npj | systems biology and applications

Article

Published in partnership with the Systems Biology Institute

https://doi.org/10.1038/s41540-024-00429-2

Peptide hemolytic activity analysis using visual data mining of similarity-based complex networks

Check for updates

Kevin Castillo-Mendieta[1], Guillermin Agüero-Chapin[2,3], Edgar A. Marquez[4], Yunierkis Perez-Castillo[5], Stephen J. Barigye[6], Nelson Santiago Vispo[7], Cesar R. García-Jacas[8] & Yovani Marrero-Ponce[9,10]

Peptides are promising drug development frameworks that have been hindered by intrinsic undesired properties including hemolytic activity. We aim to get a better insight into the chemical space of hemolytic peptides using a novel approach based on network science and data mining. Metadata networks (METNs) were useful to characterize and find general patterns associated with hemolytic peptides, whereas Half-Space Proximal Networks (HSPNs), represented the hemolytic peptide space. The best candidate HSPNs were used to extract various subsets of hemolytic peptides (scaffolds) considering network centrality and peptide similarity. These scaffolds have been proved to be useful in developing robust similarity-based model classifiers. Finally, using an alignment-free approach, we reported 47 putative hemolytic motifs, which can be used as toxic signatures when developing novel peptide-based drugs. We provided evidence that the number of hemolytic motifs in a sequence might be related to the likelihood of being hemolytic.

Peptides are relatively small chains of amino acids (AAs) that can be chemically synthesized or purified from living organisms[1]. Our own bodies naturally produce peptides that carry out several critical physiological functions including healing, defense against infections or as chemical messengers[2,3]. Currently, peptides are becoming highly relevant in medical applications as they have shown to exhibit not only promising therapeutic activities such as antimicrobial, antifungal, antiviral, antiparasitic and anticancer but also due to their interesting pharmacological characteristics such as high efficacy, target selectivity and good tolerability[4–6]. Peptide drugs were reported to have sales of more than $70 billion in 2019[7] and in the last decades, they have gained more attention as potential therapeutic drugs than antibodies and small-molecule-based drugs[8,9].

Diseases such as fibrosis, asthma and cancer are treated using peptide-based therapies[2,3]. For instance, the synthetic peptide Leuprolide has been successfully used to treat prostate and breast cancers by acting as an agonist of the gonadotropin-releasing hormone[10]. In addition to Leuprolide, the current 6.0 version of the DrugBank database reports 46 other peptide-based drugs (length ≤ 100 AAs) that have been approved (accessed on May 30, 2024, SM1.5)[11]; however, these numbers are quite low compared with the several thousand potential therapeutic peptides that have been identified[12]. This concerning low proportion of peptide drugs on the market is partially explained by the short half-life, lability during storage, poor oral bioavailability and undesirable toxicity that peptides usually have[2,6,9]. Mainly, peptide-associated hemolysis is perhaps one of the main drawbacks of these potential therapeutic drugs[4] since the products released after the lysis of red blood cells (RBCs) can lead to systemic inflammation and widespread tissue damage[13].

Currently, there are many datasets available containing information about hemolytic peptides. The main databases include: i) Hemolytik[14], with more than 2000 experimentally validated hemolytic peptides; ii) Database of

[1]School of Biological Sciences and Engineering, Yachay Tech University, Urcuquí, Ecuador. [2]CIIMAR—Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Terminal de Cruzeiros do Porto de Leixões, Porto, Portugal. [3]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal. [4]Grupo de Investigaciones en Química y Biología, Departamento de Química y Biología, Facultad de Ciencias Básicas, Universidad del Norte, Universidad del Norte, Barranquilla, Colombia. [5]Bio-Chemoinformatics Research Group and Escuela de Ciencias Físicas y Matemáticas. Universidad de Las Américas, Quito, Ecuador. [6]Departamento de Química Física Aplicada, Facultad de Ciencias, Universidad Autónoma de Madrid (UAM), Madrid, Spain. [7]Clinical Biotec, Madrid 28029, Spain. [8]Investigador por México, Consejo Nacional de Humanidades, Ciencias y Tecnologías (Conahcyt), 03940 Ciudad de Mexico, Mexico. [9]Facultad de Ingeniería, Universidad Panamericana, Augusto Rodin 498, 03920, Ciudad de México, CDMX, México. [10]Universidad San Francisco de Quito (USFQ), Grupo de Medicina Molecular y Traslacional (MeM&T), Colegio de Ciencias de la Salud (COCSA), Escuela de Medicina, Edificio de Especialidades Médicas; and Instituto de Simulación Computacional (ISC-USFQ), Diego de Robles y vía Interoceánica, Quito, Pichincha, Ecuador. ✉e-mail: gchapin@ciimar.up.pt; ymarrero77@yahoo.es; ymarrero@usfq.edu.ec

Antimicrobial Activity and Structure of Peptides (DBAASP v3)[15], with more than 11321 entries showing information on hemolytic and cytotoxic activities of antimicrobial peptides (AMPs); and iii) *StarPepDB*[16] which is a graph-based database that contains 45120 peptides with annotated activities retrieved from multiple sources, from which 2004 are hemolytic peptides[16]. In the last decade, some efforts have been made to utilize the information from these databases and predict the hemolytic activity of peptides using machine learning (ML) algorithms[1,2,4–6,8,9,17,18]. However, to our knowledge, no effort has been made to explore the feature space of hemolytic peptides using network science to elucidate the defining characteristics that make certain potential therapeutic peptides hemolytic.

Network science has been previously applied to successfully model many real-world systems[19]. For instance, the “small world-model” introduced by Watts and Strogatz[20] has helped to understand the way different areas of the brain communicate with each other[21]; and more recently, during the Covid-19 pandemic, network science concepts were applied to develop strategies to lower the spread of infection[22,23]. Concerning therapeutic peptides, network science has been recently used to explore the chemical space and build prediction models for tumor-homing peptides[24] and antiparasitic peptides[25], having promising results.

Hence, following the same approach, this report aims to get insight into the chemical space of hemolytic peptides from the *StarPepDB* using network science and visual (interactive) data mining. Useful information can be retrieved from this strategy, including identifying/delineating the structural diversity among hemolytic peptides, most central and atypical peptides (singletons), the relationship between hemolysis and certain therapeutic activities, and identifying motifs related to hemolytic activity which can be useful when designing therapeutic peptide drugs[26,27]. Moreover, relatively small subsets of hemolytic peptides can be extracted for further studies. These subsets (called “scaffolds”) have the advantage of representing the whole chemical space of hemolytic peptides but just using a fraction of the nodes of the complex network of hemolytic peptides[28].

Here, we describe for the first time the use of Half-Space Proximal Networks (HSPNs) to represent the chemical space of hemolytic peptides; such networks have been previously used only to explore the antiparasitic peptide space[25]. These networks possess many advantages as they generate highly connected but sparse networks that contain the minimum spanning tree as a sub-graph[26,29]. Moreover, these networks do not strictly need a pairwise similarity threshold (t) between peptides for the construction of informative networks, as is the case of Chemical Space Networks (CSNs) described in other studies[24,25]. Nevertheless, despite a cutoff value t is not mandatory for HSPNs, it might affect the representativeness of the scaffolds. Hence, we compared HSPNs without a cutoff value (namely $t = 0.00$) with networks constructed using the same parameters but generated with their optimal similarity cutoff value. Other comparative analyses were also conducted to study the HSPN construction and visualization phase involving the use of distance metrics and centrality measures, respectively, while for extracting a representative subset of hemolytic peptides from the HSPNs, different centrality measures and global and local alignments were also evaluated.

Although, previous studies based on network science have employed the Euclidean distance as the default similarity measure metric; it is suggested that the use of different (dis)similarity measures allows the codification of orthogonal information. Hence it should not be assumed that only one measure is the best suited for calculating the similarity between objects, especially in high dimensional space[30–32]. For this reason, we evaluated five different two-way (dis)similarity measures for constructing such HSPNs: 1) Angular Separation, 2) Bhattacharyya, 3) Chebyshev, 4) Euclidean and 5) Soergel. Finally, by using community information from these networks and an alignment-free method for motif discovery, we reported new putative motifs that hallmark hemolytic peptides along with their further enrichment on external datasets to validate their significance.

Results

The overall workflow consists of four stages: (i) Metadata network visual mining, (ii) HSPNs generation and analysis, (iii) scaffold extraction and exploration, (iv) motif discovery and enrichment (Fig. 1). The first step involves the generation of metadata networks (METNs) and exploration of critical features related to hemolytic peptides. The second step consists in building HSPNs that represent the chemical space of hemolytic peptides retrieved from *StarPepDB*. Then the best HSPN candidates were selected based on global network descriptors for further analysis. In the third step, representative subsets (scaffolds) from the best HSPN candidates, built up with the optimal t value, and from their respective networks with cutoff $t = 0.00$ were extracted by using sequence alignment and centrality information from each peptide in the graph. Finally, the last step consists in proposing new putative hemolytic motifs by using an alignment-free approach and by comparing them with reported hemolytic motifs using benchmark datasets (enrichment analysis) to further select the most representative ones. All the steps of this section were performed using the StarPep toolbox, aided with in-house python scripts and the SeqKit toolkit[33].

Metadata networks (METNs)

METNs are graphs that use metadata information (e.g., origin, target, activity) from the hemolytic peptides reported in the *StarPepDB* (refer to the “Materials and Methods” section for a more detailed description). Betweenness Centrality[34] was employed as a measure the relevance of the

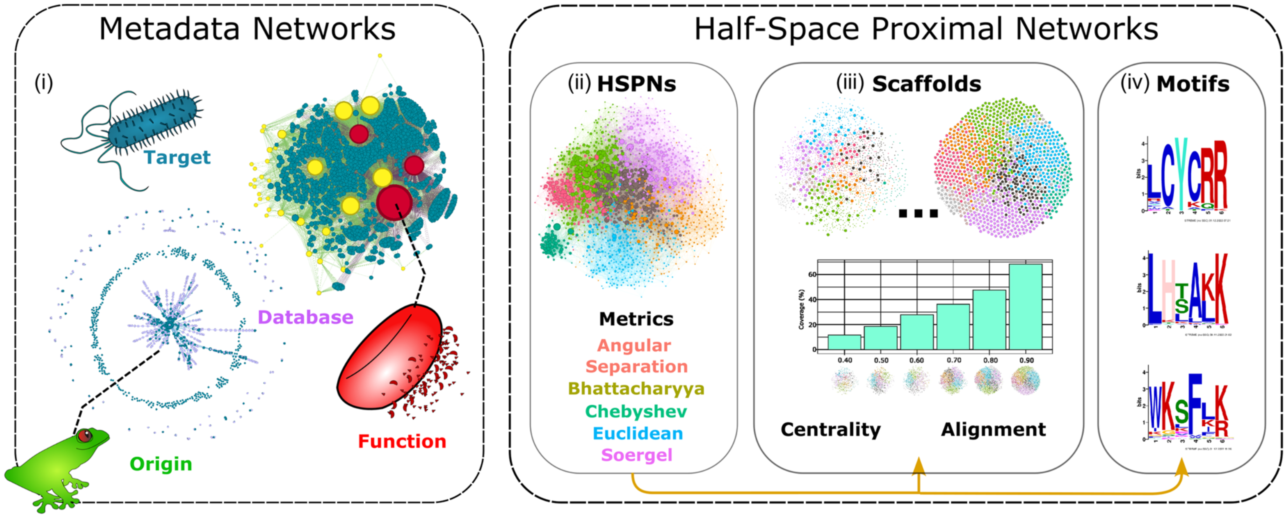


Fig. 1 | Workflow overview of the experimental procedure. Figure created with Inkscape[74].

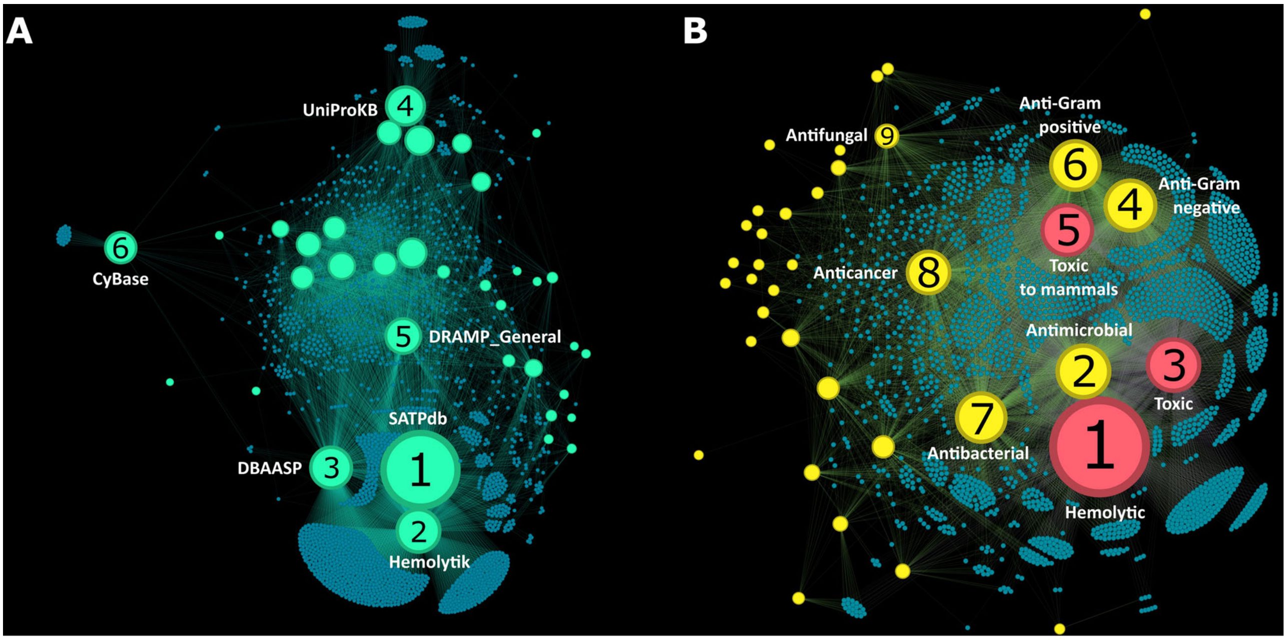


Fig. 2 | Metadata networks (METNs) of Database and Function. A *Database METN* describes the source databases from which hemolytic peptide from the *StarPepDB* has been retrieved. Aquamarine nodes represent the databases whereas blue-green nodes represent hemolytic peptides. The six most central databases were numbered according to their betweenness centrality rank: 1. SATPdb, 2. Hemolytik, 3. DBAASP, 4. UniProtKB, 5. DRAMP_General, 6. CyBase. **B** *Function METN* describes the functions associated with hemolytic peptides. Yellow nodes represent the functions reported for these peptides (red nodes are also metadata nodes but are related to hemolytic activity: "toxic", "toxic to mammals" and "hemolytic"). Blue-green nodes represent hemolytic peptides. The nine most central peptide functions were numbered according to their betweenness centrality rank: 1. hemolytic, 2. antimicrobial, 3. toxic, 4. anti-Gram negative, 5. toxic to mammals, 6. anti-Gram positive, 7. antibacterial, 8. antifungal, 9. anticancer. These networks were visualized in Gephi[70] using Force Atlas 2 layout[67] and edited with Inkscape[74].

nodes in the graphs. Four types of METNs were constructed: Database, Function, Origin and Target.

Database METN. Most hemolytic peptides of the *StarPepDB* come from the SATPdb[12], Hemolytik[14], DBAASP[15], UniProt[35], DRAMP[36] and CyBase[37] databases that are the six most central nodes in Fig. 2A. Most peptides are shared by SATPdb, Hemolytik, DBAASP and DRAMP, whereas CyBase contains more unique sets of peptides. It might be because CyBase mainly focuses on collecting information about specific types of proteins, cyclic proteins which have shown to possess important advantages such as higher stability and binding affinity compared with linear peptides[38].

In addition, SATPdb has the highest betweenness centrality and node degree value since it is connected to 1817 hemolytic peptides. On the contrary, the databases having the least number of hemolytic peptides are NeuroPep[39], Defensins[40] and Bagel 2[41] which have node degrees of 4, 2, and 1, respectively. Overall, the *Database METN* can be helpful when searching for the most important databases regarding peptide hemolytic activity as well as the most unique and most specialized databases.

Function METN. When designing therapeutic drugs, understanding other activities associated with hemolytic peptides can be a good starting point for inferring possible mechanisms of action or chemical characteristics of peptides that might be related not only to certain therapeutic activity but also with hemolysis. A *Function METN* can be a fast and easy approach to tackle this question by using the StarPep toolbox. Figure 2B shows a *Function METN* of the 2004 hemolytic peptides reported in the *StarPepDB*. Evidently, the most central activities are "hemolytic", "toxic" and "toxic to mammals" since the peptides of study are hemolytic and the metadata nodes are hierarchically related (colored red in Fig. 2B with centrality ranks: 1, 3, and 5, respectively). However, most of these peptides are also related to antimicrobial activity and hierarchically related metadata: antibacterial, anti-Gram positive, anti-Gram negative, antifungal, etc. In fact, these metadata comprise the nine most central nodes in the *Function METN*.

Since the main target of AMPs is the bacterial cell membrane which is disrupted by several reported modes of action[42], it might be feasible that similar modes of action can also target and disrupt human cells, specifically RBCs. Many studies have proposed that due to the positive charge of many AMPs, they can selectively disrupt negatively charged membranes of bacteria while not affecting the neutral membranes of mammals[43,44]. However, it has been demonstrated that several AMPs (some with high antimicrobial activity) can also disrupt mammalian cells as well, causing hemolysis in RBCs[42,45]. In fact, *Function METN* shows that 94.46% of the 2004 peptides that comprise the hemolytic space, have both antimicrobial and hemolytic activity.

Origin METN. This type of METN helps to easily identify the origin of hemolytic peptides, whether they are synthetic or isolated from living organisms. Figure 3A shows the complete *Origin METN* in the dashed box. The central part of the METN was zoomed in and depicted in the center of Fig. 3A. Looking at the complete *Origin METN* three distinctive regions can be observed, an outer ring, a middle ring and a central network. The outer ring represents peptides isolated from living organisms but have not been chemically synthesized. For instance, the peptide StarPep_06954[46] whose metadata origin node corresponds to only *Caenorhabditis elegans*. The middle ring represents peptides with nodes of degree zero.

On the other hand, the central network shows peptides that have only synthetic origin (the most central blue-green nodes) and peptides isolated from living organisms that have also been chemically synthesized (nodes connected to the central violet metadata node and connected to radial violet nodes). Radial violet nodes connected in a chain-like way represent hierarchical taxonomic ranks that are related to species from which a particular peptide was obtained. For instance, the subsequent metadata nodes are connected in the following manner Urochordata->Ascidiacea->Pleurogona->Stolidobranchia->Pyuridae->*Halocynthia*->*Halocynthia*

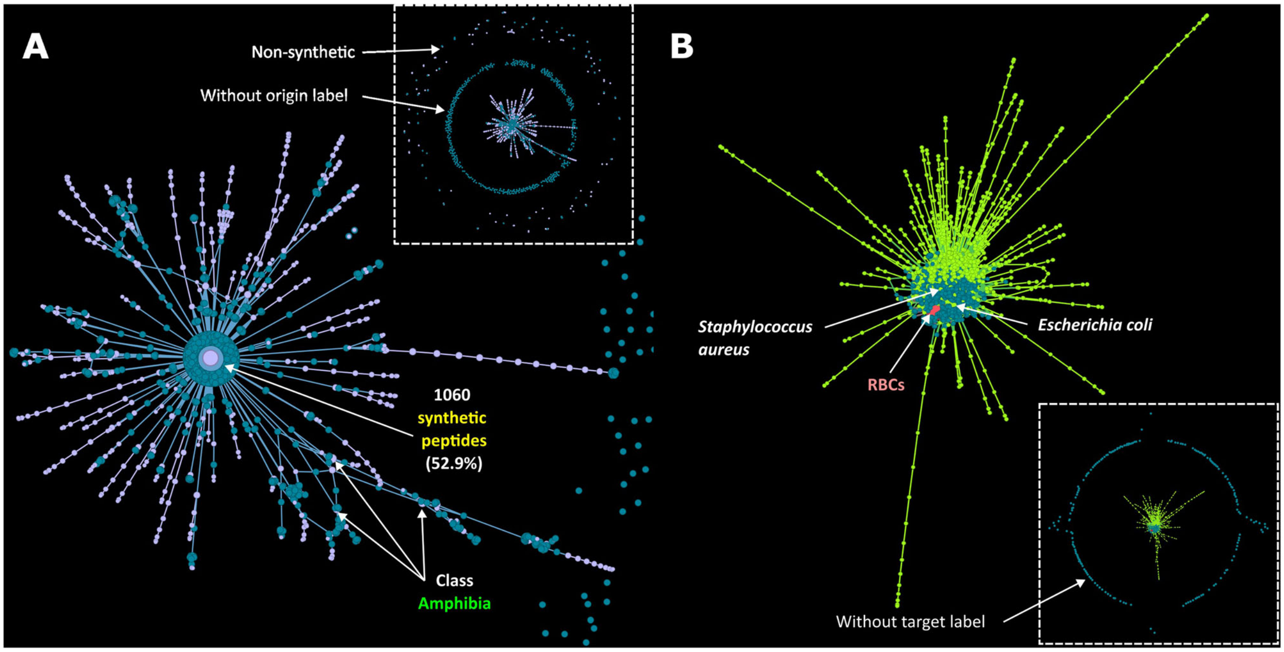


Fig. 3 | Metadata networks (METNs) based on Origin and Target. A *Origin METN* describes the origin of the hemolytic peptides (e.g., synthetic, isolated from *Halocynthia aurantium*, etc.). The dashed box represents the whole *Origin METN* whereas the bigger figure represents the central part of the *Origin METN* that was zoomed in for a better visualization. Blue-green nodes represent peptides while violet nodes represent the origin of the peptides. **B** *Target METN* describes the target of the hemolytic peptides (e.g., RBCs, Gram-positive bacteria, etc.) which is useful information when exploring associations between therapeutic and hemolytic activities. The dashed box represents the whole *Target METN* whereas the bigger figure represents the central part of the *Target METN* that was zoomed in for a better visualization. Blue-green nodes represent peptides whereas green nodes represent the reported target of the peptides. These networks were visualized in Gephi[70] using Force Atlas 2 layout[67] and edited with Inkscape[74].

aurantium. The *H. aurantium* metadata node is then connected to 6 peptide nodes isolated from that species.

Over half of the hemolytic peptides (1060) are of synthetic construct, whereas the rest are isolated from various organisms. Of the top 20 most central origin metadata nodes (synthetic construct not included), half of them belong to the class Amphibia. This is expected because most of the hemolytic peptides in the *StarPepDB* are antimicrobial (Fig. 2B) and a significant part of them have been isolated from frogs and toads since it has been known that they can produce broad-spectrum AMPs in their granular glands in the skin as a defense strategy[47–49].

Target METN. An outer ring and a central network can be observed in this METN (Fig. 3B). The outer ring of peptides seen in the dashed box are peptides that do not have a metadata node related to a target. This metadata network works in the same fashion as the *Origin METN*, where chain-like nodes represent the hierarchical taxonomic ranks, but instead of representing the origin of the peptide, it displays the target of the peptide i.e., the species/cell type in which a certain peptide activity has been evaluated. Evidently, the main target is human erythrocytes (colored red in Fig. 3B) since we are exploring the hemolytic peptide space. Other central targets include *Escherichia coli*, *Staphylococcus aureus*, *Pseudomonas aeruginosa*, *Bacillus subtilis* and *Candida albicans*. They are among the six most central metadata nodes in this METN. It shows that several of the hemolytic peptides have been evaluated as potential AMPs in important human pathogens such as *P. aeruginosa* which has become a real concern in hospital-acquired infections due to drug-resistance appearance[50].

GraphML files of METNs and the descriptor information from each node are available at SM2.

Half-space proximal networks (HSPNs)

The HSPN is a special type of network that was employed to represent the chemical space of hemolytic peptides based on sequence-based molecular descriptors (refer to the "Materials and Methods" section). The properties of the HSPNs were studied based on their global network parameters consisting of the number of edges, modularity, density, average clustering coefficient (ACC), number of communities and singletons, among others. Such statistics can provide a good picture of the topology of the graphs and help selecting networks with the cutoff t that better projects the chemical space of hemolytic peptides.

Our results are consistent with another study that showed that there was little change in the global network parameters when networks are created within the cutoff t range 0.00–0.45[25]. This is because of the highly low number of edges that are removed within this range. In fact, on average, the number of removed edges at $t = 0.50$ correspond to the 1.9% of the initial edges when $t = 0.00$ (See SM3.6).

Moreover, it can be observed that networks generated by different metric measures address differently the similarity between peptides (Fig. 4). Based on their behavior, the networks used in this study can be roughly grouped into three classes: Class I: Angular Separation; Class II: Bhattacharyya, Euclidean and Soergel; and Class III: Chebyshev. The influence of the metric measure in the global parameters of the networks is provided below. All global network parameters calculated for each metric are provided in SM3.

Modularity. This is a measure of network connectivity which indirectly represents how well-defined communities are in the graph and is associated with the number of communities. Graphs generated with Angular Separation (AS) initially possess higher modularity values compared to the other metrics; however, the modularity keeps relatively low at higher t values (0.550 at $t = 0.95$) whereas the other four metrics increase their modularity to values near 1. On the other hand, Chebyshev (Ch) networks show the lowest modularity at low cutoff values, but then it increases to high values comparable with Soergel (So), Euclidean (Eu) and Bhattacharyya (Bh). So- and Eu-derived networks have quite similar behavior in the whole range of t values, whereas Bh networks initially

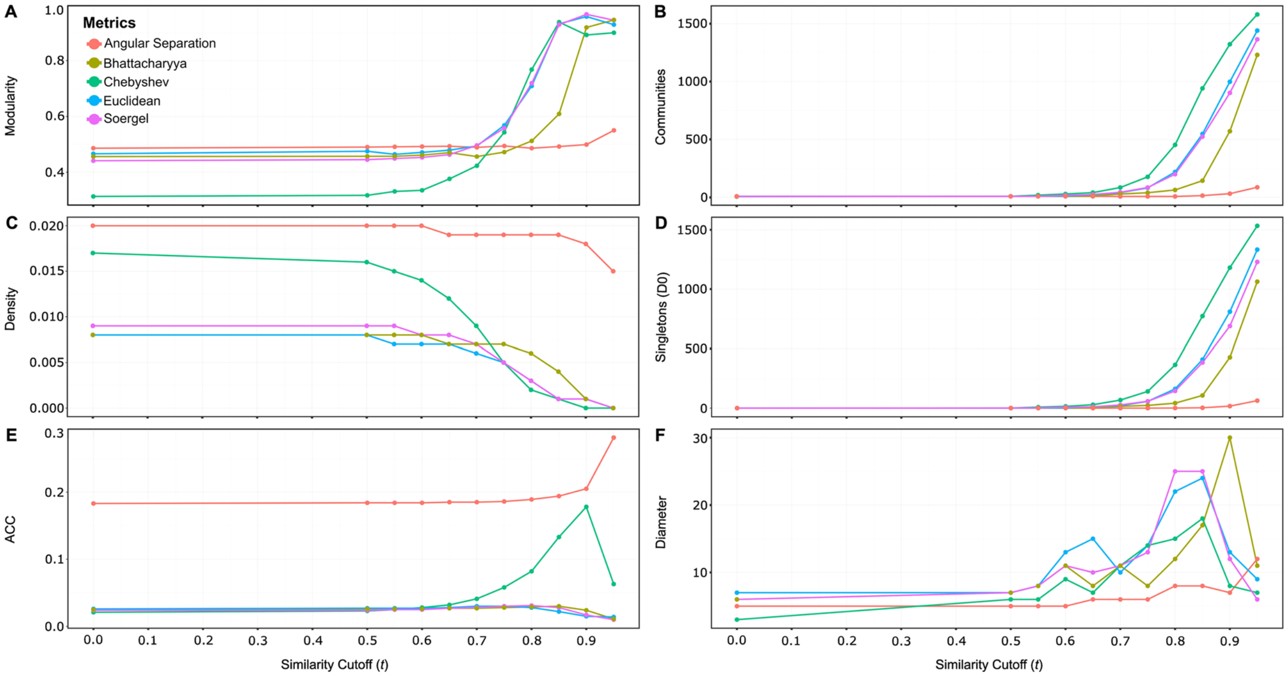


Fig. 4 | Global network parameters of HSPNs created with different metrics and similarity cutoff values *t*. The properties of the HSPNs were analyzed based on their global network parameters, including **A** modularity, **B** number of communities, **C** density, **D** singletons (atypical sequences or outliers), **E** average clustering coefficient (ACC), and **F** diameter. These parameters provide a comprehensive overview of the graph topology, aiding in the selection of networks with the optimal cutoff t for accurately representing the chemical space of hemolytic peptides, as well as facilitating comparisons between different metric measures. ACC average clustering coefficient. This figure was created with ggplot2 R package[75] and edited with Inkscape[74].

behave like Eu and So networks, but then diverge at $t = 0.70$ (Fig. 4A). An adequate selection of modularity is important since highly sparse networks with an elevated number of communities would not provide useful information as several resulting communities would be just artifacts.

Density. It shows the ratio between the edges present in the network and the maximum number of possible edges. Similarity networks have been shown to have an inversely proportional relationship between similarity threshold (t) and density[26,51]. The same pattern is observed for all metrics, but with some notable variations. Here, we can identify three behaviors according to the three classes of metrics. AS networks have the highest density in the entire range of t, whereas Class II metrics (i.e., Bh, Eu, and So) have the lowest density until $t = 0.70$. On the other hand, Ch networks not only have an intermediate initial density but also show the biggest variation of density along the whole range of t (Fig. 4C). In order to select adequate networks, we should choose graphs that are neither too dense nor too sparce since the former would hamper retrieval of useful information whereas the latter would lose information[52]. Density values below 0.20 are desired as they allow us to properly understand the network while preserving high modularity. Particularly, HSPNs are suited because they have the intrinsic characteristic of showing low densities. In fact, the highest density value in this study corresponds to 0.020 (0.00_AS network).

Average clustering coefficient (ACC). This measures the connectivity of the network, and it has been previously studied on molecular similarity networks varying the similarity cutoff t. One study showed that the ACC maximum peak correlates with the best clustering outcome and is a good indicator for finding the appropriate value of t[51]. In our study, three behaviors related to the metric class can be observed again. AS networks have the highest ACCs in the whole range of t with their local maximum at $t = 0.95$. On the other hand, Ch networks start with very low ACCs and get increased at $t = 0.65$ reaching their maximum peak at $t = 0.90$. Finally, Class II metrics have the lowest ACCs in the entire range of t with their maximum peaks at 0.70 (Eu), 0.80 (So) and 0.85 (Bh) (Fig. 4E).

Communities and singletons. The number of communities determined with the Louvain method, the number of singletons D0 (nodes of degree zero) and the number of singletons GC (nodes disconnected from the giant component) were calculated to select the networks with the most reasonable values of these parameters. When $t = 0.00$, HSPNs have the minimum spanning tree as a subgraph, this implies that at this t value all nodes are connected. In other words, no singletons D0 nor singletons GC are found. Regarding the number of communities at $t = 0.00$, all metric networks showed similar values (on average 8 communities). At higher t values, the number of communities and singletons D0 increase dramatically for all the metric networks, except for AS networks (Fig. 4B–D). This is expected as more edges are removed, more nodes are isolated, and now singletons are counted within the communities. Hence, an appropriate t value should be selected that comprises an equilibrium between singletons (atypical peptides) and communities that reflect a real chemical relationship.

Other global network parameters were also calculated to characterize the networks, such as the diameter of the graph (Fig. 4F), the average path length and average degree (See SM3.6). To find the best t value for each metric network, we should look for a compromise between the best parameter value for each descriptor i.e., networks with low density, with neither too many clusters (<20) nor too many singletons (~15–30), retaining high ACC and high modularity. The global descriptors of the selected networks with their best cutoff value t and their respective networks constructed with $t = 0.00$ (10 networks in total) are shown in Table 1.

Finally, we calculated the probability of k (also known as the degree distribution) for each of the selected networks (Fig. 5). Overall, all networks show a right-skewed bell-shaped distribution with high probability of intermediate node degrees. Evidently, plots on the left ($t = 00$) show a probability of zero for singletons ($k = 0$) whereas plots on the right (best

Table 1 | Global network parameters of HSPNs with their best *t* values and their corresponding network at *t* = 0.00

No	Metrics	Cutoff (t)	Edges	Modularity	Density	ACC	Clusters (no D0)	Singletons (D0)	Diameter
1	Angular Separation	0.00	26471	0.490	0.020	0.183	8	0	5
2		0.90	25065	0.499	0.018	0.205	15	17	7
3	Bhattacharyya	0.00	10555	0.456	0.008	0.025	8	0	6
4		0.75	9364	0.472	0.007	0.028	17	23	8
5	Chebyshev	0.00	22431	0.313	0.017	0.021	7	0	3
6		0.65	16809	0.376	0.012	0.032	12	29	7
7	Euclidean	0.00	10498	0.466	0.008	0.026	9	0	7
8		0.70	8482	0.494	0.006	0.030	20	21	10
9	Soergel	0.00	12077	0.441	0.009	0.024	8	0	6
10		0.70	9521	0.496	0.007	0.028	17	27	11

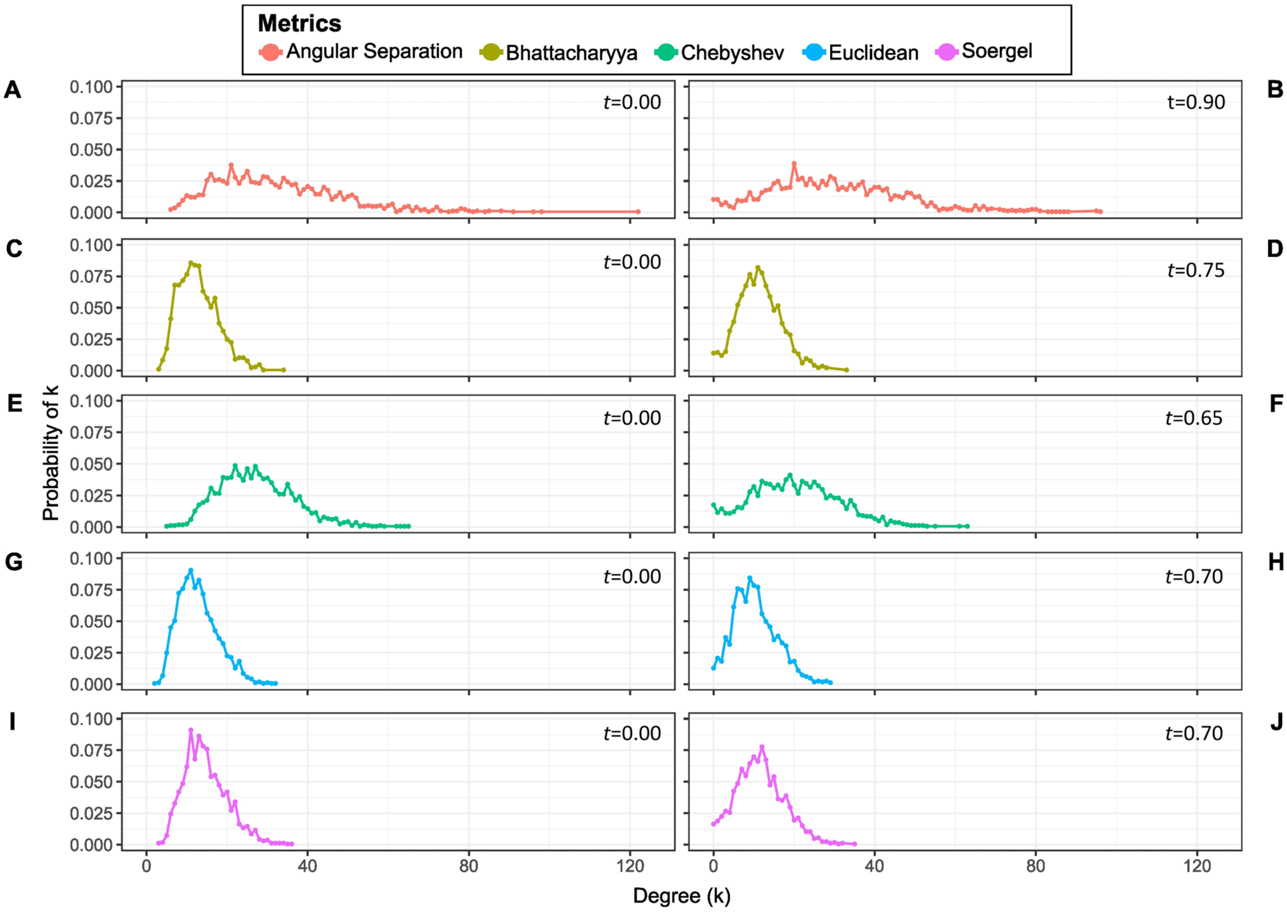


Fig. 5 | Probability of *k* (degree distribution) of the HSPNs with cutoff *t* = 0.00 (left) and with the best cutoff *t* (right) presented in Table 1. The average degree is presented next to the name of the corresponding network. **A** 0.00_AS: 32.15. **B** 0.90_AS: 30.44. **C** 0.00_Bh: 12.82. **D** 0.75_Bh: 11.37. **E** 0.00_Ch: 27.24. **F** 0.65_Ch: 20.31. **G** 0.00_Eu: 12.75. **H** 0.70_Eu: 10.30. **I** 0.00_So: 14.67. **J** 0.70_So: 11.56. This figure was created with ggplot2 R package[75] and edited with Inkscape[74].

value *t*) tend to have a higher probability when *k* = 0. In addition, plots with the best *t* value have smaller maximum degrees (as well as the average degree) compared with same-metric networks at *t* = 0.00. Thus, when comparing networks with the same metric but varying the cutoff value (*t* = 0.00 *vs*. best cutoff *t*), it seems both retain a similar degree distribution. However, when comparing networks with different metrics we can get marked differences. AS networks tend to have a wider distribution range and a higher average degree whereas Ch networks show intermediate values, and networks constructed with Class II metrics show a similar distribution shape among them and have the lowest distribution ranges and average degrees of all metrics. Figure 6 shows the graphical representation of the 10 selected HSPNs.

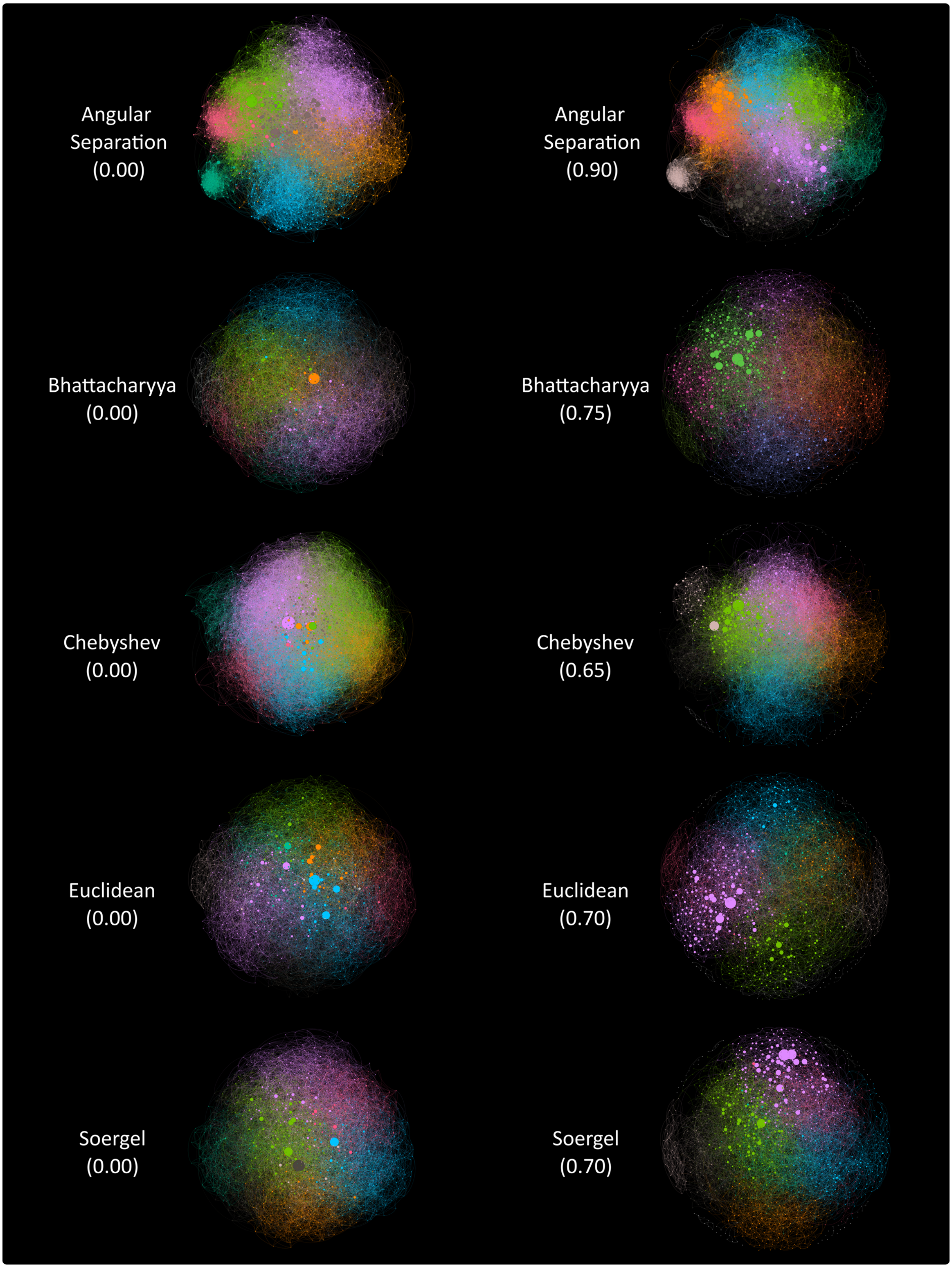


Fig. 6 | Graphical representation of HSPNs with $t = 0.00$ (left) and networks with the best t value for each metric (right). Node colors represent communities of peptides, and the size of the node represents the HB centrality value. Layout: Fruchterman-Reingold[69]. Networks were created with StarPep toolbox[26], visualized in Gephi[70] and edited with Inkscape[74].

HSPNs scaffolds

A total of 240 scaffolds were extracted from the 10 HSPNs (SM4.1). To better understand the effect of the centrality measure, type of alignment and cutoff value s when constructing the scaffolds, several pairwise similarity comparisons between scaffolds were carried out using the Jaccard similarity coefficient (JSC)[53].

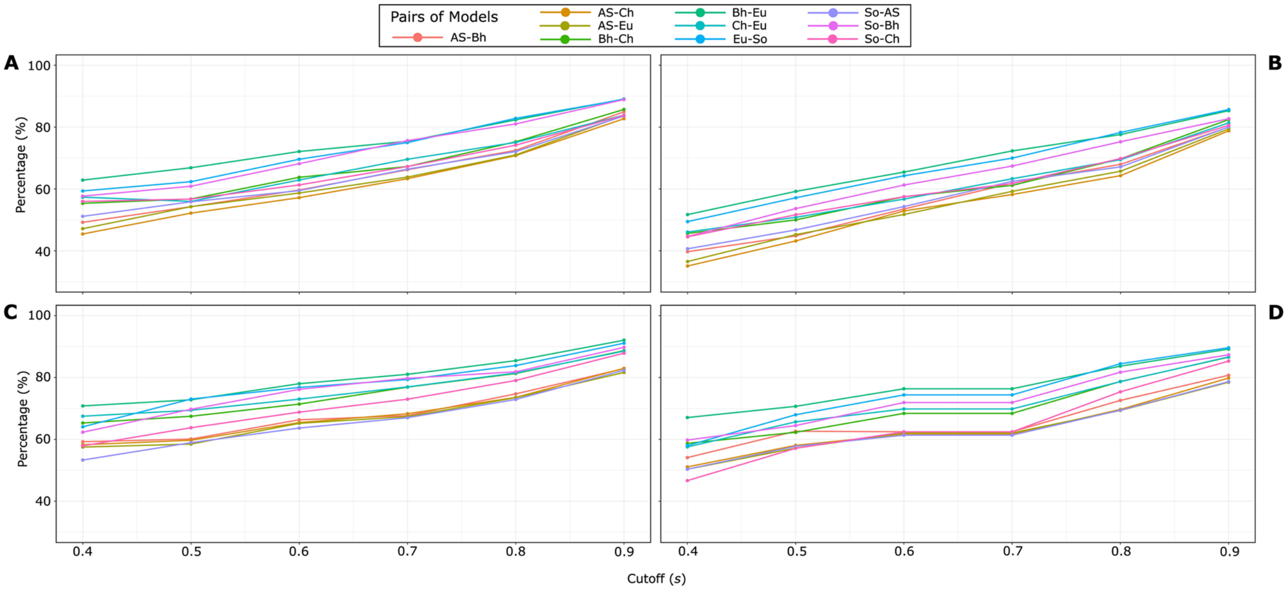


Fig. 7 | Pairwise Jaccard similarity coefficient (JSC) between scaffolds from networks constructed with different metrics when $t = 0.00$. A, B HB centrality. **C, D** HC centrality. **A, C** Global alignment. **B, D** Local alignment. The cutoff *s* represents the similarity cutoff applied to extract the scaffolds whereas the percentage in the *y*-axis represents the percentage of the JSC, which is the number of common peptides between a pair of scaffolds with respect to the union of the peptides of these scaffolds. The higher the percentage, the higher the number of common peptides between pairs of scaffolds. This figure was created with ggplot2 R package[75] and edited with Inkscape[74].

Metric comparison

We compared the type of metric measure used to build the parental networks of the scaffolds. For this comparison, scaffolds ($t = 0.00$) built with the same combinations of centrality, alignment, and cutoff *s* but with different metrics were evaluated (SM 4.3.1). Each pair of scaffolds is represented as a point in Fig. 7.

In all plots of Fig. 7 when $s \geq 0.60$, all scaffold pairs constructed with Class II metrics (i.e., Bh, Eu, So) show the highest similarity percentage compared with the pairs from other combination of metrics. Moreover, scaffold pairs in which one of them is extracted by the AS metric show the smallest similarity percentage at almost any cutoff value *s*. On the contrary, scaffolds selected with Ch metric have an intermediate similarity percentage when compared with scaffolds extracted by other metrics.

These results agree with the previous result which showed that the five metrics tend to have three types of behavior (three classes of metrics). The density (Fig. 4C) and the degree distribution (Fig. 5) of the networks with different metrics are the global descriptors most correlated with the results from the percentage similarity among scaffolds. Thus, it is possible to reduce the number of highly similar scaffolds by using only those HSPNs with the metrics that mostly differ in the global network parameters. In this case, Class II metrics: Bh, Eu, and So are the metric measures with the most similar behavior since they produce similar networks and scaffolds. Therefore, it was decided to conduct the following analyses using only one of the metrics of Class II: Euclidean. This metric was chosen since it is the default metric used in other studies[24,25], and it would be advantageous to compare its performance with the other metrics not previously used in this type of study. Overall, this step allowed us to reduce the redundancy in the scaffold representativity from 240 to 144 scaffolds (SM 4.2).

Cutoff comparison. A cutoff value *t* is not mandatory when constructing HSPNs since at $t = 0.00$, these networks already have low densities under 0.20. However, the topology, characterized by global network features, tends to vary when varying *t* as was demonstrated in the “Half-Space Proximal Networks” section. Thus, it is important to evaluate the effect of selecting a cutoff value (or not) when constructing representative scaffolds of the chemical space. The JSC was calculated between pairs of scaffolds extracted by using the same metric but at different cutoff values ($t = 0.00$ vs. best *t* value), see SM4.3.2 (Fig. 8).

A marked difference was observed when these scaffold pairs were constructed with different types of centralities. Scaffolds constructed with HB centrality (Fig. 8A, B) tend to have more unique peptides at low *s* values and the number of common peptides between scaffold pairs tend to increase when *s* increases. A similar pattern was observed in Fig. 7. However, when the same scaffolds are constructed replacing HB centrality with HC centrality all scaffold pairs tend to share more than 89.50% of peptides regardless of the value of *s* (SM4.3.2.2) (Fig. 8 C, D). Furthermore, the same patterns are preserved when any alignment type is applied. Hence, when generating scaffolds using HC centrality, it is unnecessary to first find the best *t* value for the parental networks since similar scaffolds will be obtained using networks with $t = 0.00$.

Alignment comparison. A clear pattern can be observed when extracting scaffold either using global or local alignment (Figs. 7, 8). In general, local alignment tends to discriminate more strongly at low *s* values than global alignments. Hence, scaffold pairs extracted with local alignment at such low *s* values have a lower similarity percentage than the analog scaffold pairs extracted using global alignment.

In addition, when comparing the similarity percentage of scaffold pairs extracted using the same parameters but differing the alignment type, the same behavior was observed independently of the metric, type of centrality or the *t* value used, see Fig. 9. Scaffold pairs differing only in their alignment type tend to have a low percentage of similarity at low s values, which might indicate that these methods capture the similarity between peptides differently. However, when analyzing the proportion of unique peptides between these scaffold pairs, scaffolds extracted using local alignment are practically a subset of scaffolds extracted when using global alignment. In fact, the average number of unique sequences in local scaffolds when comparing them with their global counterparts at any cutoff *s* is 16.19 (SM4.3.3). An example is provided for the scaffold pairs: 0.00_AS_HB_G_0.40 and 0.00_AS_HB_L_0.40 (Fig. 10).

Centrality comparison. Pairwise comparisons of the scaffolds constructed using the same parameter but changing the centrality measure

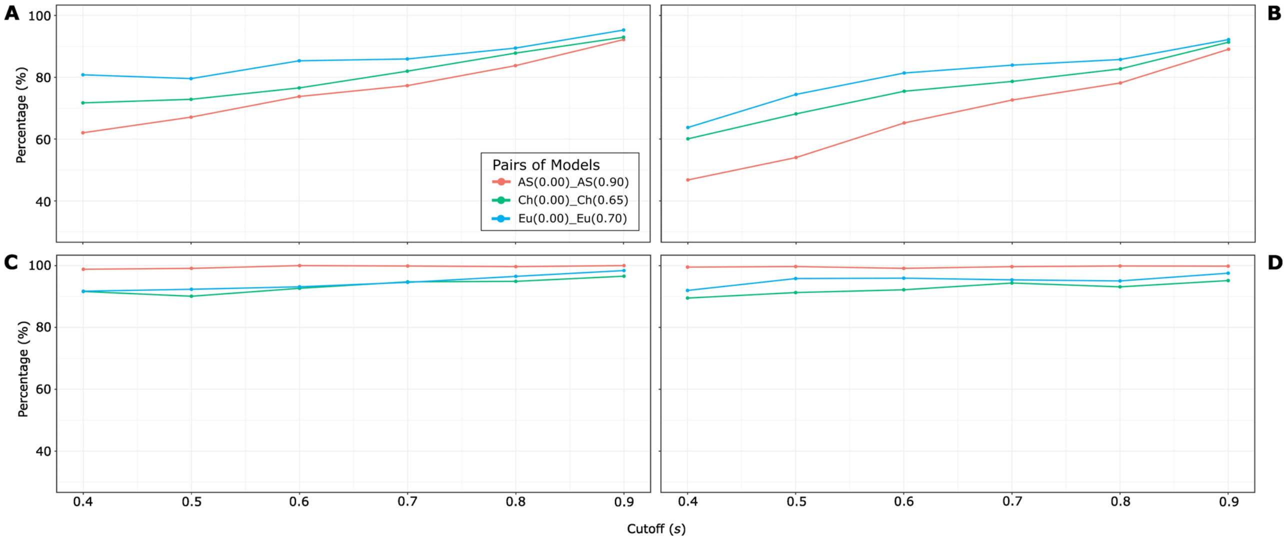


Fig. 8 | Pairwise Jaccard similarity coefficient (JSC) between scaffolds from networks constructed with the same metric but differing their *t* values ($t = 0.00$ vs. best *t* value). **A**, **B** HB centrality. **C**, **D** HC centrality. **A**, **C** Global alignment. **B**, **D** Local alignment. This figure was created with ggplot2 R package[75] and edited with Inkscape[74].

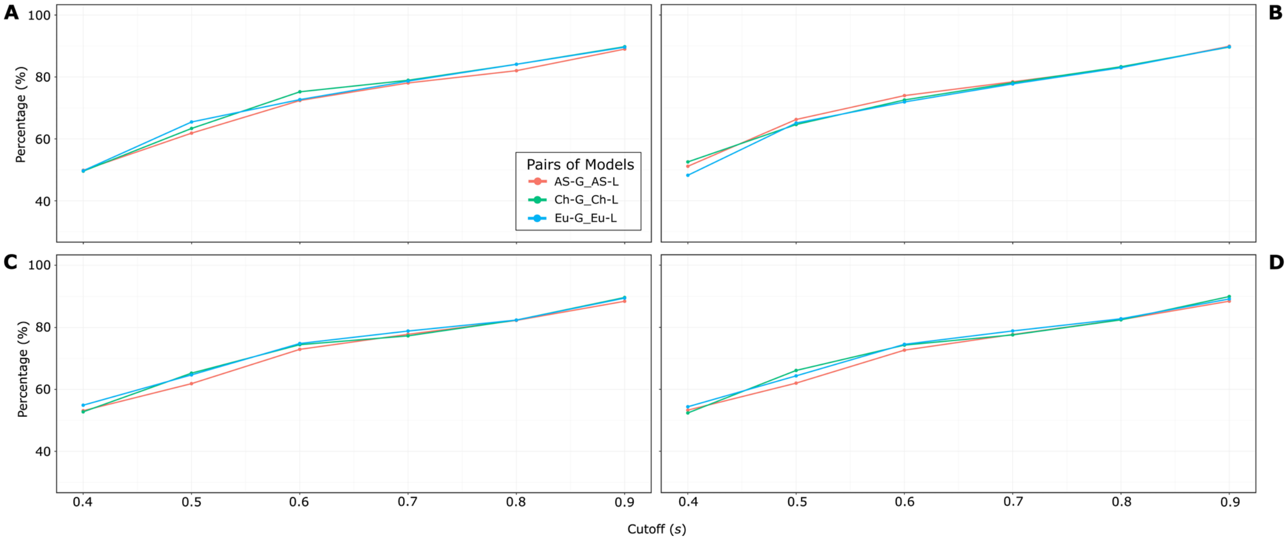


Fig. 9 | Pairwise Jaccard similarity coefficient (JSC) between scaffolds from networks constructed with the same metric but differing alignment type. **A**, **B** HB centrality. **C**, **D** HC centrality. **A**, **C** networks with $t = 0.00$. **B**, **D** networks with best cutoff *t*: AS (0.90), Ch (0.65), Eu (0.70). This figure was created with ggplot2 R package[75] and edited with Inkscape[74].

show a trend like the pairwise comparisons presented before (SM4.3.4). This implies that the type of centrality used to extract the scaffold will affect the sequences that are removed/retained, especially at low *s* values.

On the other hand, when comparing centrality measures, JSC between scaffold pairs extracted from networks with best *t* value tend to be higher than JSC from scaffold pairs from networks with $t = 0.00$. This pattern is clearer at low *s* values (Fig. 11).

All scaffolds presented in this section can be used in many applications. For instance, they can be used as training datasets for both ML-based and Multi-Query Similarity Searching (MQSS) prediction models of hemolytic peptides. In fact, a recent study demonstrated that MQSS models based on the scaffolds identified in this study outperformed state-of the art ML-based model classifiers[54]. The advantage of using these scaffolds is that they store information of central and important peptides as well as outliers or atypical hemolytic peptides while avoiding overrepresentation of certain peptide classes (sampling bias). Each scaffold bears a unique type and amount of information of the hemolytic peptide space and one scaffold can be more suitable than another depending on the scaffold's use. Scaffolds extracted at low cutoff *s* values tend to cover fewer peptides of the original space, whereas higher *s* values capture more information of the space, but peptide overrepresentation might be present. Figure 12 depicts an example of the scaffold coverage when varying the cutoff *s*.

Hemolytic motif discovery and enrichment

Motif discovery. Peptides from each community were used as input sequences to uncover new hemolytic motifs within the communities' diversity by means of STREME, an alignment-free method[55–61]. Table 2 shows a sample of the 42 new motifs discovered using clusters of HSPNs ($t = 0.00$) created with different metrics. 12 motifs were found from 6 clusters of the network 0.00_AS, 14 motifs were discovered from 4 clusters of the network 0.00_Ch and 16 motifs from 5 clusters were discovered using the network 0.00_Eu. The three metrics commonly detected only four motifs: GLP, MFTKL, ERBADE and VCTRN. It is worth mentioning that several other motifs were similar but not identical such as: GLP/GLPV or VGGTCN/GGTCN. In addition, 15 motifs were discovered without considering the community diversity by using all 1647 hemolytic peptides as input sequences. All these motifs were grouped as HSPNs motifs. After removing duplicated motifs, 50 HSPNs motifs were discovered (SM5.1.2).

Two previous reports on ML models for predicting hemolytic activity of peptides have also reported hemolytic motifs, namely: HemoPI[8] and HAPPENN[1]. HemoPI reported 21 motifs extracted using MERCI software that were enriched in positive sequences from HemoPI-1 and HemoPI-2 datasets, whereas HAPPENN motifs resulted by looking for the 20-top motifs found exclusively in the positive dataset of HAPPENN. No HSPN-derived motifs were found among the reported ones. To generate a unique list of non-redundant hemolytic motifs, HSPNs motifs were combined with the previously reported ones resulting in 91 putative motifs. Then similar motifs were combined into consensus motifs resulting in 57 non-redundant motifs (SM5.1.2 and SM5.1.3).

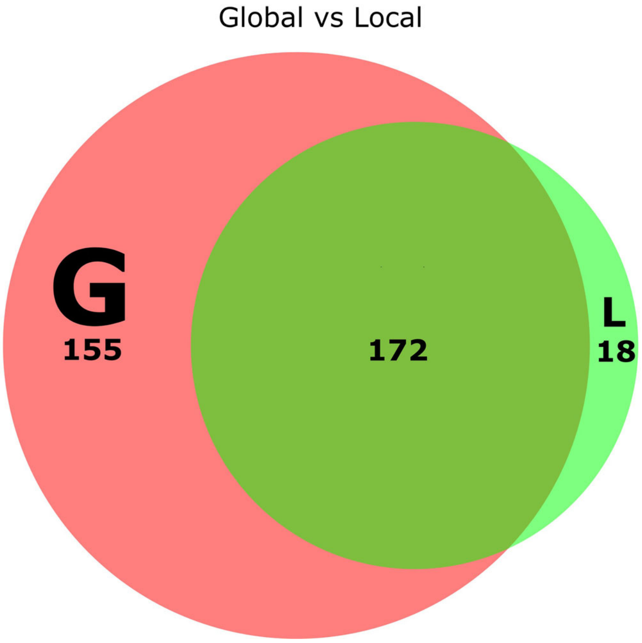


Fig. 10 | Size comparison of scaffold pairs generated from the network 0.00_AS. Pink area (G) represents the peptide sequences unique to the scaffold 0.00_AS_HB_**G**_0.40, green area (L) represents the sequences unique to the scaffold 0.00_AS_HB_**L**_0.40. The intersection of pink and green represents the number of common peptides between these two scaffolds. The area-proportional Venn diagram was created using DeepVenn[76] and edited with Inkscape[74].

Motif enrichment. To identify and validate the most representative hemolytic motifs and remove some artifacts from the 57 potential hemolytic motifs, we conducted enrichment analyses using SEA method on three different datasets: *HemoPI-1*, *StarPepDB* and *Big-Hemo* (SM5.2). Motifs not reported as significant in at least one dataset were removed. The resulting 47 hemolytic motifs sorted by the average enrichment ratio of all datasets are presented below (newly discovered motifs by HSPNs are shown in red): **MFTLK, ALKAIS, GTCN, WKSFJK, VCGETC, WKK**, AKKAL, **GETCV**, CYCR, LKKL, CVCV, **ISWIK**, RFC, **LHTA[KL], FLHSAK, CSW**, LWKT, **FLGTI, GAVLKV, PGC**, KKILG, KITK, **KHI**, LGKL, KWK, VNWK, **K[GT]AGK, VCT**, ALW, **SWP**, HIF, **LLKK, [VI]LDTJ**, CRR, **KLL, JGKL**, FKK, GAIA, VLK, **GLP**, PKIF, GKEV, **GTIS**, AAAK, GCS, **IAS**, MAL (Table 3).

These motifs might be involved in the mechanisms of action of hemolytic peptides as well as antimicrobial activity, but further studies are needed to corroborate this assumption. Another possible use of these motifs can be as a toxic signature, where proteins containing some of these motifs could be attributed to a relatively high hemolytic activity in comparison with proteins with few or nonhemolytic motifs. Table 4 shows an example of three pairs of peptides whose hemolytic activity is related to the number of hemolytic motifs present in their sequences.

We decided to further explore this hypothesis by comparing the relation between the number of hemolytic motifs in a peptide and its likelihood of being hemolytic (SM5.3). To obtain the predicted hemolytic activity of a peptide, we used the consensus of two different model classifiers that were identified in a previous report to have a robust performance after a multiple comparison[54]. This experiment was carried out using three datasets: *antibacterial*, *antiviral* and *FDA-approved.* All three datasets have peptides with lengths up to 100 AAs.

For the *antibacterial* and *antiviral* datasets, a general pattern can be identified. Most peptides without any of the reported hemolytic motifs tend to be non-hemolytic. When peptides have one or more motifs, peptides are mostly hemolytic. The hemolytic/non-hemolytic ratio gets more pronounced with the increase of the number of motifs. Interestingly, peptides with a high number of motifs were predicted to be exclusively hemolytic

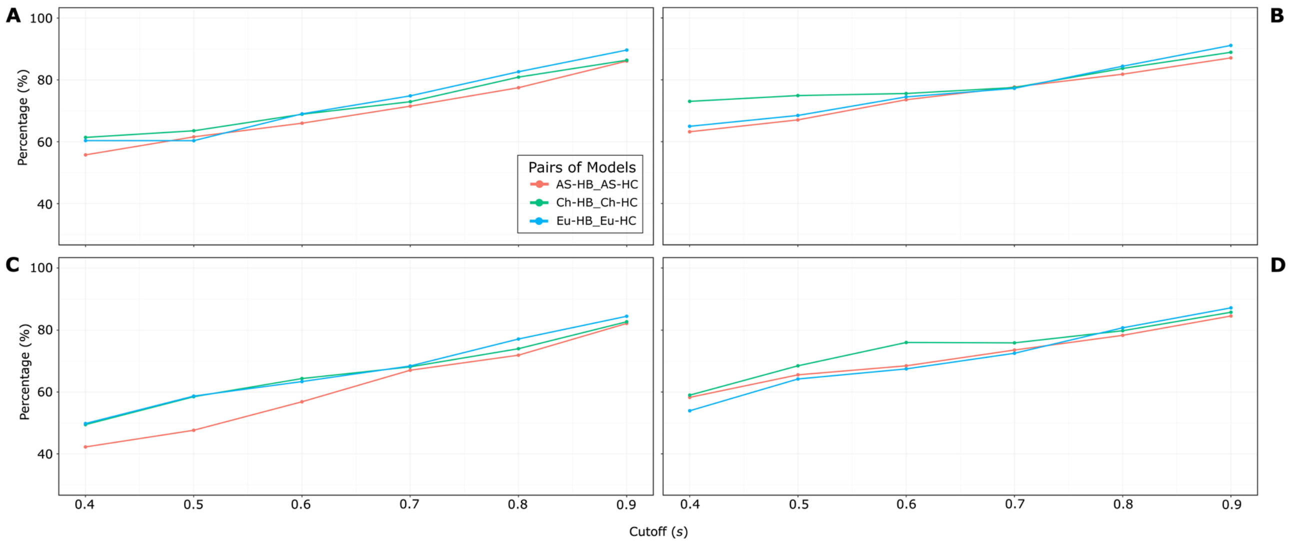


Fig. 11 | Pairwise Jaccard similarity coefficient (JSC) between scaffolds from networks constructed with the same metric but differing the centrality type. **A**, **B** Global alignment. **C**, **D** Local alignment. **A**, **C** networks with $t = 0.00$. **B**, **D** networks with best cutoff t: AS (0.90), Ch (0.65), Eu (0.70). This figure was created with ggplot2 R package[75] and edited with Inkscape[74].

peptides. Nevertheless, it is worth noting that only a few peptides actually contained a high number of these motifs (Fig. 13).

The activity of the peptide might be another aspect to consider when conducting this type of analysis. For example, antibacterial peptides without any reported motif tend to be non-hemolytic, but a high number of peptides without motifs are also predicted to be hemolytic. On the other hand, antiviral peptides without any motifs are almost exclusively non-hemolytic. Therefore, the absence of reported hemolytic motifs does not imply that the peptides are not hemolytic. The same is true when peptides contain one or more hemolytic motifs; it does not necessarily mean they are hemolytic, but the higher the number of motifs, the higher the possibility that peptides are hemolytic (Fig. 13).

When the same analysis was conducted in 49 FDA-approved peptides, a major difference was observed. Only three peptides were predicted as hemolytic[54], Glatiramer acetate (Th1113/ seq_32) contained one hemolytic motif, whereas two other peptides did not report any hemolytic motif, namely Lucinactant (Th1146/ seq_41) and Gramicidin D (Th1024/ seq_8). More importantly, the trend was opposed to what was found in the *antibacterial* and *antiviral* datasets. The ratio between hemolytic/non-hemolytic peptides containing more than one motif was inverted, i.e., hemolytic peptides with one or two motifs were scarce compared to non-hemolytic peptides containing the same number of motifs. Another important detail is that in these sequences, in spite of having the same maximum length (100 AAs) as the *antibacterial* or *antiviral datasets*, peptides from the *FDA-approved* dataset displayed at maximum two reported motifs (Fig. 13C, F). This result agrees with the fact that approved peptides have to be safe and to avoid being hemolytic/toxic, unless the application mode does not directly interact with the bloodstream as is the case of Gramicidin D, a commercial antibacterial drug that is only applied on the skin because of its high hemolytic activity[62].

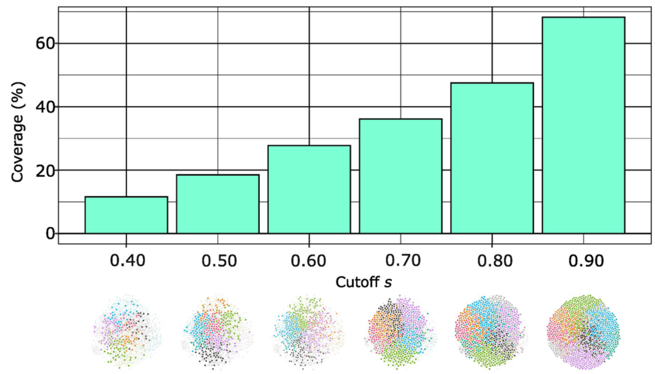
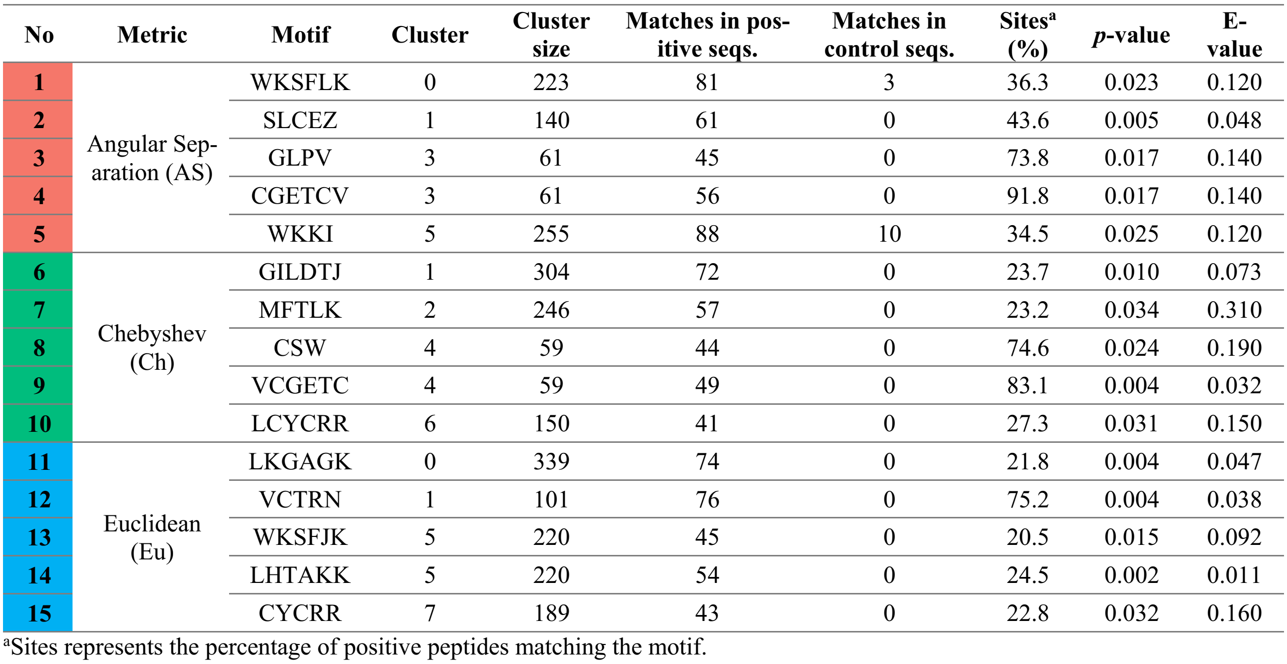


Fig. 12 | Barplot showing the coverage of the scaffolds 0.00_AS_HB_L at different *s* values. Scaffold representations are shown below their cutoff *s* values. This figure was created with ggplot2 R package[75] and edited with Inkscape[74].

Discussion

Positive endpoints are commonly evaluated in peptides to better understand a specific therapeutic activity; however, getting insight into negative endpoints such as hemolysis should be equally important. In this study, the exploration of the chemical space of hemolytic peptides through a synergic combination of network science and interactive data mining resulted in an easy and feasible way to get more insight into the features that characterize this type of peptides. METNs helped elucidate that almost all hemolytic peptides (94.46%) have also antimicrobial activity. The fact that most central nodes in the Target METN are related to human pathogens further supports the claim that antimicrobial activity and hemolytic activity might be related. Furthermore, The Database METN shows that hemolytic peptides can be identified in datasets specialized in hemolysis (Hemolytik) but can be also found in datasets specialized in therapeutic peptides (SATPdb) or in antimicrobial peptide databases (DBAASP). A high degree of hemolytic peptide redundancy has been also found between these databases. Regarding the origin of hemolytic peptides, over half of them come from synthetic constructs; however, many other peptides were isolated from living organisms, being the class Amphibia one of the most important ones. METNs were also useful in identifying missing metadata information regarding the origin and target of many hemolytic peptides.

Table 2 | Motifs discovered by STREME using the community information from the HSPNs created using Angular Separation, Chebyshev, and Euclidean metrics with $t = 0.00$



No	Metric	Motif	Cluster	Cluster size	Matches in positive seqs.	Matches in control seqs.	Sites[a] (%)	*p*-value	E-value
1	Angular Separation (AS)	WKSFLK	0	223	81	3	36.3	0.023	0.120
2		SLCEZ	1	140	61	0	43.6	0.005	0.048
3		GLPV	3	61	45	0	73.8	0.017	0.140
4		CGETCV	3	61	56	0	91.8	0.017	0.140
5		WKKI	5	255	88	10	34.5	0.025	0.120
6	Chebyshev (Ch)	GILDTJ	1	304	72	0	23.7	0.010	0.073
7		MFTLK	2	246	57	0	23.2	0.034	0.310
8		CSW	4	59	44	0	74.6	0.024	0.190
9		VCGETC	4	59	49	0	83.1	0.004	0.032
10		LCYCRR	6	150	41	0	27.3	0.031	0.150
11	Euclidean (Eu)	LKGAGK	0	339	74	0	21.8	0.004	0.047
12		VCTRN	1	101	76	0	75.2	0.004	0.038
13		WKSFJK	5	220	45	0	20.5	0.015	0.092
14		LHTAKK	5	220	54	0	24.5	0.002	0.011
15		CYCRR	7	189	43	0	22.8	0.032	0.160

[a] Sites represents the percentage of positive peptides matching the motif.

[a] Sites represents the percentage of positive peptides matching the motif.

Table 3 | Hemolytic motifs that have all their *E*-value ranks less than 37 sorted by their average enrichment ratio of the three datasets: *HemoPI-1*, *StarPepDB* and *Big-Hemo*

		HemoPI-1			*StarPepDB*			*Big-Hemo*		
No.	**Motif**	**ER[a]**	**E-value**	**Rank[b]**	**ER[a]**	**E-value**	**Rank[b]**	**ER[a]**	**E-value**	**Rank[b]**
1	**ALKAIS**	3.66	1.92E-09	36	40.10	3.53E-21	35	3.32	8.48E-12	9
2	**WKSFJK**	19.20	2.80E-40	2	5.06	3.55E-158	1	4.94	6.01E-22	3
3	AKKAL	16.10	7.19E-29	11	3.33	6.66E-101	4	1.55	6.30E-04	24
4	LKKL	12.60	1.44E-31	4	3.62	2.65E-130	2	1.68	1.37E-08	12
5	**ISWIK**	7.86	5.69E-19	19	6.19	3.45E-59	15	2.51	4.45E-05	22
6	**LHTA[KL]**	3.94	1.90E-13	25	8.25	4.73E-27	29	3.74	1.76E-11	10
7	**FLHSAK**	7.04	1.82E-11	29	5.69	2.10E-45	21	1.95	1.14E-03	26
8	LWKT	7.25	4.60E-31	6	2.35	5.59E-55	18	3.50	2.59E-10	11
9	**FLGTI**	6.94	1.41E-14	22	2.15	1.18E-21	33	3.88	5.70E-24	2
10	KKILG	6.71	1.61E-26	13	3.29	3.56E-77	11	1.85	1.36E-07	13
11	KITK	6.99	5.48E-26	15	2.48	1.22E-57	16	2.05	1.68E-01	36
12	LGKL	5.47	1.14E-29	7	2.17	5.13E-87	8	3.34	5.48E-12	8
13	KWK	4.84	2.02E-31	5	3.97	1.22E-55	17	1.98	1.79E-07	15
14	**KGAGK**	5.13	2.35E-27	12	2.66	2.25E-43	22	2.81	2.43E-14	4
15	**SWP**	4.56	5.42E-26	14	3.76	7.67E-35	26	1.98	5.44E-03	28
16	**LLKK**	4.31	1.88E-34	3	3.82	1.35E-126	3	1.60	1.18E-01	35
17	**[VI]LDTJ**	3.02	4.39E-10	33	2.15	1.05E-40	23	4.27	1.58E-24	1
18	JGKL	4.07	1.38E-29	8	2.32	8.01E-90	7	1.71	2.12E-07	17
19	VLK	3.00	8.34E-17	20	2.06	9.64E-64	14	1.88	1.52E-07	14
20	PKIF	2.89	1.05E-14	21	2.19	3.22E-46	20	1.47	4.79E-03	27

[a] *ER* enrichment ratio.
[b] SEA returns motifs ranked according to their associated *E*-value.
For complete information about all motifs see SM5.2. Motifs highlighted in red are newly reported hemolytic motifs.

Table 4 | An example of the use of hemolytic motifs as toxic signatures

No.	**Sequence**	**Length**	**No. Motifs**	**Consensus Motifs**	**Hemolytic Activity**	**Ref**
1	RGLRRLGRKIAHGVKKYGPTVKRIKRKA	28	0		Not active at 100 μM	[62]
2	KWKSFLKTFKSAAKTVLHTALKAISS	28	4	WKSFJK, LHTA[KL], KWK, ALKAIS	50% hemolysis at 16 μM	[63]
3	MASRAARLAARLARLALRAL	20	0		1% hemolysis at 92.95 μM	[64]
4	ALWMTLLKKVLKAAAKAALN	20	4	LLKK, VLK, AAAK, ALW	50% hemolysis at 5 ± 1 μM	[65]
5	KRLFRRWQWRMKKY	14	0		Not active up to 100 μM	[66]
6	WCYCRRRFCVCVGR	14	3	RFC, CYCR, CRR	> 50% hemolytic at 44.3 μM	[67]

The table shows pairs of hemolytic peptides with similar lengths showing different hemolytic activities based on the number of hemolytic motifs. Motifs in red are newly reported hemolytic motifs.

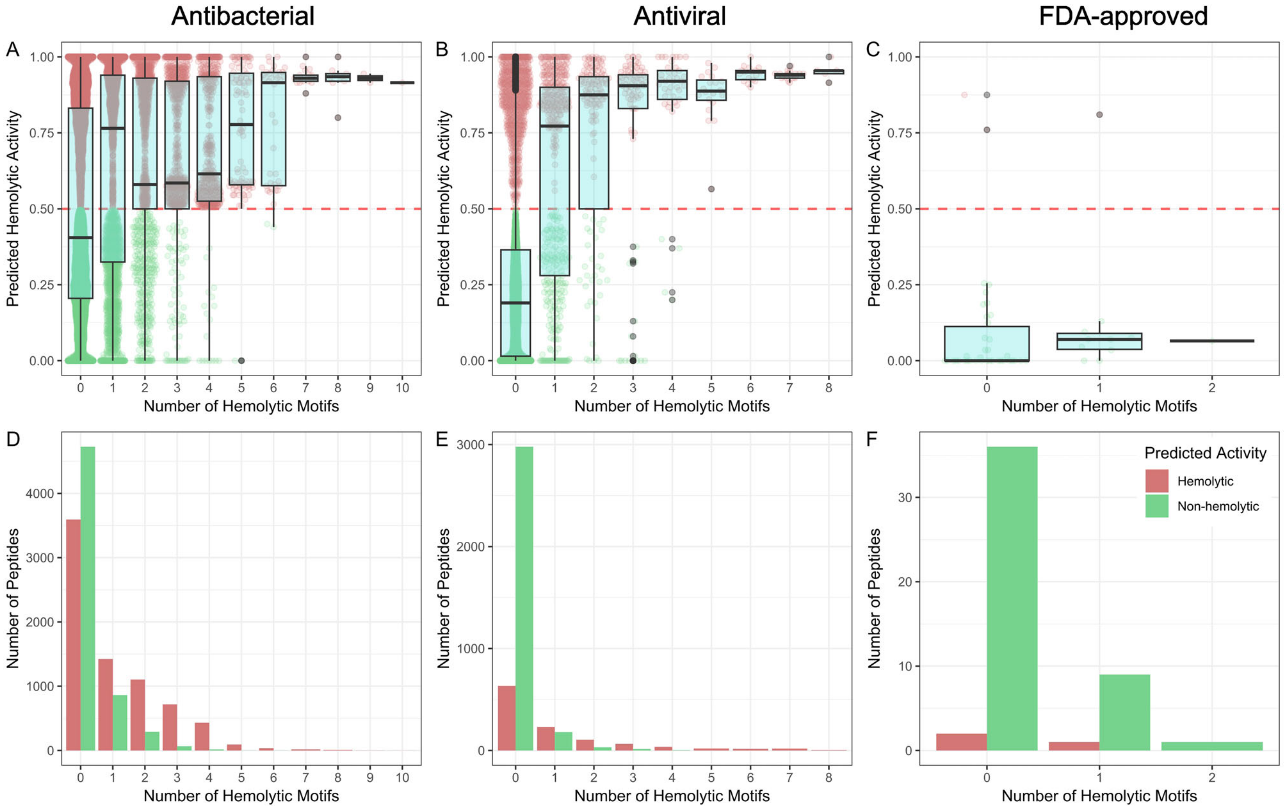


Fig. 13 | Relation between the number of hemolytic motifs and the hemolytic activity predicted using two model classifiers: "SVM + Motif (HemoPI-1) based"[8] and "MQSSM-I1"[54]. This analysis was performed in three datasets: Antibacterial, Antiviral and FDA-approved. **A**, **B** and **C** represent boxplots displaying the number of hemolytic motifs in a peptide and the predicted hemolytic activity obtained using the model classifiers. **D**, **E** and **F** show the absolute frequency of hemolytic and non-hemolytic peptides for each group (number of hemolytic motifs). This figure was created with ggplot2 R package[75] and edited with Inkscape[74].

On the other hand, HSPNs resulted in a valuable representation of the chemical space of hemolytic peptides. The use of five different (dis)similarity measures to construct different HSPNs gave rise to in the identification of three different classes of measures that interpreted differently peptide relatedness (Class I: AS; Class II: Bh, Eu, So; Class III: Ch). In fact, measures within class II tend to produce similar behaviors in the network topology when varying the cutoff *t*, similar degree distributions in the best HSPNs, and they also produce similar subsets in the scaffold extraction process. Scaffolds extracted from the ten different HSPNs were further compared using the JSC. We found that the use of a cutoff *t* in the construction of HSPN is not mandatory, especially when using HC centrality to score node relevance within the networks. The use of either global or local alignment, on the other hand, might result in different scaffold representations, especially using local alignment at low cutoff *s* values. It is worth noting that scaffolds extracted using the same parameters but only varying the alignment type resulted in global scaffolds with more peptides than local scaffolds, and most peptides from the local scaffold were already included in the global scaffold. Additionally, a general trend was found regarding the similarity between scaffolds while varying the cutoff *s*: the higher the cutoff value, the higher the similarity between scaffolds. This exploration process was essential because it helps understand the effect of varying each parameter in the scaffold extraction process which can then be used to generate only scaffolds with unique representations. Scaffolds are indeed useful in many applications including the study of a simplified representation of the hemolytic peptide space that retains as much information as possible. Nevertheless, one of the most useful applications is the use of scaffolds as queries to construct MQSS model classifiers[54].

The use of HSPNs using a diverse set of (dis)similarity measures also contributed to the discovery of new motifs highly associated with hemolytic activity using an alignment-free approach. Form the 47 motifs identified more than half of them (53.19%) were not previously reported. More importantly, we provided a statistical approach to assess the enrichment ratio of motifs in three different datasets containing positive (hemolytic) and negative (non-hemolytic) peptides, which was not available to the previously reported motifs. Finally, we found a positive association between the number of hemolytic motifs that can be found in a sequence and the likelihood of a peptide being hemolytic. Although the probability of finding a peptide with more than one motif is reduced as the number of motifs increases, it is shown here that hemolytic peptides are the dominant peptides that contain at least one hemolytic motif compared to their non-hemolytic counterparts. This trend, however, was not observed in the *FDA-approved* dataset because it contains peptides that have been selected and approved for having a low toxic profile.

Overall, this method provides a new alternative to the discovery, study and repurposing of promising peptides with therapeutic activity and low toxicity. However, as a new method, there are still some processes that need further explored and enhanced. The topology of the peptide space highly depends on the number of peptides, type of network and the (dis)similarity metric used. Here we only explored HSPNs along with five metrics, but it would be interesting to test using different networks like CSNs and see if the information retrieved from these networks resembles each other or not. Moreover, the addition of newly reported hemolytic peptides will positively enrich the peptide space. It would be also recommended to use in the future networks that consider the potency of hemolysis to discover properties and motifs that might be exclusive to highly hemolytic peptides which are more concerning when designing therapeutic peptides. However, this approach can be hampered by the lack of a clear consensus about the criteria to classify peptides with high vs low hemolytic activity.

Finally, the comparison between the number of reported hemolytic motifs and the likelihood of a peptide being hemolytic is limited by the fact that predictions of peptide hemolytic activity were employed instead of experimentally validated results. It was conducted this way as most peptides in the datasets used in this experiment did not report experimental data about hemolytic activity, but we wanted to study the overall patterns in the entire antibacterial, antiviral and FDA-approved datasets. However, this was still a good approximation as the model classifiers used here were reported to be reliable[54].

Materials and methods

Datasets

The datasets used in this study to construct METNs, HSPNs, generate hemolytic motifs, and subsequently conduct motif enrichment analyses are described in Supplementary Material (file SM1). The usage of each dataset in this report is detailed below.

- *StarPepDB*. It is a graph database embedded in the StarPep toolbox that consists of 45,120 peptides with annotated activities retrieved from 40 bioactive databases and other sources[16]. A sub-dataset consisting of 2004 hemolytic peptides was extracted from this database to generate HSPNs, METNs and discover new hemolytic motifs. The complete *StarPepDB* was also used in the motif enrichment process to help find the most representative hemolytic motifs. Furthermore, two additional datasets were obtained from the *StarPepDB*, namely *antibacterial* (13,399 sequences) and *antiviral* (4345 sequences). In these two datasets, peptides with non-standard AAs and with length greater than 100 AAs were removed. *Antibacterial* and *antiviral* datasets were used to study the relation between the number of hemolytic peptides and the likelihood of being hemolytic (SM1.1).
- *HemoPI-1*. It encompasses 552 experimentally validated highly hemolytic peptides (positive) and 552 random peptides extracted from Swiss-Prot (negative)[8]. This dataset was only used in motif enrichment analysis (SM1.2).
- *Big-Hemo*. It is a non-redundant combination of several datasets that contain either hemolytic or highly hemolytic peptides as positive samples and non-hemolytic or low hemolytic peptides as negative samples. The datasets used to generate the *Big-Hemo* dataset are HemoPI-2 Main and Validation[8], HemoPI-3 Main and Validation[8], HAPPENN[1], HLPred-Fuse Layer 2 Training and Independent datasets[4] and HemoNet[9]. To construct *Big-Hemo*, only positive samples labeled as “highly hemolytic” were retrieved from these datasets to handle the problem of lack of agreement and standardization at considering when a peptide is hemolytic or not, and the way of measuring this property, respectively[1,42]. Although HAPPENN dataset contains positive samples not labeled as highly hemolytic, its positive samples were also included in *Big-Hemo* to gain more diversity and a better representation of hemolytic peptides. Thus, this dataset was addressed to evaluate whether our novel motifs are enriched in highly hemolytic peptides, which are more concerning when designing therapeutic peptides. In addition to redundancy removal, peptides containing ‘X’ several times in a sequence and Nphe or Nleu in their sequences were also discarded. The resulting *Big-Hemo* dataset contains 2196 highly hemolytic peptides. Like *HemoPI-1* dataset, *Big-Hemo* was also used for motif enrichment analysis (SM1.3).
- *FDA-approved*. This dataset contains 47 FDA-approved peptide-based drugs retrieved from ref. 63. Peptides with non-standard AAs and with length greater than 100 AAs were removed. Two FDA-approved drugs consisted of two peptide sequences each (Th1027: seq_9 and seq_10; and Th1041: seq_13 and seq_14); hence 49 peptides were retrieved (SM1.4). This dataset was used to study the relation between the number of hemolytic motifs and the likelihood of being hemolytic in approved peptides. The *FDA-approved* dataset was used instead of the information reported on DrugBank 6.0[11] because many peptide sequences were not available on there.

Network generation and analysis

Metadata Networks (METNs). A Metadata Network (METNs) is an unweighted pseudo-bipartite graph defined as $F = (V, E)$, where $E(F)$ is the set of edges of the graph and $V(F)$ is the set of nodes or vertices which comprises two classes: hemolytic peptides and metadata information (e.g., origin and function of peptides). In these networks, peptide nodes are adjacent to their corresponding metadata nodes. For instance, if a peptide is hemolytic, an edge will connect this peptide node to the hemolytic metadata node. However, METNs are not fully bipartite graphs since in the last ones the nodes belonging to the same class cannot be adjacent[64], whereas METNs can set edges within the metadata class as long as one node is hierarchically related to another. For instance, for the “Function” metadata, “Toxic to mammals” is hierarchically connected to “Hemolytic”, thus an edge connects these two metadata nodes.

StarPep toolbox allows the easy construction of METNs, which helps to get insight into the related data associated with the hemolytic peptides. A *Database METN*, for instance, shows the databases where each hemolytic peptide has been reported by connecting each peptide node to its corresponding database nodes. This information is useful to get an overview of the most populated databases with hemolytic peptides, to analyze peptide redundancy in different databases and to detect what peptides are uniquely reported dataset, etc. Hence, we created four METNs based on different metadata information: database, function, origin, and target. The peptide class of $V(F)$ was the set of 2004 hemolytic peptides from *StarPepDB*.

Half-Space Proximal Networks (HSPNs). HSPNs, are weighted graphs defined as $G = (V, E)$ where $V(\mathrm{G})$ represents the set of nodes (hemolytic peptides) and $E(G)$ represents the set of edges. The nodes are characterized by vectors whose components are values of sequence-based molecular descriptors (MDs), whereas the edges link nodes in a pairwise manner following the subsequent steps:

- A (dis)similarity measure is calculated for each pair of nodes using the vectors of peptide features. Then these values are normalized (min-max normalization). This forms a symmetric similarity matrix M of size $n \times n$ where n represents the number of hemolytic peptides and $M_{i,j}$ represents the similarity score between the nodes $V_i(G)$ and $V_j(G)$, being 1 the highest similarity value and 0 the lowest. Then a rule called Half-Space Proximal (HSP) test[29] is applied to construct the HSPN, which is a strongly connected but sparse network[26], that preserves the number of nodes while containing a relatively low number of edges compared to the counterparts, CSNs[26].
- Finally, a threshold or cutoff value t can be applied to the weighted edges to further reduce the density of the graph by removing edges whose similarity values are lower than t. This helps to study the topology of the resulting graphs and subsequently find the best representative network of the chemical space occupied by hemolytic peptides. It is worth mentioning that for the construction of HSPNs, using a t value is not mandatory.

HSPNs were constructed as follows. From the 45120 peptides found in *StarPepDB*[16], 2004 peptides with known hemolytic activity were retrieved using the query option of StarPep toolbox[26]. Redundancy in the peptide sequences was removed using Smith-Waterman local alignment[65] and BLOSUM-62 substitution matrix[66] considering at least 98% sequence identity, resulting in 1647 peptides (SM1.1.3). Then MDs were calculated for each peptide sequence and an unsupervised feature selection was performed, removing near constant peptide features using Shannon entropy (threshold 10%), whereas redundant features were removed using Spearman correlation coefficient (threshold 0.8%). Then all the remaining peptide features were selected for generating the networks. See reference[26] for a detailed description of the peptide feature extraction method.

Regarding the (dis)similarity measures, HSPNs were constructed using Angular Separation (AS), Bhattacharyya (Bh), Chebyshev (Ch), Euclidean (Eu), and Soergel (So) measures. Their formulae and properties are stated in Table 5. We tested several measures since previous studies demonstrated

Table 5 | (Dis)Similarity Measures used to Construct HSPNs

Measure	Formula[a]	Range[b]	Average	Range
Angular Separation/ [1-Cosine (Ochiai)] (AS)	$d_{XY} = 1 - Cos_{XY}$ where, $Cos_{XY} = \frac{XY}{\|\boldsymbol{X}\|\|\boldsymbol{Y}\|} = \frac{\sum_{j=1}^{h} x_j y_j}{\sqrt{\sum_{j=1}^{h} x_j^2 \sum_{j=1}^{h} y_j^2}}$	$[0, 2]$		
Bhattacharyya (Bh)	$d_{XY} = \sqrt{\sum_{j=1}^{h}\left(\sqrt{x_j} - \sqrt{y_j}\right)^2}$	$[0, \infty)$	$\bar{d} = \frac{d_{XY}}{\sqrt{h}}$	$[0, \infty)$
Chebyshev/ Lagrange (Ch)	$d_{XY} = \max\left\{\left\|x_j - y_j\right\|\right\}$	$[0, \infty)$	$\bar{d} = \frac{d_{XY}}{h^{1/p}}$	$[0, \infty)$
Euclidean (Eu)	$d_{XY} = \left(\sum_{j=1}^{h}\left\|x_j - y_j\right\|^2\right)^{\frac{1}{2}}$			
Soergel (So)	$d_{XY} = \frac{1}{h}\sum_{j=1}^{h}\frac{\|x_j - y_j\|}{\max\{x_j, y_j\}}$	$[0, 1]$	$\bar{d} = \frac{d_{XY}}{h}$	$\left[0, \frac{1}{h}\right]$

[a] The variables x_j and y_j are the values of the molecular descriptor j of the peptides m and n, respectively. Peptides m and n are represented by the feature vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$. The h value is the number of peptide's features. The p value in the Euclidean metric is 2, whereas for Chebyshev $p \to \infty$.

[b] "*Range*" refers to "range" and not to "rank" and is defined as Range = max{x_j} – min {x_j}.

that different distance measures can codify orthogonal information; thus, not necessarily the Euclidean distance might be the best suited for a specific application[30,31].

In addition, to explore the behavior of HSPNs when varying the value of t, 11 different cutoffs were applied for each metric: 0.00 and from 0.50 to 0.95 in steps of 0.50, resulting in a total of 55 HSPNs available at SM3 (i.e., 11 networks for each metric). We applied these cutoffs, since a previous study showed that when constructing HSPNs, most of the global parameters barely changed when the similarity cutoff t ranges between 0.00 and 0.45[25].

Finally, since several combinations will be generated in the following steps, we will use the following notation when referring to a specific network: "cutoff (t)_metric". For instance, for a network generated with a t = 0.00 using the metric Angular Separation, its corresponding name will be: 0.00_AS.

Network visualization. For METNs, Betweenness Centrality[34] was calculated and the size of metadata nodes was proportionally projected according to the corresponding centrality value. *Database*, *Function*, *Origin* and *Target* METNs were visualized by coloring their metadata nodes: aquamarine, yellow, light violet and green, respectively. Metadata nodes related to hemolytic activity (i.e., toxic, toxic to mammals, hemolytic, Red Blood Cells) were colored red. On the other hand, all peptide nodes had the same size and were colored blue green for all the METNs. *Database* and *Function* METNs displayed their most central metadata nodes numbered. Finally, Force Atlas 2 layout algorithm was always used to visualize METNs[67].

For HSPNs, the nodes were clustered using the Louvain method[68], and the Hub-Bridge centrality (HB) measure was calculated for each node. Finally, to better visualize the networks, we colored the nodes according to the cluster they belong to, and the node size was set to be proportional to its HB centrality value using the Bezier interpolator. Finally, we applied the Fruchterman Reingold layout algorithm[69]. The resulting METNs and HSPNs were exported as GraphML files and further visualized with Gephi 0.9.7[70].

Selection of the best HSPNs. Using Gephi 0.9.7, the following global network parameters were retrieved for each HSPN: number of edges, modularity, density, average clustering coefficient (ACC), number of clusters/communities, singletons GC (nodes disconnected from the giant component), singletons D0 (nodes of degree zero), diameter, average path length, average degree and the probability of k (degree distribution). These features were used to study the behavior of the networks and select the best representations for each (dis)similarity measure (five networks in total). The best networks with their optimal cutoff value t were then used for scaffold extraction.

HSPNs scaffold extraction and analysis

This step aims to retrieve representative subsets of the hemolytic peptide space. The five best networks and the five networks with $t = 0.00$ (one for each similarity measure) were used to build the scaffolds. The following steps were applied for each of these networks:

The selected HSPNs were generated again, but now only the corresponding cutoff value t for each network was applied. We also calculated the Harmonic centrality (HC) for each node. After that, we applied the scaffold extraction method (integrated into the StarPep toolbox), which retrieves the most central and unique hemolytic peptides by ranking each peptide in decreasing order regarding their centrality and then redundant sequences were removed as follows: if a pair of sequences have a percentage identity higher than a certain cutoff value s, the least central peptide of the pair will be removed (please, do not confuse cutoff value t with cutoff value s. The former was used to construct networks whereas the latter was used for scaffold extraction). Finally, the resulting scaffolds of peptide sequences were exported as fasta files. This method generally assures extraction of the most representative peptides from all the centrality ranges but removes sequence redundancy.

Following the same notation for HSPNs, for naming the scaffolds, we inherit the name of the parent network followed by the centrality measure, alignment type, and cutoff s value: "cutoff t_metric_centrality_alignment_cutoff s". For instance, a scaffold extracted from the network 0.00_AS using harmonic centrality, local alignment, and a cutoff s = 0.80 would be named as: 0.00_AS_HC_L_0.80.

In this experiment, we varied the type of centrality measure, the sequence alignment type and the cutoff value of percentage identity s. We used Hub-Bridge (HB) or Harmonic centrality (HC), and Needleman-Wunsch global alignment (G)[71] or Smith-Waterman local alignment (L)[65] were used for sequence comparison, both with BLOSUM-62 substitution matrix. Moreover, we tested various cutoff values s ranging from 0.40 to 0.90 in steps of 0.10. As a result, for each of the ten selected networks, we generated 24 different scaffolds using the combinations described above. In total 240 scaffolds were obtained (see SM4).

Scaffold comparison by metric. For all the scaffolds generated from networks where the cutoff $t = 0.00$, the JSC was calculated between scaffold pairs created with the same parameters but differing in their metric. For instance, 0.00_**AS**_HB_G_0.40 *vs*. 0.00_**Bh**_HB_G_0.40. JSC is defined as the number of elements of the intersection of sets A and B divided by the number of elements of the union of those sets[53]. We calculated this distance to assess the similarity between scaffold pairs generated with different parameters.

Scaffold comparison by cutoff *t*, type of alignment and centrality measure. Using scaffolds constructed from orthogonal metrics, we compared the effect of varying the *t* value when generating networks. For this task we calculated the JSC between scaffold pairs created with the same parameters but differing in their *t* value. For instance, **0.00**_AS_HB_G_0.40 *vs.* **0.90**_AS_HB_G_0.40. Similarly, the same approach was followed to evaluate the effect of the type of alignment and centrality measure in the representativity of the scaffolds.

All pairwise scaffold comparisons were conducted using SeqKit toolkit to extract peptide IDs and then the relation between sets was obtained using https://bioinformatics.psb.ugent.be/webtools/Venn/.

Motif discovery and enrichment

Motif discovery. Motif discovery was performed using the alignment-free method STREME (short for Sensitive, Thorough, Rapid Enrichment Motif Elicitation), which finds ungapped motifs enriched in input sequences compared to control sequences providing a statistical significance for each motif[55]. To generate a diversity of potential new hemolytic motifs, we employed the community information of HSPNs with cutoff $t = 0.00$ generated with the metrics: Angular Separation, Chebyshev and Euclidean. The following steps were performed for each of the networks:

- Using the StarPep toolbox we extracted the sequences of peptides belonging to each cluster (community) and saved them as fasta files. Then these files were used as input sequences for motif discovery. For control sequences, we let the method use shuffled input sequences. Since our peptides contain non-standard AAs, we provided a customized alphabet (SM5.1.1). Motifs ranging from three to six letters, at least 20% present in the input sequences and with a *p* value lower than 0.05 were retrieved.
- Similarly, the same steps were applied to retrieve motifs from the file containing 1647 non-redundant hemolytic peptides from *StarPepDB*. Then, motifs resulting from this process were combined with hemolytic motifs reported in the literature[1,8], and duplicated motifs were removed (see SM5.1.2).

Motif enrichment. Motif Enrichment was conducted using SEA (Simple Enrichment Analysis) from the MEME suite (https://meme-suite.org/meme/tools/sea)[72]. Hemolytic motifs identified in this study and motifs reported in the literature[1,8] were employed to assess whether they are enriched in benchmark databases. These motifs were evaluated in *HemoPI-1*, *StarPepDB* and *Big-Hemo* datasets. Sequences labeled as non-hemolytic were used as control sequences in enrichment analysis on *HemoPI-1*; sequences not having “hemolytic” metadata were used as control when the *StarPepDB* was used. For enrichment analysis in the *Big-Hemo* dataset, input sequences were shuffled by the SEA algorithm and used as control sequences. In addition, sequences with length less than three AAs were discarded (SM5.1.4). Finally, those motifs that are statistically significant in all three datasets were kept.

To study the relation between the number of hemolytic motifs in a peptide and the likelihood of being hemolytic, we conducted a motif count using FIMO (Find Individual Motif Occurrences)[73] on the datasets: ABPs, AVPs and FDA-approved. Those datasets were filtered to only contain peptides with up to 100 AAs. The number of hemolytic motifs contained in each peptide of a dataset was obtained using this approach. Additionally, using the consensus between two different robust model classifiers[54], MQSSM-I1[54] and “SVM + Motif (HemoPI-1) based” (HemoPI-1)[8], we predicted the peptide hemolytic activity. A threshold of 0.50 was used to consider a peptide as hemolytic.

Data availability

The starPep toolbox software and the respective user manual, are freely available online at http://mobiosd-hub.com/starpep. All underlying code and installation files are accessible through GitHub (GitHub - Grupo-Medicina-Molecular-y-Traslacional/StarPep: StarPep toolbox: a software for studying the antimicrobial chemical space with newtork science tools and similarity searching models) under the Apache 2.0 license.

Received: 29 June 2023; Accepted: 22 August 2024;
Published online: 04 October 2024

References

1. Timmons, P. B. & Hewage, C. M. HAPPENN is a novel tool for hemolytic activity prediction for therapeutic peptides which employs neural networks. *Sci. Rep.* **10**, 10869 (2020).
2. Win, T. S. et al. HemoPred: a web server for predicting the hemolytic activity of peptides. *Future Med. Chem.* **9**, 275–291 (2017).
3. Xiao, Y.-F. et al. Peptide-based treatment: a promising cancer therapy. *J. Immunol. Res.* **2015**, e761820 (2015).
4. Hasan, M. M. et al. HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. *Bioinformatics* **36**, 3350–3356 (2020).
5. Kumar, V., Kumar, R., Agrawal, P., Patiyal, S. & Raghava, G. P. S. A method for predicting hemolytic potency of chemically modified peptides from its structure. *Front Pharm.* **11**, 54 (2020).
6. Plisson, F., Ramírez-Sánchez, O. & Martínez-Hernández, C. Machine learning-guided discovery and design of non-hemolytic peptides. *Sci. Rep.* **10**, 16581 (2020).
7. Wang, L. et al. Therapeutic peptides: current applications and future directions. *Sig Transduct. Target Ther.* **7**, 1–27 (2022).
8. Chaudhary, K. et al. A web server and mobile app for computing hemolytic potency of peptides. *Sci. Rep.* **6**, 22843, https://doi.org/10.1038/srep22843 (2016).
9. Yaseen, A., Gull, S., Akhtar, N., Amin, I. & Minhas, F. HemoNet: predicting hemolytic activity of peptides with integrated feature learning. *J. Bioinform. Comput. Biol.* **19**, 2150021, https://doi.org/10.1142/S0219720021500219 (2021).
10. Wilson, A. C., Vadakkadath Meethal, S., Bowen, R. L. & Atwood, C. S. Leuprolide acetate: a drug of diverse clinical applications. *Expert Opin. Investig. Drugs* **16**, 1851–1863 (2007).
11. Knox, C. et al. DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Res.* **52**, D1265–D1275 (2024).
12. Singh, S. et al. SATPdb: a database of structurally annotated therapeutic peptides. *Nucleic Acids Res.* **44**, D1119–D1126 (2016).
13. Van Avondt, K., Nur, E. & Zeerleder, S. Mechanisms of haemolysis-induced kidney injury. *Nat. Rev. Nephrol.* **15**, 671–692 (2019).
14. Gautam, A. et al. Hemolytik: a database of experimentally determined hemolytic and non-hemolytic peptides. *Nucleic Acids Res.* **42**, D444–D449 (2014).
15. Pirtskhalava, M. et al. DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Res.* **49**, D288–D297 (2021).
16. Aguilera-Mendoza, L. et al. Graph-based data integration from bioactive peptide databases of pharmaceutical interest: toward an organized collection enabling visual network analysis. *Bioinformatics* **35**, 4739–4747 (2019).
17. Salem, M., Keshavarzi Arshadi, A. & Yuan, J. S. AMPDeep: hemolytic activity prediction of antimicrobial peptides using transfer learning. *BMC Bioinform.* **23**, 389 (2022).
18. Sharma, R. et al. EnDL-HemoLyt: ensemble deep learning-based tool for identifying therapeutic peptides with low hemolytic activity. *IEEE J. Biomed. Health Inform.* 1–11 (2023).
19. Vespignani, A. Twenty years of network science. *Nature* **558**, 528–529 (2018).
20. Watts, D. J. & Strogatz, S. H. Collective dynamics of ‘small-world’ networks. *Nature* **393**, 440–442 (1998).
21. Sporns, O., Chialvo, D. R., Kaiser, M. & Hilgetag, C. C. Organization, development and function of complex brain networks. *Trends Cogn. Sci.* **8**, 418–425 (2004).

22. Roy, S., Cherevko, A., Chakraborty, S., Ghosh, N. & Ghosh, P. Leveraging network science for social distancing to curb pandemic spread. *IEEE Access* **9**, 26196–26207 (2021).
23. Roy, S., Biswas, P. & Ghosh, P. Effectiveness of network interdiction strategies to limit contagion during a pandemic. *IEEE Access* **9**, 95862–95871 (2021).
24. Romero, M. et al. A novel network science and similarity-searching-based approach for discovering potential tumor-homing peptides from antimicrobials. *Antibiotics* **11**, 401 (2022).
25. Ayala-Ruano, S. et al. Network science and group fusion similarity-based searching to explore the chemical space of antiparasitic peptides. *ACS Omega* **7**, 46012–46036 (2022).
26. Aguilera-Mendoza, L. et al. Automatic construction of molecular similarity networks for visual graph mining in chemical space of bioactive peptides: an unsupervised learning approach. *Sci. Rep.* **10**, 18074 (2020).
27. Aguilera-Mendoza, L. et al. StarPep toolbox: an open-source software to assist chemical space analysis of bioactive peptides and their functions using complex networks. *Bioinformatics* **39**, btad506 (2023).
28. Agüero-Chapin, G. et al. Emerging computational approaches for antimicrobial peptide discovery. *Antibiotics* **11**, 936 (2022).
29. Chavez, E. et al. Half-space proximal: a new local test for extracting a bounded dilation spanner of a unit disk graph. In Proceedings of the Principles of Distributed Systems; Anderson, J. H., Prencipe, G., Wattenhofer, R., Eds. pp. 235–245 (Springer, 2006).
30. Aggarwal, C. C. Hinneburg, A. & Keim, D. A. On the surprising behavior of distance metrics in high dimensional space. In Proceedings of the Database Theory — ICDT 2001; Van den Bussche, J., Vianu, V., Eds.; pp. 420–434 (Springer, 2001).
31. Marrero-Ponce, Y. et al. Optimum search strategies or novel 3D molecular descriptors: is there a stalemate? *Curr. Bioinform.* **10**, 533–564 (2015).
32. Miranda-Quintana, R. A., Bajusz, D., Rácz, A. & Héberger, K. Differential consistency analysis: which similarity measures can be applied in drug discovery? *Mol. Inform.* **40**, 2060017, https://doi.org/10.1002/minf.202060017 (2021).
33. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLOS ONE* **11**, e0163962 (2016).
34. Brandes, U. A faster algorithm for betweenness centrality. *J. Math. Sociol.* **25**, 163–177 (2001).
35. UniProt Consortium UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res* **51**, D523–D531 (2023).
36. Fan, L. et al. DRAMP: a comprehensive data repository of antimicrobial peptides. *Sci. Rep.* **6**, 24482 (2016).
37. Wang, C. K. L., Kaas, Q., Chiche, L. & Craik, D. J. CyBase: a database of cyclic protein sequences and structures, with applications in protein discovery and engineering. *Nucleic Acids Res.* **36**, D206–D210 (2008).
38. Katsara, M. et al. Round and round we go: cyclic peptides in disease. *Curr. Med. Chem.* **13**, 2221–2232 (2006).
39. Wang, Y. et al. NeuroPep: a comprehensive resource of neuropeptides. *Database* **2015**, bav038 (2015).
40. Seebah, S. et al. Defensins knowledgebase: a manually curated database and information source focused on the defensins family of antimicrobial peptides. *Nucleic Acids Res.* **35**, D265–D268 (2007).
41. de Jong, A., van Heel, A. J., Kok, J. & Kuipers, O. P. BAGEL2: mining for bacteriocins in genomic data. *Nucleic Acids Res.* **38**, W647–W651 (2010).
42. Greco, I. et al. Correlation between hemolytic activity, cytotoxicity and systemic in vivo toxicity of synthetic antimicrobial peptides. *Sci. Rep.* **10**, 13206 (2020).
43. Shai, Y. Mechanism of the binding, insertion and destabilization of phospholipid bilayer membranes by alpha-helical antimicrobial and cell non-selective membrane-lytic peptides. *Biochim Biophys. Acta* **1462**, 55–70 (1999).
44. Matsuzaki, K. Why and how are peptide-lipid interactions utilized for self defence? *Biochem Soc. Trans.* **29**, 598–601 (2001).
45. Saviello, M. R. et al. New insight into the mechanism of action of the temporin antimicrobial peptides. *Biochemistry* **49**, 1477–1485 (2010).
46. Kato, Y. et al. Abf-1 and Abf-2, ASABF-type antimicrobial peptide genes in caenorhabditis elegans. *Biochem J.* **361**, 221–230 (2002).
47. Conlon, J. M. The therapeutic potential of antimicrobial peptides from frog skin. *Rev. REs. Med. Microbiol.* **15**, 17 (2004).
48. Conlon, J. M. et al. A family of Brevinin-2 peptides with potent activity against pseudomonas aeruginosa from the skin of the Hokkaido Frog, Rana Pirica. *Regul. Pept.* **118**, 135–141 (2004).
49. Wang, H. et al. Molecular cloning and characterization of antimicrobial peptides from skin of the broad-folded Frog, Hylarana Latouchii. *Biochimie* **94**, 1317–1326 (2012).
50. Bassetti, M., Vena, A., Croxatto, A., Righi, E. & Guery, B. How to manage pseudomonas aeruginosa infections. *Drugs Context* **7**, 212527 (2018).
51. Zahoránszky-Kőhalmi, G., Bologa, C. G. & Oprea, T. I. Impact of similarity threshold on the topology of molecular similarity networks and clustering outcomes. *J. Cheminform.* **8**, 16 (2016).
52. Coscia, M. The Atlas for the Aspiring Network Scientist (Michele Coscia, 2021).
53. Reina, D. G., Toral, S. L., Johnson, P. & Barrero, F. Improving discovery phase of reactive ad hoc routing protocols using Jaccard distance. *J. Supercomput.* **67**, 131–152 (2014).
54. Castillo-Mendieta, K.; et al. Multiquery similarity searching models: an alternative approach for predicting hemolytic activity from peptide sequence. *Chem. Res. Toxicol*. https://doi.org/10.1021/acs.chemrestox.3c00408 (2024).
55. Bailey, T. L. STREME: accurate and versatile sequence motif discovery. *Bioinformatics* **37**, 2834–2840 (2021).
56. Shin, S. Y. et al. Structure-activity analysis of SMAP-29, a sheep leukocytes-derived antimicrobial peptide. *Biochem. Biophys. Res. Commun.* **285**, 1046–1051 (2001).
57. Sun, S. et al. Specificity and mechanism of action of alpha-helical membrane-active peptides interacting with model and biological membranes by single-molecule force spectroscopy. *Sci. Rep.* **6**, 29145, https://doi.org/10.1038/srep29145 (2016).
58. Dykes, G. A., Aimoto, S. & Hastings, J. W. Modification of a synthetic antimicrobial peptide (ESF1) for improved inhibitory activity. *Biochem. Biophys. Res. Commun.* **248**, 268–272 (1998).
59. Feder, R., Dagan, A. & Mor, A. Structure-activity relationship study of antimicrobial dermaseptin S4 showing the consequences of peptide oligomerization on selective cytotoxicity. *J. Biol. Chem.* **275**, 4230–4238 (2000).
60. Nikawa, H., Fukushima, H., Makihira, S., Hamada, T. & Samaranayake, L. P. Fungicidal effect of three new synthetic cationic peptides against *Candida Albicans*. *Oral. Dis.* **10**, 221–228 (2004).
61. Langham, A. A. et al. Correlation between simulated physicochemical properties and hemolycity of protegrin-like antimicrobial peptides: predicting experimental toxicity. *Peptides* **29**, 1085–1093 (2008).
62. Gramicidin, D. Available online: https://go.drugbank.com/drugs/DB00027 (accessed 5 April 2023).
63. Usmani, S. S. et al. THPdb: database of FDA-approved peptide and protein therapeutics. *PLoS One* **12**, e0181748 (2017).
64. Diestel, RGraph theory 5th *Graduate Texts in Mathematics, Vol 173*. 5th (Springer: 2017).
65. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
66. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *PNAS* **89**, 10915–10919 (1992).
67. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi Software. *PLOS ONE* **9**, e98679 (2014).

68. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).
69. Graph Drawing by Force-directed Placement—Fruchterman—1991—Software: Practice and Experience—Wiley Online Library Available online: https://onlinelibrary.wiley.com/doi/10.1002/spe.4380211102 (accessed 13 Feb. 2023).
70. Bastian, M., Heymann, S. & Jacomy, M. Gephi: an open source software for exploring and manipulating networks. *Proc. Int. AAAI Conf. Web Soc. Media* **3**, 361–362 (2009).
71. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
72. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME suite. *Nucleic Acids Res.* **43**, W39–W49 (2015).
73. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
74. Inkscape. Inkscape Project 2023.
75. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis; Use R!; 1st ed*. (Springer, 2009). ISBN 978-0-387-98141-3.
76. Hulsen, T. DeepVenn—a Web Application for the Creation of Area-Proportional Venn Diagrams Using the Deep Learning Framework Tensorflow.Js https://doi.org/10.48550/arXiv.2210.04597 (2022).

Acknowledgements

Y.M.-P. thanks to the program “Profesor invitado” for a post-doctoral fellowship to work at Universidad Panamericana (UP) in 2023, to the support from USFQ Collaboration Grant 2021-23 (Project ID17886) and Med Grant 2023 (Project ID17601). G.A.-C. was supported by national funds through FCT—Foundation for Science and Technology within the scope of UIDB/04423/2020 and UIDP/04423/2020.

Author contributions

K.C.-M was involved in all experiments and writing of the manuscript. G.A.-C. and Y.M.-P. worked mainly on the conceptualization, formal analysis, supervision, validation, writing and reviewing of the manuscript. E.A.-M., C.R.G.-J., and Y.P.-C. worked/supervised data curation and analyses of METNs and HSPNs representing the hemolytic peptide space. S.J.-B. and N.S.-V. were responsible for the scaffold extraction and selection as well as the new hemolysis-related motifs discovery. All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41540-024-00429-2.

Correspondence and requests for materials should be addressed to Guillermin Agüero-Chapin or Yovani Marrero-Ponce.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.