/D5SC04465K

4.3 Visual Representation of the Chemical Space of Peptides

Based on the advances in molecular representations of peptides and the development of novel bio-chemoinformatics descriptors, recent visual representations have enabled illustrating efficiently and intuitively peptidic structure-function relationships and activity-based clusters. This has allowed the creation of intuitive representations to make smart decisions about prospective evaluations for peptide-based compounds.[141,142] For example, advances in the visualization of atom connectivity similarity-activity relationships have opened new horizons for the optimization process of non-canonical peptides.[143]

4.4 Mapping Bioactive Peptides in Chemical Space

An illustrative example of the chemical space mapping to decode complex biological properties is shown in Figure 3. This landscape offers the possibility to study systematically underexplored peptides, like de-extinted peptides, *i.e.*, peptides from the “extinctome” (the proteomes of extinct organisms), which recently has covered particular relevance to developing novel antimicrobial agents.[144,145] Figure 3 remarks on the flexibility of the chemical space techniques (*e.g.*, using network-based approaches) to identify rapid structure-property relationships in peptides. For example, Figure 3A illustrates scaffold relationships between each pair of ancient and hemolytic (or non-hemolytic) peptides, and Figure 3B illustrates chemical space localizations ancient and anti-MRSA (Methicillin-resistant *Staphylococcus aureus*) peptides. Thus, remarks about how it is possible to establish structure-hemolytic relationships in peptides and quickly identify the potential anti-MRSA activity of underexplored peptides. Here, it is possible to identify that peptide 1 (Figure 3C) could have anti-MRSA activity without hemolytic side effects.

View Article Online
DOI: 10.1039/D5SC04465K

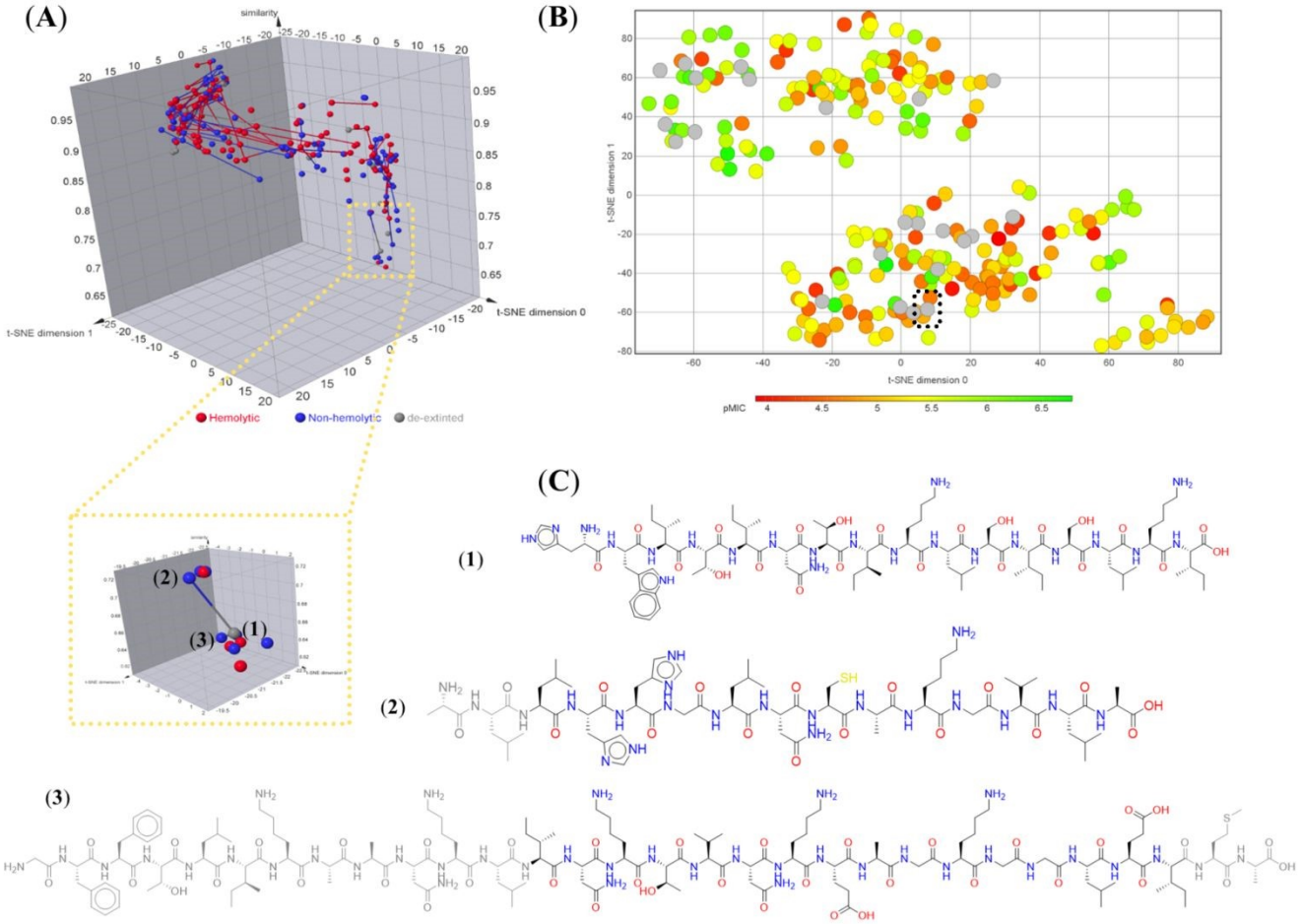

Figure 4. Chemical space representations to study ancient peptides. (**A**) Structure-activity relationships based on scaffold similarity of representative peptides with hemolytic and non-hemolytic activity. Chemical space based on coordinates and the structure of molecular scaffolds was constructed using a network-based protocol implemented in KNIME software and visualized using Datawarrior software, version 06.04.01.[132, 133,146] Each compound is represented by dots, and their distinctive biological property is represented by different colors. Finally, each compound was linked with other molecules that share a scaffold relationship; (**B**) Visual representation of the chemical space of 223 anti-MRSA non-canonical and canonical peptides. The visual representation was constructed by dimensional reduction (t-SNE coordinates) of the ECPF4 fingerprint. Each data point in the graph represents a peptide, and the color in a continuous scale represents the activity values in pMIC from low (red) to high (green). Data points in grey represent peptides with unknown pMIC value (ancient peptides). The dotted line illustrates the chemical space localization of **1** (panel A); (**C**) Representative examples of peptides are illustrated in panel A of this figure (**1**, HWITINTIKLSISLKI; **2**, ALLHHGLNCAKGVLA; **3**, GFFTLIKAANKLINKTVNKEAGKGGLEIMA). The atoms colored in grey in each peptidic structure represent the not-aligned moiety with the peptide **1.**[147, 148]

4.5 Development of Machine Learning Models

Recent advances in machine and deep learning technologies have led to the creation of network-based chemical space representations based on peptidic sequences, which opens up new perspectives on the

View Article Online
DOI: 10.1039/D5SC04465K

smart identification of “privileged” scaffolds and representative conserved moieties against specific chemical and biological endpoints. For example, solubility, type of tridimensional folding, cell permeability, bioactivity, toxicity, and hemolysis.[149-151]

The integration of high-throughput screening (HTS) technologies has transformed the early stages of peptide discovery. By enabling the rapid and parallel evaluation of thousands of compounds, HTS platforms generate extensive multidimensional datasets that can be mined to extract patterns linking peptide features with biological responses. These large-scale datasets constitute a critical substrate for the training and validation of AI-based models capable of predicting pharmacologically relevant properties.[152] Additionally, the implementation of robotic laboratory systems has further advanced this paradigm by ensuring the precision, scalability, and reproducibility of experimental workflows.[152] Automated liquid-handling robots and integrated analytical instruments can execute complex assay cascades with minimal human supervision, thereby reducing experimental bias and facilitating the generation of standardized, high-quality data amenable to algorithmic modeling.[153] Complementary to experimental acceleration, virtual screening protocols have become an indispensable component of contemporary peptide discovery pipelines. Virtual screening based on molecular docking, pharmacophore modeling, and quantitative structure–activity relationship (QSAR) analyses enable the rapid triage of chemical libraries, prioritizing peptides with predicted high affinity against specific receptors, favorable physicochemical properties, and adequate ADMET or sensory profiles.[154,155] Finally, molecular dynamics and quantum chemistry simulations provide a complementary, physics-based dimension to data-driven modeling, which allow the integration of conformational free-energy landscapes, binding stability, and interaction fingerprints—into deep learning frameworks, enhances the capacity to predict structure–function relationships at atomic resolution level.[156] These hybrid approaches bridge mechanistic simulation and statistical inference, enabling a more comprehensive understanding of molecular determinants underlying pharmacological efficacy and selectivity, as well as the decoding of properties with application beyond pharmacological disciplines.[157-159]

The emergence of AI-powered pharmaceutical laboratories is a paradigm shift toward closed-loop, autonomous discovery ecosystems. These laboratories combine robotic automation with AI to establish self-optimizing experimental frameworks in which predictive models continuously learn from experimental

View Article Online
DOI: 10.1039/D5SC04465K

feedback. Such adaptive systems can autonomously design, execute, and analyze experiments, thereby expediting the identification of lead peptides and optimizing their chemical space exploration.[160]

4.6 Expansion

Exploring peptide chemical space has emerged as a strategy to identify novel bioactive compounds with enhanced stability, specificity, and pharmacokinetic profiles. Recent advancements have focused on expanding this chemical space through various methodologies, including the design of synthetic combinatorial libraries, incorporation of non-canonical amino acids, the generation of "small-peptidic chemical chimeras", and generative (*de novo*) computational approaches.[161] These efforts aim to traverse previously underexplored regions of peptide chemical space, facilitating the discovery of compounds with unique biological activities, higher selectivity, and improved drug-like properties.[162-164]

4.7 Peptide Enumeration and *de Novo* Generation

Peptide enumeration and *de novo* generation have generated important contributions in modern peptide science, particularly with the advent of generative models capable of producing diverse and biologically relevant molecules. These approaches leverage advanced machine learning architectures, such as deep generative models and language-based neural networks, to systematically explore and expand the peptide chemical space beyond known structures and/or sequences. For instance, deep generative models can be trained to create novel candidates with desired biological or physicochemical properties, offering a powerful alternative to traditional combinatorial enumeration methods.[165] Other protocols now incorporate multi-objective optimization strategies, enabling simultaneous design for multiple properties such as structural stability, bioactivity, and membrane permeability, allowing the decodification of sequence-structure-function relationships;[166,167] Collectively, these methods represent a paradigm shift toward intelligent, data-driven design of peptide therapeutics, enabling the rapid identification of high-potential candidates with tunable properties. To this end, Geylan et. al introduced a novel approach, PepINVENT, designed to expand the landscape of peptide therapeutics by incorporating both natural and non-natural amino acids into *de novo* peptide design. PepINVENT enables the generation of peptides with enhanced properties such as binding affinity, plasma stability, and membrane permeability, which are crucial for

View Article Online
DOI: 10.1039/D5SC04465K

therapeutic efficacy. This example integrates reinforcement learning algorithms to create and navigate novel regions of the peptide chemical space, incorporating multi-objective optimization strategies inspired by a holistic molecular design approach.[134]

The systematic enumeration of peptide sequences within defined molecular property ranges, such as quantitative estimate of drug-likeness (QED) and toxicity, has become a focal point in computational peptide design. Artificial intelligence algorithms now integrate predictive models that assess key physicochemical properties, enabling the generation of peptide libraries tailored to specific therapeutic profiles, generating the first generation of "focused peptide libraries".[168,169] These types of libraries are characterized by their enriched content of biologically relevant and synthetically feasible sequences, typically constrained by user-defined filters such as solubility, stability, net charge, hydrophobicity, and low off-target toxicity. Unlike random or exhaustive combinatorial libraries, focused peptide libraries are constructed to maximize the probability of bioactivity while minimizing redundancy and undesired pharmacokinetic features.[170,171] In practice, this allows researchers to streamline screening efforts by working with a smaller, high-quality subset of candidates more likely to translate into therapeutic success. Furthermore, the incorporation of domain-specific constraints, such as protease resistance, membrane permeability, organ targeting, and endpoint-specificity into the library generation process allows these datasets to be aligned with specific applications.[172]

On the other hand, maximizing coverage and diversity in peptide chemical space is essential for discovering novel peptides, particularly those residing in underrepresented or unexplored regions. Strategies to enhance this exploration often involve generative and evolutionary algorithms designed to produce peptide libraries with broad structural and functional diversity. For example, Capecchi et al. proposed using genetic algorithms to populate peptide space by generating over one million unique sequences, revealing that evolutionary computation can effectively sample distant regions of sequence space that are inaccessible through traditional design methods.[17] In addition, sequence-based deep generative models have emerged as powerful tools for learning complex peptide sequence patterns while ensuring the generation of novel candidates with diverse scaffolds and biological potential.[173] These approaches collectively contribute to a more comprehensive exploration of peptide chemical space, supporting the discovery of functionally rich and previously overlooked molecules.

View Article Online
DOI: 10.1039/D5SC04465K

5. Perspectives and Outlook on Peptide Design

One of the major challenges in peptide design (Table 3) is the limited availability of standardized and curated datasets encompassing both canonical (cAAs) and non-canonical amino acids (ncAAs). Existing repositories often lack comprehensive structural, physicochemical, and bioactivity data, which restricts the exploration of peptides beyond conventional pharmacological applications. To address this gap, the development of open-access, high-quality peptide databases is critical. Such repositories should integrate data from diverse contexts, including nanomaterials, diagnostics, and biosensors, thereby enabling broader applications and more informed peptide design strategies. Additionally, the necessity to develop more efficient metrics and approaches which allow the systematic study of peptides continues to be one of the great contemporary challenges in drug design.[174,175]

Another significant hurdle involves incomplete information regarding peptide chirality and monomer configuration, particularly at asymmetric centers such as sulfoxides or hydroxyproline. Systematic annotation protocols incorporating chiral descriptors, for instance through MAP4c fingerprints,[176] could substantially improve the interpretability and predictive accuracy of *in silico* models. Ensuring detailed stereochemical information is essential for accurately modeling peptide folding, activity, and interactions.

The expansion of peptide chemical space remains a key opportunity, especially through the inclusion of ncAAs, post-translational modifications, and backbone alterations. Generating libraries of peptides with modified backbones, peptidomimetics, and macrocycles could reveal novel folding motifs and unique biological properties, opening new avenues for therapeutic and functional applications.

Synthetic challenges also persist, particularly for peptides containing ncAAs or chiral sulfur atoms, which are often difficult to incorporate with precision. Advancements in automated solid-phase synthesis and biotechnological platforms can facilitate the stereochemically defined incorporation of these building blocks, enabling the efficient production of complex peptide structures.[177]

Finally, the application of AI in peptide design presents both opportunities and challenges. Current predictive models are often incomplete or non-integrative, relying separately on structure-based or ligand-based approaches. Integrating these models within deep learning frameworks, such as graph neural networks or attention-based transformers, could generate consensus predictions with improved

View Article Online
DOI: 10.1039/D5SC04465K

interpretability.[178] Furthermore, employing explainable AI techniques will allow researchers to uncover key structural determinants that govern peptide function, stability, and folding, ultimately enhancing the rational design of bioactive peptides.[179]

Table 3. Future challenges and opportunities in peptide design.

Category	Challenge	Perspectives
Data availability	Lack of standardized and curated peptide datasets containing information about canonical (cAAs) and non-canonical amino acids (ncAAs).	Development of open-access, high-quality peptide repositories integrating structural, physicochemical, and bioactivity data across pharmacological and non-pharmacological contexts *e.g.*, nanomaterials, diagnostics, and biosensors.
	Incomplete information on peptide chirality and monomer configuration, including asymmetric centers (*e.g.*, sulfoxides, hydroxyproline).	Implementation of systematic annotation protocols including chiral descriptors *e.g.*, MAP4c fingerprint[176] to improve the interpretability and predictive accuracy of *in silico* models.
Data analysis	The quantity and diversity of available compounds are constantly and rapidly increasing.	The development of metrics and tools that enable the massive analysis of peptides will allow the correct identification of complex patterns associated with their biological activity.
Chemical Space Study	Limited exploration of the chemical space derived from ncAAs, post-translational modifications, and backbone alterations.	Generation of peptide libraries encompassing modified backbones, peptidomimetics, and macrocycles to discover new folding patterns and biological properties.
	Identify representative descriptors of the great chemical, physical, and biological diversity of peptides with cAA and ncAA.	The construction of complementary representations (*e.g.*, based on chemical multiverses) will facilitate the study of peptides from different perspectives. [180]
Synthesis of Canonical and Non-Canonical Peptides	Synthetic constraints for peptides containing ncAAs or chiral sulfur atoms.	Development of automated solid-phase synthesis and biotechnological platforms enabling the precise incorporation of stereochemically defined ncAAs.

View Article Online
DOI: 10.1039/D5SC04465K

AI-based Modeling Approaches	Fragmented use of structure-based or ligand-based predictive models without integration of multimodal data.	Coupling of structure- and ligand-based models with deep learning frameworks *e.g.*, graph neural networks, attention-based transformers, to generate consensus and interpretable predictions.
	Low interpretability of deep learning representations of peptide activity and folding.	Application of explainable AI techniques to uncover key structural determinants driving peptide function and stability.
	The activity of a peptide must be explained from a holistic perspective that integrates chemical, physical, and/or biological data.	The use of AI-based tools that allow for the correct fusion and interpretation of different types of data will aid in the correct decoding of peptide activity.

6. Conclusions

The peptide chemical space represents a vast and continuously expanding landscape shaped by the remarkable structural and functional diversity of peptides. This diversity arises not only from canonical amino acid sequences but also from the incorporation of non-canonical residues and post-translational modifications, which collectively generate an immense array of molecular architectures. Accurately navigating and characterizing the peptide chemical space demands a comprehensive suite of molecular descriptors, ranging from sequence-based and connectivity-driven features to 3D structural and physicochemical representations. The availability and integration of diverse descriptors are essential for designing, analyzing, and predicting peptide behavior across different applications. Furthermore, the growing number and complexity of peptide datasets[181-183] underscore the critical need for robust chemoinformatics and bioinformatics methodologies, whose synergy enables a deeper understanding of peptide function, structure–property (activity) relationships, and the rational design and generation of novel peptide libraries. Moving forward, the development of unified, open-source frameworks and consensus-driven computational standards will be pivotal in capturing the full extent of peptide chemical space and leveraging it for innovation in therapeutics, materials science, nutrition, cosmetics, and beyond.

View Article Online
DOI: 10.1039/D5SC04465K

Associated content

Data availability

Data sets are available in GitHub server at: https://github.com/EdgL2/PepChemSpace

Author contributions

Edgar López-López, Jean Paul Sánchez, Massyel S. Martinez-Cortés: investigation, writing - original draft, visualization, formal analysis, writing - review & editing; Cesar de la Fuente-Nunez: conceptualization, writing - review & editing. José L. Medin-Franco: conceptualization, investigation, resources, writing - review & editing, supervision.

Author information

Edgar López-López - *DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico*; *Department of Chemistry and Graduate Program in Pharmacology, Center for Research and Advanced Studies of the National Polytechnic Institute, Section 14-740, 07000 Mexico City, Mexico;* orcid.org/0000-0002-7422-6059; Email: elopez.lopez@cinvestav.mx

Jean Paul Sánchez - *DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico*; Email: jeanpaul.sac@comunidad.unam.mx

Massyel S. Martinez-Cortés - *DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico*; Email: mass.bqd@gmail.com

Cesar de la Fuente-Nunez - *Machine Biology Group, Departments of Psychiatry and Microbiology, Institute for Biomedical Informatics, Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania; Philadelphia, Pennsylvania, United States of America; Departments*

View Article Online
DOI: 10.1039/D5SC04465K

of Bioengineering and Chemical and Biomolecular Engineering, School of Engineering and Applied Science, University of Pennsylvania; Philadelphia, Pennsylvania, United States of America; Department of Chemistry, School of Arts and Sciences, University of Pennsylvania; Philadelphia, Pennsylvania, United States of America; Penn Institute for Computational Science, University of Pennsylvania; Philadelphia, Pennsylvania, United States of America; orcid.org/0000-0002-2005-5629; Email: cfuente@upenn.edu

José L. Medina-Franco - *DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico*; orcid/0000-0003-4940-1107; Email: jose.medina.franco@gmail.com; medinajl@unam.mx

Notes

Cesar de la Fuente-Nunez is a co-founder of, and scientific advisor, to Peptaris, Inc., provides consulting services to Invaio Sciences, and is a member of the Scientific Advisory Boards of Nowture S.L., Peptidus, European Biotech Venture Builder, the Peptide Drug Hunting Consortium (PDHC), ePhective Therapeutics, Inc., and Phare Bio. All other authors declare no competing financial interests.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

ELL thanks the Secretaria de la Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) for the PhD holder scholarship 762342 (No. CVU: 894234). We also thank the Direction General de Cómputo y de Tecnologías de la Información y Comunicación (DGTIC), UNAM, for the computational resources to use Miztli supercomputer at UNAM under project LANCAD-UNAM-DGTIC-335. Cesar de la Fuente-Nunez holds a Presidential Professorship at the University of Pennsylvania, is a recipient of the Langer Prize by the AIChE Foundation, and acknowledges funding from the IADR Innovation in Oral Care Award, the Procter & Gamble Company, United Therapeutics, a BBRF Young Investigator Grant, the Nemirovsky Prize, Penn Health-Tech Accelerator Award, the Dean's Innovation Fund from the Perelman School of

View Article Online
DOI: 10.1039/D5SC04465K

Medicine at the University of Pennsylvania, the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM138201, and the Defense Threat Reduction Agency (DTRA; HDTRA1-22-10031, HDTRA1-21-1-0014, and HDTRA1-23-1-0001).

References

1. J. L. Medina-Franco, A. L. Chávez-Hernández, E. López-López and F. I. Saldívar-González, *Molecular Informatics*, 2022, **4**, 2200116.
2. A. M. Virshup, J. Contreras-García, P. Wipf, W. Yang and D. N. Beratan, *Journal of the American Chemical Society*, 2013, **135**, 7296–7303.
3. J.-L. Reymond, *Journal of Cheminformatics*, 2025, **17**, 6.
4. B. J. Pepe-Mooney and R. Fairman, *Current Opinion in Structural Biology*, 2009, **19**, 483–494.
5. M. D. T. Torres, J. Cao, O. L. Franco, T. K. Lu and C. de la Fuente-Nunez, *ACS Nano*, 2021, **15**, 2143–2164.
6. P. E. Saw, X. Xu, S. Kim and S. Jon. *Accounts of Chemical Research*, 2021, **54**, 3576–3592.
7. I. W. Hamley, *Chemical Reviews*, 2017, **117**, 14015–14041.
8. K. Sato, *Journal of Agricultural and Food Chemistry*, 2018, **66**, 3082–3085.
9. Y. Tang, T. Nie, L. Zhang, X. Liu and H. Deng, *Cosmetics*, 2025, **12**, 107.
10. M. Muttenthaler, G. F. King, D. J. Adams and P. F. Alewood, *Nature Reviews Drug Discovery*, 2021, **20**, 309–325.
11. Y. Zhao, *Chemistry – A European Journal*, 2018, **24**, 14001–14009.
12. A. Capecchi and J.-L. Reymond, *Medicine in Drug Discovery*, 2021, **9**, 100081.
13. S. B. Kent, *Journal of Peptide Science*, 2025, **31**, e70013.
14. L. Chen, X. Xin, Y. Zhang, S. Li, X. Zhao, S. Li and Z. Xu, *Molecules*, 2023, **28**, 6745.
15. J. L. Hickey, D. Sindhikara, S. L. Zultanski and D. M. Schultz, *ACS Medicinal Chemistry Letters*, 2023, **14**, 557–565.
16. M. Orsi and J. Reymond, *Molecular Informatics*, 2025, **44**, e202400186.
17. A. Capecchi, A. Zhang and J.-L. Reymond, *Journal of Chemical Information and Modeling*, 2019, **60**, 121–132.

View Article Online
DOI: 10.1039/D5SC04465K

18. J. L. Warthen and M. J. Lueckheide, *Biomacromolecules*, 2024, **25**, 6923–6935.
19. C. M. Li, P. Haratipour, R. G. Lingeman, J. J. P. Perry, L. Gu, R. J. Hickey and L. H. Malkas, *Cells*, 2021, **10**, 2908.
20. O. A. Musaimi, D. AlShaer, B. G. de la Torre and F. Albericio, *Pharmaceuticals*, 2025, **18**, 291.
21. I. S. Johnson, *Science*, 1983, **219**, 632–637.
22. L. Diao and B. Meibohm, *Clinical Pharmacokinetics*, 2013, **52**, 855–868.
23. P. K. Lund, *Regulatory Peptides*, 2005, **128**, 93–96.
24. L. B. Knudsen and J. Lau, *Frontiers in Endocrinology*, 2019, **10**, 155.
25. L. Anthony and P. U. Freda, *Current Medical Research and Opinion*, 2009, **25**, 2989–2999.
26. J. Pless, *Journal of Endocrinological Investigation*. 2005, **28**,1–4.
27. E. W. Iepsen, S. S. Torekov and J. J. Holst, *Expert Review of Cardiovascular Therapy*, 2015, **13**, 753–767.
28. V. Iglesias, O. Bárcenas, C. Pintado.-Grima, M. Burdukiewicz and S. Ventura, *FEBS Open Bio*, 2025, **15**, 254-268.
29. T. D. Gladwell, *Clinical Therapeutics*, 2002, **24**, 38–58.
30. M. Coppens, J. W. Eikelboom, D. Gustafsson, J. I. Weitz and J. Hirsh, *Circulation Research*, 2012, **111**, 920–929.
31. R. M. Scarborough, *American Heart Journal*, 1999, **138**, 1093–1104.
32. S. Yuan, Y. Q. Luo, J. H. Zuo, H. Liu, F. Li and B. Yu, *European Journal of Medicinal Chemistry*, 2021, **215**, 113284.
33. K. Clément, E. van den Akker, J. Argente, A. Bahm, W. K. Chung, H. Connors, K. De Waele, I. S. Farooqi, J. Gonneau-Lejeune, G. Gordon, K. Kohlsdorf, C. Poitou, L. Puder, J. Swain, M. Stewart, G. Yuan, M. Wabitsch and P. Kühnen, *Lancet Diabetes & Endocrinology*, 2020, **8**, 960–970.
34. L. Costa and C. Fernandes, *Drugs and Drug Candidates*, 2024, **3**, 311–327.
35. S. Li, Y. Li, Y. Li, Y. Wu, Q. Wang, L. Jin and D. zhang, *International Journal of Molecular Sciences*, 2023, **24**, 8642.
36. T. L. Silva, G. G. Barbosa, C. J. C. de Santana, P. M. G. Paiva, M. S. Castro and T. H. Napoleão, *Chemistry*, 2024, **6**, 333–344.

View Article Online
DOI: 10.1039/D5SC04465K

37. L. Sukmarini, *Molecules*, 2022, **27**, 2619.

38. V. Gogineni and M. T. Hamann, *Biochimica et Biophysica Acta. General subjects*, 2018, **1862**, 81–196.

39. Y. Lee, C. Phat. and S. C. Hong, *Peptides*, 2017,**95**, 94–105.

40. F. Yang and Y. Ma, *Amino acids*, 2024, **56**, 1–12.

41. A. Qureshi, *Discover Viruses*, 2025, **2**, 1–13.

42. Y. Yang, L. Huang, Z. Huang, Y. Ren, Y. Xiong, Z. Xu and Y Chi, *Critical Reviews in Food Science and Nutrition*, 2024, **65**, 3186–3207.

43. A. A. Zaky, J. Simal.-Gandara, J. B. Eun, J. H. Shim and A. M. Abd El.-Aty, *Frontiers in Nutrition*, 2022, **8**, 815640.

44. R. Kumari, S. Sanjukta, D. Sahoo and A. K. Rai, *Systems Microbiology and Biomanufacturing*, 2021, **2**, 1–13.

45. M. Chalamaiah, W. Yu, W and J. Wu, *Food Chemistry*, 2018, **245**, 205–222.

46. A. Ashraf, Y. Guo, T. Yang, A. S. ud Din, K. Ahmad, K. Li and ou H. Hou, *Journal of Agricultural and Food Chemistry*, 2025, **73**, 1000–1013.

47. L. Gu, Y. Tang, J. Zhang, N. Tao, X. Wang, L. Wang and C. Xu, *Foods*, 2025,**14**, 406.

48. Y. Eilam, H. Khattib, N. Pintel and D. Avni, *Global Challenges*, 2023, **7**, 2200177.

49. K. Chen, D. Ma, X. Yang, P. Liu, J. Wang and W. Liao, *Food Frontiers*, 2024, **5**, 2483–2497.

50. A. Pintea, A. Manea, C. Pintea, R.-A. Vlad, M. Bîrsan, P. Antonoaea, E. M. Rédai and A. Ciurba, *Biomolecules*, 2025, **15**, 88.

51. C. L. Burnett. W. F. Bergfeld, D. V. Belsito, R. A. Hill, C. D. Klaassen, D. C. Liebler, J. G. Marks, R, C. Shank. T. J. Slaga, P. W. Snyder, M. Fiume and B. Heldreth, *International Journal of Toxicology*, 2023, **42**, 102S–113S.

52. L. Wang, Z. Wu, X. Wang, X. Wang, J. Mao, Y. Yan, L. Zhang and Z. Zhang, *Journal of Peptide Science*, 2025, **31**, e3668.

53. N. S. Murthy, E. Tavasoli, M. T. Mammone and N. Karaman.-Jurukocska, *International Journal of Cosmetic Science*, 2025, **47**, 554-562.

54. D. J. Rodi, A. S. Soares and L. Makowski, *Journal of Molecular Biology*, 2002, **322**, 1039–1052.

View Article Online
DOI: 10.1039/D5SC04465K

55. N. van Walraven, R. J. FitzGerald, H. J. Danneel and M. Amigo-Benavent, *Peptides*, 2025, **193**, 171440.
56. M. M. Abeer, S. Trajkovic and D. J. Brayden, *Biomedicine & Pharmacotherapy*, 2021, **144**, 112275.
57. R. Liu, A. M. Enstrom and K. S. Lam, *Experimental Hematology*, 2003, **31**, 11–30.
58. T. Uhlig, T. Kyprianou, F. G. Martinelli, C. A. Oppici, D. Heiligers, D. Hills, X. R. Calvo and P. Verhaert, *EuPA Open Proteomics*, 2014, **4**, 58–69.
59. K. R. Clauser, P. Baker and A. L. Burlingame, *Analytical Chemistry*, 1999, **71**, 2871–2882.
60. J. Kim, H. Kim and S. B. Park, *Journal of the American Chemical Society*, 2014, **136**, 14629–14638.
61. A. Flissi, E. Ricart, C. Campart, M. Chevalier, Y. Dufresne, J. Michalik, P. Jacques, C. Flahaut, F. Lisacek, V. Leclère and M. Pupin, *Nucleic Acids Research*, 2020, **48**, D465–D469.
62. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Research*, 2000, **28**, 235–242.
63. F. Desiere, E. W. Deutsch, N. L. King, A. I. Nesvizhskii, P. Mallick, J. Eng, S. Chen, J. Eddes, S. N. Loevenich and R. Aebersold, *Nucleic Acids Research*, 2006, **34**, D655–D658.
64. The UniProt Consortium, *Nucleic Acids Research*, 2025, **53**, D609–D617.
65. C. Quiroz, Y. B. Saavedra, B. Armijo-Galdames, J. Amado-Hinojosa, Á. Olivera-Nappa, A. Sanchez-Daza and D. Medina-Ortiz, *Database*, 2021, **2021**, baab055.
66. P. J. Linstrom and W. G. Mallard, *Journal of Chemical and Engineering Data*, 2001, **46**, 1059–1063.
67. R. Vita, N. Blazeska, D. Marrama, IEDB Curation Team Members, S. Duesing, J. Bennett, J. Greenbaum, M. De Almeida Mendes, J. Mahita, D. K. Wheeler, J. R. Cantrell, J. A. Overton, D. A. Natale, A. Sette and B. Peters, *Nucleic Acids Research*, 2025, **53**, D436–D443.
68. L. Lautenbacher, P. Samaras, J. Muller, A. Grafberger, M. Shraideh, J. Rank, S. T. Fuchs, T. K. Schmidt, M. The, C. Dallago, H. Wittges, B. Rost, H. Krcmar, B. Kuster, M. Wilhelm, *Nucleic Acids Research*, 2022, **50**, D1541–D1552.
69. Y. Perez-Riverol, C. Bandla, D. J. Kundu, S. Kamatchinathan, J. Bai, S. Hewapathirana, N. S. John, A. Prakash, M. Walzer, S. Wang and J. A. Vizcaíno, *Nucleic Acids Research*, 2025, **53**, D543–D553.
70. S. Singh, K. Chaudhary, S. Kumar Dhanda, S. Bhalla, S. Sadullah Usmani, A. Gautam, A. Tuknait, P. Agrawal, D. Mathur and G. P. S. Raghava, *Nucleic Acids Research*, 2016, **44**, D1119–D1126.

View Article Online
DOI: 10.1039/D5SC04465K

71. T. Ma, Y. Liu, B. Yu, X. Sun, H. Yao, C. Hao, J. Li, M. Nawaz, X. Jiang, X. Lao, H. Zheng, *Nucleic Acids Research*, 2025, **53**, D403–D410.
72. M. Pirtskhalava, A. A. Amstrong, M. Grigolava, M. Chubinidze, E. Alimbarashvili, B. Vishnepolsky, A. Gabrielian, A. Rosenthal, D. E. Hurt and M. Tartakovsky, *Nucleic Acids Research*, 2021, **49**, D288–D297.
73. U. Gawde, S. Chakraborty, F. Hanif Waghu, R. Shankar Barai, A. Khanderkar, R. Indraguru, T. Shirsat and S. Idicula-Thomas, *Nucleic Acids Research*, 2023, **51**, D377–D383.
74. P. M. Martins, L. H. Santos, D. Mariano, F. C. Queiroz, L. L. Bastos, I. D. S. Gomes, P. H. C. Fischer, R. E. O. Rocha, S. A. Silveira, L. H. F. de Lima, M. T. Q. de Magalhães, M. G. A. Oliveira and R. C. de Melo-Minardi, *BMC Bioinformatics*, 2021, **22**, 1, 2021.
75. S. Bhalla, R. Verma, H. Kaur, R. Kumar, S. S. Usmani, S. Sharma and G. P. Raghava, *Scientific Reports*, 2017, **7**, 1511.
76. Z. Wen, J. He, H. Tao and S. Y. Huang, *Bioinformatics*, 2019, **35**, 175–177.
77. N. Zhu, F. Dong, G. Shi, X. Lao, and H. Zheng, *Scientific Data*, 2022, **9**, 187.
78. M. Chauhan, A. Gupta, R. Tomer and G. P. Raghava, *Database*, 2025, **2025**, baaf030.
79. J. Yu, L. Wang, X. Kong, Y. Cao, M. Zhang, Z. Sun, Y. Liu, J. Wang, B. Shen, X. Bo, and J. Feng, *Frontiers in Bioengineering and Biotechnology*, 2022, **10**, 819583.
80. G. Wang, C. Schmidt, X. Li and Z. Wang, *Nucleic Acids Research*, 2025, gkaf860.
81. M. Wang, L. Wang, W. Xu, Z. Chu, H. Wang, J. Lu, Z. Xue, and Y. Wang, *Journal of Molecular Biology*, 2024, **436**, 168416.
82. M. Novković, J. Simunić, V. Bojović, A. Tossi and D. Juretić, *Bioinformatics*, 2012, **28**, 1406–1407.
83. D. Mathur, S. Prakash, P. Anand, H. Kaur, P. Agrawal, A. Mehta, R. Kumar, S. Singh and G. P. S. Raghava, *Scientific Reports*, 2016, **6**, 36617.
84. S. Caboche, M. Pupin, V. Leclere, A. Fontaine, P. Jacques and G. Kucherov, *Nucleic Acids Research*, 2007, **36**, D326–D331.
85. S. D. H. Nielsen, N. Liang, H. Rathish, B. J. Kim, J. Lueangsakulthai, J. Koh, Q. Yunyao, N. Søren Drud-Heydary, and D. C. Dallas, *Critical Reviews in Food Science and Nutrition*, 2023, **64**, 11510–11529.

View Article Online
DOI: 10.1039/D5SC04465K

86. S. S. Usmani, G. Bedi, J. S. Samuel, S. Singh, S. Kalra, P. Kumar, A. Arora Ahuja, M. Sharma, et al., *PLOS ONE,* 2017, **12**, e0181748.

87. P. Minkiewicz, A. Iwaniak and M. Darewicz, *Applied Sciences*, 2022, **12**, 7204.

88. J.-L. Reymond and M. Awale, *ACS Chemical Neuroscience*, 2012, **3**, 649–657.

89. C. Petrou and Y. Sarigiannis, in *Peptide Applications in Biomedicine, Biotechnology and Bioengineering*, Elsevier, 2018, pp. 1–21.

90. M. Paradís-Bas, J. Tulla-Puche and F. Albericio, *Chemical Society Reviews*, 2016, **45**, 631–654.

91. C. Wynne and R. B. P. Elmes, *Sensors & Diagnostics*, 2024, **3**, 987–1013.

92. K. Z. Q. Zhou and R. Obexer, *Israel Journal of Chemistry*, 2024, **64**, e202400006.

93. Y. Ding, J. P. Ting, J. Liu, S. Al-Azzam, P. Pandya and S. Afshar. *Amino Acids*, 2020, **52**, 1207–1226.

94. R. Gillane, D. Daygon, Z. G. Khalil and E. Marcellin, *Frontiers in Bioengineering and Biotechnology*, 2024, **12**, 1468974.

95. M. Fleck and A. M. Petrosyan, (2014). Amino Acid Structures. In Salts of Amino Acids: Crystallization, Structure and Properties (pp. 21–82). Cham: Springer International Publishing.

96. A. Galbiati, A. Zana, C. Borsari, M. Persico, S. Bova, O. Tkachuk, A. I. Corfu, L. Tamborini, N. Basilico, C. Fattorusso, S. Bruno, S. Parapini and P. Conti, *Molecules*, 2023, **28**, 3172.

97. P. M. Fischer, *Current Protein and Peptide Science*, 2003, **4**, 339–356.

98. S. Lohan, A.G. Konshina, E. H. M. Mohammed, N. M. Helmy, S. K. Jha, R. Kumar Tiwari, I. Maslennikov, R. G. Efremov and K. Parang, *npj Antimicrobials and Resistance,* 2025, **3**, 56.

99. M. W. Ishaq, P. Farzeen, L. R. Vaughn, D. J. Stone, S. A. Deshmukh and C. E. Callmann, *ACS Central Science*, 2025, **11**, 1573–1580.

100. R. S. Bon and H. Waldmann, *Accounts of Chemical Research*, 2010, **43**, 1103–1114.

101. R. Saklani and A. J. Domb, *ACS Omega*, 2024, **9**, 17726–17740.

102. L. Chen, X. Xin, Y. Zhang, S. Li, X. Zhao, S. Li and Z. Xu, *Molecules*, 2023, **28**, 6745.

103. H. Zou, L. Li, T. Zhang, M. Shi, N. Zhang, J. Huang and M. Xian, *Biotechnology Advances*, 2018, **36**, 1917–1927.

104. T. G. Castro, M. Melle-Franco, C. E. A. Sousa, A. Cavaco-Paulo and J. C. Marcos, *Biomolecules*, 2023, **13**, 981.

View Article Online
DOI: 10.1039/D5SC04465K

105. M. Rother and J. A. Krzycki, *Archaea*, 2010, **2010**, 1–14.
106. T. Lugtenburg, A. Gran-Scheuch and I. Drienovská, *Protein Engineering, Design and Selection*, 2023, **36**, gzad003.
107. S. Wang, A. O. Osgood and A. Chatterjee, *Current Opinion in Structural Biology*, 2022, **74**, 102352.
108. F. Ardito, M. Giuliani, D. Perrone, G. Troiano and L. L. Muzio, *International Journal of Molecular Medicine*, 2017, **40**, 271–280.
109. M. R. Larsen and A. V. G. Edwards, in *Encyclopedia of Cell Biology*, Elsevier, 2016, pp. 177–186.
110. J. M. Lee, H. M. Hammarén, M. M. Savitski and S. H. Baek, *Nature Communications*, 2023,**14**, 201.
111. R. Matsumoto, D. Wang, M. Blonska, H. Li, M. Kobayashi, B. Pappu, Y. Chen, D. Wang and X. Lin, *Immunity*, 2005, **23**, 575–585.
112. M. L. Newby, J. D. Allen and M. Crispin, *Biotechnology Advances*, 2024, **70**, 108283.
113. B. Emenike, O. Nwajiobi and M. Raj, *Frontiers in Chemistry*, 2022, **10**, 868773.
114. S. Bondalapati, M. Jbara and A. Brik, *Nature Chemistry*, 2016, **8**, 407–418.
115. S. Duengo, M. I. Muhajir, A. T. Hidayat, W. J. A. Musa and R. Maharani, *Molecules*, 2023, **28**, 8017.
116. S. S. Kulkarni, J. Sayers, B. Premdjee and R. J. Payne, *Nature Reviews Chemistry*, 2018, **2**, 0122.
117. A. Saha, H. Suga and A. Brik, *Accounts of Chemical Research*, 2023, **56**, 1953–1965.
118. O. Harel and M. Jbara, *Molecules*, 2022, **27**, 4389.
119. O. M. Lateef, M. O. Akintubosun, O. T. Olaoba, S. O. Samson and M. Adamczyk, *International Journal of Molecular Sciences*, 2022, **23**, 938.
120. J. Xie and P. G. Schultz, *Nature Reviews Molecular Cell Biology*, 2006, **7**, 775–782.
121. X. Jia, Y. K.-Y. Chin, A. H. Zhang, T. Crawford, Y. Zhu, N. L. Fletcher, Z. Zhou, B. R. Hamilton, M. Stroet, K. J. Thurecht and M. Mobli, *Communications Chemistry*, 2023, **6**, 48.
122. M. A. Siani, D. Weininger and J. M. Blaney, *Journal of Chemical Information and Computer Sciences*, 1994, **34**, 588–593.
123. W. L. Chen, B. A. Leland, J. L. Durant, D. L. Grier, B. D. Christie, J. G. Nourse and K. T. Taylor, *Journal of Chemical Information and Modeling*, 2011, **51**, 2186–2208.
124. A. Capecchi, D. Probst and J. L. Reymond, *Journal of Cheminformatics*, 2020, **12**, 43.
125. A. Shendre, N. K. Mehta, A. S. Rathore, N. Kumar, S. Patiyal and G. P. Raghava, ArXiv preprint,

View Article Online
DOI: 10.1039/D5SC04465K

2025, 2505.03403.

126. D. Rogers and M. Hahn, *Journal of Chemical Information and Modeling*, 2010, **50**, 742–754.

127. T. Zhang, H. Li, H. Xi, R. V. Stanton and S. H. Rotstein, *Journal of Chemical Information and Modeling*, 2012, **52**, 10, 2796–2806.

128. Biochemfusion, Protein Line Notation (PLN), http://www.biochemfusion.com/doc/Biochemfusion_PLN_1.4_spec.pdf, accessed December 1st, 2025.

129. P. Zhou, Y. Zhou, S. Wu, B. Li, F. Tian and Z. Li, *Chinese Science Bulletin*, 2006, **51**, 524–529.

130. A. Capecchi, M. Awale, D. Probst, and J. L. Reymond, *Molecular Informatics*, 2019, **38**, 1900016.

131. C. Dietz and M. R. Berthold, (2016). KNIME for Open-Source Bioimage Analysis: A Tutorial. In: De Vos, W., Munck, S., Timmermans, JP. (eds) Focus on Bio-Image Informatics. Advances in Anatomy, Embryology and Cell Biology, vol 219. Springer, Cham. https://doi.org/10.1007/978-3-319-28549-8_7

132. T. Sander, J. Freyss, M. Von Korff and C. Rufener, *Journal of Chemical Information and Modeling*, 2015, **55**, 460–473.

133. E. López-López, J. J. Naveja and J. L. Medina-Franco, *Expert Opinion on Drug Discovery*, 2019, **14**, 335–341.

134. G. Geylan, J. P. Janet, A. Tibo, J. He, A. Patronov, M. Kabeshov, W. Czechtizky, F. David, O. Engkvist and L. De Maria, *Chemical Science*, 2025, **16**, 8682–8696.

135. NovoPro, PepSMI: Convert Peptide to SMILES string, https://www.novoprolabs.com/tools/prot-sol, accessed December 1st, 2025.

136. H. Nielsen, F. Teufel, S. Brunak and G. von Heijne, (2024). SignalP: the evolution of a web server. In Protein Bioinformatics (pp. 331–367). New York, NY: Springer US.

137. R. Gurdeep Singh, A. Tanca, A. Palomba, F. Van der Jeugt, P. Verschaffelt, S. Uzzau, S., L. Martens, P. Dawyndt and B. Mesuere, *Journal of Proteome Research*, 2019, **18**, 606–615.

138. T. Fox, M. Bieler, P. Haebel, R. Ochoa, S. Peters and A. Weber, *Journal of Chemical Information and Modeling*, 2022, ***62***, 3942–3947.

139. J. H. Jensen, T. Hoeg-Jensen and S. B. Padkjær, *Journal of Chemical Information and Modeling*, 2008, **48**, 2404–2413.

View Article Online
DOI: 10.1039/D5SC04465K

140. W. Zhou, Y. Liu, Y. Li, S. Kong, W. Wang, B. Ding, J. Han, C. Mou, X. Gao and J. Liu, *Patterns*, 2023, **4**, 100702.

141. C. Li, L. Mora and F. Toldrá, *Food Chemistry*, 2022, **375**, 131673.

142. Z. Chen, X. Liu, P. Zhao, C. Li, Y. Wang, F. Li, T. Akutsu, C. Bain, R. B. Gasser, J. Li, Z. Yang, X. Gao, L. Kurgan and J. Song, *Nucleic Acids Research*, 2022, **50**, W434–W447.

143. E. López-López, O. Robles, F. Plisson and J. L. Medina-Franco, *Digital Discovery*, 2023, **2**, 1494–1505.

144. J. R. M. A. Maasch, M. D. T. Torres, M. C. R. Melo and C. de la Fuente-Nunez, *Cell Host & Microbe*, 2023, **31**, 1260-1274.e6.

145. F. Wan, M. D. T. Torres, J. Peng and C. de la Fuente-Nunez, *Nature Biomedical Engineering*, 2024, **8**, 854–871.

146. E. López-López and J. L. Medina-Franco, *Biomolecules*, 2023, **13**, 176.

147. K. O. Osiro, A. Gil-Ley, F. C. Fernandes, K. B. S. de Oliveira, C. de la Fuente-Nunez and O. L. Franco, *Microbial Cell*, 2025, **12**, 1.

148. K. Chaudhary, R. Kumar, S. Singh, A. Tuknait, A. Gautam, D. Mathur, P. Anand, G. C. Varshney and G. P. S. Raghava, *Scientific Reports*, 2016, **6**, 22843.

149. L. Aguilera-Mendoza, Y. Marrero-Ponce, C. R. García-Jacas, E. Chavez, J. A. Beltran, H. A. Guillen-Ramirez and C. A. Brizuela, *Scientific Reports*, 2020, **10**, 18074.

150. E. C. L. de Oliveira, K. Santana, L. Josino, A. H. Lima e Lima and C. de Souza de Sales Júnior, *Scientific Reports*, 2021, **11**, 7628.

151. C. Guntuboina, A. Das, P. Mollaei, S. Kim and A. Barati Farimani, *The Journal of Physical Chemistry Letters*, 2023, **14**, 10427–10434.

152. N. M. Tripathi and A. Bandyopadhyay, *European Journal of Medicinal Chemistry*, 2022, **243**, 114766.

153. A. Stephenson, L. Lastra, B. Nguyen, Y. J. Chen, J. Nivala, L. Ceze and K. Strauss, *ACS Synthetic Biology*, 2023, **12**, 3156–3169.

154. C. Qian, B. Niu, R. B. Jimenez, J. Wang and M. Albarghouthi, *Journal of Pharmaceutical and Biomedical Analysis*, 2021, **198**, 113988.

155. K. N. Amarasinghe, L. De Maria, C. Tyrchan, L. A. Eriksson, J. Sadowski and D. Petrovic, *Journal of*

View Article Online
DOI: 10.1039/D5SC04465K

Chemical Information and Modeling, 2022, **62**, 2999–3007.

156. W. Zhao, L. Su, S. Huo, Z. Yu, J. Li and J. Liu, *Food Science and Human Wellness*, 2023, **12**, 89–93.
157. L. Chang, A. Mondal, B. Singh, Y. Martínez-Noa and A. Perez, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2024, **14**, e1693.
158. N. van Hilten, J. Methorst, N. Verwei and H. J. Risselada, *Science Advances*, 2023, **9**, eade8839.
159. G. Rosenman, P. Beker, I. Koren, M. Yevnin, B. Bank-Srour, E. Mishina and S. Semin, *Journal of Peptide Science*, 2011, **17**, 75–87.
160. F. X. B. de Thé, C. Baudier, R. A. Pereira, C. Lefebvre, P. Moingeon and Patrimony Working Group, *Drug Discovery Today*, 2023, **28**, 103772.
161. S. Zhai, T. Liu, S. Lin, D. Li, H. Liu, X. Yao and T. Hou, *Drug Discovery Today*, 2025, **30**, 104300.
162. S. Chen, T. Lin, R. Basu, J. Ritchey, S. Wang, Y. Luo, X. Li, D. Pei, L. B. Kara and X. Cheng, *Nature Communications*, 2024, **15**, 1611.
163. Y. Wu, J. Williams, E. D. D. Calder and L. J. Walport, *RSC Chemical Biology*, 2021, **2**, 151–165.
164. M. Bartoloni, R. U. Kadam, J. Schwartz, J. Furrer, T. Darbre and J. L. Reymond, *Chemical Communications*, 2011, **47**, 12634-12636.
165. F. Wan, D. Kontogiorgos-Heintz and C. de la Fuente-Nunez, *Digital Discovery*, 2022, **1**, 195–208.
166. H. Yu, R. Wang, J. Qiao and L. Wei, *Journal of Chemical Information and Modeling*, 2023, **64**, 316–326.
167. L. Wang, Y. Liu, X. Fu, X. Ye, J. Shi, G. G. Yen, Q. Zou, X. Zeng and D. Cao, *Journal of Medicinal Chemistry*, 2025, **68**, 8346–8360.
168. H. Khabbaz, M. H. Karimi-Jafari, A. A. Saboury and B. BabaAli, *BMC Bioinformatics*, 2021, **22**, 1–11.
169. J.-H. Wang and T.-Y. Sung, *ACS Omega*, 2024, **9**, 32116–32123.
170. K. Bozovičar, B. Jenko Bizjan, A. Meden, J. Kovač and T. Bratkovič, *Scientific Reports*, 2021, **11**, 11650.
171. C. Sohrabi, A. Foster and A. Tavassoli, *Nature Reviews Chemistry*, 2020, **4**, 90–101.
172. M. Muttenthaler, G. F. King, D. J. Adams and P. F. Alewood, *Nature Reviews Drug Discovery*, 2021, **20**, 309–325.
173. M. Mardikoraem, Z. Wang, N. Pascual and D. Woldring, *Briefings in Bioinformatics*, 2023, **24**,

View Article Online
DOI: 10.1039/D5SC04465K

bbad358.

174. M.D.T Torres, E.F. Brooks, A. Cesaro, H. Sberro, M. O. Gill, C. Nicolaou, A. S. Bhatt and C. de la Fuente-Nunez, *Cell*, 2024, **187**, 5453–5467.

175. K. L. Perez, E. López-López, F. Soulage, E. Felix, J. L. Medina-Franco and R. A. Miranda-Quintana, *Journal of Chemical Information and Modeling,* 2025, **65**, 6788–6796.

176. M. Orsi and J. L. Reymond, *Journal of Cheminformatics*, 2024, **16**, 53.

177. S. B. Kent, *Journal of Peptide Science*, 2025, **31**, e70013.

178. J. Bajorath, A. L. Chávez-Hernández, M. Duran-Frigola, E. Fernández-de Gortari, J. Gasteiger, E. López-López, G. M. Maggiora, J. L. Medina-Franco, O. Méndez-Lucio, J. Mestres, R. A. Miranda-Quintana, T. I. Oprea, F. Plisson, F. D. Prieto-Martínez, R. Rodríguez-Pérez, P. Rondón-Villarreal, F. I. Saldívar-Gonzalez, N. Sánchez-Cruz and M. Valli, *Journal of Cheminformatics*, 2022, **14**, 82.

179. E. López-López and J. L. Medina-Franco, *Drug Discovery Today*, 2024, **29**, 104046.

180. J. L. Medina-Franco, A. L. Chávez-Hernández, E. López-López and F. I. Saldívar-González, *Molecular Informatics*, 2022, **41**, 2200116.

181. F. Wan, M. D. T. Torres, J. Peng and C. de la Fuente-Nunez, *Nature Biomedical Engineering*, 2024, **8**, 854–871.

182. C. D. Santos-Júnior, M. D. T. Torres, Y. Duan, A. Rodríguez Del Río, T. S. B. Schmidt, H. Chong, A. Fullam, M. Kuhn, C. Zhu, A. Houseman, J. Somborski, A. Vines, X. M. Zhao, P. Bork, J. Huerta-Cepas, C. de la Fuente-Nunez and L. P. Coelho, *Cell,* 2024, **187**, 3761–3778.

183. X. M. Zhao, P. Bork, J. Huerta-Cepas, C. de la Fuente-Nunez and L. P. Coelho, *Cell,* 2024, **187**, 3761–3778.

View Article Online
DOI: 10.1039/D5SC04465K

Data sets are available in GitHub server at: https://github.com/EdgL2/PepChemSpace