PCCP
Physical Chemistry Chemical Physics

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: E. Cristhian Lima de Oliveira, J. A. S. Feio, G. Pereira Coelho, L. Diniz Do Nascimento, A. De Spiegeleer, C. Sales, A. H. Lima, C. M. F. Rodrigues, E. Wynendaele, B. De Spiegeleer and K. Costa, *Phys. Chem. Chem. Phys.*, 2026, DOI: 10.1039/D5CP04611D.

DOI: 10.1039/D5CP04611D

Navigating in the Chemical Space of Peptides: Computational Strategies and Molecular Features to Unveil Their Functional and Drug-like Properties

Ewerton Cristhian Lima de Oliveira[1,*] Juliana Auzier[2,*], Gabriel Pereira Coelho[3], Lidiane Diniz do Nascimento,[3] Anderson Henrique Lima e Lima,[4] Caio Marcos Flexa Rodrigues,[3] Anton De Spiegeleer[6,7], Evelien Wynendaele[5,7], Claudomiro Sales[1,*], Bart De Spiegeleer[5,7,α], Kauê Santana[2,β]

1 Instituto Tecnológico Vale, 66055-090 Belém, Pará, Brasil.

2 Laboratório de Inteligência Computacional e Pesquisa Operacional, Campos Belém, Instituto de Tecnologia, Universidade Federal do Pará, 66075-110 Belém, Pará, Brasil

3 Laboratório de Simulação Computacional, Campus Santarém, Instituto de Biodiversidade, Universidade Federal do Oeste do Pará, 68.040-070, Santarém, Pará, Brasil.

4 Laboratório de Planejamento e Desenvolvimento de Fármacos. Instituto de Ciências Exatas e Naturais, Universidade Federal do Pará, 66075-110, Belém, Pará, Brasil.

5 Drug Quality and Registration (DruQuaR) group, Faculty of Pharmaceutical Sciences, Ghent, University, Ottergemsesteenweg 460, B-9000 Ghent, Belgium.

6 Department of Geriatrics, Faculty of Medicine and Health Sciences, Ghent University Hospital, Ghent, Belgium.

7 Translational Research in Immunosenescence, Gerontology and Geriatrics (TRIGG) group, Ghent University Hospital, Ghent, Belgium.

α, β Corresponding authors:

α: Bart De Spiegeleer (Bart.DeSpiegeleer@ugent.be); β: Kauê Santana (kaue.costa@ufopa.edu.br)

ORCID of the authors:

Juliana Auzier: 0009-0009-9781-5911, Gabriel Pereira Coelho: 0009-0004-6029-5557, Lidiane Diniz: 0000-0003-1370-4472, Ewerton de Oliveira: 0000-0002-2338-7178, Anderson Henrique Lima e Lima: 0000-0002-8451-9912, Caio Marcos Flexa Rodrigues: 0009-0006-9456-2718, Anton De Spiegeleer: 0000-0002-3681-2807, Evelien Wynendaele: 0000-0003-0969-7580. Claudomiro Sales: 0000-0002-2735-1383, Bart De Spiegeleer: 0000-0001-6794-3108, Kauê Santana: 0000-0002-2735-8016.

View Article Online
DOI: 10.1039/D5CP04611D

Abstract

Peptides, short chains of amino acids linked by peptide bonds, typically ranging from 2 to 50 residues, are fundamental to diverse biological processes and represent a valuable source for the development of novel bioactive compounds. In this work, we provide a comprehensive and conceptual overview of approaches to exploring the peptide chemical space. We emphasize intrinsic challenges in their chemical space investigation, particularly the complex interplay among peptide conformation, bioactivity, and bioavailability, as well as the role of sequence- and structure-derived molecular features in elucidating structure-activity relationships. Furthermore, we examine computational strategies, such as dimensionality reduction techniques, machine learning models, and similarity-based complex networks for classifying and characterizing this chemical space. Finally, we underscore the importance of interdisciplinary frameworks in advancing peptide research, highlighting how integrative approaches can uncover intersections of bioactivity across different peptide classes and leverage alternative chemical spaces to optimize and characterize peptide structures.

Keywords: chemical space, peptide prediction, molecular descriptors compendium, molecular fingerprints, structure-activity relationships

View Article Online
DOI: 10.1039/D5CP04611D

Introducing the Main Concepts of Chemical Spaces and Their Applications to Peptide Science

The term ‘chemical space’ is used in two complementary ways. In a broad conceptual sense, it refers to the theoretical universe of chemically feasible molecules. In practical chemoinformatics, however, chemical space is operationally defined for a given dataset by representing molecules as vectors in a multidimensional descriptor or fingerprint space, where inter-compound distances reflect similarity relationships within the chosen representation.[1,2]. The choice of representation, i.e., how the peptides are encoded into a form that computers can handle, influences which analyses are efficiently possible.[3] These molecular descriptors or fingerprints encode structural and physicochemical properties, enabling visualization (often via 2D/3D projections) and analysis of structure–property/activity relationships relevant to bioactivity, encompassing pharmacodynamics and pharmacokinetics.[4] In this work, we use ‘chemical space’ primarily in this operational, dataset-driven sense, as it enables visualization and analysis of structure–property/activity relationships for peptide sets.

Determining the chemical space of molecules aims to classify compounds, identify potential bioactive molecules, design and improve lead candidates, and understand their molecular properties.[5,6] Independent of the method, mapping the chemical space can significantly enhance the efficiency of novel discoveries across various areas of chemistry, including chemical synthesis[7–9], quantum chemistry,[10] materials science,[11–13] and drug discovery.[5,14,15]

The chemical space of compounds can be explored by analyzing libraries of organic molecules using two-dimensional (2D) or three-dimensional (3D) visual representations of multidimensional descriptor spaces plotted in Cartesian coordinates, which often require dimensionality-reduction or clustering techniques.[16,17]

Figure 1 illustrates the application of a computational workflow to map the chemical space of peptides. Chemical space mapping integrates heterogeneous peptide inputs — either sequence-based representations (primary structure) or 3D structural models (tertiary structure) — to derive informative molecular descriptors. The visual exploration of peptide chemical space can be performed

View Article Online
DOI: 10.1039/D5CP04611D

using computational tools that efficiently provide a visual correlation between molecules exhibiting similar chemical and/or functional properties[18–20]. The most straightforward approach to accomplish this task involves the development of pipelines that integrate: the calculation of molecular descriptors using chemical packages, such as RDKit[21], iFeature[22,23], and Mordred[24]; the application of feature correlations tests; and finally clustering methods, such as K-means and density-based spatial clustering of applications with noise (DBSCAN); as well as dimensionality reduction techniques including t-distributed stochastic neighbor embedding (t-SNE)[25] and uniform manifold approximation and projection (UMAP)[26]; and finally the graphical representations for the 2D and 3D visualization of the numerical representations generated from these projections.[18,20]

View Article Online
DOI: 10.1039/D5CP04611D

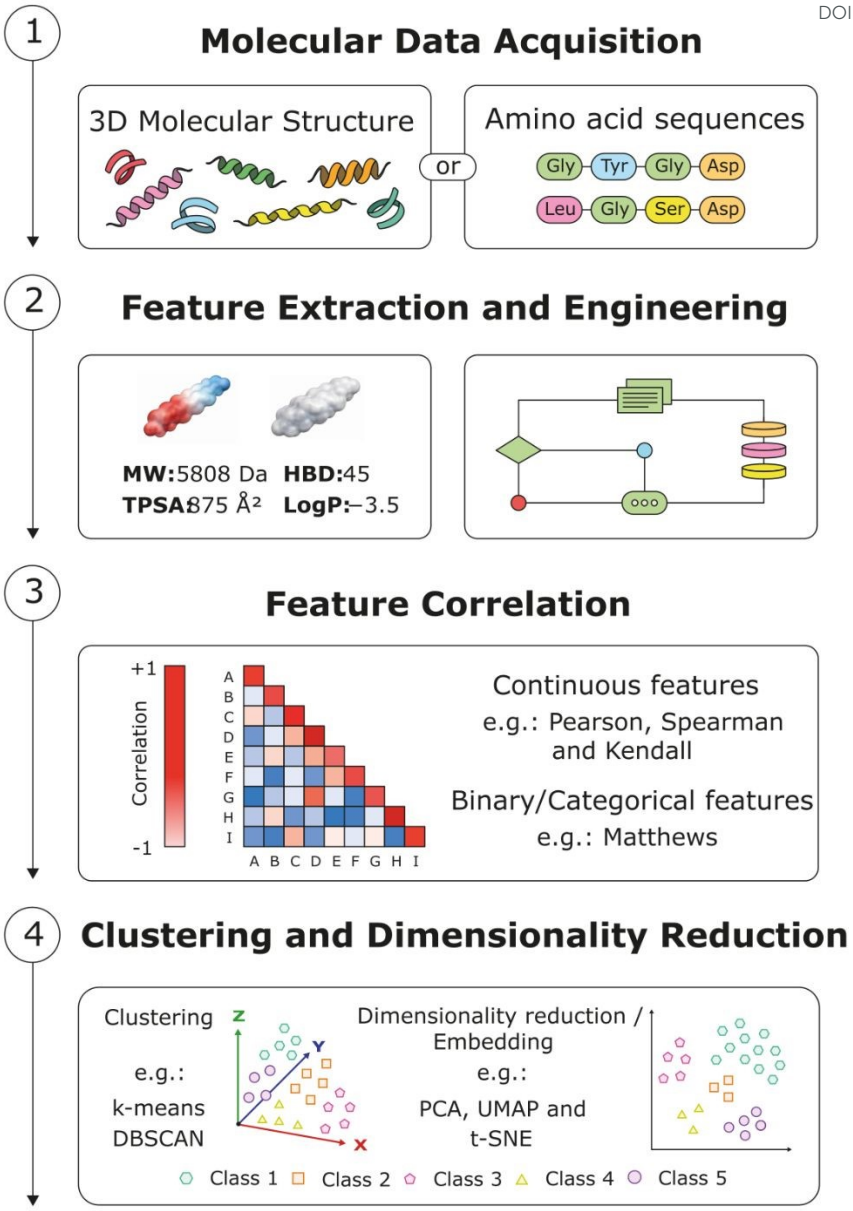

Figure 1. Conceptual framework to navigate the peptide chemical space.

Several open-source computational tools have been developed to facilitate the exploration of molecular chemical space based on molecular descriptor calculations. ChemPlot is a Python-based tool designed for both static and interactive visualization of chemical space in molecular datasets, encompassing dimensionality reduction techniques and similarity computations.[27] TMAP is a tool developed for Python and employed to visualize high-dimensional chemical space by organizing molecules into a minimum spanning tree structure through molecular fingerprint comparisons.[28] Similarly, KNIME, an open-source software developed for data science and visual analytics, in which computational

View Article Online
DOI: 10.1039/D5CP04611D

workflows are structured as flowchart-based pipelines, has dedicated extensions to support chemical space visualization and other cheminformatics applications.[29]

Figure 2 illustrates an example of how the chemical space of blood-brain barrier penetrating peptides (B3PPs) and quorum-sensing peptides (QSPs) can be represented in a 2D chart with the results of the dimensionality reduction provided by PCA and the clustering of the peptides predicted by the K-means algorithm.

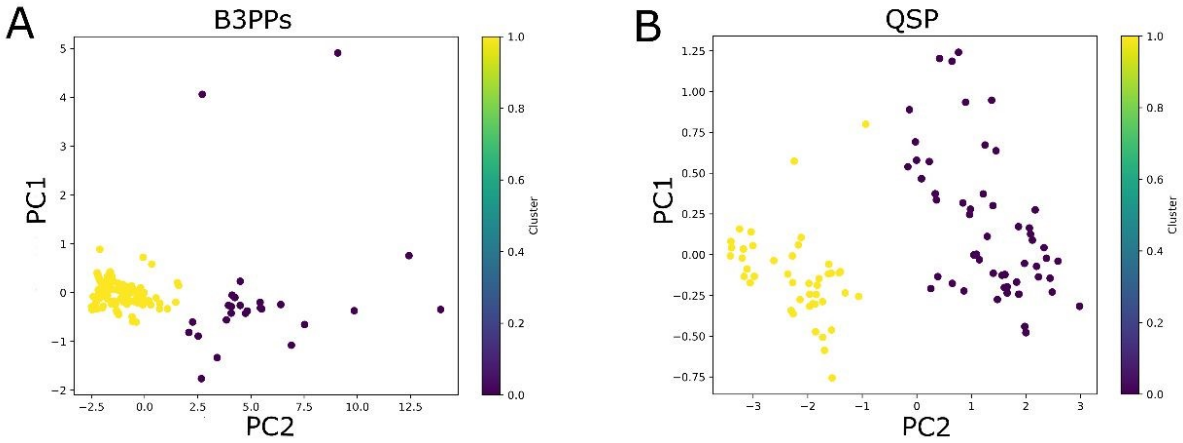

Figure 2. Comparison of the chemical space of B3PPs (panel A) and QSPs (panel B) using PCA for dimensionality reduction and K-means to cluster peptides.

Note: The molecular descriptors used to investigate the B3PPs included topological polar surface area (tPSA), oxygen plus nitrogen atoms (O+N), coefficient of lipophilicity (logP), nitrogen (N), oxygen (O), hydrogen bond acceptor (HBA), and hydrogen bond donor (HBD). The molecular descriptors used to investigate the QSPs included the logP, tPSA, HBA, and HBD.

In contrast to coordinate-based representations, chemical space networks (CSNs) have been introduced for chemical space analyses, allowing the exploration of molecular properties without reducing dimensionality[30–33]. Various similarity-based complex networks, including half-space proximal networks (HSPNs), metadata networks, and CSNs, have been utilized to study the bioactivity of compounds and their associated chemical space. Figure 3 represents an overview of the applications of these methods in peptide science.

View Article Online
DOI: 10.1039/D5CP04611D

The similarity-based complex networks are graphical representations of the chemical space of peptides, where nodes represent the peptides and the edges between two nodes denote their pairwise similarity or dissimilarity relationships in the space[34]. The distance between the compounds is often measured using similarity (or dissimilarity) distance metrics, such as Euclidean, Manhattan, Tanimoto, and Soergel coefficients. In these networks, the relevance of the elements is investigated using centrality measures (betweenness, closeness, and edge betweenness), as well as global network properties and their corresponding global measures, such as modularity, connectivity, density, and size[35,36]. For example, the StarPep Toolbox is a platform to explore the chemical space of antimicrobial peptides (AMPs) through molecular network-based representations and similarity-search methods to support peptide drug repurposing, as well as the development and optimization of novel sequences[37]. Recently, antiviral peptides (AVPs) were mapped into a chemical space using HSPNs and contextualized with metadata networks using the StarPep toolbox. The analyses revealed eight chemically distinct, biologically coherent AVP communities without fixed similarity thresholds and linking them to origins, functions, and viral targets through metadata networks. The authors performed a centrality-guided scaffold extraction, which revealed four non-redundant subsets suitable for modeling and multi-query searches. The mapping of motifs against non-AVP datasets indicated that motif burden correlates with higher predicted AVP probabilities, with peptides carrying four to five motifs achieving the highest scores across independent predictors, suggesting that the motif-driven design is an interesting strategy to expand AVP chemical space[33].

Recently, similarity-based complex networks and machine-learning algorithms have been used to map the landscape of some classes of peptides [38–40]. Half-space proximal networks, metadata networks, and chemical space networks are examples of computational methods that leverage graph theory to analyze and explore relationships among chemical entities based on a given molecular property[41,42]. These methods aim to simplify and analyze the vast complexity of chemical data, associating this information with the desired properties. For example, Ayala-Ruano et al. (2022) applied network analyses and similarity-based searching to explore the chemical space of antiparasitic

View Article Online
DOI: 10.1039/D5CP04611D

peptides, addressing the challenge of discovering new therapeutic peptides from a vast chemical space. The authors utilized HSPNs, CSNs, and metadata networks to identify central peptides and to perform multi-query similarity searches against the StarPepDB database. Despite achieving strong predictive performance (Matthews Correlation Coefficient values ranging from 0.834 to 0.965), challenges remain in the high sequence diversity of peptides, the need for effective filtering to eliminate toxic candidates, and the reliance on computational methods that may not fully capture the complexity of peptide interactions and biological functions[38]. Similarly, a study conducted by Castillo-Mendieta et al. (2024) used chemical space complex networks to map the chemical space of hemolytic peptides and to enhance the design of safe peptide-based therapeutics. By analyzing a database of 2,004 hemolytic peptides, the authors identified 12 consensus hemolytic motifs. They developed multi-query similarity searching models that outperformed the existing machine learning models in predicting hemolytic activity[39].

View Article Online
DOI: 10.1039/D5CP04611D

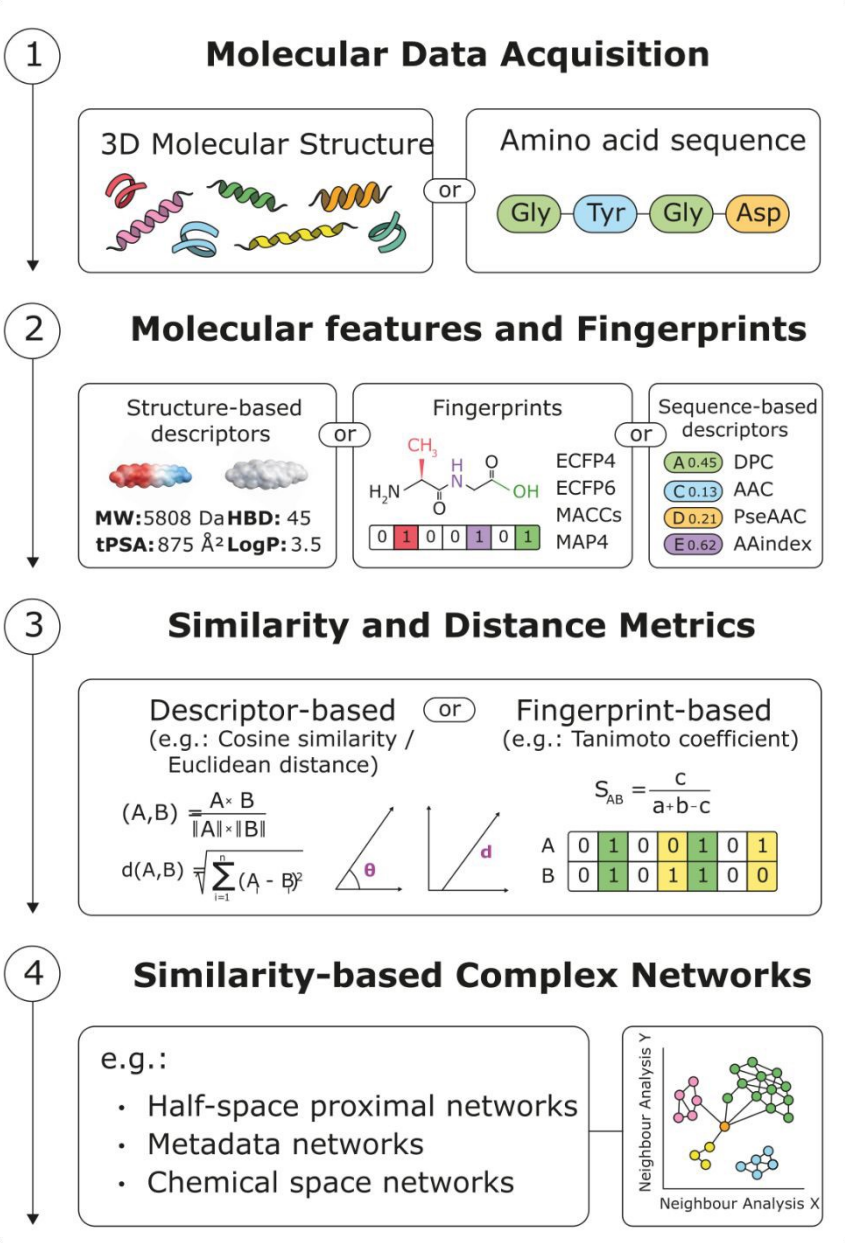

Figure 3. Schematic workflow that represents the peptide functional spaces mapping using similarity-based complex networks.

In another study, Wang et al. (2024) improved a computational framework using a reinforcement learning (RL)-driven generative model integrated with graph attention mechanisms, which captured the connectivity structure between amino acid residues in peptides and used it to guide the search for optimal peptide sequences. The algorithm incorporates bioactivity and ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties, ensuring that the generated peptides meet drug-like criteria[43].

View Article Online
DOI: 10.1039/D5CP04611D

In contrast to the consensus chemical space, the concept of the chemical multiverse was introduced as a collection of multiple chemical spaces, each defined by different descriptors. This concept emphasizes that there is no single chemical space; rather, various representations of the same set of molecules can yield distinct chemical spaces[44]. The chemical multiverse allows for a more comprehensive analysis of compound datasets using multiple descriptors, which can capture different aspects of molecular structures and properties. This approach contrasts with the idea of a consensus chemical space widely applied in medicinal chemistry, which attempts to combine various descriptors into a single representation, potentially losing valuable information in the process. The chemical multiverse does not rely on multiple descriptor combinations; instead, it consists of various alternative graphical representations of the chemical space, incorporating different molecular properties derived from the investigated compounds, such as molecular fingerprints, structure-based or sequence-based descriptors[44]

The description of the chemical space of peptides has faced significant challenges, partly due to their chemical structure, vast physicochemical properties, and their intrinsic polymer-like features related to the amide backbone repetition, which tends to mask the prediction of their bioactive properties, hindering the analysis of their properties in the chemical space[45,46]. In addition, it has been demonstrated that some unrelated classes of peptides show unexplained intersections between their bioactivities[47,48], thus evidencing that some previous distinct chemical spaces exhibit molecular similarities that must be better explored to find possible intersections in the intervals of the molecular descriptors applied to characterize them.

The chemical space usually contains some distinct clusters named '*constellations*', which are populated by molecules with specific properties that can be identified using scaffold-based analysis due to the presence of a common structural core[49,50]. The Murcko framework has been widely applied to investigate the structural core of drugs, revealing structural information and distinguishing molecules by their ring systems, linkers, and side chain atoms[51]. However, Murcko frameworks can only represent molecules containing ring systems; therefore, acyclic (linear) peptides are usually omitted from these analyses[50].

View Article Online
DOI: 10.1039/D5CP04611D

Moreover, peptides are characterized by their large molecular size and shape, as well as by the presence of polar groups, which usually put them beyond the conventional predictors of drug-likeness for molecules[52,53] and impose more complexity in evaluating their conformational changes and pharmacophore properties[54].

Scrambled peptides contain similar amino acid composition (AAC) and sequence length; however, they could acquire different conformations, which confer them different biological activities[55,56]. The study of scrambled peptides has demonstrated some of these peculiarities of peptides compared to other small molecules.

Revealing the Complexity Behind the Chemical Space of Peptides: Investigating the Interplay Between Conformation, Bioactivity, and Bioavailability

The chemical space of peptides encompasses a multidimensional subset of molecular properties linked to functional and drug-like characteristics[57], due to their unique physicochemical and structural properties that differentiate them from traditional small molecules used in the discovery and design of new drugs [15,58,59]. These properties present a central challenge in characterizing a general chemical space of peptides, as they are crucial for predicting their biological, pharmacokinetic, and pharmacodynamic properties[60]. This comprehension aids medicinal chemists in accurately classifying peptides and identifying suitable applications.

Chemically modified residues that undergo post-translational modifications are likely to be extensive. The peptide chemical space thus encompasses a wide array of changes that can significantly alter the properties and functions of peptides [61]. Understanding the chemical space of peptides also involves the comprehension of the three-dimensional conformations adopted by these molecules in the solvation medium, as their geometries are closely related to the mechanisms of action associated with membrane permeability and stereoselectivity against the molecular target[60,62,63]. The conformation of peptides refers to the spatial arrangement of their atoms adopted due to the rotation

View Article Online
DOI: 10.1039/D5CP04611D

around a single bond over time. It intrinsically depends on the peptide sequence and the external environment, and it is related to the formation of the secondary structure propensity. Peptides dynamically adopt a collection of conformations distributed across a free energy landscape, with their occurrence governed by Boltzmann-weighted probabilities[64].

Some conformational adaptability in peptides can also modulate their bioactivity, because changes in the conformational ensemble alter the positions of pharmacophoric side chains and the population of binding-competent states, thereby affecting potency, selectivity, and the recognition to the target [65–67]. This is consistent with observations that scrambled variants of peptides with similar composition may adopt distinct conformations and exhibit different biological activities[56]. As the secondary structure is a determinant of peptide bioactivity, some strategies have been developed to impose constraints on their conformation to control their bioactivity[68,69]. For example, some peptide design strategies for AMPs include the incorporation of restrictors, such as lactam and disulfide bridges, which act as conformational inducers (promoting β-like structures) and enhance resistance to protease degradation[70,71].

Two classes of peptides that naturally cross biological barriers illustrate the relevance of conformational changes: the cell-penetrating peptides (CPPs) and blood-brain barrier-penetrating (B3PPs).[72,73] Both classes exhibit a conformational characteristic that influences their ability to cross these barriers, which is called chameleonic properties. This conformational property refers to their ability to change conformation in response to environmental conditions, particularly to expose or hide polar groups when crossing biological membranes[74,75]. Some well-reported examples of chameleonic properties of peptides include cyclosporin A[75] and some of its derivatives[76] (Figure 4, panel A), as well as some cyclic peptides[75,77]. This property has significant implications for their chemical space and the overall functionality of biomembrane-penetrating peptides. By altering their conformation, these peptides can effectively navigate through the hydrophobic core of cell membranes, improving their bioavailability[77].

Some computational models may struggle to accurately represent the conformational variability of peptides, posing challenges for predicting their

View Article Online
DOI: 10.1039/D5CP04611D

pharmacokinetic properties[58]. For example, the calculation of topological polar surface area (tPSA) does not depend on the three-dimensional characteristics of the molecules, and it has been widely applied to correlate with the hydrogen bond pattern of molecules in the aqueous phase[78]. This property has been associated with the prediction models of solubility and passive diffusion through cell membranes[79–82]. Elevated tPSA values are associated with complexation with water molecules and increased molecular volume, which can hinder membrane permeability[83]. Typically, the penetration of compounds across cell membranes is restricted when tPSA exceeds 140 Å² [84]. However, higher values are generally acceptable for macrocyclic peptides (tPSA = 220 Å²) and peptides exhibiting chameleonic properties (tPSA = 280 Å²)[81,85]. The Molecular 3D PSA (MPSA) has emerged as a more accurate measure of compound solubility and membrane permeability than the tPSA, as it considers the three-dimensional conformation of the compound in a given environment[86]. The tPSA, however, does not depend on the tridimensional structure and could reach satisfactory prediction, especially when associated with the molecular weight of compounds[87]. The macrocycle peptide cyclosporine A is an example of a natural peptide that exhibits chameleonic activity. Cyclosporin A has a high tPSA value of 279 Å² and an MPSA value equal to 105 Å² with approximately 62% of its PSA concealed in nonpolar environments[87,88].

The permeability of peptides into cell membranes could be significantly influenced by their secondary structure[89]. Studies have demonstrated the impact of peptide conformation on arginine-mediated internalization in cell membranes[90]. Similarly, other studies have shown that for CPPs, some helices stabilized by hydrocarbon cross-links can effectively enter the cells[91,92]. The N-methylation of cyclic peptides has been widely reported as an interesting strategy to improve the permeability in cell membranes[79,93]. White et al. (2011), for example, demonstrated that methylated analogues generated by on-resin N-methylation improved their membrane intake. The Figure X, panel (B) shows the regioselective backbone N-methylation of a cyclic hexapeptide scaffold cyclo[Leu¹, D-Leu², Leu³, Leu⁴, D-Pro⁵, Tyr⁶] (compound 1) and its trimethylated analogue (compound 3) generated by on-resin N-methylation. In compound 3, D-

View Article Online
DOI: 10.1039/D5CP04611D

Leu², Leu³, and Tyr⁶ are N-methylated (Me), a pattern associated with markedly improved passive permeability in the study by White et al. 2011[93].

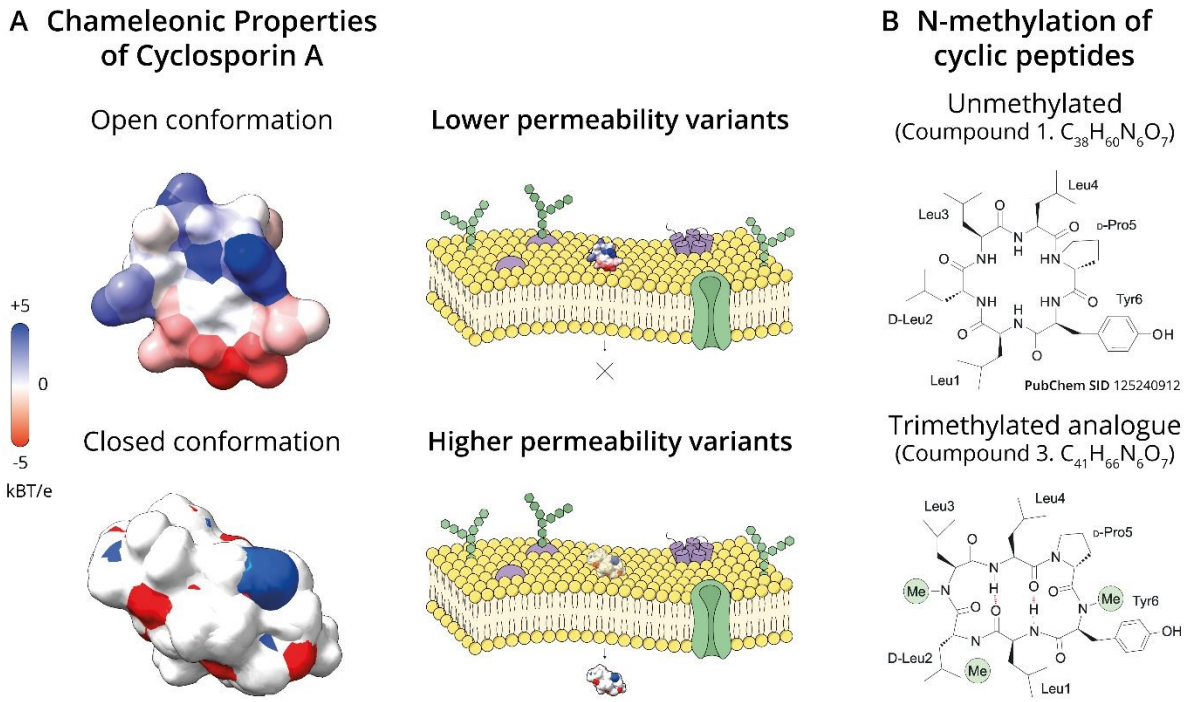

Figure 4. Chameleonicity and backbone N-methylation as determinants of passive membrane permeability in macrocyclic peptides. Panel (A) Cyclosporin A, a chameleonic molecule, shown in representative open (greater polar surface exposed) and closed (polar surface partially buried) conformations, with electrostatic potential maps (scale in kBT/e) that illustrate environment-dependent exposure/burial of polar groups through conformational switching and intramolecular hydrogen bonding. Center: conceptual schematic of passive diffusion through a lipid bilayer: lower-permeability variants tend to retain solvent-exposed hydrogen-bond donors and acceptors, whereas higher-permeability variants more effectively mask polarity. Panel (B) Parent scaffold cyclo[Leu¹, D-Leu², Leu³, Leu⁴, D-Pro⁵, Tyr⁶] (compound 1) and the corresponding trimethylated analogue (compound 3), bearing backbone N-methyl groups at D-Leu², Leu³, and Tyr⁶ (Me)—a modification pattern associated with higher permeability, as reported by White et al. (2011).[93]

View Article Online
DOI: 10.1039/D5CP04611D

Currently, most *in silico* models—including cheminformatics filters and machine-learning approaches used in peptide science—assume that these molecules cross biological membranes primarily via passive diffusion, implying that membrane penetration occurs mainly through biophysical interactions between the peptide's structure and the membrane[58]. Consequently, active transport pathways, like receptor-mediated transcytosis, active influx transport, and carrier-mediated transcytosis, are often overlooked in the design of these models, primarily due to the intricate binding processes of membrane proteins associated with the conformational accommodation related to receptor binding[58,94]. Furthermore, some cheminformatics filters that use a set of molecular descriptors and their intervals to characterize drug-like molecules, such as the BOILED-Egg model and Lipinski's Rule of Five, often fail to accurately predict the permeability of peptides due to their distinct chemical space[82,95]. These classical rules and empirical models were largely developed and calibrated using small molecules and therefore delineate a region of property space enriched for passive permeability and oral bioavailability[79,95]. When applied to peptides, these filters often fail to accurately predict peptide permeability and overall drug-like behavior because peptides typically present higher molecular weight, larger polar surface area, multiple hydrogen-bond donors/acceptors, and higher conformational flexibility — features that shift them outside conventional small-molecule boundaries.[83,95] Therefore, applying small-molecule drug-likeness rules can artificially truncate peptide chemical space and may lead to misleading conclusions to their bioactivity or bioavailability.

Moreover, some peptide classes can partially overcome these restrictions through structural adaptations that are not captured by simple 2D descriptors used in the cheminformatic filters. For example, macrocyclic and "chameleonic" peptides can conceal polar surface area via intramolecular hydrogen bonding and environment-dependent conformational changes, thereby improving passive permeability despite high tPSA values. In addition, descriptors that incorporate 3D conformation may be more informative than purely 2D metrics for some peptide subclasses[62,74].

Considering these limitations, efforts have been made to accurately represent their geometries by analyzing the rotamers and possible

View Article Online
DOI: 10.1039/D5CP04611D

conformational changes of peptides to enhance the prediction of their molecular activities [96,97].

An overview of molecular descriptors applied to peptides is demonstrated in Figure 5.

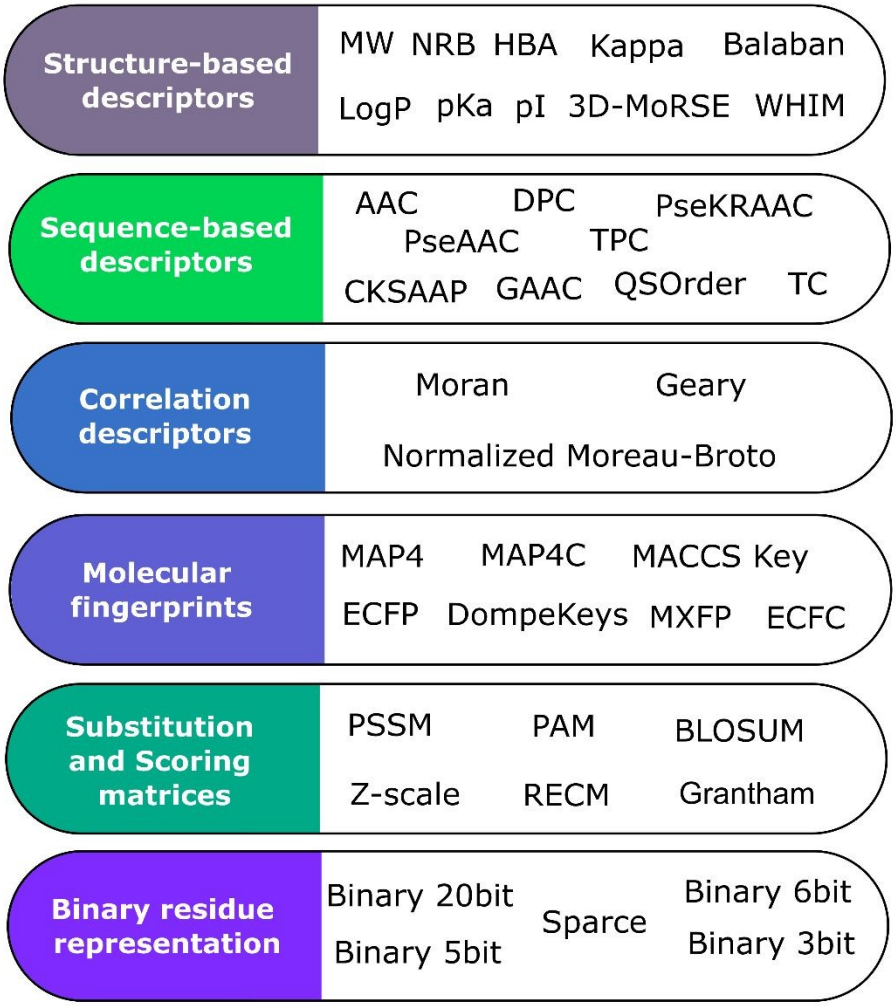

Figure 5. Classes of molecular descriptors commonly used to represent peptide sequences and structures and to analyze peptide structure–activity relationships.

Structure-based Features to Describe the Chemical Space of Peptides

Many chemical properties of peptides, particularly topological, stereochemical, electronic, aromatic, and physicochemical features, can only be

View Article Online
DOI: 10.1039/D5CP04611D

calculated by examining their atomic structure. This is particularly true for synthetically modified peptides or those with post-translational modifications, where chemical changes to the amino acid side chains (such as the addition or removal of functional groups) create new molecular functions and properties[5,98]. These descriptors can be calculated from the 2D or 3D representation of peptides in structural data files, such as SDF, CDX, and MDL Molfiles.

Many descriptors widely used in peptide science and medicinal chemistry are computed from 2D connectivity or composition and therefore do not require a specific 3D conformer (e.g., MW, HBA/HBD counts, tPSA, atom counts, fragment-based LogP, and 2D fingerprints). In contrast, geometry-, surface-, and shape-based properties are conformation-dependent and should be computed from an explicit 3D structure or a conformational ensemble (e.g., SASA/3D PSA such as MPSA, WHIM, GETAWAY, and 3D-MoRSE descriptors). The 2D representation is often sufficient to calculate most descriptors since these descriptors capture features like atomic constitution, the presence of specific chemical groups, physicochemical properties, and molecular topology. The selection of the most appropriate molecular descriptors usually depends on the type of peptide under investigation and its associated biological activity[58].

Descriptors used in large peptide libraries are broadly borrowed from medicinal chemistry, originally designed for small molecules, where they proved useful in predicting drug-likeness and bioavailability. Key examples include tPSA, MW, HBA, and HBD, number of aromatic rings (NAR), the fraction of sp^3-hybridized carbon atoms (Fsp^3), lipophilicity calculated by the logarithm of 1-octanol/water partition coefficient (LogP), number of chiral centers (NCC), and the number of rotatable bonds (NRB) [79,84,99,100]. Additional descriptors capture structural features, such as secondary structure composition, ionization state, topology, shape, and hydrophilicity.

The importance of these descriptors lies in their connection to bioavailability, i.e., how efficiently a compound can dissolve in aqueous environments and cross biological membranes. Two widely used descriptors are LogP and tPSA, which reflect lipophilicity and polar surface area, respectively [82,83]. Similarly, intrinsic solubility can be explored through lipophilicity (e.g., LogP

View Article Online
DOI: 10.1039/D5CP04611D

and the logarithm of the distribution partition coefficient (LogD) at pH 7.4), structural constitution (e.g., NAR), and molecular flexibility (e.g., Fsp^3, NCC, and NRB)[53,79,81,82,99,101].

Molecular descriptors associated with the ionization state are informative of aqueous solubility (hydrophilicity) and include the isoelectric point (pI) and the logarithmic value of the acid dissociation constant (pKa). Molecular hydrophobicity (lipophilicity) is usually quantified by the calculations of LogP values, and alternative computational methods, such as XLogP3[102], CLogP, and ALogP[103] calculate some of its derivative values. The ALogP has demonstrated superior predictive accuracy for peptides compared with other calculated logP values [104].

Certain molecular descriptors that capture the shape and topology of peptides can be linked to their stereoselectivity towards molecular receptors. These descriptors reflect the spatial rearrangement and connectivity of the molecules. Examples include Kier's Kappa indices[105], Balaban indices[106], Burden eigenvalues[107], and Randić shape indices[108]. Other descriptors focus on molecular complexity. For instance, Basak's Indices provide a numerical representation of structural features such as connectivity, branching, and overall topology[109].

Among these shape- and topology-oriented metrics, the Kappa descriptors stand out as topological indices derived from the hydrogen-suppressed molecular graph. They quantify molecular shape and the intricacy of branching by considering paths of different lengths through the structure, comparing the observed branching to an idealized linear or maximally branched reference. Because they are sensitive to the global shape rather than the size of the molecule, Kappa indices offer a refined picture of molecular flexibility and compactness. In QSAR applications, they are particularly useful for highlighting steric and topological features that influence biological activity and receptor interaction.[110]

Another important topological descriptor known for its low degeneracy and strong correlation with physicochemical properties is the Balaban index. It is calculated from the distance matrix of the molecular graph, integrating

View Article Online
DOI: 10.1039/D5CP04611D

information from distance sums and the number of edges. By encoding details related to cyclicity, branching, and structural compactness while remaining largely independent of molecular size, the Balaban index stands out at distinguishing structural isomers, which is especially valuable in studies involving complex ring systems and detailed structure–property relationships.[111]

To study the geometry of compounds, researchers often rely on 3D molecular descriptors, which require the explicit 3D conformation of a molecule to capture the spatial arrangement of its atoms. These descriptors are widely used in structure-activity relationship (SAR) analyses. Notable examples include WHIM (Weighted Holistic Invariant Molecular descriptors), GETAWAY (GEometry, Topology, and WAter Accessibility),[112] and 3D-MoRSE (Molecular Representation of Structures based on Electronic diffraction)[113].

The 3D-MoRSE encodes fundamental 3D atomic coordinates of the molecular structure using a fixed-size vector, drawing on concepts akin to electron diffraction. The 3D-MoRSE descriptors differ from purely topological indices because they incorporate three-dimensional structure along with electronic information. They are generated from simplified theoretical scattering curves—conceptually similar to electron diffraction—using atomic coordinates and weighted properties such as mass or partial charges. The resulting values represent the distribution of electron density across a range of scattering angles. In doing so, the 3D-MoRSE descriptors capture steric effects, electrostatic interactions, and other conformation-dependent features that cannot be inferred from 2D graph topology alone.[113]

The WHIM (Weighted Holistic Invariant Molecular descriptors)[114] is a set of 3D molecular descriptors that capture the spatial arrangement of atoms in a molecule.[114] The WHIM descriptors [91] are also based on 3D atomic coordinates but summarize them through a weighted principal component analysis. It's used a covariance matrix built from atomic positions, with weights that may reflect atomic mass, polarizability, electronegativity, or other physicochemical attributes. Because they derive from molecular invariants, WHIM descriptors remain consistent under translation and rotation of the molecule, providing a global

View Article Online
DOI: 10.1039/D5CP04611D

statistical representation of size, shape, symmetry, and atom distribution relative to three orthogonal axes.[115]

The GETAWAY (GEometry, Topology, and Atom-Weights AssemblY) is a 3D-molecular geometry provided by the molecular influence matrix and atom-relatedness by a molecular topology using different atomic weighting[112]. The GETAWAY descriptors complement these approaches by integrating geometric features with molecular topology through the influence matrix and various atom-relatedness measures. Together, this family of 3D descriptors enables characterization of steric, electronic, and conformational aspects that strongly influence molecular recognition processes.[112]

Measures of molecular complexity lack a universal concept, but they have been frequently associated with synthetic accessibility and, in the context of drug design, with the specificity to a molecular target[116]. Currently, different descriptors capture various aspects of molecular complexity, and their complementary use may provide an overview of this molecular property, such as topological and physicochemical descriptors (e. g.: NCC and Fsp^3)[99,101] and some substructure-based descriptors (number of rings, unsaturations, and heteroatoms). For peptides, the difficulty of synthesis has been associated with long amino acid chains and functional groups associated with their side chains[117]. A non-exhaustive list of the most applied structure-based descriptors to investigate peptides is presented in Table 1.

Table 1. Molecular descriptors derived from the peptide structures which are applied to evaluate their intrinsic properties.

Structure-based Descriptors	Peptide information
Bounds count, heavy atoms count, atoms type count (e.g, N-, O-, C-, S), NRB.	Atomic constitution
NPA, NG, NNCAA	Presence of molecular groups
MW, NHA, VdW, MSA	Molecular size
LogP, Log D (pH 7.4), XLogP3, ALogP, CLogP, LogKow	Lipophilicity/hydrophobicity

View Article Online
DOI: 10.1039/D5CP04611D

Kappa indices (Kappa1, Kappa2, and Kappa3), Burden eigenvalues, Basak's Indices, Balaban index, Weiner index, Randić indices, WHIM indices[114]	Molecular shape and topology
Fsp^3, NRB	Molecular flexibility
tPSA, MPSA, PSA	Polar surface (polarity)
WHIM indices[114], 3D-MoRSE[113], GETAWAY [112]	Molecular geometry
NCC[99], Fsp^3 [99,101], CC[118],	Molecular complexity
HBA, HBD, tPSA, net charge, pKa, LogP, XLogP3, ALogP, CLogP, LogKow, SASA	Hydrophilicity (aqueous solubility)
pKa, pI	Ionization state
Number of α-helices, number of β-sheets, number of coils	Secondary structure
Eisenberg scale[119]	Hydrophobic moment

Notes: CC: number of chiral carbons; Fsp^3: fraction of sp^3-hybridized carbon atoms[101]; GETAWAY: GEometry, Topology, and Atom-Weights AssemblY[112]; HBA: hydrogen bond acceptors, HBD: hydrogen bond donors; LogP: 1-octanol/water partition coefficient; 3D-MoRSE: 3D molecular representations of structure based on electron diffraction; MPSA: Molecular 3D polar surface area; MW: molecular weight; MSA: molecular surface area; NCC: number of chiral centers; NPA: number of primary amino groups ($-NH_2$); NHA: number of heavy atoms, NAR: number of aromatic rings, NG: number of guanidine groups; NNCAA: number of negatively charged amino acid groups; NRB: number of rotatable bonds; pKa: logarithm of the acid dissociation constant, pI: isoelectric point; tPSA: topological polar surface area[78]; PSA: polar surface area; VdW: Van der Waals Volume; SASA: solvent accessible surface area; XLOGP3: LogP estimated from the atom/fragment contribution values[120]; WHIM: Weighted Holistic Invariant Molecular descriptors[114].

Sequence-based Features to Describe the Chemical Space of Peptides

While all peptide properties ultimately originate from the primary structure (amino acid composition and sequence order) and the chemical modifications, only a subset of properties can be directly computed from the sequence[121–123]. Accordingly, sequence-based descriptors are associated with information calculated from the primary structure of an oligopeptide sequence, assessing the amino acid composition, sequence motifs, amino acid arrangements, and their physicochemical patterns, including the presence of hydrophilic and hydrophobic

View Article Online
DOI: 10.1039/D5CP04611D

regions, and are especially valuable when tertiary structure information is unavailable[58]. Some encoders, scoring, and substitution matrices have been applied to categorize peptides into distinct classes, providing a statistical description of their sequences, and filling a critical gap when tertiary structure information is absent[58,124,125]. Most of these sequence-based descriptors were developed to classify and analyze protein sequences, but have also been applied for peptides.[15,126,127] These descriptors could be calculated from the string representation of amino acids represented by the FASTA[128], PLN, BILN[129], and HELM[130] file formats that contain the peptide's primary structure[131].

Peptides contain both hydrophilic and hydrophobic regions, often influenced by the relative abundance of specific amino acid residues, which in turn shape their molecular mechanisms of action. A notable example is the CPPs, which are typically enriched in lysine and arginine residues. This composition accounts for their cationic or amphipathic nature at physiological pH, as well as their water solubility and cell membrane permeability[132,133]. Studies have demonstrated that incorporating arginine into cyclic peptides and protein surfaces enhances cell penetration[134,135]. Adapting the amino acid structure of a peptide might influence their biological activity and drug-like properties[5]. Cyclization and N-methylation are examples of chemical modifications that shift the peptides' positions within the chemical space, enhancing their potential bioavailability and biological activity[5]. For example, a study revealed significant overlaps between the chemical space of synthetic linear and cyclic pentapeptides containing N-methylation and some FDA-approved peptide drugs. Some studies have shown that machine-learning algorithms applied to predict some classes of peptides that employ an optimized integration of sequence- and structure-based descriptors in the feature composition achieve greater accuracy than those relying solely on sequence- or structure-based descriptors[15,136–138].

The sequence-based properties have also been used to design novel domains and modulate switchable properties of peptides. Self-assembling peptides (SAPs) are short polypeptide chains that, in an aqueous solution, can spontaneously organize themselves into complex, well-ordered, and stable nano- and meso-structures through the formation of non-covalent interactions[139,140], thus forming versatile building blocks which have been extensively studied to

View Article Online
DOI: 10.1039/D5CP04611D

create stimuli-sensitive supramolecular systems[141–143]. The amino acid sequence composition and the orientation of the amino acids of SAPs could play a critical role in driving the self-assembling properties. For example, in the SAP sequence, aromatic amino acids, such as phenylalanine, tyrosine, and tryptophan, contribute mostly to aggregation through π-stacking as the main driving force for self-assembly[144]. On the other hand, the presence of histidine, serine, and threonine amino acids has highly polarizable side chains, and thus, these peptide structures could promote aggregation through hydrogen bonding formation. Some SAP structures are characterized by their amphiphilicity, meaning their sequences contain hydrophilic and hydrophobic domains that facilitate self-assembly in aqueous solutions, forming non-covalent interactions between amino acid residues[145].

Peptide molecules that self-assemble into peptide nanofibers are primarily amphiphilic molecules. These consist of hydrophilic heads containing active peptide segments, hydrophobic tails with alkyl chains, and several amino acids between these two regions, creating enough space to prevent spontaneous aggregation after introducing negative charges[146,147]. Amphipathic peptides are more likely to self-assemble into amyloid-like β-sheet fibrils when their primary sequence shows a pattern of alternating hydrophobic and hydrophilic amino acids. These fibrils form a bilayer structure comprising two β-sheets that align to conceal the hydrophobic side chains within the bilayer's interior. In contrast, the hydrophilic side chains remain exposed on the surface of the bilayer[148,149]. Recently, a study demonstrated that the SAP sequence significantly influences structural sensitivity to supramolecular polymerization pathways, affecting the resulting polymers' structural and functional properties [150]. Yuan et al. (2022) demonstrated that the order of amino acids in the sequences AAEE and AEAE (A and E represent alanine and glutamic acid, respectively) impacts the driving forces involved in peptide polymerization, which directly correlates with mechanical properties and bioactivity[150].

Some computational models have been developed using a combination of sequence-based and structure-based descriptors to predict the bioactivity of peptides. These models have improved performance compared to algorithms relying solely on one class of molecular descriptors[15,94,124,136,137]. For example,

View Article Online
DOI: 10.1039/D5CP04611D

Rajput et al. (2015) analyzed the QSPs according to their amino acid composition, residue position, physicochemical properties, and sequence motifs and identified that some aromatic residues, such as tryptophan, tyrosine, and phenylalanine play an important role in their characterization, as well as positional preferences of residues, such as serine at the N-terminal end and phenylalanine at the C-terminal end so that these sequence-based properties could be used for their identification [124]. Physicochemical properties, such as aromaticity, molecular weight, and secondary structure, contribute to QSP identification[124]. Recent approaches utilize propensity score representation learning to extract and combine the propensities of amino acids and dipeptides[151].

Sequence-based descriptors provide a quantitative framework for analyzing peptides, capturing features such as amino acid composition, positional distribution, and sequence arrangement patterns. These descriptors enable the application of statistical and machine learning approaches to uncover structure-activity/property relationships in peptide research[127,152–154]. In this context, we focus on molecular descriptors and scoring matrices that support large-scale, data-driven analyses. Several Python libraries, widely used for big data applications, offer built-in tools for calculating such descriptors, such as BioPython[155], RDkit[21], Mordred[24], and PyBioMed[156]. Additionally, there are tools focused on peptide analysis, such as PepFun[157], iFeature[23], iFeatureOmega[22], Peptide.py (https://pypi.org/project/peptides/)[158], and PepFuNN[159].

Some sequence-based descriptors evaluate the amino acid or k-mers constitution, providing information about their relative abundance or scarcity of amino acids, such as AAC[160], dipeptide composition (DPC)[161], tripeptide composition (TPC)[152], and terminus composition (TC)[162]. K-mers are substrings created by moving a window of length *k* along the sequence at a set interval. In addition to the constitution, they reflect the overall frequency of these amino acids.

Other sequence-based encoders offer information about group- and gap-based amino acid rearrangements. The group-based amino acid descriptors aim to mitigate the high-dimensional data derived from the existence of 20 amino acids, so this class of encoders groups or reduces the amino acid compositions

View Article Online
DOI: 10.1039/D5CP04611D

to investigate the peptide sequences. High-dimensional data can lead to overfitting, compromising the prediction accuracy of the models when the number of features exceeds the number of independent samples[163]. Thus, these descriptors extract characteristics that better reflect the relationships of groups of amino acid residues in the sequence. The group-based amino acid composition descriptors, for example, include the grouped tripeptide composition (GTPC)[164], grouped dipeptide composition (GDPC)[164], pseudo K-tuple reduced amino acid composition (PseKRAAC)[165], and the grouped amino acid composition (GAAC)[23]. In contrast, the gap-based amino acid descriptors create bi-mers from peptide sequences using various gap sizes, and subsequently analyze the distribution of the resulting gap-based bi-mers. These sequence-based descriptors include composition of *k*-spaced amino acid pairs (CKSAAP)[166], and adaptive skip dipeptide composition (ASDC)[167,168].

In addition to these descriptors, some libraries extract from the sequence the AAindex, a curated, literature-derived database that compiles numerical indices describing physicochemical properties of amino acids. In its core component (AAindex1), each “amino acid index” represents a single property as a set of 20 numerical values, one per standard amino acid, enabling sequences to be converted into quantitative property profiles[169].

Several substitution and scoring matrices have also been developed to represent the variability, the physicochemical properties, and substitution patterns of polypeptide sequences, including position-specific scoring matrix (PSSM), residue pairwise energy content matrix (RECM)[170], Z-scale[171], and BLOcks Substitution Matrix (BLOSUM).[172] The BLOSUM and PAM matrices, for example, are derived from oligopeptide sequence alignments, and both are commonly used as encoders to characterize peptide sequences based on their evolutionary substitution profiles (Table 3)[173], showing variations depending on the identities of the pre-computed datasets. For example, BLOSUM matrices come in different versions, such as BLOSUM50, BLOSUM62, and BLOSUM80, created using the observed frequencies of amino acids in peptide sequences. The 62% identity threshold (BLOSUM62) is widely used for peptide and protein sequence characterization[174,175]. In contrast, the position-specific scoring matrix (PSSM), residue pairwise energy content matrix (RECM), and Z-scale[171] are

View Article Online
DOI: 10.1039/D5CP04611D

classified as scoring matrices applied for amino acid sequences (see Table 3). The z-scale, for example, is an amino acid descriptor set used to numerically represent the physicochemical, hydrophobic, and polar properties of amino acids in protein or peptide sequences. This matrix is derived from a PCA of various amino acid and physicochemical properties, reducing them into a few orthogonal components[171].

While the content of the secondary structure is dependent on the conformation of the peptide, and is more accurately calculated using information derived from the three-dimensional structure, several sequence-based prediction methods have been developed that demonstrate promising results for predicting secondary features and classifying oligopeptide sequences[176–178]. For example, Zhang et al. (2011) developed a transition probability matrix to represent secondary structures[178], and Dai et al. (2013) introduced a statistical position-based feature of secondary structural elements to predict the structural classes of oligopeptide sequences[177]. The secondary structure elements content (SSEC), for instance, is a molecular descriptor calculated from the primary structure predicted by the PSIPRED V4.0 and provides the content of three types of secondary structure elements[23].

The correlation encoder quantifies the relationship between amino acids by calculating correlation coefficients that reflect differences in the molecular descriptors that reveal information about hydrophobicity, hydrophilicity, mass, shape, topology, constitution, etc. These descriptors reveal how specific properties of amino acids are interrelated to the sequence. Moran[179], Normalized Moreau-Broto, and Geary[179] are autocorrelation descriptors that uses eight amino acid indexes by default for peptide sequences, according to the following: the DAYM780201 represents the the residue substitution profile, the CHOC760101 represents the residue accessible surface area in tripeptide, the CIDH920105 represents the normalized average hydrophobicity scales, the BHAR880101 represents the average flexibility indices, CHAM820101 represents the polarizability parameter, CHAM820102, represents the free energy in water, the BIGC670101 represents the volume of the residue, and the CHAM810101 the steric parameter[23,180].

View Article Online
DOI: 10.1039/D5CP04611D

Binary encoders are descriptors that transform amino acid sequences into statistical vectors, with each amino acid encoded as a 20-dimensional binary vector consisting of 0s and 1s. The binary representation is subdivided into 3, 5, 6, and 20 bits, and they represent some groups of amino acids of the sequence depending on their physicochemical properties[181]. For example, the binary 6-bit uses a six-element amino-acid groups {*e*1, *e*2, *e*3, *e*4, *e*5, *e*6} to encode the oligopeptide sequence, where *e*1 ∈ {H, R, K}, *e*2 ∈ {D, E, N}, *e*3 = C, *e*4 ∈ {S, T, P, A, G}, *e*5 ∈ {M, I, L, V}, *e*6 ∈ {F, Y, W}. These groups capture conservative substitutions that can occur over evolutionary time. They function as equivalence classes grouping amino acids by similarity, and their definitions are based on PAM-based relationships. Then, each group is represented by a 6-dimensional binary vector, e.g., *e*1 is encoded by (100000), *e*2 is encoded by (010000), and so on[22]. In the sparse encoding approach, each peptide sequence is mapped to a fixed-length vector of 100 positions, corresponding to the maximum sequence length stored in the database. A reference list containing the 20 standard amino acids plus one additional symbol for gaps or empty positions is used. Each amino acid is converted into a one-hot vector of length 21, where a single element indicating its position in the list is set to “1”, and the remaining elements are set to “0”. Consequently, every position in the 100-length sequence corresponds to a 21-dimensional vector. This representation ensures that each amino acid is uniquely identified by its position within the encoding space[181].

A list of molecular descriptors derived from the sequence is described in Table 2. A list of applied scoring and substitution matrices is described in Table 3. A list of autocorrelation descriptors associated with the amino acid indices is presented in Table 4.

View Article Online
DOI: 10.1039/D5CP04611D

Table 2. List of some molecular descriptors derived from the sequence (primary structure) that are applied to the peptide analyses.

Sequence-based Descriptors	Peptide information
Amino acid composition (AAC)[160]	Frequencies of the 20 types of native amino acids present over the peptide sequence
Pseudo-amino acid composition (PseAAC)[182]	Frequencies of the discrete sequence correlation factors and the twenty components of the conventional amino acid composition
Amphiphilic Pseudo-Amino Acid Composition (APAAC)	Frequencies of the discrete sequence correlation factors related to the hydrophobicity and hydrophilicity
Dipeptide composition (DPC) [161]	Frequencies of 400 types of dipeptides present over the sequence
Tripeptide composition (TPC)[152]	Frequencies of 8,000 types of tripeptides present over the sequence
Grouped amino acid composition (GAAC) [23]	Frequencies of five groups of amino acids based on their physicochemical properties: negative charge (D, E), positive charge (H, R, K), aromatic group (F, Y, W), aliphatic group (A, G, I, L, M, V), and uncharged (C, N, P, Q, S, T).
Terminus composition (TC) [162]	Frequencies of amino acids and dipeptides for 5, 10, and 15 residues present at the N- and C-terminus of the peptide sequence.
Composition of *k*-spaced amino acid pairs (CKSAAP)[166]	Frequencies of 400 types of residue pairs separated by k other amino acids ($k = 1$, 2, 3) within a sequence or sequence fragment.
CTDT (Composition/Transition/Distribution)	Distribution of amino acid composition patterns linked to specific chemical, physical, or structural properties within the peptide sequence. The Composition (C) refers to the amino acid composition in sequence, the Transition (T) corresponds to changes among three patterns: neutral, hydrophobic, and polar, and the Distribution (D) refers to the pattern of distribution of these properties over the sequence.
Pseudo K-tuple reduced amino acids composition (PseKRAAC)[165]	Frequencies of the 16 types of reduced K-tuple pseudo amino acids calculated from the sequence-order information for all dipeptides and the correlation between nth nearest residue.

View Article Online
DOI: 10.1039/D5CP04611D

Adaptive skip dipeptide composition (ASDC)[167]	Frequencies of amino acid pairs separated by a variable (adaptive) number of intervening residues.
Quasi-sequence-order descriptors (QSOrder)[183]	Frequencies of the amino acid sequence orders calculated using the sequence-order-coupling numbers that reflect the interactions between amino acids at various ranks of proximity. The coupling factor used to calculate these numbers is based on the physicochemical distance between amino acids, which considers properties like hydrophobicity, hydrophilicity, side-chain volume, and polarity.
Secondary structure elements content (SSEC)	Number of α-helices, β-sheets, and coils
Shannon information entropy	Scoring value that measures the degree of variability at a specific amino acid position in a multiple sequence alignment
AAindex[169]	Compilation of literature-reported scales that quantify physicochemical tendencies of the standard amino acids. Each scale corresponds to one property and is encoded as a 20-element numeric vector, assigning a specific value to each amino acid
GRAVY index [184, *]	Hydropathic character
FLEX index[185,**]	Structural Flexibility

Notes: ** GRAVY: Grand Average of Hydropathy, corresponds to the value of the hydropathic index calculated by the Kyte-Doolittle method using the peptide sequence; ** FLEX index: corresponds to the structural flexibility calculated from the peptide sequence according to the Vihinen et al. 1994.

View Article Online
DOI: 10.1039/D5CP04611D

Table 3. Substitution matrices are derived from multiple amino acid sequence alignments and represent the substitution patterns in polypeptide sequences verified over evolution. The scoring matrices are a subclass of biological matrices derived from diverse data that represent position-specific variability, physicochemical properties, or pairwise energy content for each amino acid.

Scoring and Substitution Matrices	Peptide information
Position-specific scoring matrix (PSSM)[186]	Scoring matrix containing the likelihood of each amino acid at a specific position in a peptide sequence. It is derived from multiple sequence alignments, aiding the identification of conserved regions.
Residue pairwise energy content matrix (RECM)[170]	Scoring substitution 20 × 20 matrix containing residue pairwise energy for 20 standard amino acids derived from the primary structure of 674 proteins.
BLOcks Substitution Matrices (BLOSUM)[172, *]	Substitution 20 × 20 matrix based on observed substitutions in conserved blocks, with a threshold of identity*.
Grantham Distance Matrix[187]	Scoring substitution 20 × 20 matrix that incorporate residue substitution frequencies that better correspond to the overall chemical differences including composition, polarity, and molecular volume.
Point Accepted Mutation (PAM) matrices (also named Dayhoff matrices)	Substitution 20 x 20 matrix where each entry in the matrix represents the likelihood of one amino acid being replaced by another through accepted mutations over a specified evolutionary period
Z-scale[171]	Scoring 87 × 26 matrix applied for amino acid sequences, where the 87 rows correspond to the different amino acids (including 20 standard amino acids plus many non-coded or unusual ones) and 26 columns correspond to different physicochemical descriptor scores.

Note: * Different identity thresholds can be applied in BLOSUM to characterize peptide sequences, with 62% typically used in most alignments.

View Article Online
DOI: 10.1039/D5CP04611D

Table 4. List of autocorrelation descriptors associated with eight molecular properties related to the amino acids.

Autocorrelation Descriptors	Equations
Moran*	$I(d) = \frac{\frac{1}{N-d}\sum_{i=1}^{N-d}(P_i - \overline{P})(P_{i+d} - \overline{P})}{\frac{1}{N}\sum_{i=1}^{N-d}(P_i - \overline{P})^2}$, $d = 1,2,3,\ldots, nlag$ $\overline{P} = \frac{\sum_{r=1}^{20} P_r}{20}$, $\sigma = \sqrt{\frac{1}{20}\sum_{r=1}^{20}(P_i - \overline{P})^2}$
Geary**	$C(d) = \frac{\frac{1}{2(N-d)}\sum_{i=1}^{N-d}(P_i - P_{i+d})^2}{\frac{1}{N-1}\sum_{i=1}^{N}(P_i - \overline{P})^2}$, $d = 1,2,3,\ldots, nlag$
Normalized Moreau-Broto autocorrelation ***	$ATS(d) = \frac{AC(d)}{N-d}$, $d = 1,2,3,\ldots, nlag$ $AC(d) = \sum_{i=1}^{N-d} P_i \times P_{i+d}, d = 1,2,3,\ldots, nlag$

Notes:

* *I*(*d*) is the Moran autocorrelation, *d* is the lag of the autocorrelation, *nlag* is the maximum value of the lag (default value: 30), P_i and P_{i+d} are the properties of the amino acids at positions *i* and *i* + *d*, respectively. $\overline{P}$ is the average of the considered property *P* over the entire sequence of length *N*.

** *C (d)* is the Geary autocorrelation, *d, P, Pi*, and P_{i+d}, *nlag*, and N have the same definitions as defined for Moran.

*** ATS (*d*) is the Normalized Moreau-Broto autocorrelation, AC(*d*) is the Moreau-Broto autocorrelation, *d, P, Pi*, and P_{i+d}, nlag, and N have the same definitions as defined for Moran.

View Article Online
DOI: 10.1039/D5CP04611D

Molecular Fingerprints for Peptides

The selection of an appropriate molecular representation and the molecular properties most correlated with the investigated set of compounds plays a crucial role in analyzing structure-property relationships and exploring broader chemical space[188]. Several computational strategies have emerged to capture peptides' chemical and functional space[6,15,189,190]. These approaches have been applied to different bioactive peptide classes[189–191], peptides approved for human use [57], as well as molecules derived from peptides (peptide-like molecules, e.g, peptoids)[46].

Molecular fingerprints provide a cost-efficient computational method for analyzing large compound libraries, due to their compact representation of complex molecular structures[6,192], which justifies their integration with computationally demanding virtual screening techniques[2,193]. Molecular fingerprints serve as representations of a chemical structure that encode the presence or the absence of a particular molecular feature[193,194]. These types of molecular representation are essential for analyzing large chemical libraries and comparing their structures using quantitative assessment of pairwise similarity [194]. Currently, six main categories of molecular fingerprints are used to describe molecules: (1) descriptor-based, (2) substructure-based, (3) pharmacophore-based, (4) path-based (or hashed), (5) string-based, and (6) circular fingerprints[195].

Descriptor-based fingerprints use molecular features derived from physicochemical properties, such as the Van der Waals Surface Area (VSA) fingerprint. The substructure-based fingerprints are used to identify the presence of specific substructures, including functional groups and rings of certain sizes. This class includes the MACCS (Molecular ACCess System) key fingerprint[196]. The pharmacophore fingerprints encode the pharmacophore groups present in molecules, and this class characterizes the interaction of the molecules with the protein environment. Belonging to this class is the MXFP, an atom-pair fingerprint that describes molecular shape and pharmacophores[192]. The path-based (or hashed) fingerprints identify all types of subgraphs, including linear subgraphs representing the shortest paths between atom pairs and circular fingerprints that

View Article Online
DOI: 10.1039/D5CP04611D

capture the neighborhoods of bonded atoms, hashing them inside a fixed-size vector. Atom-pairs are a subclass of path-based that describes a molecule by analyzing all possible triplets present in two atoms and the shortest path that connects them[197]. These fingerprints include the E3FP[198], ECFP, and MAP4[6]. The string-based fingerprints create molecular representations by analyzing the SMILES string of a compound rather than its graphical representation. Finally, the circular fingerprints decompose the analyzed compound into various fragments, similar to substructure-based fingerprints. However, instead of depending on predefined structural patterns, they dynamically generate these fragments from the molecular graph of each compound.

Currently, most virtual screening strategies or chemical space mapping of compounds applied in drug discovery use the MACCS key fingerprint, Morgan fingerprint - commonly referred to as the ECFP fingerprint [199], and MinHashed fingerprint MHFP6[200]. Nevertheless, these molecular fingerprints usually struggle to accurately capture the overall characteristics of molecules, including their size and shape. Additionally, they are inadequate at recognizing structural variations that could be significant in larger molecules, such as distinguishing between linkers of varying lengths, identifying scrambled peptide sequences with the same amino acid composition and sequence length, or differentiating between regioisomers[52].

A pharmacophore-based fingerprint derived from the 2D structure of peptides, termed 2DP, was developed to encode the molecular shape and pharmacophore properties of peptides. This fingerprint represents the peptides' topology as a graph where nodes correspond to α-carbon atoms and edges represent bonds between them. This fingerprint captures key molecular features, including the number of hydrophobic, positively charged, negatively charged, and total non-hydrogen atoms in each residue. Distances between atom pairs are calculated along the shortest path in the peptide's topology, and Gaussian functions centered on these distances are used to generate a 136-dimensional chemical space. This fingerprinting method enables the exploration of peptides with unknown or flexible 3D structures, making it particularly suited for studying unconventional topologies like bicyclic peptides[201].

View Article Online
DOI: 10.1039/D5CP04611D

It has been demonstrated that some 2D fingerprints can effectively distinguish between peptide-like molecules with varying degrees of biological activity. Eckert and Bajorath (2007) found that Molprint2D performed best in recovering active molecules with strong peptide character. However, the property descriptor-based fingerprint excelled in identifying compounds with lower peptide character, indicating its utility in transitioning from peptide-like compounds to non-peptide alternatives[45]. Capecchi et al. (2020) developed the MAP4 which represents the relationships between pairs of atoms in a molecule, considering their types and the topological distance. This fingerprint was designed to handle large and complex molecules, such as peptides, proteins, and peptide-like compounds, while maintaining computational efficiency [52]. Recently, Capecchi and Reymond (2021) used a genetic algorithm with the molecular fingerprint MAP4 to represent the chemical space of peptides, organizing them by sequence and size. The chemical space represents 40,531 peptides from eleven open-access peptide and peptide-containing databases, and the map obtained categorizes the peptides by activity type, indicating that the majority of the peptides in the investigated databases, comprising 17,260 sequences, or 43% of the total, are classified as antimicrobial and anticancer [6]. The Reymond group also developed MAP4C, a chiral adaptation of the MAP4 fingerprint, to analyze the stereochemical properties of large molecules, such as peptides. This fingerprint generates MinHashes derived from character strings encoding the SMILES representations of all pairs of circular substructures with diameters of up to four bonds and the shortest topological distance between their central atoms. The MAP4C incorporates Cahn-Ingold-Prelog (CIP) annotations (R, S, r, or s) for chiral atoms at the center of circular substructures, uses a question mark for undefined stereocenters, and includes *cis-trans* information for double bonds when specified. In non-stereoselective virtual screening approaches, MAP4C performs slightly better than the achiral MAP4, ECFP, and AP fingerprints[202]. To evaluate the chemical space of antimicrobial peptides (AMPs) and identify new candidates with therapeutic potential, Orsi et al. (2024)[190] integrated cheminformatics, ligand-based virtual screening, and machine-learning techniques. Virtual peptide libraries, including bicyclic and dendritic structures, were constructed and analyzed using molecular fingerprints, such as MAP4 and its chiral variant, MAP4C. These fingerprints measure molecular similarities and

View Article Online
DOI: 10.1039/D5CP04611D

facilitate the visualization of the chemical space through dimensionality reduction methods, including PCA. The ligand-based virtual screening was employed to prioritize AMP candidates based on their similarity or diversity to known bioactive molecules, significantly enhancing the efficiency and success rate compared to random selection. Furthermore, machine- learning models, such as support vector machines and recurrent neural networks, were trained on experimental AMP datasets to predict antimicrobial activity and toxicity, aiding in identifying promising peptides for experimental validation[190].

Some molecular fingerprints applied for peptide analyses are described in Table 5. We focused on explaining the most commonly used fingerprints, which are applied for the analysis of peptides and are usually accessible on the most used C++, Java, and Python libraries, such as Scikit-fingerprints[203], RDKit[21], iFeatureOmega[22], and Open Babel[204].

Table 5. List of molecular fingerprints applied to analyze peptides' chemical space and their respective description, category, and implemented open-source libraries or developer website source.

Molecular Fingerprint	Description	Category	Implemented libraries or Websites
MinHashed atom-pair up to a diameter of four bonds fingerprint (MAP4)[52]	The circular substructures (radii r = 1 and r = 2) around each atom in an atom pair are represented as two SMILES pairs linked by the topological distance between the central atoms. These atom-pair molecular shingles are hashed and then undergo MinHashing.	Circular and path-based fingerprint (subclass atom-pair)	scikit-fingerprints GitHub (https://github.com/reymond-group/map4)
Chiral MAP4 (MAP4C)[6]	Chiral representation of the MAP4 fingerprint	String-based and path-based fingerprint	GitHub (https://github.com/reymond-group/mapchiral)
Molecular ACCess System (MACCS Key)	Consists of a fixed-length bit vector, typically 166 bits. Encodes molecules'	Substructure fingerprint	RDKit, OpeBabel, and scikit-fingerprints

View Article Online
DOI: 10.1039/D5CP04611D

	predefined substructures or functional groups, such as rings, bonds, and specific atom types.		
Extended Connectivity Fingerprint 6/4 (ECFP6, ECFP4)	Encodes the environment of each atom circularly, capturing information about the atom and its neighboring atoms up to a specified radius.	Path-based fingerprint	RDKit, scikit-fingerprints, OpenBabel, and iFeatureOmega
Extended-connectivity Count fingerprint (ECFC6)	A variant of ECFPs that not only indicates the presence of specific substructures but also counts the occurrences of each substructure within the molecule. The "6" refers to the maximum diameter considered during the fingerprint generation.	Path-based fingerprint	RDKit, scikit-fingerprints, and OpenBabel
DompeKeys[205]	Set of substructure-based fingerprint descriptors designed to encode patterns of functional groups and chemical features within molecular structures.	Substructure fingerprint	Developer website (https://dompekeys.exscalate.eu),
MolPrint2D[206]	Encodes molecular structures by representing the atom environment up to a specific distance. It generates exhaustive lists of substructures surrounding each atom, which are then indexed for similarity comparison.	Circular fingerprint	OpenBabel
Macromolecule eXtended FingerPrint (MXFP)[192]	A 217-dimensional fuzzy fingerprint representing atom pairs from seven pharmacophore groups, which is ideal for comparing large molecules and facilitating scaffold hopping.	Pharmacophore fingerprint	Developer GitHub (https://github.com/markusorsi/mxfp_python)

View Article Online
DOI: 10.1039/D5CP04611D

Embeddings for Peptides

Embeddings are continuous numerical representations of discrete elements useful to compress conformation, shape, physicochemical features, and context-dependent exposure of polar and hydrophobic groups from amino-acid sequences or molecular graph-based representations related to peptides in a meaningful way for modern machine and deep learning algorithms[207–209]. These embeddings can be derived from learned encoders, such as graph neural networks (GNNs), language models (LMs), and autoencoders (AEs), which extract informative features from raw sequence or structural data during training[210]. Embeddings aim to produce representations of molecules in which similarity relationships and structural patterns become computationally exploitable.[207,211] The embeddings support the training of machine-learning models, clustering methods, and the performance of similarity-driven analysis[131,212,213].

In peptide modeling, embeddings are generated through neural or deep-learning encoders trained on sequence and/or structural data. These encoders learn internal representations that capture regularities in biochemical composition, residue interactions, structural motifs, and context-dependent effects[207,209]. The resulting latent space reflects patterns discovered from the data and organizes peptides according to shared properties and structural similarity.[214,215] A well-trained embedding preserves chemically relevant information while structuring peptides in a way that facilitates similarity analysis, interpolation, clustering, screening, and predictive modeling[210,216]. A schematic overview of an embedding-based workflow applied to peptide sciences is shown in Figure 6.

View Article Online
DOI: 10.1039/D5CP04611D

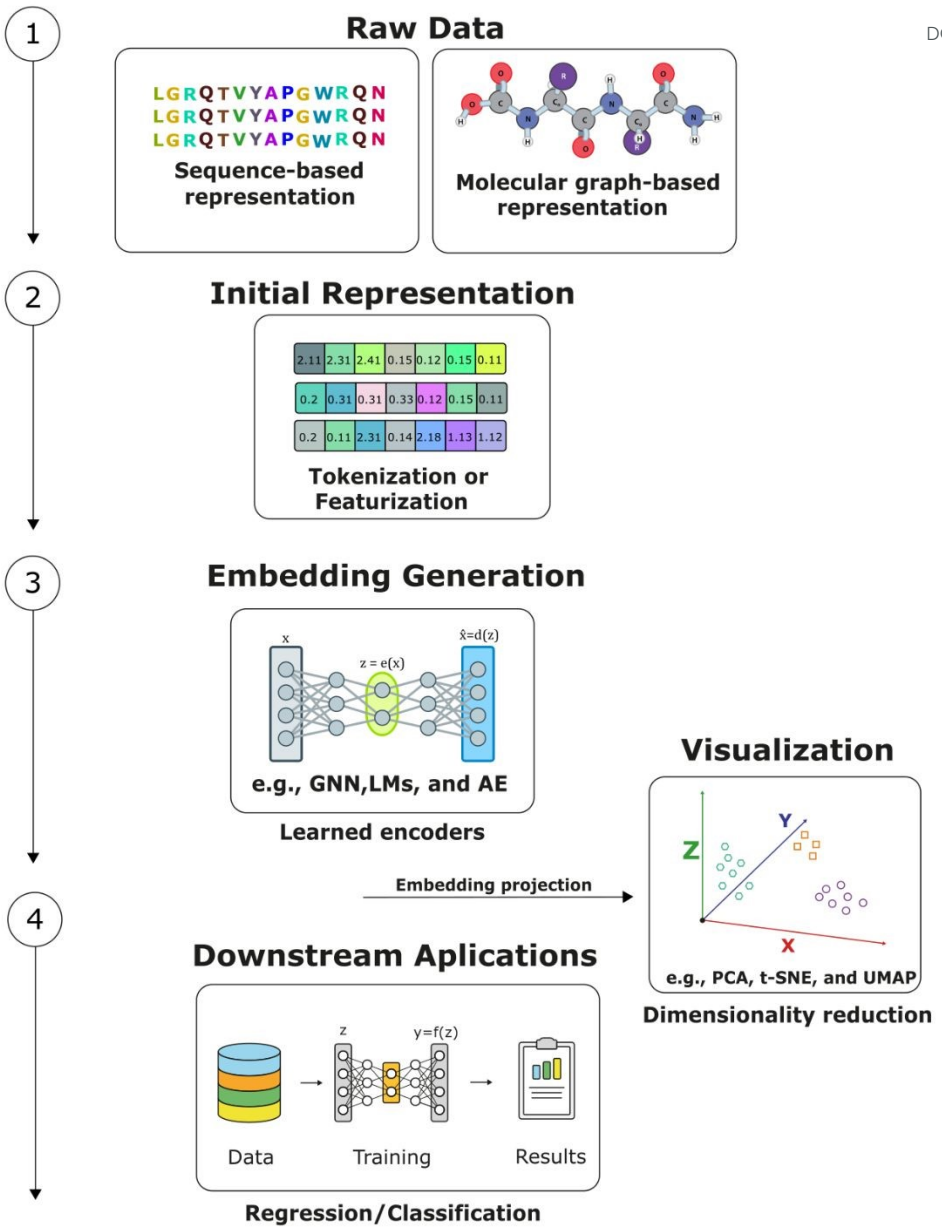

Figure 6. Raw peptide data, represented either as sequences or molecular graphs, are first converted into initial machine-readable representations through tokenization or featurization. Learned encoders, such as graph neural networks (GNNs), language models (LMs), and autoencoders (AEs), are then used to generate embeddings that capture relevant peptide features in a compact latent space. These embeddings can subsequently support downstream tasks, including regression and classification, and can be projected into lower-dimensional spaces for visualization using methods such as PCA, t-SNE, and UMAP.

View Article Online
DOI: 10.1039/D5CP04611D

The choice of representation of peptides is especially important because these molecules can be encoded at different hierarchical levels. A growing body of peptide cheminformatics literature emphasizes that amino acid-based representations often better reflect the functional building blocks of peptides than purely atom-level descriptions, since peptide activity is frequently driven by residue identity, order, and context[131]. Accordingly, residue-level notations such as FASTA[128], PLN[217], HELM[130], and BILN[129] are highly relevant for embedding workflows. Although FASTA provides a simple encoding for canonical peptide sequences, the PLN extends the representation to a broader range of modified peptides, and HELM offers a richer formalism for complex biomolecules, including cyclic, branched, and crosslinked peptides, while BILN improves the human readability of HELM-derived peptide descriptions[128,130].

A good encoding scheme preserves the relevant chemical content while producing inputs that are compatible with modern deep-learning pipelines[218]. For residue-based representations, the usual workflow begins by tokenizing the sequence into amino acids or modified monomers, followed by conversion into machine-readable vectors. In the simplest case, one-hot encoding represents each residue by a sparse binary vector, typically defined over the alphabet of canonical amino acids, although the vocabulary can be expanded to include non-canonical residues and common chemical modifications. More informative sequence encoders instead map each token to a dense learned vector, allowing the model to capture contextual dependencies, long-range interactions, and position-dependent effects across the peptide chain. In parallel, some peptide-focused pipelines use property-informed residue encodings, in which each amino acid is represented by physicochemical descriptors such as hydrophobicity, charge, steric parameters, or polarity-related indices, thereby injecting biochemical priors that can be especially helpful when data are limited or when interpretability is desired.

Atom-based molecular representation formed by strings of characters, such as tokenized SMILES, peptide/biopolymer notations like CHUCKLES (monomer-sequence SMILES translation), or robust grammars such as SELFIES, a typical workflow starts by converting the string into tokens and then mapping those tokens to numerical vectors suitable for neural networks[3]. These

View Article Online
DOI: 10.1039/D5CP04611D

tokens are processed by neural architectures that can capture contextual dependencies along the sequence, such as long-range residue interactions and position-dependent effects. Through training, the model internalizes statistical regularities of sequence composition and structural tendencies, yielding embeddings that reflect both local and global sequence organization (Figure 7, panel A). Neural architectures specifically designed for graph processing operate directly on these topological structures, learning embeddings that encode local chemical environments as well as global connectivity patterns[219]. These representations incorporate structural constraints, stereochemical relationships, and bond-level information, making them particularly suitable for computational tasks. In the one-hot encoding, each symbol in the amino acid alphabet is represented by a sparse binary vector with a single “1” at the index of that symbol and “0” elsewhere, and each peptide is represented by a vector of length N, where N is often the 20 canonical amino acids, but it can be expanded to include non-canonical residues and common chemical modifications. In parallel, peptide-focused pipelines often incorporate property-informed encodings, where each residue is mapped to a vector of physicochemical descriptors (e.g., hydrophobicity, charge-related indices, steric parameters, etc). These encodings inject biochemical priors that can be helpful when data are limited or when interpretability of residue contributions is desired[220,221]. For graph-based representations of peptides, commonly derived from structural formats such as SDF, MOL, MOL2, and CDX, the atoms are represented as nodes and the bonds as edges, the encoding typically requires building an adjacency matrix to capture atomic connectivity and a node feature matrix to describe atom-level attributes; in many cases, edge features are added as well to encode bond properties such as type, order, or aromaticity (Figure 7, panel B)[212].

View Article Online
DOI: 10.1039/D5CP04611D

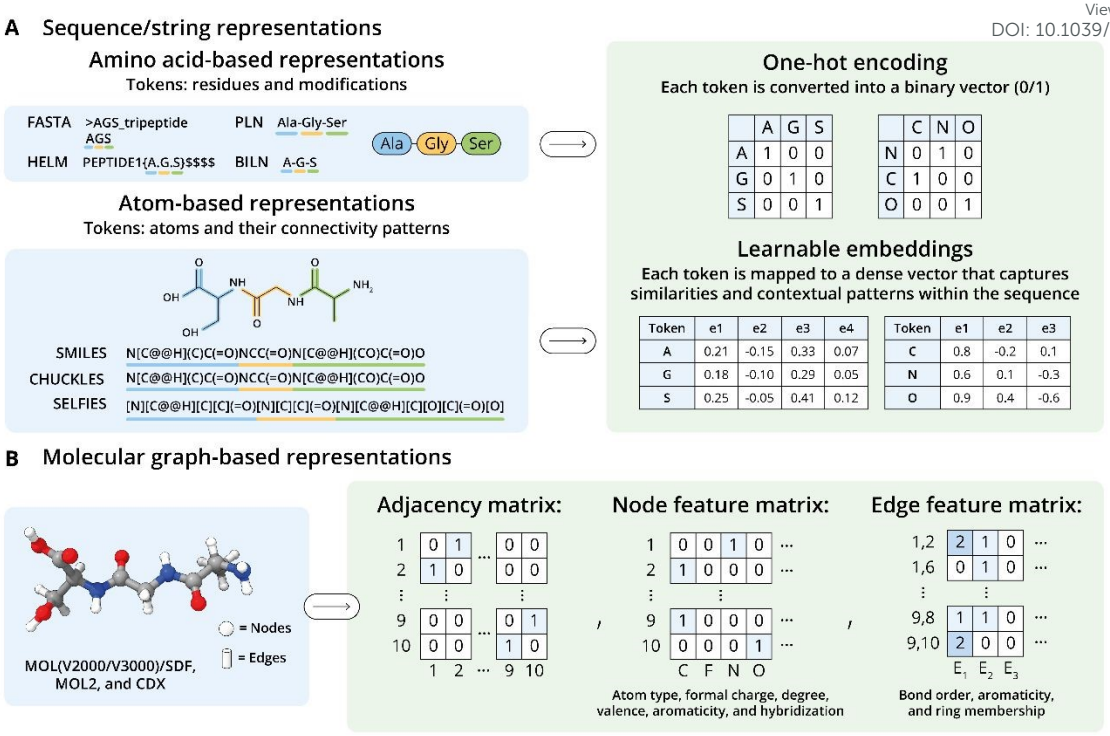

Figure 7. Overview of sequence- and graph-based representations used to generate embeddings for peptides. (A) Sequence- and string-based representations. Peptides can be represented using amino acid-based notations, in which tokens correspond to residues and possible modifications, or atom-based chemical string representations, in which tokens encode atoms and their connectivity patterns. These discrete tokens can then be transformed into numerical representations through one-hot encoding or dense learned embeddings. (B) Molecular graph-based representations. Chemical structures derived from formats such as MOL, SDF, MOL2, and CDX can be converted into molecular graphs, in which atoms are represented as nodes and bonds as edges. These graphs are typically described by an adjacency matrix together with node and edge feature matrices containing molecular information, such as atom type, formal charge, degree, valence, aromaticity, hybridization, bond order, and ring membership.

For interpretability and exploratory analysis, low-dimensional projection techniques such as PCA, UMAP, or t-SNE may be applied to the higher-dimensional embeddings (latent feature space). These methods are used solely for visualization purposes, enabling qualitative assessment of similarity relationships, neighborhood structures, and clustering tendencies[222,223]. They do

View Article Online
DOI: 10.1039/D5CP04611D

not define the embedding itself; rather, they provide a reduced-dimensional view of the latent space. The chemically meaningful representation remains in the higher-dimensional latent space produced by the trained neural encoder[222,223].

Recent computational workflows increasingly rely on embeddings from learned representations rather than other classes of descriptors, because embeddings can encode sequence context in a way that better reflects both biological function and indirectly structural constraints[214,224]. This shift was catalyzed by large protein language models (pLMs)[225] such as ESM-1b[207] and the Rostlab ProtTrans family (e.g., ProtBERT[226] and ProtT5[209]), which transform a protein or peptide sequence into dense vectors that summarize informative patterns across the entire chain. In practice, these models are pretrained with self-supervised objectives, learning the likelihood of residues given their surrounding context and producing contextualized embeddings at the residue and/or sequence level[227]. Because pLMs are trained on massive sequence collections such as UniRef50[228], the resulting embeddings can be reused as general-purpose features to be implemented in different computational tasks[229,230]. Importantly, recent studies indicate that pLM-derived embeddings can also be effective for peptides, frequently matching or outperforming traditional representations based on composition and physicochemical descriptors in predictive modeling and similarity analyses[230,231]. Recently, specific learned representations for peptides have been developed to predict peptide properties, including PeptideCLM[229] and Multi-Peptide.[232]

The Chemical Spaces Overlappings and Molecular Determinants of Peptide Bioactivities

Studies have revealed unexpected overlaps in the bioactivities of certain unrelated peptide classes,[47,48,57], thus suggesting that previously distinct chemical spaces may share molecular similarities. QSPs, CPPs, and B3PPs are examples of such peptide classes.

Quorum-sensing peptides are signaling molecules that enable communication within bacterial communities and coordinate their behavior based on population density. The QSPs regulate various physiological activities, including biofilm formation and virulence factor production. These peptides play

View Article Online
DOI: 10.1039/D5CP04611D

a crucial role in this communication, often functioning as autoinducers that bind to specific receptors on neighboring cells, triggering a cascade of gene expression changes[47,124]. Several studies have focused on analyzing their molecular properties to create prediction models of these peptides[124,151]. However, these molecules have also been shown to have selective BBBP properties[48] as well as to interact with mammalian cells, selectively promoting cancer metastasis [173,233], influencing immune[234], and muscle[235] cells. According to Wynendaele et al. (2015), the chemical space of quorum-sensing peptides is divided into three main clusters, as indicated by analyses of principal components. The peptide size and compactness comprise the first cluster. The descriptors that illustrate these characteristics include the radial distribution function (RDF), Burden eigenvalues (BEH, BEL), Randic shape indices, autocorrelation descriptors (ATS, GATS), weighted holistic invariant molecular (WHIM) descriptors, Balaban index, and the lopping centric index. Furthermore, the chemical space is also influenced by lipophilicity and hydrophilicity, evaluated through LogP values, tPSA, and the counts of HBD and HBA, along with connectivity indices that account for peptide cyclization and descriptors related to HOMA, AROM, and ARR aromaticity, which define the second principal component. The third principal component is characterized by S-evaluating descriptors representing thiol groups, thiolactones, or disulfides, while the fourth principal component emphasizes the presence and frequency of nitrogen bonds (N–N, N–O, and N–C). As a result, peptides high in cysteine and methionine cluster together, whereas those with basic amino acids and amides, such as asparagine form another cluster. Investigating the brain influx and efflux properties of three chemically diverse QSPs, Wynendaele et al. (2015) identified, according to clustering of the PCA results, three peptides named PhrCACET1, BIP-2, and PhrANTH2. These QSPs were investigated using a multiple-time regression technique in an *in vivo* mouse model (ICR-CD-1) to assess blood-brain transfer characteristics. The authors discovered that these peptides show blood-brain barrier (BBB) permeation, as well. The PhrCACET1 exhibited a notably high initial influx into the mouse brain (Kin = 20.87 µl/(g×min)), whereas the brain penetrabilities of BIP-2 and PhrANTH2 were determined to be low (Kin = 2.68 µl/(g×min)) and very low (Kin = 0.18 µl/(g×min)), respectively [48]. These

View Article Online
DOI: 10.1039/D5CP04611D

findings directly implicate the chemical characterization of peptide space and demonstrate the existence of an intersection not been characterized previously.

Similarly, the CPPs have been identified with blood-brain barrier permeation[47]. For example, de Oliveira et al. (2021) identified that CPPs possess higher MW, tPSA, and NRB values compared to clinically approved peptides, suggesting that their mechanisms of membrane penetration may involve processes beyond passive diffusion, such as pore formation or endocytosis. Additionally, their findings emphasize the importance of molecular flexibility and specific structural features, such as hydrogen bond patterns and the presence of aromatic rings, in influencing the permeability of these peptides, which could be related to their stereoselectivity. Regarding the B3PPs, Cavaco et al. (2024) recently identified key molecular determinants for peptides effectively crossing the BBB: a slightly hydrophobic nature, with a mean hydrophobic residue content of approximately 35%; a small size, with an average molecular weight of 2,046 g mol^{-1}; few or no aromatic residues, indicated by an average molar absorptivity of 3790 $M^{-1} \cdot cm^{-1}$ at 280 nm, which corresponds to 1–2 tyrosine or 0–1 tryptophan residues; and a slightly cationic charge, with an average net charge of +2. The study emphasizes that not all CPPs can function as B3PPs, as the overlap between these two families is minimal. Experimental validation demonstrated that four newly identified B3PPs exhibited high translocation abilities *in vitro* and greater brain accumulation *in vivo* than established B3PPs, highlighting the importance of specific physicochemical characteristics for effective brain targeting [236].

Despite the cell membrane and BBB having highly diverse functional and chemical compositions, and various molecular mechanisms of permeation being described for molecules into these membranes, most predictive models typically attribute passive transport through the membrane as the most important mechanism[58]. Furthermore, the biophysical interaction with these membranes has been explored as a key factor for permeation. Therefore, it is interesting to note that some molecular features usually applied to predict the CPPs have also been pointed out as relevant to predicting B3PPs. For example, Dichiara et al. (2019) established a set of chemical descriptors to facilitate the successful prediction of BBB permeation. They evaluated statistically 328 compounds,

View Article Online
DOI: 10.1039/D5CP04611D

correlating their experimental in vivo LogBB values with various computed descriptors. They constructed contingency tables, calculated observed and expected distributions, and analyzed the relationships between descriptors and BBB permeation. The authors identified a significant influence of nine specific physicochemical properties on BBB permeation, including polar surface area, nitrogen and oxygen count, LogP, nitrogen count, LogD, oxygen count, ionization state, hydrogen bond acceptors, and hydrogen bond donors[237].

Despite both classes representing distinct biological activity, a previous study showed that chemically distinct CPPs named pVEC, SynB3, Tat 47-57, transportan 10 (TP10), and TP10-2 exhibit varying abilities to enter the BBB. Specifically, Tat 47-57, SynB3, and pVEC demonstrated significantly high rates of unidirectional influx, whereas the transportan variants displayed minimal to low brain penetration[47].

Final Considerations

Exploring chemical space and the concept of a chemical multiverse for peptides can provide a robust framework for understanding the diverse properties and biological activities of these biomolecules, thereby opening new avenues for the identification of novel chemical entities in screening strategies as well as for the design of new bioactive peptides [14,238].

The concept of chemical space gains particular significance when applied to peptides because their amino acid sequence intrinsically encodes and influences physicochemical and structural properties such as solubility, hydrophobicity, folding patterns, and three-dimensional conformation, which ultimately shape their biological activities. The unique characteristics of peptides, especially their conformational flexibility, further underscore the complexity of their chemical space, as these molecules can adopt distinct conformational states depending on the environment, which is often crucial for their biological function. This flexibility contributes to their ability to interact with diverse biological targets and, in some cases, to penetrate biological barriers. The choice of molecular representation strongly determines which aspects of peptide behavior become computationally accessible. Classical molecular descriptors and fingerprints

View Article Online
DOI: 10.1039/D5CP04611D

remain essential for interpretable, scalable, and cost-effective analyses, particularly in virtual screening strategies. However, recent advances in machine learning have expanded this landscape by enabling embeddings derived from learned encoders, such as GNNs, LMs, and AEs, which can extract informative features directly from raw sequence or structural data. These learned representations organize peptides in latent spaces where similarity, clustering patterns, and predictive relationships become more readily exploitable, thus providing a powerful complement to conventional descriptor-based strategies.

In addition, peptide bioactivity should not be interpreted independently of conformation and context. Features such as chameleonicity, secondary-structure propensity, backbone flexibility, polarity masking, and membrane-interaction mechanisms reinforce the notion that peptide function emerges from a dynamic relationship between structure and environment. This complexity also helps explain why apparently unrelated peptide classes may partially overlap in chemical space and bioactivity, revealing intersections that are biologically meaningful and potentially useful for peptide discovery, repurposing, and design. Moreover, the overlap among distinct peptide classes with pleiotropic activities, such as QSPs, CPPs, and B3PPs, suggests shared regions of chemical space that warrant further investigation and may reveal new peptide functions as well as biotechnological and therapeutic opportunities.

Declaration of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Author Contributions

ECLO, LDN, AHLL, CS, BDS, and KS conceptualized the study. ECLO, JA, GPC, LDN, AHLL, CMFR, ADS, EW, CS, BDS, and KS contributed to data curation, analysis, interpretation of the results, and scientific discussion. ECLO, JA, GPC, LDN, AHLL, and CMFR contributed to the visualization and preparation of figures, tables, and graphical materials. KS coordinated the study, supervised the project, wrote the manuscript, and critically reviewed the text. CMFR contributed to

View Article Online
DOI: 10.1039/D5CP04611D

manuscript improvement, revision of visual materials, and scientific refinement of the work. ADS, EW, LDN, CS, and BDS contributed to the critical review of the manuscript and the intellectual improvement of the study. All authors reviewed, edited, and approved the final version of the manuscript.

Funding Declaration

The research was funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES (ROR ID: 00x0ma614). K.S. is grateful for the National Council for Scientific and Technological Development (CNPq, grant numbers: 408367/2024-5 and 442559/2025-9), a Brazilian funding agency, for the financial support of the study.

Acknowledgments

The article processing charge for the publication of this research was funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES/Brazil (ROR ID: 00x0ma614). The authors have assigned the Creative Commons CC BY license to any accepted article version for open access. K.S. is grateful for the National Council for Scientific and Technological Development (CNPq, grant numbers: 408367/2024-5 and 442559/2025-9), a Brazilian funding agency, for the financial support of the study. The authors would like to thank the illustrator Miguel Silva for his countless revisions, which were essential to achieving the technical and scientific level of detail presented in the illustrative schemes.

References

(1) Medina-Franco, J. L.; Saldívar-González, F. I. Cheminformatics to Characterize Pharmacologically Active Natural Products. *Biomolecules* **2020**, *10* (11), 1566. https://doi.org/10.3390/biom10111566.

(2) Santana, K.; do Nascimento, L. D.; Lima e Lima, A.; Damasceno, V.;

Nahum, C.; Braga, R. C.; Lameira, J. Applications of Virtual Screening in Bioprospecting: Facts, Shifts, and Perspectives to Explore the Chemo-Structural Diversity of Natural Products. *Front. Chem.* **2021**, *9*. https://doi.org/10.3389/fchem.2021.662688.

(3) David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular Representations in AI-Driven Drug Discovery: A Review and Practical Guide. *J. Cheminform.* **2020**, *12* (1), 56. https://doi.org/10.1186/s13321-020-00460-5.

(4) Reymond, J.-L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48* (3), 722–730. https://doi.org/10.1021/ar500432k.

(5) Díaz-Eufracio, B. I.; Palomino-Hernández, O.; Houghten, R. A.; Medina-Franco, J. L. Exploring the Chemical Space of Peptides for Drug Discovery: A Focus on Linear and Cyclic Penta-Peptides. *Mol. Divers.* **2018**, *22* (2), 259–267. https://doi.org/10.1007/s11030-018-9812-9.

(6) Capecchi, A.; Reymond, J.-L. L. Peptides in Chemical Space. *Med. Drug Discov.* **2021**, *9*, 100081. https://doi.org/10.1016/j.medidd.2021.100081.

(7) Schwaller, P.; Vaucher, A. C.; Laplaza, R.; Bunne, C.; Krause, A.; Corminboeuf, C.; Laino, T. Machine Intelligence for Chemical Reaction Space. *WIREs Comput. Mol. Sci.* **2022**, *12* (5). https://doi.org/10.1002/wcms.1604.

(8) Wang, M.; Li, S.; Wang, J.; Zhang, O.; Du, H.; Jiang, D.; Wu, Z.; Deng, Y.; Kang, Y.; Pan, P.; Li, D.; Wang, X.; Yao, X.; Hou, T.; Hsieh, C.-Y. ClickGen: Directed Exploration of Synthesizable Chemical Space via Modular Reactions and Reinforcement Learning. *Nat. Commun.* **2024**, *15* (1), 10127. https://doi.org/10.1038/s41467-024-54456-y.

(9) Kim, H.; Ryu, S.; Jung, N.; Yang, J.; Seok, C. CSearch: Chemical Space Search via Virtual Synthesis and Global Optimization. *J. Cheminform.* **2024**, *16* (1), 137. https://doi.org/10.1186/s13321-024-00936-8.

(10) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in

View Article Online
DOI: 10.1039/D5CP04611D

Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6* (12), 2326–2331. https://doi.org/10.1021/acs.jpclett.5b00831.

(11) Gorai, P.; Parilla, P.; Toberer, E. S.; Stevanović, V. Computational Exploration of the Binary A 1 B 1 Chemical Space for Thermoelectric Performance. *Chem. Mater.* **2015**, *27* (18), 6213–6221. https://doi.org/10.1021/acs.chemmater.5b01179.

(12) Mroz, A. M.; Posligua, V.; Tarzia, A.; Wolpert, E. H.; Jelfs, K. E. Into the Unknown: How Computation Can Help Explore Uncharted Material Space. *J. Am. Chem. Soc.* **2022**, *144* (41), 18730–18743. https://doi.org/10.1021/jacs.2c06833.

(13) Tudi, A.; Li, Z.; Xie, C.; Baiheti, T.; Tikhonov, E.; Zhang, F.; Pan, S.; Yang, Z. Functional Modules Map of Unexplored Chemical Space: Guiding the Discovery of Giant Birefringent Materials. *Adv. Funct. Mater.* **2024**, *34* (51). https://doi.org/10.1002/adfm.202409716.

(14) Di Bonaventura, I.; Baeriswyl, S.; Capecchi, A.; Gan, B.-H.; Jin, X.; Siriwardena, T. N.; He, R.; Köhler, T.; Pompilio, A.; Di Bonaventura, G.; van Delden, C.; Javor, S.; Reymond, J.-L. An Antimicrobial Bicyclic Peptide from Chemical Space against Multidrug Resistant Gram-Negative Bacteria. *Chem. Commun.* **2018**, *54* (40), 5130–5133. https://doi.org/10.1039/C8CC02412J.

(15) de Oliveira, E. C. L.; Santana, K.; Josino, L.; Lima e Lima, A. H.; de Souza de Sales Júnior, C. Predicting Cell-Penetrating Peptides Using Machine Learning Algorithms and Navigating in Their Chemical Space. *Sci. Rep.* **2021**, *11* (1). https://doi.org/10.1038/s41598-021-87134-w.

(16) Barazorda-Ccahuana, H. L.; Juárez-Mercado, K. E.; Medina-Franco, J. L.; Chavez-Fumagalli, M. A. Visualizing and Analyzing the Chemical Space of Natural Product Databases for Drug Discovery. *J. Vis. Exp.* **2024**, No. 211. https://doi.org/10.3791/66349.

(17) Saldívar-González, F. I.; Lenci, E.; Trabocchi, A.; Medina-Franco, J. L. Exploring the Chemical Space and the Bioactivity Profile of Lactams: A Chemoinformatic Study. *RSC Adv.* **2019**, *9* (46), 27105–27116.

View Article Online
DOI: 10.1039/D5CP04611D

https://doi.org/10.1039/C9RA04841C.

(18) López-López, E.; Sánchez-Castañeda, J. P.; Martinez-Cortés, M. S.; de la Fuente-Nunez, C.; Medina-Franco, J. L. Exploring and Expanding the Chemical Multiverse of Peptides. *Chem. Sci.* **2026**, *17* (3), 1461–1479. https://doi.org/10.1039/D5SC04465K.

(19) Digiesi, V.; de la Oliva Roque, V.; Vallaro, M.; Caron, G.; Ermondi, G. Permeability Prediction in the Beyond-Rule-of 5 Chemical Space: Focus on Cyclic Hexapeptides. *Eur. J. Pharm. Biopharm.* **2021**, *165*, 259–270. https://doi.org/10.1016/j.ejpb.2021.05.017.

(20) Pelton, J. M.; Hochuli, J. E.; Sadecki, P. W.; Katoh, T.; Suga, H.; Hicks, L. M.; Muratov, E. N.; Tropsha, A.; Bowers, A. A. Cheminformatics-Guided Cell-Free Exploration of Peptide Natural Products. *J. Am. Chem. Soc.* **2024**, *146* (12), 8016–8030. https://doi.org/10.1021/jacs.3c11306.

(21) Lovrić, M.; Molero, J. M.; Kern, R. PySpark and RDKit: Moving towards Big Data in Cheminformatics. *Mol. Inform.* **2019**, *38* (6), 1800082. https://doi.org/10.1002/minf.201800082.

(22) Chen, Z.; Liu, X.; Zhao, P.; Li, C.; Wang, Y.; Li, F.; Akutsu, T.; Bain, C.; Gasser, R. B.; Li, J.; Yang, Z.; Gao, X.; Kurgan, L.; Song, J. IFeatureOmega: An Integrative Platform for Engineering, Visualization and Analysis of Features from Molecular Sequences, Structural and Ligand Data Sets. *Nucleic Acids Res.* **2022**, *50* (W1), W434–W447. https://doi.org/10.1093/nar/gkac351.

(23) Chen, Z.; Zhao, P.; Li, F.; Leier, A.; Marquez-Lago, T. T.; Wang, Y.; Webb, G. I.; Smith, A. I.; Daly, R. J.; Chou, K.-C.; Song, J. IFeature : A Python Package and Web Server for Features Extraction and Selection from Protein and Peptide Sequences. *Bioinformatics* **2018**, *34* (14), 2499–2502. https://doi.org/10.1093/bioinformatics/bty140.

(24) Moriwaki, H.; Tian, Y.-S. S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminform.* **2018**, *10* (1), 4. https://doi.org/10.1186/s13321-018-0258-y.

(25) Van Der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach.*

View Article Online
DOI: 10.1039/D5CP04611D

Learn. Res. **2008**, *9*, 2579–2625.

(26) McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **2018**, *3* (29), 861. https://doi.org/10.21105/joss.00861.

(27) Cihan Sorkun, M.; Mullaj, D.; Koelman, J. M. V. A.; Er, S. ChemPlot, a Python Library for Chemical Space Visualization**. *Chemistry–Methods* **2022**, *2* (7). https://doi.org/10.1002/cmtd.202200005.

(28) Probst, D.; Reymond, J.-L. Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees. *J. Cheminform.* **2020**, *12* (1), 12. https://doi.org/10.1186/s13321-020-0416-x.

(29) Sosnin, S. Chemical Space Visual Navigation in the Era of Deep Learning and Big Data. *Drug Discov. Today* **2025**, *30* (7), 104392. https://doi.org/10.1016/j.drudis.2025.104392.

(30) Castillo-Mendieta, K.; Marrero-Ponce, Y.; Márquez, E. A.; García-Giménez, J. L.; Antunes, A.; Agüero-Chapin, G. Mapping the Antibiofilm Peptide Space with Similarity Networks and Curated Negative Sets. *ACS Omega* **2025**, *10* (49), 60457–60476. https://doi.org/10.1021/acsomega.5c07679.

(31) Zhang, B.; Vogt, M.; Maggiora, G. M.; Bajorath, J. Design of Chemical Space Networks Using a Tanimoto Similarity Variant Based upon Maximum Common Substructures. *J. Comput. Aided. Mol. Des.* **2015**, *29* (10), 937–950. https://doi.org/10.1007/s10822-015-9872-1.

(32) Burri, C.; Brun, R. Eflornithine for the Treatment of Human African Trypanosomiasis. *Parasitol. Res.* **2003**, *90 Supp 1*, S49-52. https://doi.org/10.1007/s00436-002-0766-5.

(33) de Llano García, D.; Marrero-Ponce, Y.; Agüero-Chapin, G.; Rodríguez, H.; Ferri, F. J.; Márquez, E. A.; Mora, J. R.; Martinez-Rios, F.; Pérez-Castillo, Y. Mapping the Chemical Space of Antiviral Peptides with Half-Space Proximal and Metadata Networks Through Interactive Data Mining. *Computers* **2025**, *14* (10), 423. https://doi.org/10.3390/computers14100423.

(34) Zwierzyna, M.; Vogt, M.; Maggiora, G. M.; Bajorath, J. Design and Characterization of Chemical Space Networks for Different Compound Data Sets. *J. Comput. Aided. Mol. Des.* **2015**, *29* (2), 113–125. https://doi.org/10.1007/s10822-014-9821-4.

(35) Castillo-Mendieta, K.; Agüero-Chapin, G.; Marquez, E. A.; Perez-Castillo, Y.; Barigye, S. J.; Vispo, N. S.; García-Jacas, C. R.; Marrero-Ponce, Y. Peptide Hemolytic Activity Analysis Using Visual Data Mining of Similarity-Based Complex Networks. *npj Syst. Biol. Appl.* **2024**, *10* (1), 115. https://doi.org/10.1038/s41540-024-00429-2.

(36) Maggiora, G. M.; Bajorath, J. Chemical Space Networks: A Powerful New Paradigm for the Description of Chemical Space. *J. Comput. Aided. Mol. Des.* **2014**, *28* (8), 795–802. https://doi.org/10.1007/s10822-014-9760-0.

(37) Aguilera-Mendoza, L.; Ayala-Ruano, S.; Martinez-Rios, F.; Chavez, E.; García-Jacas, C. R.; Brizuela, C. A.; Marrero-Ponce, Y. StarPep Toolbox : An Open-Source Software to Assist Chemical Space Analysis of Bioactive Peptides and Their Functions Using Complex Networks. *Bioinformatics* **2023**, *39* (8). https://doi.org/10.1093/bioinformatics/btad506.

(38) Ayala-Ruano, S.; Marrero-Ponce, Y.; Aguilera-Mendoza, L.; Pérez, N.; Agüero-Chapin, G.; Antunes, A.; Aguilar, A. C. Network Science and Group Fusion Similarity-Based Searching to Explore the Chemical Space of Antiparasitic Peptides. *ACS Omega* **2022**, *7* (50), 46012–46036. https://doi.org/10.1021/acsomega.2c03398.

(39) Castillo-Mendieta, K.; Agüero-Chapin, G.; Mora, J. R.; Pérez, N.; Contreras-Torres, E.; Valdes-Martini, J. R.; Martinez-Rios, F.; Marrero-Ponce, Y. Unraveling the Hemolytic Toxicity Tapestry of Peptides Using Chemical Space Complex Networks. *Toxicol. Sci.* **2024**, *202* (2), 236–249. https://doi.org/10.1093/toxsci/kfae115.

(40) Agüero-Chapin, G.; Antunes, A.; Mora, J. R.; Pérez, N.; Contreras-Torres, E.; Valdes-Martini, J. R.; Martinez-Rios, F.; Zambrano, C. H.; Marrero-Ponce, Y. Complex Networks Analyses of Antibiofilm Peptides: An Emerging Tool for Next-Generation Antimicrobials' Discovery. *Antibiotics*

2023, *12* (4), 747. https://doi.org/10.3390/antibiotics12040747.

View Article Online
DOI: 10.1039/D5CP04611D

(41) Kunimoto, R.; Bajorath, J. Combining Similarity Searching and Network Analysis for the Identification of Active Compounds. *ACS Omega* **2018**, *3* (4), 3768–3777. https://doi.org/10.1021/acsomega.8b00344.

(42) Lo, Y.-C.; Senese, S.; Li, C.-M.; Hu, Q.; Huang, Y.; Damoiseaux, R.; Torres, J. Z. Large-Scale Chemical Similarity Networks for Target Profiling of Compounds Identified in Cell-Based Chemical Screens. *PLOS Comput. Biol.* **2015**, *11* (3), e1004153. https://doi.org/10.1371/journal.pcbi.1004153.

(43) Wang, Q.; Hu, X.; Wei, Z.; Lu, H.; Liu, H. Reinforcement Learning-Driven Exploration of Peptide Space: Accelerating Generation of Drug-like Peptides. *Brief. Bioinform.* **2024**, *25* (5). https://doi.org/10.1093/bib/bbae444.

(44) Medina-Franco, J. L.; Chávez-Hernández, A. L.; López-López, E.; Saldívar-González, F. I. Chemical Multiverse: An Expanded View of Chemical Space. *Mol. Inform.* **2022**, *41* (11). https://doi.org/10.1002/minf.202200116.

(45) Eckert, H.; Bajorath, J. Exploring Peptide-Likeness of Active Molecules Using 2D Fingerprint Methods. *J. Chem. Inf. Model.* **2007**, *47* (4), 1366–1378. https://doi.org/10.1021/ci700086m.

(46) Orsi, M.; Reymond, J. Navigating a 1E+60 Chemical Space of Peptide/Peptoid Oligomers. *Mol. Inform.* **2024**. https://doi.org/10.1002/minf.202400186.

(47) Stalmans, S.; Bracke, N.; Wynendaele, E.; Gevaert, B.; Peremans, K.; Burvenich, C.; Polis, I.; De Spiegeleer, B. Cell-Penetrating Peptides Selectively Cross the Blood-Brain Barrier In Vivo. *PLoS One* **2015**, *10* (10), e0139652. https://doi.org/10.1371/journal.pone.0139652.

(48) Wynendaele, E.; Verbeke, F.; Stalmans, S.; Gevaert, B.; Janssens, Y.; Van De Wiele, C.; Peremans, K.; Burvenich, C.; De Spiegeleer, B. Quorum Sensing Peptides Selectively Penetrate the Blood-Brain Barrier. *PLoS One* **2015**, *10* (11), e0142071.

View Article Online
DOI: 10.1039/D5CP04611D

https://doi.org/10.1371/journal.pone.0142071.

(49) Naveja, J. J.; Medina-Franco, J. L. Finding Constellations in Chemical Space Through Core Analysis. *Front. Chem.* **2019**, *7*. https://doi.org/10.3389/fchem.2019.00510.

(50) Langdon, S. R.; Brown, N.; Blagg, J. Scaffold Diversity of Exemplified Medicinal Chemistry Space. *J. Chem. Inf. Model.* **2011**, *51* (9), 2174–2185. https://doi.org/10.1021/ci2001428.

(51) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887–2893. https://doi.org/10.1021/jm9602928.

(52) Capecchi, A.; Probst, D.; Reymond, J.-L. One Molecular Fingerprint to Rule Them All: Drugs, Biomolecules, and the Metabolome. *J. Cheminform.* **2020**, *12* (1), 43. https://doi.org/10.1186/s13321-020-00445-4.

(53) Doak, B. C.; Over, B.; Giordanetto, F.; Kihlberg, J. Oral Druggable Space beyond the Rule of 5: Insights from Drugs and Clinical Candidates. *Chem. Biol.* **2014**, *21* (9), 1115–1142. https://doi.org/10.1016/j.chembiol.2014.08.013.

(54) Bockus, A. T.; McEwen, C. M.; Lokey, R. S. Form and Function in Cyclic Peptide Natural Products: A Pharmacokinetic Perspective. *Curr. Top. Med. Chem.* **2013**, *13* (7), 821–836. https://doi.org/10.2174/1568026611313070005.

(55) Sunde, H.; Ryder, K.; Bekhit, A. E.-D. A.; Carne, A. Analysis of Peptides in a Sheep Beta Lactoglobulin Hydrolysate as a Model to Evaluate the Effect of Peptide Amino Acid Sequence on Bioactivity. *Food Chem.* **2021**, *365*, 130346. https://doi.org/10.1016/j.foodchem.2021.130346.

(56) Mishra, B.; Lakshmaiah Narayana, J.; Lushnikova, T.; Zhang, Y.; Golla, R. M.; Zarena, D.; Wang, G. Sequence Permutation Generates Peptides with Different Antimicrobial and Antibiofilm Activities. *Pharmaceuticals* **2020**, *13* (10), 271. https://doi.org/10.3390/ph13100271.

(57) Gevaert, B.; Stalmans, S.; Wynendaele, E.; Taevernier, L.; Bracke, N.; D';Hondt, M.; De Spiegeleer, B. Exploration of the Medicinal Peptide Space. *Protein Pept. Lett.* **2016**, *23* (4), 324–335. https://doi.org/10.2174/0929866523666160215162326.

(58) de Oliveira, E. C. L.; da Costa, K. S.; Taube, P. S.; Lima, A. H.; Junior, C. de S. de S. Biological Membrane-Penetrating Peptides: Computational Prediction and Applications. *Front. Cell. Infect. Microbiol.* **2022**, *12*. https://doi.org/10.3389/fcimb.2022.838259.

(59) Sánchez-Navarro, M.; Teixidó, M.; Giralt, E. Jumping Hurdles: Peptides Able to Overcome Biological Barriers. *Acc. Chem. Res.* **2017**, *50* (8), 1847–1854. https://doi.org/10.1021/acs.accounts.7b00204.

(60) Bock, J. E.; Gavenonis, J.; Kritzer, J. A. Getting in Shape: Controlling Peptide Bioactivity and Bioavailability Using Conformational Constraints. *ACS Chem. Biol.* **2013**, *8* (3), 488–499. https://doi.org/10.1021/cb300515u.

(61) Ramazi, S.; Zahiri, J. Post-Translational Modifications in Proteins: Resources, Tools and Prediction Methods. *Database* **2021**, *2021*. https://doi.org/10.1093/database/baab012.

(62) Miao, J.; Descoteaux, M. L.; Lin, Y. S. Structure Prediction of Cyclic Peptides by Molecular Dynamics + Machine Learning. *Chem. Sci.* **2021**, *12* (44), 14927–14936. https://doi.org/10.1039/d1sc05562c.

(63) Petkov, P.; Lilkova, E.; Ilieva, N.; Litov, L. Self-Association of Antimicrobial Peptides: A Molecular Dynamics Simulation Study on Bombinin. *Int. J. Mol. Sci.* **2019**, *20* (21), 5450. https://doi.org/10.3390/ijms20215450.

(64) Allison, J. R. Computational Methods for Exploring Protein Conformations. *Biochem. Soc. Trans.* **2020**, *48* (4), 1707–1724. https://doi.org/10.1042/BST20200193.

(65) Siligardi, G.; Drake, A. F. The Importance of Extended Conformations and, in Particular, the P II Conformation for the Molecular Recognition of Peptides. *Biopolymers* **1995**, *37* (4), 281–292.

View Article Online
DOI: 10.1039/D5CP04611D

https://doi.org/10.1002/bip.360370406.

(66) Herrera-León, C.; Ramos-Martín, F.; El Btaouri, H.; Antonietti, V.; Sonnet, P.; Martiny, L.; Zevolini, F.; Falciani, C.; Sarazin, C.; D'Amelio, N. The Influence of Short Motifs on the Anticancer Activity of HB43 Peptide. *Pharmaceutics* **2022**, *14* (5), 1089. https://doi.org/10.3390/pharmaceutics14051089.

(67) Schmitt, M. A.; Weisblum, B.; Gellman, S. H. Interplay among Folding, Sequence, and Lipophilicity in the Antibacterial and Hemolytic Activities of α/β-Peptides. *J. Am. Chem. Soc.* **2007**, *129* (2), 417–428. https://doi.org/10.1021/ja0666553.

(68) Chapman, R. N.; Dimartino, G.; Arora, P. S. A Highly Stable Short α-Helix Constrained by a Main-Chain Hydrogen-Bond Surrogate. *J. Am. Chem. Soc.* **2004**, *126* (39), 12252–12253. https://doi.org/10.1021/ja0466659.

(69) Miller, S. E.; Kallenbach, N. R.; Arora, P. S. Reversible α-Helix Formation Controlled by a Hydrogen Bond Surrogate. *Tetrahedron* **2012**, *68* (23), 4434–4437. https://doi.org/10.1016/j.tet.2011.12.068.

(70) Marcelo Der Torossian, T.; Silva, A. F.; Alves, F. L.; Capurro, M. L.; Miranda, A.; Vani Xavier, O. Highly Potential Antiplasmodial Restricted Peptides. *Chem. Biol. Drug Des.* **2015**, *85* (2), 163–171. https://doi.org/10.1111/cbdd.12354.

(71) Der Torossian Torres, M.; Silva, A. F.; Alves, F. L.; Capurro, M. L.; Miranda, A.; Oliveira Junior, V. X. The Importance of Ring Size and Position for the Antiplasmodial Activity of Angiotensin II Restricted Analogs. *Int. J. Pept. Res. Ther.* **2014**, *20* (3), 277–287. https://doi.org/10.1007/s10989-014-9392-1.

(72) Milletti, F. Cell-Penetrating Peptides: Classes, Origin, and Current Landscape. *Drug Discov. Today* **2012**, *17* (15–16), 850–860. https://doi.org/10.1016/j.drudis.2012.03.002.

(73) Oller-Salvia, B.; Sánchez-Navarro, M.; Giralt, E.; Teixidó, M. Blood–Brain Barrier Shuttle Peptides: An Emerging Paradigm for Brain Delivery. *Chem. Soc. Rev.* **2016**, *45* (17), 4690–4707.

View Article Online
DOI: 10.1039/D5CP04611D

https://doi.org/10.1039/C6CS00076B.

(74) Ramelot, T. A.; Palmer, J.; Montelione, G. T.; Bhardwaj, G. Cell-Permeable Chameleonic Peptides: Exploiting Conformational Dynamics in de Novo Cyclic Peptide Design. *Curr. Opin. Struct. Biol.* **2023**, *80*, 102603. https://doi.org/10.1016/j.sbi.2023.102603.

(75) Payne, C. D.; Franke, B.; Fisher, M. F.; Hajiaghaalipour, F.; McAleese, C. E.; Song, A.; Eliasson, C.; Zhang, J.; Jayasena, A. S.; Vadlamani, G.; Clark, R. J.; Minchin, R. F.; Mylne, J. S.; Rosengren, K. J. A Chameleonic Macrocyclic Peptide with Drug Delivery Applications. *Chem. Sci.* **2021**, *12* (19), 6670–6683. https://doi.org/10.1039/D1SC00692D.

(76) Lee, D.; Choi, J.; Yang, M. J.; Park, C.-J.; Seo, J. Controlling the Chameleonic Behavior and Membrane Permeability of Cyclosporine Derivatives via Backbone and Side Chain Modifications. *J. Med. Chem.* **2023**, *66* (18), 13189–13204. https://doi.org/10.1021/acs.jmedchem.3c01140.

(77) Linker, S. M.; Schellhaas, C.; Kamenik, A. S.; Veldhuizen, M. M.; Waibl, F.; Roth, H.-J.; Fouché, M.; Rodde, S.; Riniker, S. Lessons for Oral Bioavailability: How Conformationally Flexible Cyclic Peptides Enter and Cross Lipid Membranes. *J. Med. Chem.* **2023**, *66* (4), 2773–2788. https://doi.org/10.1021/acs.jmedchem.2c01837.

(78) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43* (20), 3714–3717. https://doi.org/10.1021/jm000942e.

(79) Matsson, P.; Doak, B. C.; Over, B.; Kihlberg, J. Cell Permeability beyond the Rule of 5. *Adv. Drug Deliv. Rev.* **2016**, *101*, 42–61. https://doi.org/10.1016/j.addr.2016.03.013.

(80) Galúcio, J. M.; Monteiro, E. F.; de Jesus, D. A.; Costa, C. H.; Siqueira, R. C.; Santos, G. B. dos; Lameira, J.; Costa, K. S. da. In Silico Identification of Natural Products with Anticancer Activity Using a Chemo-Structural Database of Brazilian Biodiversity. *Comput. Biol. Chem.* **2019**, *83*,

107102. https://doi.org/10.1016/j.compbiolchem.2019.107102.

View Article Online
DOI: 10.1039/D5CP04611D

(81) Rossi Sebastiano, M.; Doak, B. C.; Backlund, M.; Poongavanam, V.; Over, B.; Ermondi, G.; Caron, G.; Matsson, P.; Kihlberg, J. Impact of Dynamically Exposed Polarity on Permeability and Solubility of Chameleonic Drugs beyond the Rule of 5. *J. Med. Chem.* **2018**, *61* (9), 4189–4202. https://doi.org/10.1021/acs.jmedchem.8b00347.

(82) Daina, A.; Zoete, V. A BOILED-Egg To Predict Gastrointestinal Absorption and Brain Penetration of Small Molecules. *ChemMedChem* **2016**, 1117–1121. https://doi.org/10.1002/cmdc.201600182.

(83) Bergström, C. A. S.; Charman, W. N.; Porter, C. J. H. Computational Prediction of Formulation Strategies for Beyond-Rule-of-5 Compounds. *Adv. Drug Deliv. Rev.* **2016**, *101*, 6–21. https://doi.org/10.1016/j.addr.2016.02.005.

(84) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45* (12), 2615–2623. https://doi.org/10.1021/jm020017n.

(85) Matsson, P.; Kihlberg, J. How Big Is Too Big for Cell Permeability? *J. Med. Chem.* **2017**, *60* (5), 1662–1664. https://doi.org/10.1021/acs.jmedchem.7b00237.

(86) Guimarães, C. R. W.; Mathiowetz, A. M.; Shalaeva, M.; Goetz, G.; Liras, S. Use of 3D Properties to Characterize Beyond Rule-of-5 Property Space for Passive Permeation. *J. Chem. Inf. Model.* **2012**, *52* (4), 882–890. https://doi.org/10.1021/ci300010y.

(87) Whitty, A.; Zhong, M.; Viarengo, L.; Beglov, D.; Hall, D. R.; Vajda, S. Quantifying the Chameleonic Properties of Macrocycles and Other High-Molecular-Weight Drugs. *Drug Discov. Today* **2016**, *21* (5), 712–717. https://doi.org/10.1016/j.drudis.2016.02.005.

(88) El Tayar, N.; Mark, A. E.; Vallat, P.; Brunne, R. M.; Testa, B.; van Gunsteren, W. F. Solvent-Dependent Conformation and Hydrogen-Bonding Capacity of Cyclosporin A: Evidence from Partition Coefficients

and Molecular Dynamics Simulations. *J. Med. Chem.* **1993**, *36* (24), 3757–3764. https://doi.org/10.1021/jm00076a002.

(89) Eiríksdóttir, E.; Konate, K.; Langel, Ü.; Divita, G.; Deshayes, S. Secondary Structure of Cell-Penetrating Peptides Controls Membrane Interaction and Insertion. *Biochim. Biophys. Acta - Biomembr.* **2010**, *1798* (6), 1119–1128. https://doi.org/10.1016/j.bbamem.2010.03.005.

(90) Appelbaum, J. S.; LaRochelle, J. R.; Smith, B. A.; Balkin, D. M.; Holub, J. M.; Schepartz, A. Arginine Topology Controls Escape of Minimally Cationic Proteins from Early Endosomes to the Cytoplasm. *Chem. Biol.* **2012**, *19* (7), 819–830. https://doi.org/10.1016/j.chembiol.2012.05.022.

(91) Bird, G. H.; Mazzola, E.; Opoku-Nsiah, K.; Lammert, M. A.; Godes, M.; Neuberg, D. S.; Walensky, L. D. Biophysical Determinants for Cellular Uptake of Hydrocarbon-Stapled Peptide Helices. *Nat. Chem. Biol.* **2016**, *12* (10), 845–852. https://doi.org/10.1038/nchembio.2153.

(92) Yamashita, H.; Oba, M.; Misawa, T.; Tanaka, M.; Hattori, T.; Naito, M.; Kurihara, M.; Demizu, Y. A Helix-Stabilized Cell-Penetrating Peptide as an Intracellular Delivery Tool. *ChemBioChem* **2016**, *17* (2), 137–140. https://doi.org/10.1002/cbic.201500468.

(93) White, T. R.; Renzelman, C. M.; Rand, A. C.; Rezai, T.; McEwen, C. M.; Gelev, V. M.; Turner, R. A.; Linington, R. G.; Leung, S. S. F.; Kalgutkar, A. S.; Bauman, J. N.; Zhang, Y.; Liras, S.; Price, D. A.; Mathiowetz, A. M.; Jacobson, M. P.; Lokey, R. S. On-Resin N-Methylation of Cyclic Peptides for Discovery of Orally Bioavailable Scaffolds. *Nat. Chem. Biol.* **2011**, *7* (11), 810–817. https://doi.org/10.1038/nchembio.664.

(94) Seixas Feio, J. A.; de Oliveira, E. C. L.; de Sales, C. de S.; da Costa, K. S.; e Lima, A. H. L. Investigating Molecular Descriptors in Cell-Penetrating Peptides Prediction with Deep Learning: Employing N, O, and Hydrophobicity According to the Eisenberg Scale. *PLoS One* **2024**, *19* (6), e0305253. https://doi.org/10.1371/journal.pone.0305253.

(95) Santos, G. B.; Ganesan, A.; Emery, F. S. Oral Administration of Peptide-Based Drugs: Beyond Lipinski's Rule. *ChemMedChem* **2016**, *11*

(20), 2245–2251. https://doi.org/10.1002/cmdc.201600288.

View Article Online
DOI: 10.1039/D5CP04611D

(96) Grambow, C. A.; Weir, H.; Cunningham, C. N.; Biancalani, T.; Chuang, K. V. CREMP: Conformer-Rotamer Ensembles of Macrocyclic Peptides for Machine Learning. *Sci. Data* **2024**, *11* (1), 859. https://doi.org/10.1038/s41597-024-03698-y.

(97) Wang, J.; Liu, Z.; Zhao, S.; Xu, T.; Wang, H.; Li, S. Z.; Li, W. Deep Learning Empowers the Discovery of Self-Assembling Peptides with Over 10 Trillion Sequences. *Adv. Sci.* **2023**, *10* (31). https://doi.org/10.1002/advs.202301544.

(98) Sarkar, S.; Gu, W.; Schmidt, E. W. Expanding the Chemical Space of Synthetic Cyclic Peptides Using a Promiscuous Macrocyclase from Prenylagaramide Biosynthesis. *ACS Catal.* **2020**, *10* (13), 7146–7153. https://doi.org/10.1021/acscatal.0c00623.

(99) Lovering, F. Escape from Flatland 2: Complexity and Promiscuity. *Medchemcomm* **2013**, *4* (3), 515. https://doi.org/10.1039/c2md20347b.

(100) Wager, T. T.; Hou, X.; Verhoest, P. R.; Villalobos, A. Moving beyond Rules: The Development of a Central Nervous System Multiparameter Optimization (CNS MPO) Approach to Enable Alignment of Druglike Properties. *ACS Chem. Neurosci.* **2010**, *1* (6), 435–449. https://doi.org/10.1021/cn100008c.

(101) Lovering, F.; Bikker, J.; Humblet, C. Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success. *J. Med. Chem.* **2009**, *52* (21), 6752–6756. https://doi.org/10.1021/jm901241e.

(102) Cheng, T.; Zhao, Y.; Li, X.; Lin, F.; Xu, Y.; Zhang, X.; Li, Y.; Wang, R.; Lai, L. Computation of Octanol-Water Partition Coefficients by Guiding an Additive Model with Knowledge. *J. Chem. Inf. Model.* **2007**, *47* (6), 2140–2148. https://doi.org/10.1021/ci700257y.

(103) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *J. Phys. Chem. A* **1998**, *102* (21), 3762–3772.

View Article Online
DOI: 10.1039/D5CP04611D

https://doi.org/10.1021/jp980230o.

(104) Thompson, S. J.; Hattotuwagama, C. K.; Holliday, J. D.; Flower, D. R. On the Hydrophobicity of Peptides: Comparing Empirical Predictions of Peptide Log P Values. *Bioinformation* **2006**, *1* (7), 237–241. https://doi.org/10.6026/97320630001237.

(105) Kier, L. B. An Index of Molecular Flexibility from Kappa Shape Attributes. *Quant. Struct. Relationships* **1989**, *8* (3), 221–224. https://doi.org/10.1002/qsar.19890080307.

(106) Randić, M.; Pompe, M. The Variable Molecular Descriptors Based on Distance Related Matrices. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (3), 575–581. https://doi.org/10.1021/ci0001029.

(107) Burden, F. R. A Chemically Intuitive Molecular Index Based on the Eigenvalues of a Modified Adjacency Matrix. *Quant. Struct. Relationships* **1997**, *16* (4), 309–314. https://doi.org/10.1002/qsar.19970160406.

(108) Gutman, I.; Furtula, B.; Katanić, V. Randić Index and Information. *AKCE Int. J. Graphs Comb.* **2018**, *15* (3), 307–312. https://doi.org/10.1016/j.akcej.2017.09.006.

(109) Sabirov, D.; Zimina, A.; Shepelevich, I. Complexity of Molecular Ensembles with Basak's Indices: Applying Structural Information Content; 2025; pp 113–121. https://doi.org/10.1007/978-3-031-67841-7_6.

(110) Andersson, P. M.; Sjöström, M.; Wold, S.; Lundstedt, T. Comparison between Physicochemical and Calculated Molecular Descriptors. *J. Chemom.* **2000**, *14* (5–6), 629–642. https://doi.org/10.1002/1099-128X(200009/12)14:5/6<629::AID-CEM606>3.0.CO;2-M.

(111) Singh, J.; Shaik, B.; Agrawal, V. K.; Khadikar, P. V. SAR Studies on β-Cell KATP Channel Openers. *Interdiscip. Sci. Comput. Life Sci.* **2012**, *4* (3), 215–222. https://doi.org/10.1007/s12539-012-0135-8.

(112) Consonni, V.; Todeschini, R.; Pavan, M. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*

(3), 682–692. https://doi.org/10.1021/ci015504a.

View Article Online
DOI: 10.1039/D5CP04611D

(113) Devinyak, O.; Havrylyuk, D.; Lesyk, R. 3D-MoRSE Descriptors Explained. *J. Mol. Graph. Model.* **2014**, *54*, 194–203. https://doi.org/10.1016/j.jmgm.2014.10.006.

(114) Todeschini, R.; Gramatica, P. The Whim Theory: New 3D Molecular Descriptors for Qsar in Environmental Modelling. *SAR QSAR Environ. Res.* **1997**, *7* (1–4), 89–115. https://doi.org/10.1080/10629369708039126.

(115) Gramatica, P. WHIM Descriptors of Shape. *QSAR Comb. Sci.* **2006**, *25* (4), 327–332. https://doi.org/10.1002/qsar.200510159.

(116) Méndez-Lucio, O.; Medina-Franco, J. L. The Many Roles of Molecular Complexity in Drug Discovery. *Drug Discov. Today* **2017**, *22* (1), 120–126. https://doi.org/10.1016/j.drudis.2016.08.009.

(117) Paradís-Bas, M.; Tulla-Puche, J.; Albericio, F. The Road to the Synthesis of “Difficult Peptides.” *Chem. Soc. Rev.* **2016**, *45* (3), 631–654. https://doi.org/10.1039/C5CS00680E.

(118) Clemons, P. A.; Bodycombe, N. E.; Carrinski, H. A.; Wilson, J. A.; Shamji, A. F.; Wagner, B. K.; Koehler, A. N.; Schreiber, S. L. Small Molecules of Different Origins Have Distinct Distributions of Structural Complexity That Correlate with Protein-Binding Profiles. *Proc. Natl. Acad. Sci.* **2010**, *107* (44), 18787–18792. https://doi.org/10.1073/pnas.1012741107.

(119) Eisenberg, D.; Wilcox, W.; McLachlan, A. D. Hydrophobicity and Amphiphilicity in Protein Structure. *J. Cell. Biochem.* **1986**, *31* (1), 11–17. https://doi.org/10.1002/jcb.240310103.

(120) Meylan, W. M.; Howard, P. H. Atom/Fragment Contribution Method for Estimating Octanol–Water Partition Coefficients. *J. Pharm. Sci.* **1995**, *84* (1), 83–92. https://doi.org/10.1002/jps.2600840120.

(121) Oeller, M.; Kang, R. J. D.; Bolt, H. L.; Gomes dos Santos, A. L.; Weinmann, A. L.; Nikitidis, A.; Zlatoidsky, P.; Su, W.; Czechtizky, W.; De Maria, L.; Sormanni, P.; Vendruscolo, M. Sequence-Based Prediction of the Intrinsic Solubility of Peptides Containing Non-Natural Amino Acids.

Nat. Commun. **2023**, *14* (1), 7475. https://doi.org/10.1038/s41467-023-42940-w.
View Article Online
DOI: 10.1039/D5CP04611D

(122) Xiong, H.; Buckwalter, B. L.; Shieh, H. M.; Hecht, M. H. Periodicity of Polar and Nonpolar Amino Acids Is the Major Determinant of Secondary Structure in Self-Assembling Oligomeric Peptides. *Proc. Natl. Acad. Sci.* **1995**, *92* (14), 6349–6353. https://doi.org/10.1073/pnas.92.14.6349.

(123) Kister, A. E.; Gelfand, I. Finding of Residues Crucial for Supersecondary Structure Formation. *Proc. Natl. Acad. Sci.* **2009**, *106* (45), 18996–19000. https://doi.org/10.1073/pnas.0909714106.

(124) Rajput, A.; Gupta, A. K.; Kumar, M. Prediction and Analysis of Quorum Sensing Peptides Based on Sequence Features. *PLoS One* **2015**, *10* (3), e0120066. https://doi.org/10.1371/journal.pone.0120066.

(125) Chen, W.; Ding, H.; Feng, P.; Lin, H.; Chou, K. C. IACP: A Sequence-Based Tool for Identifying Anticancer Peptides. *Oncotarget* **2016**, *7* (13), 16895–16909. https://doi.org/10.18632/oncotarget.7815.

(126) Bhadra, P.; Yan, J.; Li, J.; Fong, S.; Siu, S. W. I. AmPEP: Sequence-Based Prediction of Antimicrobial Peptides Using Distribution Patterns of Amino Acid Properties and Random Forest. *Sci. Rep.* **2018**, *8* (1), 1697. https://doi.org/10.1038/s41598-018-19752-w.

(127) Huang, K.-Y.; Tseng, Y.-J.; Kao, H.-J.; Chen, C.-H.; Yang, H.-H.; Weng, S.-L. Identification of Subtypes of Anticancer Peptides Based on Sequential Features and Physicochemical Properties. *Sci. Rep.* **2021**, *11* (1), 13594. https://doi.org/10.1038/s41598-021-93124-9.

(128) Pearson, W. R.; Lipman, D. J. Improved Tools for Biological Sequence Comparison. *Proc. Natl. Acad. Sci.* **1988**, *85* (8), 2444–2448. https://doi.org/10.1073/pnas.85.8.2444.

(129) Fox, T.; Bieler, M.; Haebel, P.; Ochoa, R.; Peters, S.; Weber, A. BILN: A Human-Readable Line Notation for Complex Peptides. *J. Chem. Inf. Model.* **2022**, *62* (17), 3942–3947. https://doi.org/10.1021/acs.jcim.2c00703.

(130) Zhang, T.; Li, H.; Xi, H.; Stanton, R. V.; Rotstein, S. H. HELM: A Hierarchical Notation Language for Complex Biomolecule Structure Representation. *J. Chem. Inf. Model.* **2012**, *52* (10), 2796–2806. https://doi.org/10.1021/ci3001925.

(131) Erckes, V.; Abderrahmane, M.; Jusot, M.; Steuer, C.; Ochoa, R. Peptide Cheminformatics Tools: Making Computational Tasks Accessible in Peptide Drug Discovery. *Drug Discov. Today* **2026**, *31* (2), 104612. https://doi.org/10.1016/j.drudis.2026.104612.

(132) Su, Y.; Doherty, T.; Waring, A. J.; Ruchala, P.; Hong, M. Roles of Arginine and Lysine Residues in the Translocation of a Cell-Penetrating Peptide From13C, 31P, And19F Solid-State NMR. *Biochemistry* **2009**, *48* (21), 4587–4595. https://doi.org/10.1021/bi900080d.

(133) Gräslund, A.; Madani, F.; Lindberg, S.; Langel, Ü.; Futaki, S. Mechanisms of Cellular Uptake of Cell-Penetrating Peptides. *J. Biophys.* **2011**, *2011*, 1–10. https://doi.org/10.1155/2011/414729.

(134) Liu, T.; Liu, Y.; Kao, H. Y.; Pei, D. Membrane Permeable Cyclic Peptidyl Inhibitors against Human Peptidylprolyl Isomer Ase Pin1. *J. Med. Chem.* **2010**, *53* (6), 2494–2501. https://doi.org/10.1021/jm901778v.

(135) Cronican, J. J.; Thompson, D. B.; Beier, K. T.; McNaughton, B. R.; Cepko, C. L.; Liu, D. R. Potent Delivery of Functional Proteins into Mammalian Cells in Vitro and in Vivo Using a Supercharged Protein. *ACS Chem. Biol.* **2010**, *5* (8), 747–752. https://doi.org/10.1021/cb1001153.

(136) Manavalan, B.; Subramaniyam, S.; Shin, T. H.; Kim, M. O.; Lee, G. Machine-Learning-Based Prediction of Cell-Penetrating Peptides and Their Uptake Efficiency with Improved Accuracy. *J. Proteome Res.* **2018**, *17* (8), 2715–2726. https://doi.org/10.1021/acs.jproteome.8b00148.

(137) Bernardes-Loch, R. M.; de Oliveira Almeida, G.; Brasiliano, I. T.; Meira Jr, W.; Pires, D. E. V; Baracat-Pereira, M. C.; de Azevedo Silveira, S. PerseuCPP: A Machine Learning Strategy to Predict Cell-Penetrating Peptides and Their Uptake Efficiency. *Bioinforma. Adv.* **2024**, *5* (1). https://doi.org/10.1093/bioadv/vbaf213.

(138) Charoenkwan, P.; Chumnanpuen, P.; Schaduangrat, N.; Shoombuatong, W. Stack-AVP: A Stacked Ensemble Predictor Based on Multi-View Information for Fast and Accurate Discovery of Antiviral Peptides. *J. Mol. Biol.* **2025**, *437* (6), 168853. https://doi.org/10.1016/j.jmb.2024.168853.

(139) Pugliese, R.; Gelain, F. Peptidic Biomaterials: From Self-Assembling to Regenerative Medicine. *Trends Biotechnol.* **2017**, *35* (2), 145–158. https://doi.org/10.1016/j.tibtech.2016.09.004.

(140) Hosseinkhani, H.; Hong, P.-D.; Yu, D.-S. Self-Assembled Proteins and Peptides for Regenerative Medicine. *Chem. Rev.* **2013**, *113* (7), 4837–4861. https://doi.org/10.1021/cr300131h.

(141) He, B.; Yuan, X.; Zhou, A.; Zhang, H.; Jiang, D. Designer Functionalised Self-Assembling Peptide Nanofibre Scaffolds for Cartilage Tissue Engineering. *Expert Rev. Mol. Med.* **2014**, *16*, e12. https://doi.org/10.1017/erm.2014.13.

(142) Sun, W.; Gregory, D. A.; Zhao, X. Designed Peptide Amphiphiles as Scaffolds for Tissue Engineering. *Adv. Colloid Interface Sci.* **2023**, *314*, 102866. https://doi.org/10.1016/j.cis.2023.102866.

(143) Ni, N.; Hu, Y.; Ren, H.; Luo, C.; Li, P.; Wan, J.-B.; Su, H. Self-Assembling Peptide Nanofiber Scaffolds Enhance Dopaminergic Differentiation of Mouse Pluripotent Stem Cells in 3-Dimensional Culture. *PLoS One* **2013**, *8* (12), e84504. https://doi.org/10.1371/journal.pone.0084504.

(144) Gazit, E. A Possible Role for Π-stacking in the Self-assembly of Amyloid Fibrils. *FASEB J.* **2002**, *16* (1), 77–83. https://doi.org/10.1096/fj.01-0442hyp.

(145) King, N. P.; Bale, J. B.; Sheffler, W.; McNamara, D. E.; Gonen, S.; Gonen, T.; Yeates, T. O.; Baker, D. Accurate Design of Co-Assembling Multi-Component Protein Nanomaterials. *Nature* **2014**, *510* (7503), 103–108. https://doi.org/10.1038/nature13404.

(146) Fontana, F.; Gelain, F. Probing Mechanical Properties and Failure Mechanisms of Fibrils of Self-Assembling Peptides. *Nanoscale Adv.* **2020**, *2* (1), 190–198. https://doi.org/10.1039/C9NA00621D.

(147) Chakraborty, P.; Tang, Y.; Yamamoto, T.; Yao, Y.; Guterman, T.; Zilberzwige-Tal, S.; Adadi, N.; Ji, W.; Dvir, T.; Ramamoorthy, A.; Wei, G.; Gazit, E. Unusual Two-Step Assembly of a Minimalistic Dipeptide-Based Functional Hypergelator. *Adv. Mater.* **2020**, *32* (9). https://doi.org/10.1002/adma.201906043.

(148) Lee, N. R.; Bowerman, C. J.; Nilsson, B. L. Effects of Varied Sequence Pattern on the Self-Assembly of Amphipathic Peptides. *Biomacromolecules* **2013**, *14* (9), 3267–3277. https://doi.org/10.1021/bm400876s.

(149) Zhang, S. Lipid-like Self-Assembling Peptides. *Acc. Chem. Res.* **2012**, *45* (12), 2142–2150. https://doi.org/10.1021/ar300034v.

(150) Yuan, S. C.; Lewis, J. A.; Sai, H.; Weigand, S. J.; Palmer, L. C.; Stupp, S. I. Peptide Sequence Determines Structural Sensitivity to Supramolecular Polymerization Pathways and Bioactivity. *J. Am. Chem. Soc.* **2022**, *144* (36), 16512–16523. https://doi.org/10.1021/jacs.2c05759.

(151) Charoenkwan, P.; Chumnanpuen, P.; Schaduangrat, N.; Oh, C.; Manavalan, B.; Shoombuatong, W. PSRQSP: An Effective Approach for the Interpretable Prediction of Quorum Sensing Peptide Using Propensity Score Representation Learning. *Comput. Biol. Med.* **2023**, *158*, 106784. https://doi.org/10.1016/j.compbiomed.2023.106784.

(152) Pandey, P.; Patel, V.; George, N. V.; Mallajosyula, S. S. KELM-CPPpred: Kernel Extreme Learning Machine Based Prediction Model for Cell-Penetrating Peptides. *J. Proteome Res.* **2018**, *17* (9), 3214–3222. https://doi.org/10.1021/acs.jproteome.8b00322.

(153) Yao, L.; Xie, P.; Guan, J.; Chung, C.-R.; Zhang, W.; Deng, J.; Huang, Y.; Chiang, Y.-C.; Lee, T.-Y. ACP-CapsPred: An Explainable Computational Framework for Identification and Functional Prediction of Anticancer Peptides Based on Capsule Network. *Brief. Bioinform.* **2024**, *25* (5). https://doi.org/10.1093/bib/bbae460.

(154) Zuo, Y.; Lv, Y.; Wei, Z.; Yang, L.; Li, G.; Fan, G. IDPF-PseRAAAC: A Web-Server for Identifying the Defensin Peptide Family and Subfamily

Using Pseudo Reduced Amino Acid Alphabet Composition. *PLoS One* **2015**, *10* (12), e0145541. https://doi.org/10.1371/journal.pone.0145541.

(155) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; De Hoon, M. J. L. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25* (11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163.

(156) Dong, J.; Yao, Z. J.; Zhang, L.; Luo, F.; Lin, Q.; Lu, A. P.; Chen, A. F.; Cao, D. S. PyBioMed: A Python Library for Various Molecular Representations of Chemicals, Proteins and DNAs and Their Interactions. *J. Cheminform.* **2018**, *10* (1), 16. https://doi.org/10.1186/s13321-018-0270-2.

(157) Ochoa, R.; Cossio, P. PepFun: Open Source Protocols for Peptide-Related Computational Analysis. *Molecules* **2021**, *26* (6), 1664. https://doi.org/10.3390/molecules26061664.

(158) Osorio, D.; Rondón-Villarreal, P.; Torres, R. Peptides: A Package for Data Mining of Antimicrobial Peptides. *R J.* **2015**, *7* (1), 4. https://doi.org/10.32614/RJ-2015-001.

(159) Ochoa, R.; Deibler, K. PepFuNN: Novo Nordisk Open-Source Toolkit to Enable Peptide in Silico Analysis. *J. Pept. Sci.* **2025**, *31* (2). https://doi.org/10.1002/psc.3666.

(160) Sahu, S. S.; Panda, G. A Novel Feature Representation Method Based on Chou's Pseudo Amino Acid Composition for Protein Structural Class Prediction. *Comput. Biol. Chem.* **2010**, *34* (5–6), 320–327. https://doi.org/10.1016/j.compbiolchem.2010.09.002.

(161) Park, K.-J.; Kanehisa, M. Prediction of Protein Subcellular Locations by Support Vector Machines Using Compositions of Amino Acids and Amino Acid Pairs. *Bioinformatics* **2003**, *19* (13), 1656–1663. https://doi.org/10.1093/bioinformatics/btg222.

(162) Agrawal, P.; Bhagat, D.; Mahalwal, M.; Sharma, N.; Raghava, G. P. S. AntiCP 2.0: An Updated Model for Predicting Anticancer Peptides. *Brief.*

Bioinform. **2021**, *22* (3). https://doi.org/10.1093/bib/bbaa153.

View Article Online
DOI: 10.1039/D5CP04611D

(163) Li, L. Dimension Reduction for High-Dimensional Data; 2010; pp 417–434. https://doi.org/10.1007/978-1-60761-580-4_14.

(164) Rao, B.; Zhou, C.; Zhang, G.; Su, R.; Wei, L. ACPred-Fuse: Fusing Multi-View Information Improves the Prediction of Anticancer Peptides. *Brief. Bioinform.* **2020**, *21* (5), 1846–1855. https://doi.org/10.1093/bib/bbz088.

(165) Zuo, Y.; Li, Y.; Chen, Y.; Li, G.; Yan, Z.; Yang, L. PseKRAAC: A Flexible Web Server for Generating Pseudo K-Tuple Reduced Amino Acids Composition. *Bioinformatics* **2017**, *33* (1), 122–124. https://doi.org/10.1093/bioinformatics/btw564.

(166) Chen, K.; Kurgan, L.; Rahbari, M. Prediction of Protein Crystallization Using Collocation of Amino Acid Pairs. **2007**, *355*, 764–769. https://doi.org/10.1016/j.bbrc.2007.02.040.

(167) Dai, R.; Zhang, W.; Tang, W.; Wynendaele, E.; Zhu, Q.; Bin, Y.; De Spiegeleer, B.; Xia, J. BBPpred: Sequence-Based Prediction of Blood-Brain Barrier Peptides with Feature Representation Learning and Logistic Regression. *J. Chem. Inf. Model.* **2021**, *61* (1), 525–534. https://doi.org/10.1021/acs.jcim.0c01115.

(168) Wei, L.; Zhou, C.; Su, R.; Zou, Q. PEPred-Suite: Improved and Robust Prediction of Therapeutic Peptides Using Adaptive Feature Representation Learning. *Bioinformatics* **2019**, *35* (21), 4272–4280. https://doi.org/10.1093/bioinformatics/btz246.

(169) Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. AAindex: Amino Acid Index Database, Progress Report 2008. *Nucleic Acids Res.* **2007**, *36* (Database), D202–D205. https://doi.org/10.1093/nar/gkm998.

(170) Dosztányi, Z.; Csizmók, V.; Tompa, P.; Simon, I. The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins. *J. Mol. Biol.* **2005**, *347* (4), 827–839. https://doi.org/10.1016/j.jmb.2005.01.071.

(171) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *J. Med. Chem.* **1998**, *41* (14), 2481–2491. https://doi.org/10.1021/jm9700575.

(172) Henikoff, S.; Henikoff, J. G. Amino Acid Substitution Matrices from Protein Blocks. *Proc. Natl. Acad. Sci.* **1992**, *89* (22), 10915–10919. https://doi.org/10.1073/pnas.89.22.10915.

(173) Wynendaele, E.; Debunne, N.; Janssens, Y.; De Spiegeleer, A.; Verbeke, F.; Tack, L.; Van Welden, S.; Goossens, E.; Knappe, D.; Hoffmann, R.; Van De Wiele, C.; Laukens, D.; Van Eenoo, P.; Vereecke, L.; Van Immerseel, F.; De Wever, O.; De Spiegeleer, B. The Quorum Sensing Peptide EntF* Promotes Colorectal Cancer Metastasis in Mice: A New Factor in the Host-Microbiome Interaction. *BMC Biol.* **2022**, *20* (1), 151. https://doi.org/10.1186/s12915-022-01317-z.

(174) Xu, J.; Li, F.; Li, C.; Guo, X.; Landersdorfer, C.; Shen, H.-H.; Peleg, A. Y.; Li, J.; Imoto, S.; Yao, J.; Akutsu, T.; Song, J. IAMPCN: A Deep-Learning Approach for Identifying Antimicrobial Peptides and Their Functional Activities. *Brief. Bioinform.* **2023**, *24* (4). https://doi.org/10.1093/bib/bbad240.

(175) Guan, J.; Yao, L.; Chung, C.-R.; Chiang, Y.-C.; Lee, T.-Y. StackTHPred: Identifying Tumor-Homing Peptides through GBDT-Based Feature Selection with Stacking Ensemble Architecture. *Int. J. Mol. Sci.* **2023**, *24* (12), 10348. https://doi.org/10.3390/ijms241210348.

(176) Chen, C.; Chen, L.-X.; Zou, X.-Y.; Cai, P.-X. Predicting Protein Structural Class Based on Multi-Features Fusion. *J. Theor. Biol.* **2008**, *253* (2), 388–392. https://doi.org/10.1016/j.jtbi.2008.03.009.

(177) Dai, Q.; Li, Y.; Liu, X.; Yao, Y.; Cao, Y.; He, P. Comparison Study on Statistical Features of Predicted Secondary Structures for Protein Structural Class Prediction: From Content to Position. *BMC Bioinformatics* **2013**, *14* (1), 152. https://doi.org/10.1186/1471-2105-14-152.

(178) Zhang, S.; Ding, S.; Wang, T. High-Accuracy Prediction of Protein

Structural Class for Low-Similarity Sequences Based on Predicted Secondary Structure. *Biochimie* **2011**, *93* (4), 710–714. https://doi.org/10.1016/j.biochi.2011.01.001.

(179) Sokal, R. R.; Oden, N. L.; Thomson, B. A. Local Spatial Autocorrelation in Biological Variables. *Biol. J. Linn. Soc.* **1998**, *65* (1), 41–62. https://doi.org/10.1111/j.1095-8312.1998.tb00350.x.

(180) Xiao, N.; Cao, D.-S.; Zhu, M.-F.; Xu, Q.-S. Protr/ProtrWeb: R Package and Web Server for Generating Various Numerical Representation Schemes of Protein Sequences. *Bioinformatics* **2015**, *31* (11), 1857–1859. https://doi.org/10.1093/bioinformatics/btv042.

(181) Bizzotto, E.; Zampieri, G.; Treu, L.; Filannino, P.; Di Cagno, R.; Campanaro, S. Classification of Bioactive Peptides: A Systematic Benchmark of Models and Encodings. *Comput. Struct. Biotechnol. J.* **2024**, *23*, 2442–2452. https://doi.org/10.1016/j.csbj.2024.05.040.

(182) Chou, K. C. Prediction of Protein Cellular Attributes Using Pseudo-Amino Acid Composition. *Proteins Struct. Funct. Genet.* **2001**, *43* (3), 246–255. https://doi.org/10.1002/prot.1035.

(183) Chou, K.-C. Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Order Effect. *Biochem. Biophys. Res. Commun.* **2000**, *278* (2), 477–483. https://doi.org/10.1006/bbrc.2000.3815.

(184) Kyte, J.; Doolittle, R. F. A Simple Method for Displaying the Hydropathic Character of a Protein. *J. Mol. Biol.* **1982**, *157* (1), 105–132. https://doi.org/10.1016/0022-2836(82)90515-0.

(185) Vihinen, M.; Torkkila, E.; Riikonen, P. Accuracy of Protein Flexibility Predictions. *Proteins Struct. Funct. Bioinforma.* **1994**, *19* (2), 141–149. https://doi.org/10.1002/prot.340190207.

(186) Henikoff, J. G.; Henikoff, S. Using Substitution Probabilities to Improve Position-Specific Scoring Matrices. *Bioinformatics* **1996**, *12* (2), 135–143. https://doi.org/10.1093/bioinformatics/12.2.135.

(187) Grantham, R. Amino Acid Difference Formula to Help Explain Protein

View Article Online
DOI: 10.1039/D5CP04611D

Evolution. *Science (80-.).* **1974**, *185* (4154), 862–864. https://doi.org/10.1126/science.185.4154.862.

(188) Medina-Franco, J. L.; Sánchez-Cruz, N.; López-López, E.; Díaz-Eufracio, B. I. Progress on Open Chemoinformatic Tools for Expanding and Exploring the Chemical Space. *J. Comput. Aided. Mol. Des.* **2022**, *36* (5), 341–354. https://doi.org/10.1007/s10822-021-00399-1.

(189) Wynendaele, E.; Gevaert, B.; Stalmans, S.; Verbeke, F.; De Spiegeleer, B. Exploring the Chemical Space of Quorum Sensing Peptides. *Pept. Sci.* **2015**, *104* (5), 544–551. https://doi.org/10.1002/bip.22649.

(190) Orsi, M.; Personne, H.; Bonvin, E.; Paschoud, T.; Olcay, B.; Hu, X.; Javor, S.; Reymond, J.-L. Chemical Space for Peptide-Based Antimicrobials. *Chimia (Aarau).* **2024**, *78* (10), 648–653. https://doi.org/10.2533/chimia.2024.648.

(191) Wynendaele, E.; Bronselaer, A.; Nielandt, J.; D'Hondt, M.; Stalmans, S.; Bracke, N.; Verbeke, F.; Van De Wiele, C.; De Tré, G.; De Spiegeleer, B. Quorumpeps Database: Chemical Space, Microbial Origin and Functionality of Quorum Sensing Peptides. *Nucleic Acids Res.* **2013**, *41* (D1), D655–D659. https://doi.org/10.1093/nar/gks1137.

(192) Capecchi, A.; Zhang, A.; Reymond, J.-L. Populating Chemical Space with Peptides Using a Genetic Algorithm. *J. Chem. Inf. Model.* **2020**, *60* (1), 121–132. https://doi.org/10.1021/acs.jcim.9b01014.

(193) Sánchez-Cruz, N.; Medina-Franco, J. L. Statistical-Based Database Fingerprint: Chemical Space Dependent Representation of Compound Databases. *J. Cheminform.* **2018**, *10* (1), 55. https://doi.org/10.1186/s13321-018-0311-x.

(194) Willett, P. Similarity-Based Virtual Screening Using 2D Fingerprints. *Drug Discovery Today*. 2006, pp 1046–1053. https://doi.org/10.1016/j.drudis.2006.10.005.

(195) Boldini, D.; Ballabio, D.; Consonni, V.; Todeschini, R.; Grisoni, F.; Sieber, S. A. Effectiveness of Molecular Fingerprints for Exploring the Chemical Space of Natural Products. *J. Cheminform.* **2024**, *16* (1), 35.

https://doi.org/10.1186/s13321-024-00830-3.

View Article Online
DOI: 10.1039/D5CP04611D

(196) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273–1280. https://doi.org/10.1021/ci010132r.

(197) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25* (2), 64–73. https://doi.org/10.1021/ci00046a002.

(198) Axen, S. D.; Huang, X.-P.; Cáceres, E. L.; Gendelev, L.; Roth, B. L.; Keiser, M. J. A Simple Representation of Three-Dimensional Molecular Structure. *J. Med. Chem.* **2017**, *60* (17), 7393–7409. https://doi.org/10.1021/acs.jmedchem.7b00696.

(199) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.

(200) Riniker, S.; Landrum, G. A. Open-Source Platform to Benchmark Fingerprints for Ligand-Based Virtual Screening. *J. Cheminform.* **2013**, *5* (1), 26. https://doi.org/10.1186/1758-2946-5-26.

(201) Di Bonaventura, I.; Jin, X.; Visini, R.; Probst, D.; Javor, S.; Gan, B.-H.; Michaud, G.; Natalello, A.; Doglia, S. M.; Köhler, T.; van Delden, C.; Stocker, A.; Darbre, T.; Reymond, J.-L. Chemical Space Guided Discovery of Antimicrobial Bridged Bicyclic Peptides against Pseudomonas Aeruginosa and Its Biofilms. *Chem. Sci.* **2017**, *8* (10), 6784–6798. https://doi.org/10.1039/C7SC01314K.

(202) Orsi, M.; Reymond, J.-L. One Chiral Fingerprint to Find Them All. *J. Cheminform.* **2024**, *16* (1), 53. https://doi.org/10.1186/s13321-024-00849-6.

(203) Adamczyk, J.; Ludynia, P. Scikit-Fingerprints: Easy and Efficient Computation of Molecular Fingerprints in Python. *SoftwareX* **2024**, *28*, 101944. https://doi.org/10.1016/j.softx.2024.101944.

(204) Banck, M.; Vandermeersch, T.; O'Boyle, N. M.; Hutchison, G. R.; Morley,

C.; James, C. A. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3* (1), 33. https://doi.org/10.1186/1758-2946-3-33.

(205) Manelfi, C.; Tazzari, V.; Lunghini, F.; Cerchia, C.; Fava, A.; Pedretti, A.; Stouten, P. F. W.; Vistoli, G.; Beccari, A. R. "DompeKeys": A Set of Novel Substructure-Based Descriptors for Efficient Chemical Space Mapping, Development and Structural Interpretation of Machine Learning Models, and Indexing of Large Databases. *J. Cheminform.* **2024**, *16* (1), 21. https://doi.org/10.1186/s13321-024-00813-4.

(206) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (1), 170–178. https://doi.org/10.1021/ci034207y.

(207) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; Fergus, R. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proc. Natl. Acad. Sci.* **2021**, *118* (15). https://doi.org/10.1073/pnas.2016239118.

(208) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* **2018**, *58* (1), 27–35. https://doi.org/10.1021/acs.jcim.7b00616.

(209) Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; Bhowmik, D.; Rost, B. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44* (10), 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381.

(210) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59* (8), 3370–3388. https://doi.org/10.1021/acs.jcim.9b00237.

(211) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J.

M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268–276. https://doi.org/10.1021/acscentsci.7b00572.

(212) Özçelik, R.; Brinkmann, H.; Criscuolo, E.; Grisoni, F. Generative Deep Learning for de Novo Drug Design—A Chemical Space Odyssey. *J. Chem. Inf. Model.* **2025**, *65* (14), 7352–7372. https://doi.org/10.1021/acs.jcim.5c00641.

(213) Hinton, G. E.; Roweis, S. T. Stochastic Neighbor Embedding. In *Neural Information Processing Systems*; 2002.

(214) Marquet, C.; Heinzinger, M.; Olenyi, T.; Dallago, C.; Erckert, K.; Bernhofer, M.; Nechaev, D.; Rost, B. Embeddings from Protein Language Models Predict Conservation and Variant Effects. *Hum. Genet.* **2022**, *141* (10), 1629–1647. https://doi.org/10.1007/s00439-021-02411-y.

(215) Renaud, S.; Mansbach, R. A. Latent Spaces for Antimicrobial Peptide Design. *Digit. Discov.* **2023**, *2* (2), 441–458. https://doi.org/10.1039/D2DD00091A.

(216) Gelman, S.; Johnson, B.; Freschlin, C. R.; Sharma, A.; D'Costa, S.; Peters, J.; Gitter, A.; Romero, P. A. Biophysics-Based Protein Language Models for Protein Engineering. *Nat. Methods* **2025**, *22* (9), 1868–1879. https://doi.org/10.1038/s41592-025-02776-2.

(217) Jensen, J. H.; Hoeg-Jensen, T.; Padkjær, S. B. Building a BioChemformatics Database. *J. Chem. Inf. Model.* **2008**, *48* (12), 2404–2413. https://doi.org/10.1021/ci800128b.

(218) Mastrolorito, F.; Gambacorta, N.; Ciriaco, F.; Cutropia, F.; Togo, M. V.; Belgiovine, V.; Tondo, A. R.; Trisciuzzi, D.; Monaco, A.; Bellotti, R.; Altomare, C. D.; Nicolotti, O.; Amoroso, N. Chemical Space Networks Enhance Toxicity Recognition via Graph Embedding. *J. Chem. Inf. Model.* **2025**, *65* (4), 1850–1861. https://doi.org/10.1021/acs.jcim.4c02140.

(219) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated

Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5* (9), 1572–1583. https://doi.org/10.1021/acscentsci.9b00576.

View Article Online
DOI: 10.1039/D5CP04611D

(220) Harding-Larsen, D.; Funk, J.; Madsen, N. G.; Gharabli, H.; Acevedo-Rocha, C. G.; Mazurenko, S.; Welner, D. H. Protein Representations: Encoding Biological Information for Machine Learning in Biocatalysis. *Biotechnol. Adv.* **2024**, *77*, 108459. https://doi.org/10.1016/j.biotechadv.2024.108459.

(221) Siedhoff, N. E.; Illig, A.-M.; Schwaneberg, U.; Davari, M. D. PyPEF—An Integrated Framework for Data-Driven Protein Engineering. *J. Chem. Inf. Model.* **2021**, *61* (7), 3463–3476. https://doi.org/10.1021/acs.jcim.1c00099.

(222) Zheng, X.; Tomiura, Y. A BERT-Based Pretraining Model for Extracting Molecular Structural Information from a SMILES Sequence. *J. Cheminform.* **2024**, *16* (1), 71. https://doi.org/10.1186/s13321-024-00848-7.

(223) Ochiai, T.; Inukai, T.; Akiyama, M.; Furui, K.; Ohue, M.; Matsumori, N.; Inuki, S.; Uesugi, M.; Sunazuka, T.; Kikuchi, K.; Kakeya, H.; Sakakibara, Y. Variational Autoencoder-Based Chemical Latent Space for Large Molecular Structures with 3D Complexity. *Commun. Chem.* **2023**, *6* (1), 249. https://doi.org/10.1038/s42004-023-01054-6.

(224) van Eck, J.; Gogishvili, D.; Silva, W.; Abeln, S. PLM-EXplain: Divide and Conquer the Protein Embedding Space. *Bioinformatics* **2026**, *42* (1). https://doi.org/10.1093/bioinformatics/btaf631.

(225) Leclercq, M.; Droit, A. Protein Language Models: Applications and Perspectives. *J. Proteome Res.* **2026**, *25* (2), 507–524. https://doi.org/10.1021/acs.jproteome.5c00506.

(226) Brandes, N.; Ofer, D.; Peleg, Y.; Rappoport, N.; Linial, M. ProteinBERT: A Universal Deep-Learning Model of Protein Sequence and Function. *Bioinformatics* **2022**, *38* (8), 2102–2110. https://doi.org/10.1093/bioinformatics/btac020.

(227) Pantolini, L.; Studer, G.; Pereira, J.; Durairaj, J.; Tauriello, G.; Schwede,

T. Embedding-Based Alignment: Combining Protein Language Models with Dynamic Programming Alignment to Detect Structural Similarities in the Twilight-Zone. *Bioinformatics* **2024**, *40* (1). https://doi.org/10.1093/bioinformatics/btad786.

(228) Suzek, B. E.; Huang, H.; McGarvey, P.; Mazumder, R.; Wu, C. H. UniRef: Comprehensive and Non-Redundant UniProt Reference Clusters. *Bioinformatics* **2007**, *23* (10), 1282–1288. https://doi.org/10.1093/bioinformatics/btm098.

(229) Feller, A. L.; Wilke, C. O. Peptide-Aware Chemical Language Model Successfully Predicts Membrane Diffusion of Cyclic Peptides. *J. Chem. Inf. Model.* **2025**, *65* (2), 571–579. https://doi.org/10.1021/acs.jcim.4c01441.

(230) Du, Z.; Ding, X.; Xu, Y.; Li, Y. UniDL4BioPep: A Universal Deep Learning Architecture for Binary Classification in Peptide Bioactivity. *Brief. Bioinform.* **2023**, *24* (3). https://doi.org/10.1093/bib/bbad135.

(231) Dee, W. LMPred: Predicting Antimicrobial Peptides Using Pre-Trained Language Models and Deep Learning. *Bioinforma. Adv.* **2022**, *2* (1). https://doi.org/10.1093/bioadv/vbac021.

(232) Badrinarayanan, S.; Guntuboina, C.; Mollaei, P.; Barati Farimani, A. Multi-Peptide: Multimodality Leveraged Language-Graph Learning of Peptide Properties. *J. Chem. Inf. Model.* **2025**, *65* (1), 83–91. https://doi.org/10.1021/acs.jcim.4c01443.

(233) De Spiegeleer, B.; Verbeke, F.; D'Hondt, M.; Hendrix, A.; Van De Wiele, C.; Burvenich, C.; Peremans, K.; De Wever, O.; Bracke, M.; Wynendaele, E. The Quorum Sensing Peptides PhrG, CSP and EDF Promote Angiogenesis and Invasion of Breast Cancer Cells In Vitro. *PLoS One* **2015**, *10* (3), e0119471. https://doi.org/10.1371/journal.pone.0119471.

(234) De Spiegeleer, A.; Descamps, A.; Govindarajan, S.; Coudenys, J.; Van der borght, K.; Hirmz, H.; Van Den Noortgate, N.; Elewaut, D.; De Spiegeleer, B.; Wynendaele, E. Bacterial Quorum-Sensing Peptides as Immune Modulators Present in Systemic Circulation. *Biomolecules* **2023**,

13 (2), 296. https://doi.org/10.3390/biom13020296.

View Article Online
DOI: 10.1039/D5CP04611D

(235) De Spiegeleer, A.; Descamps, A.; Wynendaele, E.; Naumovski, P.; Crombez, L.; Planas, M.; Feliu, L.; Knappe, D.; Mouly, V.; Bigot, A.; Bielza, R.; Hoffmann, R.; Van Den Noortgate, N.; Elewaut, D.; De Spiegeleer, B. Streptococcal Quorum Sensing Peptide CSP-7 Contributes to Muscle Inflammation and Wasting. *Biochim. Biophys. Acta - Mol. Basis Dis.* **2024**, *1870* (4), 167094. https://doi.org/10.1016/j.bbadis.2024.167094.

(236) Cavaco, M.; Fraga, P.; Valle, J.; Silva, R. D. M.; Gano, L.; Correia, J. D. G.; Andreu, D.; Castanho, M. A. R. B.; Neves, V. Molecular Determinants for Brain Targeting by Peptides: A Meta-Analysis Approach with Experimental Validation. *Fluids Barriers CNS* **2024**, *21* (1), 45. https://doi.org/10.1186/s12987-024-00545-5.

(237) Dichiara, M.; Amata, B.; Turnaturi, R.; Marrazzo, A.; Amata, E. Tuning Properties for Blood-Brain Barrier Permeation: A Statistics-Based Analysis. *ACS Chem. Neurosci.* **2020**, *11* (1), 34–44. https://doi.org/10.1021/acschemneuro.9b00541.

(238) Nguyen, L. T.; Haney, E. F.; Vogel, H. J. The Expanding Scope of Antimicrobial Peptide Structures and Their Modes of Action. *Trends Biotechnol.* **2011**, *29* (9), 464–472. https://doi.org/10.1016/j.tibtech.2011.05.001.

View Article Online
DOI: 10.1039/D5CP04611D

Data availability

No new primary data were generated or analysed in this study. All data discussed are available in the cited literature.

Prof. Kauê Santana, on behalf of all authors

Institute of Biodiversity

Laboratory of Computational Simulation